



Delft University of Technology

Document Version

Final published version

Licence

CC BY-SA

Citation (APA)

Sekwenz, M. T. (2026). *Evaluating Compliance in the European Union's Digital Services Act*. [Dissertation (TU Delft), Delft University of Technology]. <https://doi.org/10.4233/uuid:7d0020ee-6c15-4b49-9184-cdbc095ac275>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

EVALUATING COMPLIANCE IN THE EUROPEAN UNION'S DIGITAL SERVICES ACT

Marie-Therese Sekwenz



EVALUATING COMPLIANCE IN THE EUROPEAN UNION'S DIGITAL SERVICES ACT

EVALUATING COMPLIANCE IN THE EUROPEAN UNION'S DIGITAL SERVICES ACT

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof.dr.ir. H. Bijl,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op 14 september 2026 om 17:30 uur

door

Marie-Therese SEKWENZ

Dit proefschrift is goedgekeurd door de (co)promotores:

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
Prof.mr.dr. H. de Bruijn	Delft University of Technology, promotor
Prof.dr. B. Wagner	Delft University of Technology, copromotor

Onafhankelijke leden:

Assoc. Prof. D. Murray-Rust	Delft University of Technology
Prof.dr. M. van Eeten	Delft University of Technology
Prof.dr. T. McGonagle	Leiden University, the Netherlands
Prof.dr. L. Taylor	Tilburg University, the Netherlands
Prof.dr. S. Milan	University of Amsterdam, the Netherlands



Keywords: Digital Services Act, Platform Governance, Measuring Compliance, Auditing Online Platforms, Legal Design, Systemic Risk Assessment

Published by: TU Delft Open

Printed by: Gildeprint

Cover image: M-T. Sekwenz

Copyright © 2026 by M-T. Sekwenz

ISBN: 978-94-6518-418-0

An electronic version of this dissertation is available at

<https://repository.tudelft.nl/>

CONTENTS

Summary	xv
Samenvatting	xvii
1 Introduction	1
1.1 Background	1
1.1.1 The need for harmonizing regulation	1
1.2 Problem.	3
1.3 Research gaps.	4
1.4 Research aim and leading questions	4
1.4.1 Research Questions	4
1.4.2 Chapter-specific research questions	5
1.5 Approach	8
1.6 Outline and Contribution.	11
1.7 Research Significance and Novelty	11
I Conceptual elements of DSA compliance	15
2 The Digital Services Act: Rules, Rights, and Risks	17
2.1 Introduction	19
2.2 See-through regulation? Novel transparency mechanisms in the DSA. . .	21
2.2.1 A harmonised form of reporting through DSA transparency reports?	22
2.2.2 A place for all platform contracts – The terms and conditions database	23
2.2.3 Content-Moderation Decisions in the Statement of Reasons Database 23	
2.2.4 Data access for researchers	24
2.3 On user rights, processes and institutionalised flagging entities.	25
2.3.1 User rights and empowerment.	25
2.3.2 Internal complaint-handling and out-of-court dispute settlement.	25
2.3.3 Trusted Flaggers	26
2.3.4 Additional user empowerment measures	26
2.4 Identifying and mitigating systemic risks for intermediaries	26
2.4.1 Systemic risk assessments	26
2.4.2 Mitigation measures	27
2.4.3 Independent audits and enhanced supervision	27
2.4.4 Crisis response mechanism	27
2.4.5 Data access for vetted researchers	28
2.4.6 Conclusion of risk-based approach	28

2.5	Conclusion	28
3	Content Moderation and User Engagement	31
3.1	Introduction	33
3.1.1	Background	33
3.2	Related Work	35
3.3	The Process of Content Moderation Under the DSA.	37
3.4	Emerging Tensions and Implementation Challenges	39
3.5	Conclusion	40
4	Wicked by Design: Regulating and Auditing Platforms under the DSA	43
4.1	Introduction	45
4.2	Why Content Moderation Is a Wicked Problem	45
4.3	The DSA as a socio-technical and socio-legal system	46
4.4	Socio-legal system: private ordering and public oversight.	47
4.5	Multi-layered complexity in evaluating DSA compliance	47
4.5.1	What Is Legal Complexity?	48
4.5.2	Why Complexity Is Structural	49
4.5.3	The DSA as Legal Complexity in Action	49
4.5.4	The Limits of Simplification	50
4.5.5	The Problem of Legal Certainty	50
4.6	Sources of Socio-Technical Complexity	50
4.6.1	Fragmented Reporting Mechanisms	50
4.6.2	Socio-Technical Moderation	51
4.6.3	AI Support and Bias	51
4.6.4	Auditing at Scale	51
4.6.5	Interdependence Across Layers	52
4.7	Reframing Complexity as a Design Condition.	52
4.7.1	Implications for Actors.	53
4.8	Limitations and Future Work	53
4.9	Conclusion	53
II	Methodological frameworks for assessing DSA compliance	55
5	Holistic DSA Compliance Through Leveraging Consistency	57
5.1	Introduction	59
5.2	Methodology	60
5.2.1	Typology of inconsistency and plausible explanations	62
5.2.2	Interdisciplinary methodological framing	62
5.2.3	Case study selection: Instagram	63
5.3	Related Work	64
5.3.1	The DSA's Transparency and Due-Diligence Architecture	65
5.4	Three Levels of Consistency: A New Methodology to Evaluate the Compliance of DSA reporting mechanisms.	66
5.4.1	Internal Consistency.	67
5.4.2	External Consistency.	68
5.4.3	Historical Consistency	70

5.5	Challenges in Evaluating Consistency Across the SOR and Transparency Reports	70
5.5.1	Structural and Semantic Misalignments	71
5.5.2	Internal Inconsistencies in Quantitative Reporting.	71
5.5.3	Lack of Alignment on Illegal Content and Flagging Mechanisms . .	71
5.5.4	Misalignment of Categories Across Mechanisms.	72
5.5.5	Disjunctions Between SRAs and Transparency Reporting	72
5.5.6	Ambiguities in Advertising Content	73
5.5.7	Lack of Standardized Measures of Automated Accuracy	73
5.6	Recommendations to Evaluate Consistency Across DSA Compliance Mechanisms	74
5.7	Leveraging Consistency for Regulatory Evaluation	76
5.8	Discussion	76
5.8.1	Limitations.	79
5.9	Conclusion	80
6	Systemic Risk Assessment, Independent Audits and Methodologies	87
6.0.1	Introduction	89
6.0.2	Background	90
6.0.3	Auditing Very Large Online Platforms and Very Large Search Engines in the Digital Services Act	91
6.0.4	Mixed Methods – Combining Statistical Content Sampling Techniques and Content Analysis for Systemic Risk Assessments and Independent Audits	93
6.0.5	Ensuring Representativeness of Content Samples	98
6.0.6	Conducting Independent Audits and Systemic Risk Assessments .	99
6.0.7	Comparison of Systemic Risk Assessments and Audit Reports . . .	100
6.0.8	Conclusion and Further Research	103
7	Legal doctrinal content analysis – Developing a legal code book for advertising and electoral content	105
7.1	Introduction	107
7.2	Political Advertising and Regulation.	109
7.3	Introducing the legal provisions relevant for the context of elections and campaigning in the DSA, TTPA, and G–E–DSA	110
7.3.1	Ad Transparency and Systemic Risk in Europe's Digital Services Act	110
7.3.2	Regulation on Transparency and Targeting of Political Advertising.	112
7.3.3	The Commission's Guidelines on the Mitigation of Systemic Risks for Electoral Processes Pursuant to the DSA	113
7.4	Methodology	114
7.5	Results: Annotating, Reporting, and Transparency Obligations in Political Content and Advertising	116
7.6	Conclusion	117

8	Scaling up DSA Evaluations – Leveraging Large Language Models for Auditing online platforms under the Digital Services Act	119
8.1	Introduction	121
8.2	Related work	121
8.3	Methodology	122
8.4	Results	123
8.4.1	Breakout Sessions	123
8.4.2	Hands-On sessions: Evaluating prompt structures across teams . .	124
8.5	Discussion: Contributions to Professional Practice and Platform Oversight	125
8.6	Conclusion	126
III	Conducting empirical evaluation of DSA compliance	127
9	Evaluating the Systemic Risk of Disseminating Illegal content on Online Platforms under the Digital Services Act	129
9.1	Introduction	131
9.2	Definitions, scope, and limitations	132
9.3	Legal grounds for German platforms to moderate online content.	133
9.3.1	Overview of human rights, European provisions, and international provisions in the scope of content moderation in Germany	133
9.3.2	National provisions in Germany relevant to content moderation .	134
9.3.3	Terms of service of online platforms in Germany	135
9.4	Research design and coding methodology	135
9.4.1	Coding German law	137
9.5	Results	140
9.5.1	Empirical comparison: Platform decisions versus legal coding . .	140
9.5.2	Subjectivity and context in moderation decisions	140
9.5.3	Time and accuracy constraints.	140
9.6	Discussion	141
9.6.1	Implications for platform governance	141
9.6.2	Policy recommendations.	141
10	Evaluating the design of reporting mechanisms for illegal content in Practice: User Pathways, Interfaces, and Compliance	143
10.1	Introduction	145
10.2	Background	146
10.3	Related Work	147
10.4	Methodology	148
10.4.1	Study design	148
10.4.2	Legal design heuristic walkthrough	149
10.4.3	Recruitment	150
10.4.4	Research ethics	152
10.4.5	Data analysis.	152
10.4.6	Limitations.	152

10.5 Findings	153
10.5.1 Content reporting – a multi-actor process	153
10.5.2 Legal representation for users	154
10.6 Content reporting.	154
10.6.1 Differentiating illegal content from Terms of Service (ToS).	155
10.6.2 Content moderation and moderation decision.	156
10.6.3 Design of reporting mechanism	157
10.6.4 Stakeholder cooperation	158
10.7 Discussion	159
10.7.1 Users and the law: empowerment and burden	159
10.7.2 User action in the presence of platform opacity	161
10.8 Conclusion	162
11 Automating Digital Service Act User Support: Large Language Models as User Support for Filing Illegal Content Reports	163
11.1 Introduction and Background.	165
11.2 Related Work	168
11.3 Experimental Design and Procedure	170
11.4 Methodology	172
11.5 Results	175
11.6 Discussion	179
11.7 Conclusion	181
12 Understanding the Emerging Role of Trusted Flaggers under the EU Digital Services Act	183
12.1 Introduction	184
12.2 Background.	185
12.3 Related Work	186
12.4 Method	189
12.4.1 Semi-structured interviews	189
12.4.2 Recruitment	189
12.4.3 Participants	190
12.4.4 Participating organizations – website analysis	190
12.4.5 Interview analysis	191
12.4.6 Validation workshop	192
12.4.7 Ethical Considerations.	192
12.4.8 Limitations.	193
12.5 Results	193
12.5.1 Trusted Flagger activities.	194
12.5.2 Trusted Flagger activities and users	196
12.5.3 Platform frictions	198
12.5.4 Capacity constraints	199
12.5.5 Coordination challenges	200
12.5.6 Platform incentives	200
12.5.7 Operational ambiguity.	201
12.5.8 Accountability artifacts	201

12.5.9 Results - Workshop event	202
12.6 Discussion	204
12.6.1 Recommendations.	205
12.7 Conclusion	207
13 Assessing Fundamental Rights Impacts under the Digital Services Act of Very Large Platforms & Search Engines	209
13.1 Introduction	210
13.2 Background and Related Work	211
13.2.1 Systemic Risk Assessment under the DSA	212
13.2.2 Vulnerability as a Heightened Risk to Fundamental Rights.	214
13.2.3 Risk Assessment Frameworks and Toolkits.	215
13.3 Methodology	216
13.3.1 Author-led structured use case walkthrough	216
13.4 Findings	217
13.4.1 DSA Applicability and Context Information	218
13.4.2 Subsystem Identification.	219
13.4.3 Stakeholder Identification	220
13.4.4 Risk Identification and Assessment	220
13.4.5 Scoring, Weighing and Reporting	222
13.5 Limitations and Future Work	223
13.6 Discussion	223
13.7 Conclusion	224
14 Evaluating the Systemic Risk of Negative Effects on the Electoral Process content on Online Platforms under the Digital Services Act	229
14.1 Introduction	231
14.2 Theoretical and methodological approach	231
14.2.1 Theoretical framework.	232
14.2.2 DSA, online platforms, and systemic risks	233
14.2.3 Systemic electoral risks and categories of analysis	233
14.2.4 Research design, methodology, and case selection.	235
14.3 Applying the DSA's risk assessment and mitigation framework to the German federal elections	238
14.3.1 Dissemination of illegal content	238
14.3.2 Negative effects on electoral rights.	239
14.3.3 Disinformation	239
14.4 Conclusion	240
IV Synthesis	243
15 EU Digital Regulation as Industry Shaping Policy:	245
15.1 Introduction	247
15.2 Empirical Trends in DSA Enforcement: Insights from the SoR Transparency Database.	248
15.3 Strategic Incentives: Global Standards vs. National Fragmentation	252
15.4 Innovation Climate	254

15.5 Conclusion	255
15.6 Disclaimer	255
16 Regulating Through Evidence	257
16.1 Introduction	259
16.2 Background and Related Work	259
16.3 Method	260
16.4 Findings	260
16.5 Discussion	261
16.5.1 Implications for Key Stakeholders	264
16.6 Conclusion	265
17 Conclusion	267
17.1 Empirical findings	267
17.2 Methods and Measurements for Online Platforms, Regulators, and Audi- tors	269
17.3 Societal impact	270
17.4 Recommendations	272
17.5 Future Work and Limitations	274
17.6 Limitations	274
17.7 Future Work.	276
17.8 Final Reflections	277
Bibliography	279
A Appendix for Chapter 5	353
A.1 Overview of Key Legal References and Provisions	353
B Appendix for Chapter 6	355
B.1 Overview of Key Legal References and Provisions	355
C Appendix for Chapter 7	357
C.0.1 Code Book	357
C.0.2 Disinformation Categories	357
C.0.3 Electoral Rights Categories.	366
C.1 Regulation of political content	383
C.1.1 Regulation of online content.	383
D Appendix for Chapter 8	387
D.1 Information About Workshop Participants	387
D.2 Thematic Analysis Process and Coding Trace	387
D.2.1 Thematic Analysis Process	387
D.2.2 Detailed Coding Trace	389
E Appendix for Chapter 9	395
E.1 Codebook Legal and non-legal codes	395

F Appendix for Chapter 10	503
F.0.1 Semi-Structured Interview Guidelines	503
F.0.2 Platform Descriptors	506
F.0.3 Codebook and Thematic Findings in Themes and Subthemes . . .	507
G Appendix for Chapter 11	511
G.1 Summary of Hypothesis	511
G.2 H3: Decision time on no-error trials by assistance condition	512
G.3 Error Types	513
G.4 Pre-Study Questionnaire	513
G.4.1 Need for Cognition (5-point Likert)	513
G.4.2 Use & Apply AI (0–10 scale)	513
G.4.3 Know & Understand AI (0–10 scale)	514
G.4.4 Detect AI (0–10 scale)	514
G.4.5 AI Ethics (0–10 scale)	515
G.4.6 AI Self-Efficacy — Problem Solving (0–10 scale)	515
G.4.7 AI Self-Efficacy — Learning (0–10 scale)	515
G.4.8 AI Self-Competency — Persuasion Literacy (0–10 scale)	515
G.4.9 AI Self-Competency — Emotion Regulation (0–10 scale)	516
G.4.10 Trust in Automation (5-point Likert)	516
G.4.11 Legal Knowledge (5-point Likert)	516
G.5 Post-Study Questionnaire	516
G.5.1 Usefulness and Effectiveness (7-point Likert)	516
G.5.2 Workload (7-point Likert)	517
G.5.3 Trust, Control, and Responsibility (7-point Likert)	517
G.5.4 Demographics and Final Comments	517
G.6 Descriptives for Participants, Error Types, and Legal Conditions	518
G.7 Codebook for Explanation of the Reason Quality and Related Measures .	519
G.8 Prompting Template for AI generated suggestions for condition 2	520
G.9 Selection of and creation of reporting scenarios and legal cases relevant to online content reporting in Germany.	521
G.9.1 Screenshot of Facebook's Reporting UI for Illegal Content	522
G.9.2 Prompting and evaluation strategy.	522
G.9.3 Reporting scenarios and legal cases for reporting illegal content in Germany.	524
G.10 Additional Analyses for H4: Explanation-of-Reasons (EoR) Quality	524
G.10.1 Inter-Rater Reliability	524
G.10.2 Model Specification and Robustness	524
G.10.3 Codebook	525
G.11 Appendix: Detailed Analysis Plan	525
G.11.1 Overview.	525
G.11.2 Outcome Definitions and Coding	525
G.11.3 H1: Provision Accuracy Under AI Error.	526
G.11.4 H2: Misclassification Distance Under AI Error	526
G.11.5 H3: No-Error Trials (Correct AI Assistance).	527
G.11.6 H4: Explanation-of-Reasons Quality	527

G.11.7 H5: Error-Type Impact and Attenuation	528
G.11.8 H6: DSA-Relevant Downstream Risk Outcomes	529
G.11.9 Participants	530
H Appendix for Chapter 12	533
H.1 Interview Guideline	533
H.2 Trusted Flagger Activity in the Statement of Reasons Database	536
H.3 Main Themes Presented to Workshop participants	537
I Appendix for Chapter 13	539
I.1 Facts-as-Given Sheet (Evaluation Input Protocol)	539
I.2 Use Case Testing Walkthrough	540
I.2.1 DSA applicability.	540
I.2.2 Context information answers for the worked case	541
I.2.3 Sub-system identification and stakeholder specification.	542
I.2.4 Risk identification and assessment.	545
I.2.5 Documentation	547
J Appendix for Chapter 16	549
J.0.1 AI-assisted labelling prompt (primary target label).	549
J.0.2 Table 1. Decisions by category (Fines, Preliminary findings, Re- quests for information)	551
Acknowledgements	559
Authorship Contributions	571
List of Publications	575
About the Author	579

SUMMARY

This dissertation examines how compliance with the European Union's Digital Services Act (DSA) can be understood, measured, and evaluated. The DSA reshapes the governance of online platforms through procedural duties, systemic-risk mitigation, transparency requirements, and independent auditing. Evaluating those obligations requires interdisciplinary inquiry that connects legal reasoning with technical systems, organisational practices, and user experience.

Part I develops the conceptual foundations. It analyses the DSA's regulatory architecture, the content-moderation processes affected by the Regulation, and the socio-technical complexity created by their interaction. Compliance is treated not as a static checklist but as an interpretive and organisational practice involving multiple actors, legal obligations, technical systems, and interdependent artefacts, including transparency reports, statements of reasons, systemic-risk assessments, and audit reports.

Part II develops methods for evaluating compliance. These include a consistency-based approach for comparing reporting artefacts, mixed methods for systemic-risk assessment and independent auditing, a legal codebook for analysing political and electoral content, and an examination of large language models as possible support tools in DSA audits. These approaches are intended as operational starting points for regulators, auditors, researchers, and civil-society organisations.

Part III applies these approaches in empirical studies of platform reporting, moderation decisions, user reporting pathways, AI-supported reporting, Trusted Flaggers, and systemic risks. The studies identify recurring gaps between legal requirements, platform interfaces, and the evidence made available for evaluation. The results show why claims about compliance must remain tied to the scope, data, and methods of each evaluation.

Part IV and the conclusion synthesise these findings and examine the DSA's wider role as a governance and industry-shaping framework. They also identify external conditions that influence compliance but are not fully observable through platform reports alone, including business incentives, monetisation structures, geopolitical pressures, enforcement choices, and differences between jurisdictions. Future work should therefore test the proposed methods longitudinally and across platforms, languages, Member States, and regulatory settings, while improving secure data-access infrastructures and standards for AI-supported auditing.

Overall, the dissertation contributes concepts, methods, and empirical evidence for the evaluation of platform governance under the DSA. It argues that credible assessment depends on combining expertise from law, human-computer interaction, data science, governance studies, and AI research, and on making the assumptions, evidence, uncertainty, and limits of each compliance claim explicit.

SAMENVATTING

Dit proefschrift onderzoekt hoe de naleving van de digitaledienstenverordening van de Europese Unie (Digital Services Act, DSA) kan worden begrepen, gemeten en geëvalueerd. De DSA hervormt het bestuur van onlineplatforms door middel van procedurele verplichtingen, de beperking van systeemrisico's, transparantievereisten en onafhankelijke audits. De evaluatie van deze verplichtingen vereist interdisciplinair onderzoek waarin juridische analyse wordt verbonden met technische systemen, organisatorische praktijken en gebruikerservaringen.

Deel I ontwikkelt de conceptuele grondslagen. Het analyseert de regelgevingsarchitectuur van de DSA, de contentmoderatieprocessen waarop de verordening van invloed is en de sociaal-technische complexiteit die uit hun wisselwerking voortvloeit. Naleving wordt niet beschouwd als een statische checklist, maar als een interpretatieve en organisatorische praktijk waarbij meerdere actoren, juridische verplichtingen, technische systemen en onderling afhankelijke documenten betrokken zijn, waaronder transparantievereisten, motiveringen, systeemrisicobeoordelingen en auditverslagen.

Deel II ontwikkelt methoden voor het evalueren van naleving. Deze omvatten een op consistentie gebaseerde benadering voor het vergelijken van rapportagedocumenten, gemengde onderzoeksmethoden voor systeemrisicobeoordelingen en onafhankelijke audits, een juridisch codeboek voor de analyse van politieke en verkiezingsgerelateerde inhoud, en een onderzoek naar grote taalmodellen als mogelijke ondersteunende instrumenten bij DSA-audits. Deze benaderingen zijn bedoeld als praktische uitgangspunten voor toezichthouders, auditors, onderzoekers en maatschappelijke organisaties.

Deel III past deze benaderingen toe in empirische studies naar platformrapportages, moderatiebeslissingen, meldingsroutes voor gebruikers, AI-ondersteunde meldingen, betrouwbare flaggers en systeemrisico's. De studies brengen terugkerende discrepanties aan het licht tussen juridische vereisten, platforminterfaces en het bewijsmateriaal dat voor evaluatie beschikbaar wordt gesteld. De resultaten laten zien waarom uitspraken over naleving steeds moeten worden gekoppeld aan de reikwijdte, gegevens en methoden van iedere afzonderlijke evaluatie.

Deel IV en de conclusie brengen deze bevindingen samen en onderzoeken de bredere rol van de DSA als kader voor governance en voor de ontwikkeling van de platformsector. Zij identificeren tevens externe omstandigheden die de naleving beïnvloeden, maar niet volledig uit platformverslagen kunnen worden afgeleid, waaronder bedrijfsprikkels, verdienenmodellen, geopolitieke druk, handhavingskeuzes en verschillen tussen rechtsgebieden. Toekomstig onderzoek dient de voorgestelde methoden daarom longitudinaal en voor verschillende platforms, talen, lidstaten en regelgevingscontexten te toetsen. Daarnaast moeten veilige infrastructuren voor gegevenstoegang en normen voor AI-ondersteunde audits verder worden ontwikkeld.

In haar geheel draagt dit proefschrift concepten, methoden en empirisch bewijs aan voor de evaluatie van platformgovernance onder de DSA. Het betoogt dat een geloofwaardige

beoordeling afhankelijk is van de combinatie van expertise uit het recht, de mens-computerinteractie, datawetenschap, governanceonderzoek en AI-onderzoek. Bovendien moeten de aannames, het bewijsmateriaal, de onzekerheden en de beperkingen van iedere uitspraak over naleving expliciet worden gemaakt.

1

INTRODUCTION

1.1. BACKGROUND

Society at large is changing. This change is fuelled by digital elements influencing society [738]. Signs of such change are polarisation [483], disinformation [324], and exposure to dangerous or illegal situations online [101]. These changes are not necessarily good but pose risks to fundamental rights [192], people in vulnerable positions [350], democratic values [180], or the collective understanding of reality [697]. To mitigate the risks that these digital products and services pose platforms aim to reduce problematic content through different measures. A key measure to keep problems at bay is content moderation. A process to reduce the amount of illegal or otherwise perceived problematic content from being displayed to users eyes that at the same time strives for legitimacy [583]. This is a balancing act weighing between different aims of freedom of expression and protecting society from harm.

Therefore content moderation is a vital part of the products large online platforms offer [738] – yet however central to society still opaque [481, 721]. The need to moderate stems from frequency of illegal or harmful content, which can amount to approximately one in ten items of content uploaded[444]. Human moderators are actors in the socio-technical content-moderation process [177, 396, 669]: they typically act first, and their decisions may then be followed by interventions from other actors, such as public authorities, Member States, Trusted Flaggers (Art. 22 DSA), or other privileged entities. The human moderators are supported by automated systems like Artificial Intelligence solutions which help in scaling, but at the same time include new risks [186, 289, 476].

1.1.1. THE NEED FOR HARMONIZING REGULATION

As online platforms and services are becoming more and more important in society, the European Union has introduced to regulate content moderation by platforms in the Digital Services Act [246]; which aims to address these tensions between platforms, content, and users, through new transparency, reporting, design, and audit obligations. This creates a regulatory framework around content moderation as well as providing a new set

of rules and principles based upon the quality of content moderation practices can be evaluated. But the complexity and novelty of the regulation, as well as the socio-technical and socio-legal nature of the regulation's subject (e.g., moderation systems), make compliance difficult to evaluate in practice. This creates risks for users, regulators, and society, as risks such as harmful or illegal content may persist while legitimate expression may be restricted.

The DSA is a newly implemented regulation (entered fully into force in February 2024), and neither platforms, auditors, nor regulators have yet established methods for assessing compliance across legal, technical, and organizational layers in detail. Compliance under the DSA comes in many forms, for example platforms need to conduct systemic risk assessments about their services and processes covering a wide range of risks like illegal content, negative effects on human rights, gender-based violence, or the civic discourse and electoral process (Art. 34 (1) (a-d) DSA).

The DSA does not operate in isolation. Its interpretation and application interact with the EU Charter of Fundamental Rights [248], the General Data Protection Regulation [4], the AI Act [241], the Regulation on the Transparency and Targeting of Political Advertising [24], sector-specific rules such as the Audiovisual Media Services Directive [16], and substantive national law used to determine whether content is illegal. These regimes can create overlapping or complementary duties. This dissertation focuses on the evaluability of DSA obligations; it does not claim that a positive result under a DSA-focused method establishes compliance with every adjacent legal regime.

This includes assessing platform-specific processes and artefacts, such as the design of illegal-content reporting interfaces under Article 16 DSA. Systemic risk assessments are only one part of the compliance cycle and are subject to independent audits under Article 37 DSA. These audits provide a second-stage assessment of platforms' systemic risk processes and an additional compliance signal for enforcement.

Additionally, platforms must regularly publish transparency reports on their moderation practices and provide consistent data on their moderation decisions to the Commission's content-moderation database — the Statements of Reasons database [148].

The wide palette of novel reporting and assessment obligations created by the DSA is currently uncoordinated and unstructured, creating uncertainty about legal compliance and best practices. The creation of methodologies tailored for this complex problem domain would create value for regulators, practitioners, civil society, researchers, and platforms. Such methodologies would profit from standardization and the collection or proposal of best practices showing solutions of how to collect meaningful empirical evidence to assess compliance [692].

Since the DSA is in force, empirical methods for auditing platform compliance are needed to overcome legal uncertainties and to assess compliance structurally by leveraging on the insights of different disciplines. This call for empirical interdisciplinary research is not only based on the regulation domain of the DSA – online platforms – but also through the various (human) actors and technology-based processes underlying the functioning of content moderation and platforms services as a whole.

Measurement choices are themselves part of platform governance. Decisions about the unit of analysis, category boundaries, thresholds, time periods, and the treatment of missing data determine which risks become visible and which remain obscured. A

compliance metric should therefore be assessed not only for technical validity, but also for the institutional interests it serves, the assumptions it embeds, and the consequences of treating its output as authoritative. Throughout the dissertation, proposed methods are consequently presented as inspectable and contestable choices rather than neutral representations of compliance.

1.2. PROBLEM

Content moderation strives to prevent platform users from seeing problematic, illegal, or harmful content, while balancing the right to freedom of expression of users on platforms to ensure that they can express themselves freely online. To strike this balance, platforms develop internal policies (like Terms of Service (ToS) and Community Standards (CS)), mechanisms (reporting mechanisms, internal complaint-handling mechanisms, or out-of-court-dispute mechanisms), and reporting artifacts (transparency reports, entries for the Statement of Reasons Database, Systemic Risk Assessment Reports, or Independent Audit Reports) to follow the DSA's novel obligations creating a socio-legal system of content moderation which is implemented through socio-technical systems facilitated by the platform [328]. In doing so, they are constantly in danger of restricting too much content and harming free expression or restricting too little content and creating harm or even breaking the law [741]. This involves multi-actor coordination with different roles pushing for different rules to be enforced, a high degree of automation to deal with the large volumes of content, strict time windows for moderation action and ever-changing demands and contexts for new rules defining what content should be moderated [290].

Referring to the different areas within the process of content moderation and its complexity, the DSA wants to harmonize the rules within the European Union addressing the problem domain of moderating speech at scale [202]. The novelty of such a new regulation itself, however, might easily add even more complexity to the problem space. Key uncertainties remain about how reporting actors (such as users and Trusted Flaggers) through DSA-mandated artifacts (reporting mechanisms) relate to systemic risk-mitigating moderation practices, how systemic risk assessments can be conducted reliably for platforms and independent audits, and how users can exercise their rights meaningfully. Because of the different actors included in the various stages of content moderation, the legal and contractual variety of sources to base a moderation decision upon, the different algorithmic dilemmas, e.g. like restricting too much based on misclassification, which automated solutions bring to the table, and complexities that might be encountered in the content moderation process, the human workers, and the hard task to protect freedom of expression while restricting just the 'right' amount of speech that breaks the rules are reasons why content moderation is such a wicked problem.

The overarching problem this research project tackles, therefore, can be summarized as follows:

The assessment of DSA-compliance is challenging in many ways due to the novelty of the law, the socio-legal and socio-technical nature of the content moderation process itself, the lack of methodologies to assess systemic risks or compliance of processes and artifacts, as well as due to the multiple actors influencing content moderation of large online platforms dealing with a vast

amount of context dependent moderation decisions, based on the law and platform rules.

The DSA is central to the EU's regulatory strategy for digital spaces [658], but the detailed assessment of compliance remains unsettled. This uncertainty matters because the Commission has pursued platform-compliance investigations since the DSA became applicable [661, 665]. How such assessments should be conducted remains unclear for the Commission and Digital Services Coordinators, for providers carrying out self-assessments, and for auditing firms conducting annual independent audits of VLOPs and VLOSEs. These actors need actionable and methodologically sound approaches to DSA compliance. The issue extends beyond services offered in the EU because the DSA may have regulatory spillover effects in other jurisdictions [90]. This reinforces the need for legal certainty, reliable compliance methodologies, and enforceable approaches that contribute to safer online environments.

1.3. RESEARCH GAPS

The research problem reveals four related gaps: first, which conceptual elements constitute DSA compliance and which challenges regulators, platforms, auditors, civil society, and researchers face when assessing it; second, how methodological frameworks can operationalise DSA-specific concepts such as systemic risk; third, how such methods can be applied in practice; and fourth, which broader lessons the DSA offers for content moderation in a global regulatory context.

1.4. RESEARCH AIM AND LEADING QUESTIONS

The overall aim of this dissertation is to advance the understanding of how online platforms can comply with the European Union's Digital Services Act (DSA) and to develop robust methodological frameworks for evaluating empirically-based compliance practices. By answering the fundamental research question, this project seeks to propose new approaches for assessing compliance.

This thesis examines the different layers involved in assessing compliance with the DSA: conceptual, methodological, empirical, and synthesis. To achieve this, the thesis is structured around a series of interconnected chapters that span legal doctrine, socio-technical complexity, empirical auditing, and user-centred evaluation. It examines the DSA's regulatory foundations, obligations, and institutional mechanisms, as well as the rights and accountability regimes it establishes.

1.4.1. RESEARCH QUESTIONS

The overarching and fundamental research question guiding this project is:

How can online platforms be assessed to create DSA-compliant online spaces?

OVERVIEW OF MAIN RESEARCH QUESTIONS BY CHAPTER

This central question is addressed through a set of four part-specific research questions that build a comprehensive understanding of Digital Services Act (DSA) compliance.

- **Part 1 — What are conceptual elements relevant for DSA compliance?**
- **Part 2 — Which methodological frameworks can be used to assess DSA compliance?**
- **Part 3 — How can empirical evaluation be conducted to assess compliance in online spaces under the DSA?**
- **Part 4 — What lessons from DSA enforcement and practice are essential for robust compliance assessment?**

Part 1 examines the key conceptual elements that define DSA compliance and shape how platforms should understand their legal obligations. Part 2 identifies and evaluates methodological frameworks suitable for assessing compliance in practice. Part 3 explores empirical evaluation of different central DSA domains. Finally, Part 4 reflects on the DSA's broader global implications, distils lessons from its first two years of enforcement, offers recommendations for practitioners, and outlines directions for future research.

1.4.2. CHAPTER-SPECIFIC RESEARCH QUESTIONS

To answer the overarching question, *How can online platforms be assessed to create DSA-compliant online spaces?*, this dissertation breaks the problem into a set of interconnected chapters that together provide a modular scaffold for assessing selected dimensions of DSA compliance across Parts 1–4 (see Table 1.2 for an overview of the condensed chapter-specific research questions). Because the Digital Services Act cuts across legal doctrine, institutional design, socio-technical and socio-legal processes, platform practices, and empirical auditing requirements, evaluating compliance cannot rely on any single analytical lens. Instead, this research adopts a modular approach, examining different aspects of compliance in turn and drawing on interdisciplinary, mixed-method approaches.

Each chapter in this dissertation examines a core component of DSA compliance that must be understood before meaningful compliance assessment is possible. These research questions therefore do not stand alone; rather, they cumulatively build the conceptual, methodological, empirical, and synthesized foundations required to evaluate platforms under the DSA.

Part 1 lays the groundwork by clarifying the DSA's regulatory architecture. Research questions **RQ-A** and **RQ-B** (Chapters 2 and 3) map the institutional foundations of the DSA and analyse how the regulation structures content moderation from a process-oriented perspective. Research question **RQ-C** extends this view by examining how the DSA institutionalises socio-technical complexity and what this implies for assessing platform compliance (Chapter 4). Together, these questions define the conceptual elements necessary for understanding what DSA compliance entails.

Part 2 focuses on identifying suitable methodological frameworks for assessing compliance. Research questions **RQ-D** and **RQ-E** explore two central methodological pillars: consistency analysis across DSA reporting instruments (Chapter 5) and the role of sampling in systemic risk assessments and independent audits (Chapter 6). Research question **RQ-F** adds a domain-specific methodological contribution by developing a legal taxonomy and codebook for analysing political and electoral content under the DSA and related

instruments (Chapter 7). Research question **RQ-J** (Chapter 8) examines how large language models can support—and where they fall short in supporting—auditors and oversight actors in operationalising the assessment methodologies introduced in Chapters 6 and 7. Together, these chapters establish how DSA compliance can be methodologically measured.

Part 3 moves to empirical evaluation, examining what compliance looks like in practice. Research questions **RQ-G**, **RQ-I**, and **RQ-H** (Chapters 9, 13, and 14) analyse how three categories of systemic risk can be assessed in concrete settings: the dissemination of illegal content (Art. 34(1)(a) DSA), negative effects on fundamental rights (Art. 34(1)(b) DSA), and negative effects on the electoral process and civic discourse (Art. 34(1)(a) and (c) DSA). Together, these chapters provide a broad and in-depth empirical assessment of systemic risks under the DSA.

Chapters 10 and 11 address user-facing compliance mechanisms. Research question **RQ-K** examines how users and experts experience compliance with the design requirements of Art. 16 DSA for reporting interfaces for illegal content, while research question **RQ-L** analyses how AI assistance affects users' legal categorisation performance and what this implies for mitigation measures under Art. 35 DSA addressing risks under Art. 34(1)(a). Research question **RQ-M** (Chapter 12) focuses on the role of Trusted Flaggers as a newly institutionalised actor in DSA enforcement and compliance, particularly in addressing the systemic risk of disseminating illegal content under Art. 22 DSA. Together, these chapters provide empirical evidence of how DSA compliance is enacted in practice through interface design, human–AI decision-making, and novel enforcement roles.

Part 4 synthesises these contributions. Research question **RQ-N** (Chapter 15) examines how the DSA shapes platform governance and functions as an industry-shaping policy with effects on global regulatory standards. Research question **RQ-O** (Chapter 16) analyses what can be learned about DSA compliance from enforcement communications and actions during the first two years of the DSA's application. These chapters integrate the legal, methodological, and empirical findings into a consolidated account of what robust DSA compliance assessment requires in practice.

Taken together, these chapters create a modular scaffold for analyzing DSA compliance: defining its foundations, specifying how it can be assessed, demonstrating how it manifests empirically, and outlining broader implications for platform governance and enforcement. This layered structure is essential for answering the overarching question of *How can online platforms be assessed to create DSA-compliant online spaces?*

RQ / Chapter	Short Name	Condensed Research Question
RQ-A / Ch. 2	DSA Foundations	What regulatory foundations, obligations, and institutional mechanisms does the DSA establish for platform governance?
RQ-B / Ch. 3	Content Moderation	How does the DSA structure content moderation processes and define rights, obligations, and accountability for involved actors?
RQ-C / Ch. 4	Complexity of DSA Assessment	How does the DSA institutionalise socio-technical complexity, and how do reporting artefacts translate platform practice into compliance claims?
RQ-D / Ch. 5	Consistency	How can consistency be assessed across DSA reporting instruments, and how does this support oversight and auditing?
RQ-E / Ch. 6	Systemic Risk Assessment Methodology	Which sampling methods matter for DSA systemic risk assessments and independent audits, and how do VLOPs and VLOSEs apply them in practice?
RQ-F / Ch. 7	Creating Legal Taxonomies	What legal taxonomy and codebook enable the systematic assessment of compliance for political and electoral content under the DSA, TTPA, and G-E-DSA?
RQ-J / Ch. 8	LLM Support for Audits	How can—and where do they fail—large language models support auditors and oversight actors in systemic risk evaluations under the DSA?
RQ-G / Ch. 9	Dissemination of Illegal Content	How do platforms interpret and operationalise legal standards when assessing and mitigating the systemic risk of disseminating illegal content (Art. 34(1)(a) DSA)?
RQ-K / Ch. 10	Reporting Interface Compliance	How do users and experts experience compliance with Art. 16 DSA design requirements for reporting interfaces for illegal content, and how does interface design shape usability?
RQ-L / Ch. 11	AI Assistance and Legal Categorisation	How does AI assistance affect users' legal categorisation performance, and what does this imply for mitigation measures under Art. 35 DSA addressing Art. 34(1)(a) risks?
RQ-M / Ch. 12	Trusted Flaggers	What role do Trusted Flaggers (Art. 22 DSA) play as a novel actor in enforcement and compliance, particularly in addressing the systemic risk of disseminating illegal content?
RQ-I / Ch. 13	Fundamental Rights Risks	How can systemic risks to fundamental rights under Art. 34(1)(b) DSA be operationalised into assessable categories?
RQ-H / Ch. 14	Electoral Process Risks	What does a concrete DSA-aligned platform risk analysis of electoral-process and civic-discourse risks look like in practice (Art. 34(1)(a),(c) DSA)?
RQ-N / Ch. 15	Governance and Global Standards	How does the DSA shape platform governance and function as an industry-shaping policy, including effects on global regulatory standards?
RQ-O / Ch. 16	Enforcement Evidence	What can be learned about DSA compliance from enforcement communications and actions during the first two years of the DSA's application?

Table 1.2: Condensed overview of research questions per chapter

1.5. APPROACH

This thesis brings together methodologies from multiple disciplines to propose new ways of assessing compliance under the Digital Services Act (DSA). The project integrates these methodological building blocks in a modular, chapter-linked manner, resulting in an interdisciplinary mixed-methods framework where each method is deployed to answer a specific research question and then carried forward into later empirical applications. The research moves from establishing the legal and conceptual foundations of DSA obligations through doctrinal analysis and conceptual synthesis, to developing evaluative tools such as cross-artefact consistency analysis, sampling strategies combined with structured content analysis for systemic risk assessment, codebook-based analysis of enforcement and reporting materials, and LLM-supported approaches for scaling audit tasks. It then validates these approaches through empirical work, including platform- and jurisdiction-specific case studies, expert interviews and workshops, legal heuristic walkthroughs and expert artefact evaluations of reporting and complaint interfaces, and controlled experiments and user surveys testing user-facing compliance and AI-supported reporting workflows. By aligning the dissertation's structure with Tables 1.3 and 1.5, the thesis makes explicit how each chapter contributes to the overall argument by specifying its placement in the dissertation, its research focus, and its publication status, thereby clarifying the cumulative logic of the project and its contribution to DSA compliance scholarship.

Table 1.3: Dissertation chapters, research focus, and publication status/outlet (Parts I–II).

Part	Chapter	What it researches	Publication Venue
I. Conceptual elements	Ch. 2 The DSA: Rules, Rights, and Risks	DSA governance architecture: obligations, user/procedural rights, institutional actors, and systemic-risk logic.	Published: Book chapter in "The Digital Decade: How the EU Shapes Digitalisation Research" [658]
	Ch. 3 Content Moderation and User Engagement	Content moderation as a socio-legal process under the DSA: stages, roles, and accountability interfaces.	Accepted: Book on "Automated content moderation: power, law and the role of human decisions"
	Ch. 4 Wicked by Design	Institutionalised complexity and interdependent artefacts; limits of simplification for auditability.	Published: IRIS conference 2026 (Internationales Recht-informatik Symposium) [654]
II. Methods & frameworks	Ch. 5 Holistic compliance via consistency	Consistency method for evaluating alignment across DSA artefacts, illustrated through a platform case study.	Published ACM Journal for Responsible Computing [666]
	Ch. 6 Systemic risk & audit methodologies	Mixed-method design for Arts. 34–37 assessments/audits: sampling, indicators, evidentiary standards.	Under Review (Major revisions): ACM Journal for Responsible Computing
	Ch. 7 Legal codebook: political/electoral content	Operationalising DSA + political-advertising regimes into a codebook for electoral/systemic risk analysis.	Published: IRIS conference 2025 (Internationales Recht-informatik Symposium) [660]
	Ch. 8 Scaling up auditing with LLMs	LLM-assisted auditing workflows: benefits, limits, safeguards, and accountability implications.	Published & Best Paper Award: CHIWORK '25 [662]

Table 1.5: Dissertation chapters, research focus, and publication status/outlet (Parts III–IV).

Part	Chapter	What it researches	Publication Venue
III. Empirical evaluation	Ch. 9 Auditing illegal-content decisions	Comparing platform decisions against legally coded ground truth; explaining divergence patterns.	Published: Journal CLSR (Computer Law Security Review) [742]
	Ch. 10 Reporting mechanisms: user pathways & design	Evaluating notice-and-action usability (Art. 16): access, friction, clarity, procedural affordances.	Published: ACM FAccT 25' (Fairness, Accountability and Transparency) [668]
	Ch. 11 LLMs as support for filing reports	Survey/experiment on AI-assisted reporting (Revise and Resubmit): quality, user agency, and explanation-of-reasons.	Published: ACM FAccT 26' (Fairness, Accountability and Transparency) [657]
	Ch. 12 Trusted Flaggers in practice	How Trusted Flaggers work with platforms/DSCs: incentives, frictions, and evidence practices.	Under Review: Journal for Cyber Security [656]
	Ch. 13 Fundamental-rights impact assessment	Operationalising Art. 34(1)(b) into assessable categories and structured assessment outputs.	Under Review: ACM FAccT 27' (Fairness, Accountability and Transparency)
	Ch. 14 Electoral-process systemic risks	Applying the electoral-risk codebook and audit approach to a concrete case with measurable indicators.	Published & Best Paper Award: HICSS 23' (Hawaii International Conference on System Sciences) [444]
IV. Synthesis & contribution	Ch. 15 Regulation as industry-shaping policy	How EU digital regulation reshapes compliance work, incentives, and accountability infrastructures.	Published: Special Issue Network Law Review [653]
	Ch. 16 Regulating through evidence (incl. X fine vignette)	DSA enforcement as evidence work: translating obligations into auditable traces and remediation routines.	Under Review: ACM FAccT 27' (Fairness, Accountability and Transparency)
	Ch. 17 Conclusion	Integrated findings, contributions, limitations, and research agenda.	Not separately published

1.6. OUTLINE AND CONTRIBUTION

This dissertation advances debates on Digital Services Act (DSA) compliance by developing an interdisciplinary scaffold for assessing multiple layers of compliance through integrated conceptual analysis, methodological development, and empirical investigation. Across four parts, the chapters collectively address the overarching question of how very large online platforms can be assessed to support DSA-compliant online environments. Together, they move from clarifying the DSA's regulatory foundations, to developing tools for evaluation, to applying these tools in empirical studies, and finally to synthesising implications for platform governance.

Part 1 establishes the conceptual groundwork by unpacking the DSA's legal architecture, governance logic, and socio-technical challenges. It reframes DSA compliance not as a binary legal condition but as a multi-layered, processual, socio-technical phenomenon, distinguishing between regulatory intent, organisational interpretation, technical implementation, and user-experienced outcomes.

Part 2 develops assessment frameworks that enable practical evaluation of DSA compliance. It contributes reproducible interdisciplinary methods, including consistency analysis across DSA artefacts, approaches for evaluating systemic-risk assessment methodologies, legal codebooks, and scalable computational and LLM-supported techniques.

Part 3 applies these methods in empirical studies of real-world platform practices. It examines how platforms operationalise systemic-risk obligations, how users and expert actors such as Trusted Flaggers interact with reporting and redress mechanisms, and how compliance can be observed and evaluated across transparency reports, Statements of Reasons, and interface design.

Part 4 synthesises the findings and draws out implications for DSA governance and related EU regulatory frameworks. It translates abstract DSA duties into operational oversight questions and clarifies how systemic risk assessments, transparency obligations, and user-redress mechanisms can be evaluated in practice, supporting regulators, auditors, researchers, and civil society in enforcement and accountability work.

1.7. RESEARCH SIGNIFICANCE AND NOVELTY

The significance of this research lies in its intervention into a rapidly evolving, and increasingly high-stakes, regulatory settlement. As the Digital Services Act (DSA) moves decisively from rule-making to *active enforcement*, illustrated by the European Commission's recent enforcement practice discussed in Chapter 16, compliance has become a matter of institutional capacity, evidentiary robustness, and democratic legitimacy rather than abstract legal aspiration.

Against this backdrop, the dissertation positions methodological standard-setting as a core governance challenge: without shared evaluation logics, regulators, auditors, and civil society are left to contest compliance through partial signals, incomparable artefacts, and asymmetries of information.

The dissertation delivers robust empirical evidence on how the DSA's user-facing and platform-facing design obligations operate in practice—across notice-and-action mechanisms, Statement-of-Reason repositories, and the interfaces used by ordinary users, civil society actors, and Trusted Flaggers.

This produces an empirically anchored foundation for regulatory interpretation, audit practice, and enforceable compliance assessment, and equips both public authorities and civil society watchdogs like Trusted Flaggers with a clearer understanding of where platform governance succeeds, fails, or becomes strategically opaque.

On example of such a contribution can be found in Chapter 10 that links directly to the two platforms (Meta/Facebook and TikTok) who were found preliminary in breach with the DSA's obligations for designing reporting mechanisms under Art 16. Chapter 10 therefore provides direct links to ongoing enforcement of the DSA and introduces novel methodologies next to empirical evidence to study compliance of DSA compliant interface design.

The research makes a substantive—and practically engaged—contribution to ongoing regulatory and academic debates. It does so not only through scholarly analysis, but through direct participation in the interpretive infrastructure of EU digital governance. In particular, via co-authored feedback submitted to the European Data Protection Board on the interplay between the DSA and the GDPR—advancing concrete interpretations of Articles 16, 17, 25, 28, and 34 DSA—this dissertation helps to clarify how user-protection duties, data protection principles, and platform design obligations should be read together in enforcement and oversight practice.¹ The intervention informs wider debates on the governance of reporting infrastructures: how notice-and-action and complaint pathways can be made usable without becoming extractive; how dark patterns and minors' rights reshape the legality of interface choices; and how data minimisation and purpose limitation constrain the design and operation of reporting mechanisms for illegal content.

The author of this thesis also contributed, in collaboration with the Commission, to developing the Statement of Reasons database and its underlying classification system [6, 748]. Within the same project, the author contributed to the Commission's work on transparency-reporting rules under the DSA [230]. This collaboration informed Chapter 5 by connecting the DSA's different transparency-reporting mechanisms within a more holistic approach to compliance evaluation.

The thesis bridges disciplinary silos by targeting different core publication venues (See Tables 1.3 and 1.5). By approaching understanding DSA compliance as a problem that needs an interdisciplinary mindset and approach, this thesis successfully published work in top-tier venues for computer science (FAccT, HICSS, CHIWork) and law (CSLR, IRIS). By strategically tackling different disciplines this thesis aims at creating awareness within different communities for the DSA. Several publications within this thesis have received prizes (best paper wards) mainly in computer science venues (HICSS and CHIWork). These acts of recognition further strengthen the relevance that this work brings to other disciplines and speaks for the rigor of how the empirical legal studies are conducted that go beyond sole legal doctrinal analysis.

¹ See: Feedback to the EDPB Guidelines 3/2025 on the interplay between the DSA and the GDPR, co-authored contribution by the author [640].

The aim of creating awareness for digital regulation with an even wider regulatory focus was further strengthened by the publication of the book "The Digital Decade: How the EU shapes digitalisation research" [324] which is collecting key EU regulations like the DSA, the AI Act, and many more, co-edited by the author of this thesis.² The editorial project fosters interdisciplinary research through and with the different digital regulations. Chapter 2 of this thesis is providing the conceptual basis for understanding the key provisions and inner working of the DSA for an interdisciplinary audience.

Taken together, the dissertation establishes methodological benchmarks for the next generation of DSA compliance evaluation and delivers one of the first genuinely operational frameworks capable of supporting the enforcement and accountability. Its originality lies in fusing a deliberately plural methodological palette from legal analysis, HCI and legal design, organisational studies, and computational approaches into an evaluation architecture: one that treats compliance not as a box-ticking exercise, but as a socio-technical governance problem that must be measured, explained, and made legible across actors and artefacts. In doing so, the dissertation strengthens the conditions for assessment within the DSA, evidence and enforcement that are not only declared in law, but made actionable through verifiable practices and auditable methods.

²The edited volume *Digital Decade: How the EU Shapes Digitalisation Research* covers (via dedicated chapters) the final EU AI Act, the Digital Services Act (Regulation (EU) 2022/2065), the Digital Markets Act (Regulation (EU) 2022/1925), the Terrorist Content Online Regulation (Regulation (EU) 2021/784), the Political Advertising Regulation, the Directive on Copyright in the Digital Single Market (Directive (EU) 2019/790), the European Media Freedom Act, the Data Governance Act (Regulation (EU) 2022/868), the Open Data Directive (Directive (EU) 2019/1024), the Data Act, EU data protection law (GDPR; Regulation (EU) 2016/679), the European Health Data Space, the Cyber Resilience Act (CRA), and the NIS 2 Directive (Directive (EU) 2022/2555).

I

CONCEPTUAL ELEMENTS OF DSA COMPLIANCE

2

AN OVERVIEW

"[...]once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversations?'

Lewis Carroll, *Alice's Adventures in Wonderland*, "Down the Rabbit-Hole", p. 1

The Digital Services Act (DSA) represents a landmark legislative framework in the European Union, aimed at regulating online platforms, enhancing transparency, and mitigating systemic risks associated with digital services. The Act aligns with broader EU regulatory efforts, including the General Data Protection Regulation (GDPR) and the Artificial Intelligence (AI) Act, positioning it as a cornerstone of digital governance.

A key objective is to create a harmonized internal market that prevents regulatory fragmentation while ensuring consumer protection and fundamental rights. The DSA introduces obligations for intermediary services, including very large online platforms (VLOPs) and very large online search engines (VLOSEs). Moreover, the regulation mandates due diligence measures such as transparency reporting, algorithmic accountability, and user rights protections. Transparency mechanisms include the publication of terms and conditions databases, the Statement of Reasons repository, and advertising libraries. Moreover, the DSA enforces structured risk assessment and mitigation strategies, particularly for systemic risks such as illegal content dissemination, disinformation, and fundamental rights violations.

A core component of the DSA is its approach to content moderation, introducing user empowerment mechanisms such as Trusted Flaggers, internal complaint-handling systems, and out-of-court dispute resolution bodies. Additionally, the Regulation includes crisis response provisions enabling swift intervention by the European Commission in

This chapter is based on: Sekwenz, M.-T., & Gsenger, R. (2025). "The Digital Services Act: Online Risks, Transparency and Data Access." In *Digital Decade: How the EU Shapes Digitalisation Research*, pp. 115–140. Nomos. doi:10.5771/9783748943990-115.

extraordinary circumstances. To ensure compliance, the DSA establishes independent audit requirements and risk-based oversight mechanisms, reinforcing platform accountability. This Chapter aims to give an overview and comprehensive introduction to these provisions.

2.1. INTRODUCTION

The most important European legislative act currently regulating (large) online platforms and their content moderation systems is the Digital Services Act (DSA) [246]. The DSA is the legal update of the E-Commerce Directive [242] of 2000 and expands the original scope by going beyond the regulation only of individual rights [403, p. 552]. The wording of the Directive did not take into account the importance that social networks and online marketplaces would play in daily life as the digital economy has developed into a platform economy [623, p. 80]; furthermore, the scale of the services and the multiplication of various intermediaries needed to be considered. Differing legislative efforts of Member States led to the fragmentation of legal regulations and challenges regarding the enforcement of services that operate across borders [651, p. 233]. Moreover, the important role of algorithmic decision-making [198], disinformation [204, 384], and illegal content [183, 444] has become more evident.

The general aim of the DSA is a “safe, predictable and trusted online environment” (Art. 1 (1) DSA) through the realisation of an internal European market. Since platforms operate transnationally and Member States may have their own rules, there is a risk that the market might fragment, as occurred with the regulatory attempts of Germany [15] and Austria [616]. An internal market would enable companies to benefit from unification and allow them to innovate in a harmonised environment. Moreover, new markets can be accessed and consumers overall would have more choices [361, p. 33].

In December 2020, the DSA was presented in conjunction with the Digital Markets Act (DMA) [240], which aims to ensure a fair platform economy with a functioning internal market [538, p. 74]. In the first Chapter, the DSA determines its subject matter (Art. 1) and scope (Art. 2) and provides definitions (Art. 3). Chapter II focuses on the liability of providers and intermediary services, and Chapter III on due diligence and transparency. Chapter III consists of sections listing the specifications concerning the obligations of different intermediary services, such as online platforms or very large online platforms (VLOPs) and search engines (VLOSEs) (as defined by Art. 33(1) DSA).

Chapter IV specifies implementation, cooperation, penalties and enforcement and includes specificities about Digital Service Coordinators (DSCs) and other relevant authorities and competencies, such as the European Board for Digital Services. The Board acts as an independent advisory body for the DSCs (Arts. 61–64 DSA); DSCs are the regulatory body situated in each Member State. Member States choose these “competent authorities” (Art. 49, (2)), and the DSCs are subsequently “responsible for all matters relating to supervision and enforcement” of the DSA in the respective Member State (Art. 49 (2)).

The DSA defines its scope in Art. 2 (1), including intermediary services (Art. 3 (g) DSA), hosting services (Art. 3 (g) (iii) DSA), online platforms (Art. 3 (i) DSA), VLOPs (Art. 33(1) DSA) and VLOSEs (Art. 3 (j) DSA) that offer their services inside the European Union.

Intermediary services refers to three types of “information society services” (lit. g): first, services that are “mere conduit” (i), transmitting information by a recipient of the service. Second, “a ‘caching’ service” (ii) that includes the transmission and storage of information and third, “a ‘hosting’ service, consisting of the storage of information provided by, and the request of, a recipient of the service” (iii). An online platform is a hosting service that, “at the request of a recipient of the service, stores and disseminates

information to the public” (Art. 3 lit. i). Online search engines are also intermediary services that “allow[s] users to input queries in order to perform searches of, in principle, all websites, or all websites in a particular language, on the basis of a query on any subject in the form of a keyword, voice request, phrase or other input, and returns results in any format in which information related to the requested content can be found” (lit. j).

VLOPs and VLOSEs are defined as services and intermediaries operating in the European Union that are reported to have more than 45 million monthly active users (Art. 33 (1)). The number of users should cover at least 10% of the EU population. The number is reported by the platforms themselves, and they must provide an updated number of monthly active users “at least once every six months” (Art. 24 (2)) or “without undue delay” (Art. 24 (3)) upon receiving a request from the European Commission. The European Commission first designated platforms and search engines considered to be very large in April 2023; now, that list is frequently updated and includes platforms such as AliExpress, Google Search, Facebook, TikTok, Meta and Amazon [142]. This limit cannot be bypassed by European nation-states and no platforms with fewer monthly active users can be obliged to adhere to the risk identification and mitigation requirements [403, p. 533]. The European Commission assumes that intermediaries of this size have significant influence on the internal market and that they have sufficient resources to adhere to the Regulation [403, p. 541].

The DSA includes natural or legal persons who have the possibility of using a service (Art. 3 lit. b); such a person is referred to as a “recipient” (Art. 3 lit. b). However, the DSA also refers to persons as “users” throughout the text, and this term is preferred in this chapter as it is more commonly used. The DSA presents a legal definition that describes active users or recipients. This definition, however, is in the main text of the DSA, not in the Recitals, emphasizing the importance of the differentiation. An active user of an online platform can be classified in two ways: they can “request [...] the online platform [...] host information” (Art. 3 lit. p), meaning, for instance, uploading user-generated content to platforms or commenting on other content [403, p. 542]; or, an active recipient is a person “exposed to information hosted by the online platform and disseminated through its online interface” (Art. 3 lit. p). Therefore, receiving or consuming content on platforms without contributing or uploading content is sufficient to be considered an active user. Participation is confirmed independent of registration and includes the consumption of any content (whether visual or audio) that starts without any user involvement as soon as a website opens [403, p. 542]. An active recipient of a search engine “has submitted a query to an online search engine and been exposed to information indexed and presented on its online interface” (Art. 3 lit. q DSA). A query includes the input of terms, and if the query is completed automatically and the recipient presses enter, the input counts as active use [403, p. 542].

Intermediaries profit from the Good Samaritan Clause that limits liability and determines that they are not responsible for any content shared that might be illegal [322, p. 4]. Therefore, they are exempt from liability under certain conditions [361, p. 32f] and, moreover, are not required to participate in any monitoring activities (Art. 8 DSA).

Complementing the DMA, the DSA aims to enable citizens to exercise their fundamental rights in a safe online environment [538, p. 75]. The DSA aims to reduce risks of VLOPs and VLOSEs by establishing clear rules and transparency. In addition to a trustworthy

online environment, the DSA seeks to ensure that fundamental rights are protected and consumer protections strengthened. Lastly, the DSA aims to establish legal certainty [361, p. 33].

Similar to other European regulations, such as the General Data Protection Regulation, the AI Act, and the NIS 2 Directive, the DSA includes risk-based elements in its regulatory structure [783]. Risk detection, analysis, and evaluation are therefore part of the legislative framework, alongside internal measures, external audits, transparency requirements, and access mechanisms subject to scrutiny by the European Commission, Digital Services Coordinators, and researchers. The DSA practises forms of “enforced self-regulation” [403, p. 532], including systemic-risk assessments under Art. 34, for which platforms are initially responsible. Platform compliance is nevertheless subject to review and, ultimately, regulatory enforcement. The European Commission supervises VLOPs’ and VLOSEs’ compliance and may impose fines not exceeding 6% of a provider’s annual worldwide turnover (Art. 74(1) DSA). Under Art. 87 DSA, the Commission may adopt delegated acts that specify aspects of the DSA’s implementation, including audits [226]; transparency-reporting templates are specified through an implementing act [230].

In this Chapter, we will discuss the most important provisions and their consequences, including the due diligence of VLOPs and VLOSEs, including transparency mechanisms (section 2), user rights and processes (section 3), risk assessment, risk mitigation, and audits (section 4).

2.2. SEE-THROUGH REGULATION? NOVEL TRANSPARENCY MECHANISMS IN THE DSA

The DSA creates several new mechanisms that provide novel insights into the day-to-day decisions taken on platforms. Transparency mechanisms centrally include reports and databases to inform stakeholders. This text will therefore highlight the main tools of transparency by focusing on transparency reports (Art. 15, 24 and 42 DSA), the Terms and Conditions database (Art. 14 DSA), the Statement of Reasons (Art. 17 DSA) and the Ad Library, also referred to as the Ad Repository (Art. 39 DSA). Additional means of transparency can be found in the rules regarding recommendation systems (Art. 27 DSA), parameters on targeted advertising (Art. 26 DSA) and in the link between the Code of Practice of Disinformation and the DSA as a Code of Conduct (Art. 45–47 DSA) [402].

Flyverbom [266, p. 110–112] defined transparency as a complex process connected to the development, interpretation and aggregation of publishing information aimed at enhancing accountability, openness and trust within a certain period. The DSA itself does not form its own definition of transparency [436, p. 117] but rather reflects on it in several passages, including in Recital 49: “To ensure an adequate level of transparency and accountability, providers of intermediary services should make publicly available an annual report in a machine-readable format, in accordance with the harmonised requirements contained in this Regulation.” Such a machine-readable form of transparency could also enhance the automatism of checks and balances in an empirically-based accountability regime [547]. Conversely, Kusters and Gstrein [436] highlight the importance of the audience within the transparency regime and differentiate transparency into three layers: “The first layer of transparency involves the disclosure of information. The second

layer consists of ensuring that the information disclosed is also understandable to the broader public. Lastly, a third layer of transparency includes tailoring the explanation of information to the different types of users of the platform” [436, p. 130]. According to their case study of one VLOP, the DSA contributes to the first two layers through, for example, the provision of information in transparency reports, and to the second layer through the offer of a dashboard for the Statement of Reasons [166]; however, they found that the DSA was still lacking in the third layer of transparency. Ideally, by being able to interlock several different control mechanisms, the forms of transparency that the DSA creates can form a more solid understanding of meaningful, accountable and consistent transparency regimes [665].

2.2.1. A HARMONISED FORM OF REPORTING THROUGH DSA TRANSPARENCY REPORTS?

The DSA’s new rules on transparency reporting can be seen as a predecessor to the provisions that frame reporting under the German NetzDG or the Austrian KoPiG [40, 756, p. 627]. According to the DSA, platforms have different reporting obligations to be disclosed in an annual report – or for VLOPs and VLOSEs, in biennial reporting intervals – and such reports must be machine-readable (Art. 15(1) DSA). According to Art. 9 and 10 DSA, transparency reports must include information about orders from public authorities (for example, the police in a Member State), numbers about illegal content or median-time spans of action in response to such notices (Art. 15(1) (a) DSA).

Details provided in transparency reports include data concerning flags received from user-reporting (Art. 16 DSA) describing details of violation reasons, reports from Trusted Flaggers (who report to platforms about illegal content with increased flagging priority, see Art. 22 DSA), the moderation action set (for example, deletion or deplatforming), the automated means included in the moderation process (for instance, the use of Artificial Intelligence (AI) for detecting illegal content) and aspects of reaction time (Art. 15 (1) (b) DSA). Furthermore, details must be included on the specific purpose of the automated means used in the process, their accuracy and the possible error rate of tools like AI (Art. 15(1) (c) DSA). Article 15(1) (d) DSA specifies reporting details on the internal complaint-handling system according to Art. 20 DSA. This mechanism should enable users to question content moderation actions on platforms. The provisions of Art. 24 DSA (see Recital 65 DSA) only apply to online platforms, VLOPs and VLOSEs; these include paragraphs on the out-of-court dispute settlements (Art. 21 DSA), including the number of disputes received, the median time needed to form a decision or the decisions taken in such cases (Art. 24 (1) (a) DSA). In addition, information about malicious user behaviour, such as deplatforming [423], must be provided according to Art. 23 DSA, for example, details about the reason for suspension (Art. 24 (1) (b) DSA).

Recital 100 opens the scope for Art. 42 DSA, under which “additional transparency requirements should apply specifically to [VLOPs and VLOSEs]” such as biannual reporting obligations. Such platforms must report on the human resources used in the process of content moderation, including details about language skills, educational measures, training or support (Art. 42 (2) (a–b) DSA). Furthermore, Art. 42 DSA requires the inclusion of qualitative information – broken down to Member State levels (Art. 42 (2) (c) DSA – about the means of content moderation, such as details about the training of content

moderators or the educational measures provided to them.

2.2.2. A PLACE FOR ALL PLATFORM CONTRACTS – THE TERMS AND CONDITIONS DATABASE

According to the DSA, contractual rules governing online behaviour – found in the Community Standards of a platform – are to be provided within the terms and conditions, which are defined in Art. 3 (u) DSA as “all clauses, irrespective of their name or form, which govern the contractual relationship between the provider of intermediary services [the platform] and the recipients of the service [the user]”. These and other contractual rules for VLOPs and VLOSEs should be provided in the official languages of all Member State platforms that provide their services and include opt-out details addressed in the generalised contract according to Recital 48 DSA and Art. 14 (6) DSA.

Terms and conditions not only include norms and procedures but also the “measures, and tools” used in content moderation (Art. 14, 19 DSA). Since terms and conditions describe how to behave on platforms, these contractual amendments, also referred to as community standards or netiquette are a flexible way to adapt frameworks to new challenges, such as the COVID-19 pandemic or wars and conflicts [424] (European Commission). Article 14 DSA requires that users be informed about significant changes (2) and that information is to be provided in a machine-readable format (5). Furthermore, information for children is explicitly mentioned (3), and enforcement has to be in line with fundamental rights (4). Since February 2024, platforms have uploaded their terms and conditions and changes to a website that informs users about the current version that platforms use for content moderation [166]. For the first time, this organised database of contractual rules provides the reader with updates on new clauses, actions or exemptions.

2.2.3. CONTENT-MODERATION DECISIONS IN THE STATEMENT OF REASONS DATABASE

The Statement of Reasons database is a new measure of transparency in the DSA regulated under Art. 17 DSA. According to Recital 66, “to ensure transparency and to enable scrutiny over the content moderation decisions of the providers of online platforms and monitoring the spread of illegal content online, the Commission should maintain and publish a database which contains the decisions and statements of reasons of the providers of online platforms when they remove or otherwise restrict availability of and access to information”. This database therefore captures content moderation decisions in cases of a violation of the terms and conditions (Art. 14 DSA) or the law of a Member State (Art. 3 (h) DSA), similar to the Lumen Database, which was created at Harvard University to capture insight into the moderation process (Lumen Database, 2024). These captured content moderation actions either affect the visibility of content (Art. 17(1) (a) DSA), monetary elements (Art. 17(1) (b) DSA), suspension of the service (Art. 17(1) (c) DSA) or the suspension of an account (Art. 17(1) (d) DSA). Information in the so-called transparency database also includes content moderation decisions such as the facts upon which a decision is based, the circumstances of a case, the source of information (e.g. flagging) or the identity of the notifier (e.g. a Trusted Flagger). Additionally, information about the automated means in the process should be provided as a reference to legal or contractual grounds, as well as information about user rights (e.g. the internal complaint-handling

system according to Art. 20 DSA or out-of-court dispute settlements according to Art. 21 DSA). Since the general aim of increasing transparency is welcomed by the community, the accuracy, depth of information and completeness have been critiqued by researchers evaluating the meaningfulness of platforms' reporting practices [205, 413, 711]. Such a database is a novum to the world of online governance and opens a path for increased research on platforms to be conducted. The database includes an individual ID for each decision that can be linked to thorough investigations in conjunction with researcher data access or independent audits, and it can also link to the transparency reports of a platform to control for cross-transparency mechanisms [665].

AD LIBRARY

Another key database the DSA creates is the advertising repository, the use of which, according to Art. 39 DSA, is mandatory for VLOPs and VLOSEs [209, 389]. This database provides users with a publicly available search function and API. According to Art. 39(2) (a) DSA, the database should include information about the advertisement (name of the product/service/brand and the subject of the ad, e.g. political advertising). Furthermore, the person on whose behalf the ad is presented has to be disclosed (b–c), in addition to information about the duration of the ad presentation and display (d), targeted and untargeted groups (e–f) and the number of users for whom the ad has been displayed (g). Such information, however, should not be included in the database if the content was classified as illegal under the law of a Member State (Art. 39(3) DSA). Additionally, for the DSA, the upcoming Directive on Transparency and Targeting of Political Advertising will create a new centralised database for this specific type of online advertising at the European level [24, see Art. 13].

2.2.4. DATA ACCESS FOR RESEARCHERS

Article 40 DSA holds specific interest for researchers investigating platforms due to its provision of data access to the DSC or the Commission (Art. 40 (1)). The first part of the Article (1–3) regulates access by public authorities, whereas the second part (4–6 and 8–11) focuses on researcher access. The following section will focus on the second part of Art. 40 due to its relevance for researchers. Research access is provided for the purpose of investigating systematic online risks in order to reduce information asymmetries and support risk mitigation [403, p. 639]. Therefore, access should be constrained to data concerning the provisions of the DSA, especially understanding and identifying systemic risks according to Art. 35 (Art. 40 (4)).

The DSC can request that VLOPs and VLOSEs “explain the design, the logic, the functioning and the testing of their algorithmic systems” (Art. 40 (3)). Platform providers need to adhere to these requests “within a reasonable period” (Art. 40 (4)); however, platforms can request an amendment to the data access request within 15 days if they do not have access to that data or if the security of their service and trade secrets are endangered (Art. 40 (5) lit. a–b). The DSC will grant researchers requesting data access the “status of ‘vetted researchers’ for the specific research” (Art. 40 (8)). These researchers need to fulfil certain requirements as specified in Art. 40 (8) (lit. a–g): researchers must be part of a research organisation (lit. a), which is defined as “a university, including its libraries, a research institute or any other entity, the primary goal of which is to conduct

scientific research or to carry out educational activities involving also the conduct of scientific research” (Copyright Directive Art. 2 (1)). The organisation must be non-profit (lit. a) and operate in the public interest (lit. b). Additionally, researchers must work independently and not for commercial interests (lit. b), disclose their research funding (lit. c), protect personal data and implement measures to guarantee data security (lit. d). Furthermore, they must prove that data access is necessary for their research, that it is proportionate and will contribute to the understanding of risk mechanisms (lit. e, f). Finally, researchers must make their results publicly available (lit. g).

Research needs to “contribute to the detection, identification and understanding of systematic risks” (Art. 40(12)). According to [378], Art. 40(12) provides two functions: it protects providers against unjust access and minimises technical restrictions of data access for researchers. He argues that scraping should remain central for research aside from API access.

2.3. ON USER RIGHTS, PROCESSES AND INSTITUTIONALISED FLAGGING ENTITIES

The DSA introduces a new system of rights and procedures for users, platforms, and authorities in content moderation processes. This section highlights the main rights users gain under the DSA, describes the logic of internal complaint-handling, the structure of out-of-court dispute settlement bodies, and the creation of “Trusted Flaggers” as institutionalised entities for illegal content flagging.

2.3.1. USER RIGHTS AND EMPOWERMENT

A core innovation of the DSA is its attempt to strengthen the rights and empowerment of users when interacting with online platforms and to introduce remedies for situations where rights are violated [465]. Users have a right to “receive clear, accessible and meaningful information” regarding terms and conditions (Art. 14), and the reasoning behind any content moderation action taken against their content or account (Art. 17), including notification of the possibility to contest decisions.

The DSA also mandates platforms to provide an internal complaint-handling system (Art. 20), which must be “easy to access and user-friendly,” free of charge, and available for at least six months after the original decision. This process must allow users to challenge platform decisions related to content removal, account suspension, or any restriction of service. Furthermore, platforms are required to inform users about the outcome of the complaint, including the reasoning behind upholding or reversing the moderation decision.

2.3.2. INTERNAL COMPLAINT-HANDLING AND OUT-OF-COURT DISPUTE SETTLEMENT

If the outcome of the internal complaint-handling is unsatisfactory, users have the right to refer their case to an out-of-court dispute settlement body, as stipulated by Art. 21 DSA. These bodies must be certified by the relevant Digital Services Coordinator (DSC), function independently, and provide their services at a reasonable cost and within a reasonable time frame. Decisions by these bodies are not binding unless all parties have

agreed to be bound (Art. 21(3)), but the process must respect the rights of all parties and the principles of fairness, impartiality, and efficiency. The DSA requires platforms to inform users of this redress mechanism and to cooperate with the dispute settlement bodies.

2.3.3. TRUSTED FLAGGERS

Article 22 DSA introduces the concept of “Trusted Flaggers,” which are entities designated by the DSCs in each Member State based on their expertise, independence, and competence in identifying illegal content. Trusted Flaggers must act “with due diligence, accuracy and objectivity” and are subject to annual reporting and oversight. Platforms are required to process notices submitted by Trusted Flaggers “with priority and without delay” (Art. 22(6)). The aim is to institutionalize expertise in illegal content detection and ensure a high-quality, reliable flagging process that supports both users and regulators. The DSA also creates provisions for users or organizations whose Trusted Flagger status has been revoked, including the right to be heard and to appeal the decision (Art. 22(7)–(8)).

2.3.4. ADDITIONAL USER EMPOWERMENT MEASURES

Further empowerment provisions in the DSA include the right for users to select whether they want to receive recommendations based on profiling (Art. 38), requirements for platforms to provide meaningful explanations of the functioning of recommender systems (Art. 27), and transparency on advertising and sponsored content (Art. 26, 39). Special protections are included for minors, who must not be subject to profiling-based recommendations or advertising (Art. 28).

These mechanisms, collectively, aim to create a more balanced online environment, where users have clear rights, platforms have structured obligations, and regulatory oversight is robust and transparent.

2.4. IDENTIFYING AND MITIGATING SYSTEMIC RISKS FOR INTER-MEDIARIES

A key innovation of the DSA is the introduction of a risk-based regulatory approach, especially for Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs). These platforms must regularly assess, mitigate, and report on a range of systemic risks that may arise from the functioning and use of their services.

2.4.1. SYSTEMIC RISK ASSESSMENTS

Article 34 DSA requires VLOPs and VLOSEs to identify, analyse, and assess systemic risks stemming from the design, functioning, and use of their services. These risks are broadly defined and include:

- dissemination of illegal content;
- negative effects on the exercise of fundamental rights, such as freedom of expression, protection of personal data, non-discrimination, and the rights of children;

- negative consequences for civic discourse, electoral processes, and public security;
- actual or foreseeable negative effects on public health, minors, physical and mental well-being, or on gender-based violence.

The risk assessments must be conducted “at least once a year and prior to deploying new functionalities that may have a critical impact” (Art. 34(1)). Providers must take into account specific risks linked to the dissemination of illegal content, intentional manipulation of their service (including through automated exploitation or the creation of fake accounts), and emerging societal risks.

2.4.2. MITIGATION MEASURES

According to Article 35 DSA, based on their risk assessments, VLOPs and VLOSEs must implement “reasonable, proportionate and effective mitigation measures” tailored to the identified risks. Examples of such measures include:

- adapting content moderation processes or recommender system design;
- enhancing internal procedures for the detection and removal of illegal content;
- collaborating with Trusted Flaggers, researchers, and civil society organisations;
- limiting the display or amplification of certain content, especially when it concerns disinformation or the manipulation of electoral processes;
- strengthening transparency and user empowerment mechanisms.

Platforms must report on these mitigation measures in their transparency reports and cooperate with the European Commission and Digital Services Coordinators for oversight and enforcement (Art. 35(2)–(3)).

2.4.3. INDEPENDENT AUDITS AND ENHANCED SUPERVISION

VLOPs and VLOSEs are required to submit to independent annual audits assessing their compliance with the DSA, including the effectiveness of their risk assessments and mitigation strategies (Art. 37 DSA). These audits are conducted by external, accredited auditors and must result in a public audit report, except for information that could endanger public security or trade secrets. If significant shortcomings are found, platforms must submit an “audit implementation report” specifying the corrective actions they will take.

The European Commission is given enhanced supervisory powers for VLOPs and VLOSEs, including the right to request information, conduct inspections, and impose fines or other corrective measures (Arts. 67–72). National Digital Services Coordinators and the new European Board for Digital Services play key roles in cross-border supervision and enforcement.

2.4.4. CRISIS RESPONSE MECHANISM

Article 36 DSA introduces a specific mechanism for crisis situations, such as war, pandemic, or other emergencies that pose serious threats to public security or health. The

European Commission can require VLOPs and VLOSEs to assess whether and how their services contribute to the crisis and to implement specific mitigation measures within a set timeframe. These obligations are strictly time-limited and subject to judicial review.

2.4.5. DATA ACCESS FOR VETTED RESEARCHERS

To facilitate independent oversight and research, the DSA provides for data access for “vetted researchers” (see also Section 2.5). VLOPs and VLOSEs must make data available to these researchers for the purposes of detecting, identifying, and understanding systemic risks (Art. 40). Access is subject to safeguards concerning the protection of personal data, trade secrets, and security, and may only be granted to non-profit, public-interest researchers who meet strict independence and transparency criteria.

2.4.6. CONCLUSION OF RISK-BASED APPROACH

Through these systemic risk management obligations, the DSA aims to create a more resilient, accountable digital ecosystem. The Regulation empowers regulators, researchers, and civil society to monitor and challenge the activities of the largest platforms, while providing flexibility and proportionality for smaller actors.

2.5. CONCLUSION

The Digital Services Act introduces a groundbreaking regulatory framework that aims to enhance transparency, accountability and user protection across online platforms, with specific attention focused on Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs). This chapter has provided an overview of the DSA, one of the first efforts to regulate harmful online content and protect users’ fundamental rights online.

As discussed, the DSA’s emphasis on transparency is pivotal. The Regulation establishes multiple tools to ensure that platforms are open about their operations, including transparency reports (Art. 15, 24, 42), the Terms and Conditions database (Art. 14), the Statement of Reasons database (Art. 17) and the Ad Library (Art. 39). The novel transparency mechanisms for intermediary services include reports, online repositories (such as the Ad Library according to Art. 39) and Statements of Reason (Art. 17). Furthermore, the DSA provides rules for researchers to access platform data to research systemic risks (Art. 40).

The DSA aims to empower users through new roles and rights, including the Trusted Flaggers mechanism (Art. 22) and the internal complaint-handling system (Art. 20), which reflect the DSA’s aim to involve users more actively in content moderation processes by giving them the tools to flag illegal content and challenge platform decisions. Furthermore, the introduction of out-of-court dispute settlements (Art. 21) provides users with a structured and accessible way to seek redress when their rights have been infringed upon.

The DSA includes a crisis response mechanism (Art. 36), which allows the European Commission to require platforms to take specific steps during times of public crisis. A core innovation is the risk-based approach, requiring VLOPs and VLOSEs to assess, mitigate, and report on systemic risks in their operations and to cooperate with independent researchers and regulatory authorities.

The DSA's approach is characterized by a combination of enforced self-regulation, multi-level oversight, and user empowerment. This model aims to foster a safer, more predictable, and trusted online environment within the European Union, balancing the interests of users, platforms, and regulators. The real impact of the DSA will become clearer as implementation progresses and as platforms, users, and regulators put its provisions into practice.

3

THE PROCESS OF CONTENT MODERATION AND USER ENGAGEMENT: SHAPED BY THE DIGITAL SERVICES ACT

*Son of light, easily swayed, hearken and beware: hold firm to your bond! What you are, you
are only by contracts: limited and well defined is your power.
/ Lichtsohn du, leicht gefügter, hör' und hüte dich: Verträgen halte Treu! Was du bist, bist
du nur durch Verträge: bedungen ist, wohl bedacht deine Macht.*

Fasolt – Rheingold, Act I, Richard Wagner

This paper investigates the legal framework for content moderation introduced by the European Union's Digital Services Act (DSA). Content moderation—the process of identifying and addressing illegal or otherwise problematic content on online platforms—has traditionally been opaque and often relies on a combination of automated systems and human review. The DSA introduces a more transparent and regulated approach through notice-and-action mechanisms (Art. 16 DSA), internal complaint-handling systems (Art. 20 DSA), and out-of-court dispute settlement (Art. 21 DSA). These provisions redefine the responsibilities of online platforms and create new procedural rights for users. They also formalize institutional roles for Trusted Flaggers (Art. 22 DSA) and regulatory authorities. At the same time, legal formalization makes this multi-actor process more complex and creates operational challenges for platforms, users, expert entities, auditors, and regulators. Through doctrinal legal analysis, the paper maps DSA obligations onto distinct stages of content moderation and highlights the interplay between law, platform governance, and user protection. It aims to support regulators, civil society, platforms, and researchers in navigating and implementing this evolving framework.

¹This chapter has been accepted for publication in an edited volume organized by the Alexander von Humboldt Institute for Internet and Society (HIIG) under its “Human in the Loop” call.

3.1. INTRODUCTION

Content moderation has become a defining function of digital platform governance. As users increasingly rely on platforms like Facebook, YouTube, and TikTok to share, consume, and interact with information, questions about who controls online speech, and under what rules, have become more urgent. Platforms routinely moderate user content through automated systems and human review [186, 762], applying internal policies and legal obligations [580]. However, these moderation processes have long lacked transparency, consistency, adequate working conditions, and formal mechanisms for user recourse [195, 665, 668, 747].

The European Union's Digital Services Act (DSA), adopted in 2022 and fully applicable from 2024, introduces a new legal framework for content moderation [246]. Building on earlier national regulation in Germany and Austria, the DSA moves beyond voluntary self-regulation toward a structured system of legal obligations, procedural safeguards, and institutional oversight [106, 616]. These measures govern matters including terms and conditions, reporting mechanisms, recommender systems, online advertising, the submission of Statements of Reasons, systemic-risk assessments, and independent audits. The DSA thereby creates a multilayered moderation process that includes notice mechanisms, priority treatment for notices from Trusted Flaggers, transparency requirements, internal complaint-handling systems, and out-of-court dispute settlement [659].

The DSA provides a formal legal definition of content moderation and covers a broad range of actions, including content removal [536], demotion (reducing the reach of specific content), demonetisation [291], and account suspension [376]. Online platforms must submit qualifying moderation decisions to the Commission's Statement of Reasons database, thereby enabling greater scrutiny of how platforms moderate content under the DSA.

This paper asks: How does the Digital Services Act legally structure content moderation on online platforms, and what rights, obligations, and accountability mechanisms does it establish for the actors involved? Through doctrinal legal analysis, it maps relevant DSA provisions to stages in the moderation pathway and identifies the rights and duties of platforms, users, regulators, and third-party actors. These actors include independent auditors under Art. 37 DSA, who evaluate platforms' compliance with obligations that include identifying, assessing, and mitigating systemic risks under Arts. 34–35 DSA. By conceptualizing content moderation as a regulated process rather than a discretionary act, this analysis contributes to broader debates on the rule of law, accountability, and auditability in the digital public sphere.

3.1.1. BACKGROUND

The regulation of online content in the European Union has historically followed a light-touch approach. The cornerstone of earlier EU digital governance, the e-Commerce Directive of 2000, provided conditional liability exemptions for intermediary services [242]. This framework enabled the growth of the platform economy but left content moderation largely in the hands of private companies, broadly paralleling the intermediary-liability approach of Section 230 in the United States [720].

As digital platforms became more central to public discourse, their role as arbiters of online speech attracted increasing regulatory scrutiny. Concerns about hate speech,

disinformation, illegal content, and opaque moderation practices prompted calls for a more accountable and transparent system [200, 410, 514]. In the absence of binding EU-level standards, many platforms developed their own content policies, Community Standards, and moderation infrastructures [742], often guided by economic interests (e.g. moderation costs) and reputational concerns (e.g. the treatment of influencers or accounts with large followings) rather than public-law principles [653].

The DSA retains conditional liability exemptions for intermediary services (Arts. 4–6), building on Arts. 12–14 of the e-Commerce Directive, while Art. 8 DSA prohibits general monitoring obligations. It also imposes new procedural duties. Content moderation is no longer merely a discretionary business practice but a set of legally relevant acts that engage both platform duties and user rights [668]. The DSA's definition clarifies that content moderation affects the visibility and accessibility of information and therefore intersects with fundamental rights, including freedom of expression (Art. 11 of the Charter [248]; see also Recitals 3, 9, 22, 36, 40–41, 47, 51–52, 54, 63, 79, 81, and 86 and Arts. 14(4), 34(1)(b), and 35 DSA).

Importantly, the DSA acknowledges platforms' own terms and conditions (Art. 14 DSA), thereby regulating both compliance with law and private ordering through platform-defined moderation rules. Article 14(1) requires providers to disclose information about the policies, procedures, measures, and tools used for content moderation, including algorithmic decision-making, human review, and the rules governing internal complaint handling.

As part of the EU's broader Digital Services Package, the DSA seeks to harmonize rules for digital intermediaries while embedding fundamental rights and procedural fairness into platform operations, including through internal complaint-handling systems and out-of-court dispute-settlement bodies [245]. Rather than regulating only particular types of content, it also structures how content is moderated: providers must implement notice-and-action mechanisms (Art. 16), explain qualifying restrictions (Art. 17), and provide avenues for redress (Arts. 20–21). These duties serve the DSA's objective of establishing “harmonised rules for a safe, predictable and trusted online environment” (Art. 1).

One procedural safeguard is the requirement that mechanisms for notifying allegedly illegal content be easy to access and user-friendly [668]. Through these mechanisms, users and expert entities such as Trusted Flaggers alert providers to potentially illegal content [174]. Ordinary users are not part of the platform's moderation workforce, but their reporting behaviour can be encouraged or discouraged through interface design and incentives [438, 550]. Once a provider receives a notice or otherwise becomes aware of potentially problematic content, automated systems or human moderators may assess the content and impose a restriction, such as removing a post.

Content-moderation actions may affect content visibility, monetisation, service provision, or user accounts, as reflected in the Statement of Reasons database fields [1].

Article 3(t) DSA defines content moderation as activities undertaken by providers of intermediary services, whether automated or not, to detect, identify, and address illegal content or information incompatible with their terms and conditions. The definition covers measures affecting availability, visibility, accessibility, monetisation, service provision, and user accounts. Article 14 DSA also recognizes the role of platforms' own terms and conditions, leaving providers room to define contractual rules while subjecting their

Cluster	Attribute	Description	Categories	Ref.
Visibility Actions	decision_visibility	Restriction affecting visibility or accessibility of specific content.	Removal; Disabling access; Demotion; Age restriction; Interaction restriction; Labelling; Other	sec. 3.1
	decision_visibility_other	Description required if visibility restriction = "Other".	Platforms can define their visibility actions if not covered by a reason provided above	3.2
Monetary Actions	decision_monetary	Restriction affecting monetary payments.	Suspension; Termination; Other	sec. 3.3
	decision_monetary_other	Specification if monetary restriction = "Other".	Platforms can define their monetary actions if not covered by a reason provided above	sec. 3.4
Service Provision	decision_provision	Restriction on the provision of the service to a recipient.	Partial suspension; Total suspension; Partial termination; Total termination	sec. 3.5
Account Actions	decision_account	Restriction applied to the user's account.	Suspension; Termination	sec. 3.6

Table 3.2: Moderation actions and required Statement of Reasons database fields under Art. 17(3) DSA [1].

enforcement to transparency and fundamental-rights safeguards. Content moderation is therefore not merely a discretionary business practice but a set of legally relevant acts that engage platform duties and user rights.

Qualifying moderation actions must be submitted to the Statement of Reasons database and reported in aggregate through transparency reports. Article 15(1) requires disclosure of provider-initiated moderation measures and the use of automated tools, while Art. 15(1)(e) requires information about automated means, including their purposes, accuracy indicators, and possible error rates. For VLOPs and VLOSEs, Art. 42(2)(a) also requires information on the human resources dedicated to content moderation. Furthermore, platforms are obliged to assess systemic risks on their platforms [661] and propose mitigation measures directly naming content moderation systems (Art. 34 (2) (b)), for which the largest platforms must change the "[...] content moderation processes, including the speed and quality of processing notices related to specific types of illegal content and, where appropriate, the expeditious removal of, or the disabling of access to, the content notified, in particular in respect of illegal hate speech or cyber violence, as well as adapting any relevant decision-making processes and dedicated resources for content moderation" according to Article 35 (1) (c).

3.2. RELATED WORK

Research on content moderation and platform governance spans law, political science, communication studies, and computer science. Gillespie conceptualizes platforms as “custodians of the Internet” that exercise discretionary power over online speech while sustaining a myth of neutrality. He also emphasizes the often-hidden human labour of internal teams, crowd workers, community managers, volunteer moderators, and users, as well as the occupational risks created by exposure to harmful content and inadequate

linguistic support [287, pp. 24, 45, 111]. This governance power has prompted concern about arbitrary moderation decisions, limited transparency, and weak accountability mechanisms [307, 430, 766]. Klonick describes platforms as “new governors” with de facto authority over public discourse in the United States [430, p. 1604]. York situates platform speech governance within surveillance capitalism and wider power asymmetries, including gendered effects, the reproduction of offline repression online, and the velocity and virality of online hate [766, pp. 25, 73, 145, 183].

In legal and normative theory, scholars have asked how constitutional and administrative values such as due process, proportionality, and freedom of expression should apply to private platforms, particularly at the scale of contemporary moderation systems [200, 201]. Balkin distinguishes the “new-school” governance techniques of digital networks from earlier free-expression conflicts and highlights collateral censorship, public–private cooperation, and infrastructural restraints [60]. Celeste et al. argue that platform regulation should embed fundamental-rights, international-law, and procedural safeguards [121]. Wagner et al. operationalize related standards through a legal taxonomy and coding framework for content-moderation decisions [742].

Macdonald and Vaughan examine the difficulties moderation systems face in borderline cases that do not map neatly onto a single legal category, including terrorist content [490]. Napoli similarly identifies a democratic need to govern digital media environments affected by false or algorithmically generated news and filter bubbles [553, pp. 22–80].

Emerging legal analyses of the DSA examine key innovations in its structure. Van de Kerkhof addresses the role, legitimacy, and risks of the Trusted Flagger system, including certification, proof of expertise, the notification process, and reporting duties [723, 724]. Vinagre et al. examine recommender systems under Art. 27 DSA and the research opportunities created by DSA data-access provisions [733].

Hildebrandt examines data-driven regulation and the need for rule-of-law safeguards in automated systems, including AI and machine-learning systems used in content moderation and recommendation [359]. De Gregorio and Dunn analyse the DSA’s systemic-risk logic alongside instruments such as the General Data Protection Regulation and AI Act, emphasizing its risk-based approach to governing very large platforms [315]. Under Art. 33 DSA, platforms with at least 45 million average monthly active recipients in the Union may be designated as VLOPs and become subject to additional risk-assessment, audit, and transparency duties [659].

Beyond legal scholarship, research in computer science and human–computer interaction offers insights into the technical, operational, and ethical challenges of content moderation. Jhaver et al.’s interviews with Reddit moderators demonstrate both the limits of human moderation at scale and the role of shared tools such as AutoModerator [392]. Chandrasekharan et al. develop computational methods for studying moderation dynamics and suspension practices across Reddit communities [123]. Related work documents bias and fairness problems in language technologies, biometric systems, and image-cropping algorithms, including disparities across gender and dialect [78, 482, 607, 641]. These findings reinforce concerns about algorithmic opacity and the need for explainability and contestability, which the DSA addresses in part through Arts. 20 and 27 [769].

Other interdisciplinary studies focus on platform-regulation architectures. Gorwa

et al. provide typologies of platform-accountability mechanisms and frameworks for evaluating regulatory interventions [307]. Suzor emphasizes digital constitutionalism as a means of embedding human rights in digital infrastructures and processes and of reframing the relationship between users and platforms [686, p. 161]. This work treats content moderation as a site of democratic governance rather than merely a technical or managerial task, drawing attention to online abuse, Section 230 liability protections in the United States, and copyright notice-and-takedown procedures [686, pp. 41, 68, 92].

While prior scholarship has examined individual components of the DSA, such as transparency, automation, or dispute resolution, there remains a gap in conceptualizing the regulation as a legally structured process. This paper contributes to that gap by offering a doctrinal, process-oriented analysis that maps DSA provisions to the sequential stages of content moderation. In doing so, it integrates insights from law, technology studies, and platform accountability research to illuminate how content governance is reconfigured under the DSA's rule-of-law framework.

The DSA permits automated content moderation but imposes transparency and accountability obligations. Providers must disclose the use of automated means in Statements of Reasons and transparency reports and provide appropriate avenues for redress (see Sections 3.3 and 3.3). Research on large language models also illustrates how automated assistance may support explanation while introducing new reliability and accountability risks [664]. Content moderation should therefore be understood as a socio-technical process involving technical systems, institutional procedures, and human judgement [707].

The duties to explain decisions and disclose the use of automation respond to concerns about opaque “black-box” moderation and help users understand and contest automated decisions [186]. Despite these safeguards, challenges remain concerning explainability, accuracy across content types and languages, and over-reliance on automation [580, 631, 709, 759].

Hybrid models combining AI and human review are increasingly common, but risks of bias, false positives, and limited recourse persist. The DSA's requirements nevertheless mark an important step toward algorithmic accountability, user protection, and procedural fairness in digital governance [659].

3.3. THE PROCESS OF CONTENT MODERATION UNDER THE DSA

The Digital Services Act transforms content moderation from a fragmented, often opaque set of platform-led practices into a multistage legal process with procedural requirements and accountability mechanisms [307]. This process can be divided into *ex ante* and *ex post* moderation [287, 661]. *Ex ante* moderation encompasses measures applied before content becomes publicly available, whereas *ex post* moderation occurs after publication or notification. The latter can be divided into five stages: notification, assessment, action, redress, and transparency or documentation. Each stage assigns responsibilities to platforms, creates rights for users, and incorporates external oversight. This paper focuses primarily on *ex post* moderation because the DSA substantially formalizes these post-publication procedures [430, pp. 1636–1638]. This section therefore outlines these

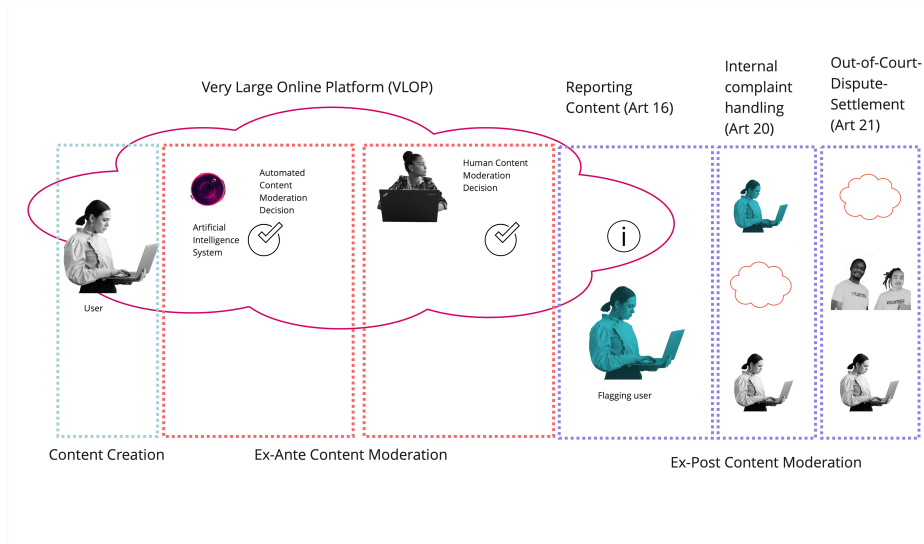


Figure 3.1: Overview of *ex ante* and *ex post* content moderation process

ex post stages and the relevant provisions and actors, as illustrated in Figure 3.1.

Notification: Notice and Action Mechanism The *ex post* moderation process begins when a provider receives a notice identifying potentially illegal content within the meaning of Art. 3(h) DSA. Article 16 requires hosting-service providers to establish mechanisms that are easy to access, user-friendly, and capable of capturing the information needed to assess a notice [668]. A notice must identify the content's exact electronic location, explain why it is alleged to be illegal, include the notifier's name and email address except in specified cases, and contain a good-faith statement. Providers must acknowledge receipt and notify the submitter of their decision. This structured mechanism differs from earlier national approaches such as Germany's NetzDG, which imposed time-bound handling duties for certain notified content [747].

Prioritisation: Role of Trusted Flaggers To support efficient enforcement, Art. 22 DSA establishes Trusted Flaggers: certified entities, including civil-society organizations, that have demonstrated expertise, competence, independence, and diligence [659]. Online platforms must give their notices priority and process them without undue delay. In practice, Trusted Flaggers may still rely on reporting interfaces whose design and process vary across platforms [668]. Trusted Flaggers must publish annual reports on their activities, while platform responses are reflected in transparency duties [724]. The mechanism may accelerate enforcement but also raises questions about transparency and accountability in the selection, funding, and monitoring of these actors [723]. As of May 2026, national authorities had appointed more than 60 Trusted Flaggers across different areas of expertise and Member States [168].

Assessment and Action by the Platform Once content is flagged, providers must assess whether it violates applicable law or their terms and conditions. Under Art. 16(5) DSA, providers must inform the notifier of their decision and the available redress options. When an online platform imposes a qualifying restriction, Art. 17 requires it to give the affected recipient a clear and specific Statement of Reasons detailing the facts and circumstances, the legal or contractual basis, the use of automation, and available avenues for redress (see Section 3.3). These duties make moderation outcomes more traceable and contestable.

Redress Mechanisms: Internal Complaint and Dispute Settlement Following an assessment, eligible users must have access to structured redress. Article 20 requires online platforms to provide a free internal complaint-handling system that is timely, non-discriminatory, and not based solely on automated means. Users may also refer disputes to certified out-of-court dispute-settlement bodies under Art. 21. These bodies must satisfy independence, expertise, accessibility, and fairness requirements. Their decisions are not binding, but the procedure does not prejudice a user's right to seek judicial redress and is intended to provide a faster, less costly alternative.

Documentation and Transparency Obligations At every stage, the DSA emphasizes documentation and transparency. Articles 15, 24, and 42 require relevant providers to publish periodic transparency reports covering matters such as notices received, moderation measures, the use and accuracy of automated means, and complaint outcomes. Article 17 also requires online platforms to submit qualifying moderation decisions to the Commission's Statement of Reasons database. Together, these mechanisms enable regulators, researchers, and civil-society actors to examine patterns in moderation grounds, measures, and outcomes.

Taken together, these stages constitute a legal moderation process grounded in quasi-administrative principles of notification, explanation, recourse, and documentation. Rather than remaining a black box, moderation under the DSA becomes a more traceable and contestable chain of decisions, creating opportunities for democratic oversight and rule-of-law accountability across different user groups.

3.4. EMERGING TENSIONS AND IMPLEMENTATION CHALLENGES

While the DSA introduces a legally grounded and procedurally rich framework for content moderation, its implementation reveals deep-seated tensions across institutional, technical, and normative dimensions. These challenges highlight the gap between regulatory intent and operational feasibility, especially in a platform ecosystem marked by asymmetries in scale, resources, and governance capacity.

First, the DSA's provisions require platforms to internalize complex legal standards—such as proportionality, justification, and user redress—within their moderation infrastructure. For large platforms, this necessitates reconfiguring automated systems and scaling up human moderation capacities to meet legal obligations for transparency, documentation, and user notifications. For smaller platforms, the administrative and technical burdens may prove prohibitive. Implementing internal complaint systems (Art. 20) or preparing

detailed transparency reports (Art. 15, 24, 42) requires legal expertise, organizational resources, and digital infrastructure that many small or medium-sized intermediaries lack. This could lead to uneven compliance and deepen market concentration, as regulatory costs disproportionately affect smaller actors.

Second, the introduction of *Trusted Flaggers* (Art. 22) creates a class of quasi-regulatory actors with enhanced procedural priority. Although intended to expedite expert notices of illegal content, the mechanism concentrates influence in a limited set of certified entities. Questions therefore arise about independence, selection criteria, and potential politicization, particularly where civil-society capacity is unevenly distributed across Member States. Platforms must also determine how priority treatment operates when notices compete for limited review capacity, creating a need for transparent and auditable prioritisation practices.

Third, a significant normative tension lies between protecting users from harmful or illegal content and safeguarding freedom of expression. The DSA's risk-based approach (Arts. 34–35 and 37) may encourage proactive moderation by VLOPs. It may also incentivise over-removal in legally grey areas where providers err on the side of caution to avoid scrutiny or sanctions. Such chilling effects can undermine pluralism and openness, particularly for marginalised or dissenting communities, including LGBTQ users [193], whose content may be disproportionately moderated [334], flagged [550], or misclassified by automated systems [78, 482].

Fourth, although the DSA introduces avenues for redress through internal complaint systems and out-of-court dispute settlement (Art. 20–21), the practical uptake and enforceability of these mechanisms remain uncertain. There is no binding requirement for platforms to comply with decisions from dispute settlement bodies, limiting their legal force. Furthermore, the lack of user awareness or legal literacy may reduce the accessibility and effectiveness of these mechanisms in practice, particularly for vulnerable or digitally marginalised users [668].

Finally, the enforcement architecture of the DSA, split between national Digital Services Coordinators and the European Commission for VLOPs, raises challenges of jurisdictional consistency, resourcing, and institutional coordination. Disparities in legal cultures, regulatory capacities, and political will across member states may lead to fragmented enforcement and a split understanding of illegal content under the DSA [742]. While the DSA aspires to harmonize digital governance across the EU, its implementation could reproduce the very fragmentation it seeks to overcome if regulatory oversight lacks coherence, interoperability, or adequate funding.

Taken together, these tensions underscore the need for sustained monitoring, capacity-building, and transnational coordination. As platforms adapt to the DSA's requirements, ongoing empirical research will be essential to evaluate how legal rights and obligations are operationalized and contested across diverse moderation contexts.

3.5. CONCLUSION

The Digital Services Act represents a transformative shift in the regulation of content moderation across the European Union. By introducing legally binding obligations for transparency, redress, and procedural fairness, the DSA moves beyond the era of voluntary compliance by platforms. It offers a multi-actor, multi-stage process of content

governance, anchored in legal certainty and fundamental rights protection. Platforms are no longer left to arbitrate speech based solely on their terms of service; instead, they are subject to clearly delineated duties concerning notice mechanisms, appeals processes, transparency reporting, and the responsible use of automated tools.

At the same time, the DSA introduces new governance actors, including Trusted Flaggers and out-of-court dispute-settlement bodies, thereby pluralizing oversight and embedding user agency in content moderation. While the framework is ambitious, its success depends on effective implementation, sufficient regulatory capacity, and good-faith compliance by providers. Further empirical research is needed to assess how these processes operate in practice, especially for smaller platforms and users in vulnerable positions. Ultimately, the DSA is a significant regulatory experiment in aligning digital platform governance with democratic values and fundamental rights.

Finally, it is crucial to recognize that the DSA is not only a regulatory framework but also an architecture for future reform. Several components of the regulation anticipate iterative improvement of content governance, embedding reform pathways directly into the legal text. The systemic risk provisions (Arts. 34–35) oblige Very Large Online Platforms (VLOPs) to continuously assess the societal effects of their content moderation systems and to identify areas where mitigation measures must evolve. Auditing obligations under Article 37 enable independent experts to uncover structural weaknesses in moderation processes, thereby generating actionable proposals for strengthening accountability, fairness, and proportionality. Moreover, the data access regime (Article 40 DSA) provides researchers and regulators the tools needed to monitor how the DSA functions in practice, creating an evidence base for future amendments.

By institutionalizing these feedback mechanisms, the DSA establishes a model for evidence-based enforcement and responsive, adaptive digital governance. The regulation couples immediate procedural duties with longer-term reform potential, allowing oversight, transparency, and user protection to evolve alongside technological change and emerging societal risks. Future research and policy work should therefore not only track compliance but also engage with these built-in reform instruments to shape EU platform governance through a societal lens.

4

WICKED BY DESIGN: REGULATING AND AUDITING PLATFORMS UNDER THE DSA

*“By the pricking of my thumbs,
Something wicked this way comes.”*

Second Witch, Macbeth, ACT IV SCENE I. – William Shakespeare

Regulating and enforcing compliance on online platforms under the EU Digital Services Act (DSA) is a paradigmatic “wicked problem.” This article argues that complexity is not a defect to be engineered away but a structural condition of the DSA’s socio-technical and socio-legal governance. We synthesize doctrinal analysis with a systems view of content moderation to show how overlapping legal norms, heterogeneous national definitions of “illegal content,” fragmented reporting artefacts (Articles 15, 17, 34, 37), and hybrid human–AI moderation generate persistent ambiguity for platforms, regulators, auditors, and users. Using reporting mechanisms as the empirical interface between law and infrastructure, we outline how transparency reports, the EU Transparency Database, systemic risk assessments, and independent audits translate practice into compliance claims, while also producing contradictions that challenge legal certainty and auditability at scale. Reframing complexity as a design condition, we propose five interlocking requirements for governing through complexity: (1) consistency analysis across reporting artefacts; (2) improved, mixed-method methodologies for systemic risk assessment; (3) independent usability audits of user reporting (Article 16); (4) AI support that scaffolds rather than substitutes human judgment, with bias, error, and contestability testing; and (5) collaborative, human-in-the-loop audit protocols (Articles 34, 37). Understanding how the DSA’s layered obligations

operate in practice is crucial, as compliance now structures much of Europe's digital public sphere. We discuss implications for law, computer science, HCI, and regulatory practice. The contribution is a conceptual and methodological blueprint that treats the DSA as a system of systems: one that can transform complexity into accountable, reproducible, and enforceable governance rather than attempting to eliminate it.

4

This chapter is based on: Sekwenz, M.-T. (2026). "Wicked by Design: Regulating and Auditing Platforms under the Digital Services Act." In *Mensch-Maschine-Kooperation im Cyberspace / Human-Machine Cooperation in Cyberspace: Proceedings of IRIS 2026*, pp. 441–450. Editions Weblaw. [Published proceedings](#).

4.1. INTRODUCTION

Online platforms structure contemporary expression, commerce, and political communication. Along-side their benefits, public concern has grown about illegal content, harmful-but-lawful speech, opaque moderation decisions, and limited user agency when accounts or posts are restricted [713]. At the same time, the Digital Services Act (DSA) does not provide a single ‘compliance model’ [246]. It distributes obligations across interfaces, organizational processes, data access, and documentation duties. In practice, compliance becomes an engineering and governance problem: teams must translate open-textured legal requirements into taxonomies, workflows, and logging infrastructures that hold up under regulatory scrutiny and external auditing. This text contributes to the ongoing discussion on how to understand DSA compliance through a de-tailed look at challenges that the regulatory domain brings with it, as well as the complex legal interplay necessary to understand for operationalization of the DSA for meaningful platform governance by answering the following research questions:

- RQ1: How does the DSA’s architecture institutionalize socio-technical and socio-legal complexity?
- RQ2: How do reporting artefacts translate platform practice into compliance claims, and where do contradictions emerge?
- RQ3: What design requirements enable governance through complexity rather than against it?

The analysis combines doctrinal legal interpretation with a socio-technical systems perspective and draws on comparative examples from ongoing DSA transparency audits, risk assessment reviews, and user interface analyses. This mixed-theoretical framing enables us to link formal regulatory text to empirical artefacts and design-level implications. The DSA, therefore, is more than a single piece of legislation: it is a system of systems.

Contribution. The paper (i) frames the DSA as a system-of-systems whose complexity is structural; (ii) surfaces common points of contradiction across transparency reports, the EU Transparency Data-base, systemic risk assessments, and audit reports; and (iii) offers a five-part governance blueprint that is actionable for regulators, platforms, auditors, and HCI/AI researchers.

4.2. WHY CONTENT MODERATION IS A WICKED PROBLEM

Content moderation is not a single task but an ongoing balancing act between freedom of expression, safety, legality, and operational constraints. Moderation decisions are made under uncertainty: the legality or harmfulness of a post depends on context [111], jurisdiction [741], language [513], and cultural meaning [669]. Following Rittel and Webber’s classic formulation, the problem is ‘wicked’ because it is ill-defined, value-laden, and continuously re-solved rather than solved once and for all [619]. The DSA’s interventions, notice-and-action, transparency, systemic risk assessment, and audits, do not remove this wickedness. Instead, they attempt to manage it by making decisions, processes, and risks visible and contestable.

Wickedness is amplified by scale. Large platforms process vast volumes of posts, comments, images, and videos under tight response expectations [290]. Strict timelines, sometimes imposed by Member State rules and sector-specific regimes, create incentives for high-throughput decisions that inevitably trade off accuracy and contextual judgment. The DSA defines content moderation broadly as activities, automated or not, aimed at detecting, identifying and addressing illegal content or information incompatible with terms and conditions, including measures that affect availability, visibility, accessibility, and monetization. This definition matters: under the DSA, moderation is simultaneously a technical pipeline, a set of organizational procedures, and a legally reviewable practice. In practice, moderation unfolds through interlinked stages: (1) creation and upload; (2) automated screening and ranking; (3) human review of flagged cases; (4) enforcement actions (e.g., removal, de-motion, demonetization); (5) user and third-party notices (Art. 16); (6) complaints and appeals (Arts. 20–21); and (7) escalation through external dispute settlement and, indirectly, regulatory oversight. Each stage produces evidence and artefacts that feed later accountability steps.

Two practical dilemmas recur across domains. First, over- and under-removal: false positives restrict lawful expression, while false negatives permit continued exposure to illegal or harmful content. Second, uneven performance: detection accuracy varies across languages, regions, and modalities (text vs. images vs. memes), often in ways that map onto existing inequalities. These dilemmas are not implementation bugs; they are structural tensions that DSA compliance must surface and manage.

4.3. THE DSA AS A SOCIO-TECHNICAL AND SOCIO-LEGAL SYSTEM

The DSA regulates platforms as socio-technical systems: multi-actor arrangements in which users, automated classifiers, human moderators, policy teams, designers, and recommender systems jointly produce governance outcomes. At the same time, it builds a socio-legal system that links private order-ing (Terms of Service and community standards) with public oversight (Digital Services Coordinators, the Commission, trusted flaggers, researchers, and independent auditors). Compliance is therefore co-constituted by infrastructure and institutions: legal norms affect how systems are designed, and system capabilities shape how norms are interpreted. Platform governance and content moderation constitute multi-actor systems involving users who create content; automated and human moderators who review it; platform designers who set technical and policy parameters; and regulators, auditors, and researchers who monitor outcomes. Singh (2014) conceptualizes socio-technical systems as “multistakeholder cyber-physical systems” connecting computational and physical resources within a societal context [675]. Chopra and Singh (2018) emphasize that such systems consist of social and technical entities whose interactions generate emergent behavior [130]. Building on Gorwa et al. (2020), content moderation can be understood as a multi-layered technical stack of data pipelines, classifiers, and feedback loops that redistribute decision-making power from humans toward automated infrastructures [306]. Gupta et al. (2022) describe this as socio-technical alignment across individual, organizational, and societal layers, the “golden

circle” of moderation [328]. At the micro-level, users and moderators interact through platform interfaces; at the meso-level, organizational rules and AI systems structure these interactions; at the macro-level, regulators and auditors feed accountability signals back into the system. Together, these layers form a continuous feedback circuit between design, moderation, and oversight.

4.4. SOCIO-LEGAL SYSTEM: PRIVATE ORDERING AND PUBLIC OVERSIGHT

The DSA institutionalizes a socio-legal architecture that links private rules with public accountability. Platforms’ Terms and Conditions (Art. 14) remain central, yet they must be applied with due regard to fundamental rights and procedural guarantees. Oversight is multi-layered: Digital Services Coordinators operate at Member State level, while the Commission exercises enhanced supervision for VLOPs and VLOSEs. Trusted flaggers and vetted researchers further expand the field of actors who can contest, inspect, or evaluate platform practices.

In this architecture, compliance is produced through documentation. The key question becomes: what evidence counts as compliance, and how is it stabilized across different artefacts and audiences?

The DSA simultaneously constructs a socio-legal architecture mediating between private platform sovereignty (contractual freedom, Terms of Service) and public democratic accountability (regulatory oversight, transparency, and audit). Cauffman and Goanta describe this as a “new order” of online [119]. Turillazzi et al. highlight its ethical and societal implications [713], while De Gregorio situates it within a broader digital-constitutional framework that transforms platforms into active compliance actors [180]. Within this domain, DSA compliance spans private and public norms:

1. Private norms such as Terms of Service (Article 14 DSA);
2. Public legal frameworks of primary EU law: including the EU Charter of Fundamental Rights, TEU , TFEU
 - Secondary EU law (Regulations): the DSA, e.g., the Terrorist Content Regulation, and the General Data Protection Regulation [243].
 - Secondary EU law (Directives): e.g. the Audiovisual Media Services Directive [16].
3. Delegated Act and Guidelines that operationalize transparency and systemic-risk requirements [226].

These artefacts form the procedural backbone through which socio-technical practices are rendered legally visible and enforceable.

4.5. MULTI-LAYERED COMPLEXITY IN EVALUATING DSA COMPLIANCE

Evaluating DSA compliance is not equivalent to checking whether a platform has ‘filed the right documents’. The core difficulty is multi-layered complexity: heterogeneity in

legal definitions, fragmentation across reporting channels, socio-technical limits of moderation, user-interface burdens, and the sheer scale of data that auditing must cover. These layers are coupled: legal ambiguity shapes compliance perception; compliance perception shapes measurement; measurement shapes audit scope; and audit scope shapes enforcement incentives. Legal systems are, at their core, social systems. They are composed of networks of institutions, courts, legislatures, regulators, and of instruments such as laws, regulations, directives, and judicial decisions. As Vivo et al. note, complexity science has long included legal systems within its scope, since law emerges from interactions among social actors, rules, and institutional feedback loops [734]. In this sense, the DSA represents not just a new law, but a complex system of governance for already complex legal and socio-technical environments. Gupta et al. describe such multi-actor governance as a form of “complex social problem organization” [328]. The DSA exemplifies precisely this: a regulatory system that must coordinate diverse institutions, interpretive logics, and technical infrastructures under conditions of constant change.

4.5.1. WHAT IS LEGAL COMPLEXITY?

Legal complexity can be understood as the difficulty of predicting, interpreting, and implementing legal norms in a multi-level system. Appermont distinguishes between two dimensions: Structural or systemic complexity, which concerns the institutional and procedural organization of the legal system; and content or rule-level complexity, which concerns the substance, density, and interaction of norms [45, 236-263]. These two dimensions together determine how navigable, coherent, and enforceable a legal framework is in practice. At the systemic level, legal complexity arises from institutional multiplicity and procedural interdependence. Appermont identifies four indicators: Number of applicable norms within a legal system; number of competent policy levels authorised to issue relevant norms; durability or longevity of those norms; and degree of internal consistency between them [45, 247]. The DSA, as an EU Regulation, embodies all four dimensions: The DSA itself is a regulation with 102 pages, 93 articles, and 156 recitals. Furthermore, it includes several other regulation, directions, and guidelines in its scope. Several policy-level actors are authorized in the compliance process, starting from platforms’ self-assessment under Article 34’s systemic risk assessment reports, to the competent authorities in member states – the Digital Service Coordinators, the trusted flaggers and their reporting processes, and the Commission. Norms are either formed in annexes to delegated regulations as a shorter form of legal guidance for compliance with the DSA itself as an EU regulation.

From a lens on consistency among them – It interacts with national laws on hate speech, defamation, and privacy; coexists with other EU instruments such as the GDPR, the Terrorist Content Regulation, and the AVMSD; and delegates significant interpretive authority to Member States and to the European Commission. Each addition increases institutional density and interpretive dependency. Schuck (1992) describes this phenomenon through four core characteristics of legal complexity: Density – the sheer volume of applicable rules; technicality – the specialized terminology and expertise required for interpretation; differentiation – the diversity of policy levels and legal sources involved; and indeterminacy – the fluidity and contestability of legal meaning [649]. These

features interact to produce practical uncertainty. Density and differentiation increase administrative burden; technicality limits accessibility; and indeterminacy complicates enforcement and auditability. As mentioned above, the DSA is a dense regulation with a large amount of norms and legal interplay between other regulations. It is a technical regulation in the sense of Schuck, since it uses not only legal terms (such as references to other regulations and legal terminology like *bona fide*), but also technical jargon (such as hosting, caching, or recommendation systems). It fulfills the criteria of differentiation since it includes primary EU law, like the Charter, in a contract between users and platforms and builds, therefore, a quasi-horizontal dimension of fundamental rights through the need for systemic risk assessments under Article 34 (1) (b) DSA. Besides, it is indeterminate in the sense that the DSA creates a very complex enforcement, reporting, and compliance process, e.g., through the different re-reporting obligations on transparency (Articles 15, 24, 42, and 17 DSA), the different audits according to Articles 34 and 37, as well as the division between member states and the Commission. All of those details create new obligations but also create many new tasks for multiple actors, or as Schuck concludes, complexity is costly: it consumes institutional resources, delays implementation, and burdens those subject to regulation [649].

4.5.2. WHY COMPLEXITY IS STRUCTURAL

Complexity is embedded in the DSA at multiple levels: legal, procedural, technical, user, and oversight. On the legal level, the Regulation defers the definition of “illegal content” to Member States, thereby generating heterogeneity through subsidiarity¹ and national traditions. The procedural level exemplifies the DSA’s multiplication of reporting artefacts: transparency reports, systemic-risk assessments, audits, and database entries—each with distinct audiences and incentives. On a technical level, platforms rely on both automated systems and human moderators, each with inherent limitations: machines lack contextual judgment, humans lack scale. At the user level, it creates rights to report or appeal, mediated by interfaces that often meet formal criteria but fail in usability and accessibility. And at an oversight level, Auditors and regulators must evaluate vast datasets, often relying on automated tools that introduce their own opacity. Complexity, therefore, is not a regulatory flaw but the operational condition under which the DSA functions.

4.5.3. THE DSA AS LEGAL COMPLEXITY IN ACTION

The Digital Services Act illustrates how structural and rule-level complexity converge. It sits at the intersection of EU harmonization and Member State subsidiarity (while it sets uniform procedural obligations, it deliberately defers substantive definitions, such as “illegal content,” to national law (Art 3 (h) DSA)). This design creates interpretive heterogeneity by intent. Moreover, the DSA multiplies compliance artefacts, transparency reports, systemic risk assessments, audit reports, and database entries, each governed by distinct procedural rules and audiences. Rather than simplifying governance, it expands the interface surface between legal, technical, and organizational systems. This deliberate layering transforms the DSA into an ecosystem for studying how legal complexity

¹ See Art. 5(3) TEU.

manifests in practice: how norms are interpreted by platforms, verified by auditors, and enforced by regulators. It also shows how socio-legal and socio-technical complexities intertwine: every reporting mechanism translates between law and infrastructure, and every audit feed legal reasoning back into system design.

4.5.4. THE LIMITS OF SIMPLIFICATION

It is tempting to treat complexity as a problem to be solved through simplification: to seek harmonized definitions of illegal content across all Member States, to insist that transparency reports “tell the full story,” or to automate away the human judgment that slows down moderation. Yet each of these strategies risks undermining the very principles the DSA is designed to protect. Complete legal harmonization would erode national autonomy and constitutional traditions; unified reporting would invite strategic manipulation of a single dataset; full automation would undermine fundamental rights by prioritizing speed over context. Simplification is misleading: it gives the illusion of certainty where none exists. It replaces one form of complexity (distributed contradictions) with another (centralized but fragile regimes), without addressing the underlying tensions. The challenge, then, is not to eliminate complexity but to govern through it.

4.5.5. THE PROBLEM OF LEGAL CERTAINTY

Even when the definition of “illegal content” is not directly at issue, the DSA introduces a different form of uncertainty through its reliance on vague, open-textured concepts. Terms such as “systemic risk” (Article 34), “reasonable assurance” (Delegated Regulation), and “easy to access” (Article 16) were deliberately chosen to provide flexibility across diverse platforms, risk domains, and national contexts. This flexibility, however, comes at the cost of enforceability. Without operational definitions, compliance depends less on legal standards than on how platforms, regulators, and auditors choose to interpret these categories in practice. Platforms, facing strong reputational and financial incentives, often exploit this vagueness. They frame risks in terms that minimize their significance, emphasizing general “communications challenges” rather than measurable electoral harms. Regulators, in turn, are left without a reliable benchmark to evaluate whether a submission represents substantive compliance or merely rhetorical performance. Legal certainty — the principle that individuals and entities should be able to understand the law and foresee its consequences — is thereby compromised.

4.6. SOURCES OF SOCIO-TECHNICAL COMPLEXITY

4.6.1. FRAGMENTED REPORTING MECHANISMS

The Digital Services Act requires platforms to generate multiple reporting artifacts: (bi-)annual transparency reports (Article 15, 24, 42), annual systemic risk assessments (Article 34), independent audit reports (Article 37), and continuous entries in the EU Transparency Database (Article 17) [148, 658]. Each artifact is meant to serve a distinct function and audience. Transparency reports provide headline figures to the public, risk assessments frame challenges for regulators, audit reports establish legal accountability, and the Transparency Database enables real-time oversight of platforms’ removal actions. In principle, this layered architecture should provide a comprehensive view of platform

governance. In practice, however, the multiplicity of artifacts generates contradictions rather than coherence [665]. Platforms strategically frame the same phenomenon in different ways depending on the reporting context: maximizing numbers in transparency reports to demonstrate action, minimizing risks in systemic assessments to avoid regulatory scrutiny, re-framing categories in audits to fit legal obligations, and drastically under-reporting in the EU Transparency Database, which covers only a fraction of removals.

4.6.2. SOCIO-TECHNICAL MODERATION

Content moderation under the DSA exemplifies the socio-technical character of enforcement. Decisions are never purely technical or purely legal: they emerge from a distributed system of automated filters, human moderators, and regulatory requirements [79]. Automated classifiers operate at scale, scanning millions of posts per day, but are challenging to verify and can still be evaded by minor alterations, such as rephrasing or re-uploading content in different formats [308]. Human moderators provide contextual judgment but work under severe time constraints and productivity quotas, with limited legal training and inconsistent access to contextual information. Together, these socio-technical conditions generate persistent tensions between efficiency, accuracy, and legality.

4.6.3. AI SUPPORT AND BIAS

The introduction of AI into user reporting systems is often framed as a technical fix to the complexity of DSA compliance. Faced with users unable to navigate dense taxonomies and moderators struggling under time constraints, platforms present AI support as the solution: faster, easier, and more consistent reporting. Yet these framing obscures deeper risks [372]. When AI is deployed as a stand-alone shortcut, it risks transforming moderation into a process that is more opaque, more biased, and less accountable – while potentially appearing formally compliant or even objective.

4.6.4. AUDITING AT SCALE

Article 37 of the DSA establishes independent audits as a cornerstone of accountability. Audits are intended to provide an external check on whether platforms' transparency reports, systemic risk assessments, and internal procedures reflect reality. Yet the effectiveness of this mechanism is constrained by scale [290]. Auditors face not just thousands but sometimes millions of entries across multiple artefacts – transparency reports, EU Transparency Database records, systemic risk assessments, and trusted flagger reports. Manually reviewing this volume is costly and, in some cases, only possible to a limited extent; the patterns are too complex, the categories are inconsistent, and the data are fragmented. Faced with this challenge, both auditors and platforms may turn to automated methods, such as large language models (LLMs), to detect anomalies or patterns that might otherwise go unnoticed [272]. The promise is efficiency and breadth of coverage. But the risk is that audits themselves become dependent on tools that are prone to error, hallucination, and opaque reasoning – creating the appearance of accountability without its substance.

4.6.5. INTERDEPENDENCE ACROSS LAYERS

The preceding sections demonstrate that legal, procedural, and technical complexity in the DSA reinforce one another rather than appearing separately. Fragmented artefacts, socio-technical constraints, and legal ambiguity together create what can be understood as a “coupled complexity.” Recognizing these interdependencies provides the rationale for reframing complexity as a condition to be designed for, rather than a defect to be engineered away.

4.7. REFRAMING COMPLEXITY AS A DESIGN CONDITION

If complexity is constitutive, then compliance must be designed for complexity, not against it. The contribution of this analysis is therefore to reframe complexity as a design condition. This requires moving from the language of failure (“fragmentation,” “contradiction,” “opacity”) to the language of governance design: how can regulatory, technical, and organizational systems be built to manage, balance, and make accountable the inherent frictions of platform governance? We propose five interlocking de-sign requirements that together form a blueprint for governing complexity:

Consistency. Substantive definitions like illegality are fragmented but can be included in more understandable guide-lines; besides, reporting categories like the Statement of Reason (SOR) entries (Art. 17) can be standardized and might provide a more consistent form of reporting across reporting mechanisms like the transparency reports and the SOR database. Platforms should be required to use common taxonomies that map national legal obligations onto a shared EU reporting schema. This enables consistency analysis and makes contradictions visible rather than hidden [665].

New auditing methodologies for systemic risk assessments. No single modality of enforcement is sufficient to accurately capture the status quo of platforms. However, improving methodologies for independent audits and systemic risk assessments that use quantitative and qualitative methods to empirically assess risk on platforms can contribute to a better understanding of the complex interplay of regulations [661].

Usability audits. Compliance with legal design obligations, such as Article 16, must be measured not by the mere presence of reporting buttons but by their actual usability and in the context of other communities, such as Human-Computer Interaction. Platforms should be subject to independent usability audits, including empirical tests of how users report illegal content, user comprehension, and legal heuristic walkthroughs with key stakeholders [668].

Support of AI. Where AI is used to assist user reporting, it must be designed to scaffold rather than replace user judgment. Evaluative models that present multiple options with pro- and con-arguments should be mandated over single-answer models that induce automation bias – if used in reporting processes for illegal content. Regulators should require independent testing of AI tools for bias, error, contestability, and usability [662].

Collaborative auditing. Auditing at scale cannot be left to either humans or machines alone. Hybrid protocols must be developed in which LLMs flag anomalies and humans verify them, with audit reports disclosing error rates, methodological steps, and verification procedures. Regulators should require methodological transparency as part of Article 37 compliance if they want to be able to use LLMs in auditing tasks in order to scale up auditing scope for their assessment [662].

4.7.1. IMPLICATIONS FOR ACTORS

Reframing complexity as a design condition has implications across disciplines. For lawyers, it means shifting focus from doctrinal certainty to interdisciplinary work, and the creation of empirical evidence of vague terms like “systemic risk.” For computer scientists, it means designing AI and audit tools that make errors visible and invite contestation, rather than nudging their users. For HCI scholars and de-signers, it means recognizing interface design as a core compliance issue, not a secondary usability problem. For regulators, it means treating enforcement as an iterative process of learning, rather than a binary check of compliance. And for users and civil society, it means demanding not only rights on paper but interfaces and accountability structures that make those rights actionable in practice.

4.8. LIMITATIONS AND FUTURE WORK

This analysis focuses on procedural and artefactual dimensions of DSA compliance and does not provide a quantitative evaluation of enforcement outcomes. Further research should link reporting data with case-level decisions, explore consistency across reporting cycles, and assess the usability of re-reporting and appeals mechanisms through empirical user studies. Developing reproducible audit methodologies, using carefully aligned mixed human–LLM protocols, represents a key direction for multiple actors involved.

4.9. CONCLUSION

The DSA is best read as an experiment in governing complexity. Its innovations – systemic risk assessments, independent audits, and user-facing rights – are all mechanisms that attempt to channel complexity into accountability. Its limitations – fragmented law, contradictory reports, complex moderation – show the enduring difficulty of doing so. The measure of its success will not be whether it eliminates complexity, but whether it can transform complexity into structures of trust, reproducibility, and enforceability. This lesson has significance beyond Europe. As other jurisdictions consider platform regulation, from the United States to Brazil and India, the DSA demonstrates both the inevitability of complexity and the possibility of governing through it. Complexity cannot be engineered away – but it can be designed for, audited, and made accountable. That is the enduring contribution of the DSA and of the model of regulatory learning it represents. The DSA is an investigation into regulatory learning under conditions of complexity. Its effectiveness will not be judged by the text of its provisions alone, but by how well platforms, regulators, auditors, users, and researchers collaborate to transform vague obligations into workable socio-technical and socio-legal systems, and how those can be feed into the compliance process that accepts degrees of legal uncertainty.

II

METHODOLOGICAL FRAMEWORKS FOR ASSESSING DSA COMPLIANCE

5

HOLISTIC DSA COMPLIANCE THROUGH LEVERAGING CONSISTENCY – THEORETICAL CONCEPTS TO EVALUATE ONLINE PLATFORMS UNDER THE DIGITAL SERVICES ACT

“Everything must be taken into account. If the fact will not fit the theory—let the theory go.”

Agatha Christie, *The Mysterious Affair at Styles*

The Digital Services Act (DSA) introduces harmonized rules for content moderation and platform governance in the European Union, mandating robust compliance mechanisms, particularly for very large online platforms and search engines. This study examined compliance with DSA requirements, focusing on Instagram as a case study. We develop and apply a multi-level consistency framework to evaluate DSA compliance. Our findings contribute to the broader discussion on empirically-based regulation, providing insight into how researchers, regulators, auditors and platforms can better utilize DSA mechanisms to improve reporting and enforcement quality and accountability. This work underscores that consistency can help detect potential compliance failures. It also demonstrates that platforms should be evaluated as part of an interconnected ecosystem rather than through isolated processes, which is crucial for effective compliance evaluation under the DSA.

This chapter is based on: Sekwenz, M.-T., Wagner, B., & de Bruijn, H. (2026). "From Reports to Reality: Testing Consistency in Instagram's Digital Services Act Compliance Data." *ACM Journal on Responsible Computing*. doi:[10.1145/3821526](https://doi.org/10.1145/3821526).

5.1. INTRODUCTION

The Digital Services Act (DSA) introduces a comprehensive regulatory framework for platform governance in the European Union, particularly targeting very large online platforms (VLOPs) and very large online search engines (VLOSEs) (Art. 33) [224]. The DSA builds on lessons from earlier national regulations (see Section 5.3). Through a suite of new compliance obligations—including transparency reports (Arts. 15, 24, 42), statements of reason (SOR) in the DSA Transparency Database (Arts. 17, 24(5)) that document all content moderation decisions taken by platforms, systemic risk assessments (Art. 34) evaluating platforms internally, independent audits (Art. 37) cross-checking systemic risk assessments by third parties, and operating advertising repositories to collect information on political content (Art. 39)—the DSA seeks to offer a multitude of new mechanisms of control to evaluate online platforms. The DSA does not move beyond transparency so much as embed transparency within a broader architecture of due diligence, documentation, risk governance, and auditing. This broader architecture is well captured by Husovec’s description of the DSA as “a digital due process regulation bundled with risk-management tools” [377, p. 1], one that “invites transparency to private decision-making and institutionalizes constant risk management” [377, p. 1]. Rather than treating individual disclosures as self-sufficient indicators of accountability, the Regulation creates a layered compliance environment in which different artefacts, including transparency reports, Statements of Reasons (SORs), systemic risk assessments, and independent audits, can be read in relation to one another. Keller usefully cautions that transparency mandates can reshape internal platform practices, impose significant compliance burdens, and push platforms toward de facto standardisation in what they track and report [420]. This broader architecture creates new opportunities for empirical scrutiny, but it also introduces new interpretive and operational challenges, because reporting outputs are shaped not only by legal obligations but also by internal logging systems, changing taxonomies, templates, and evolving implementation guidance. Leerssen therefore suggests that platform transparency should move beyond isolated disclosures and toward broader forms of observability, since these systems are sociotechnical processes whose behaviour is “emergent and constantly in flux” [461, p. 6].

We use *consistency* as a method for evaluating platform compliance, focusing on Instagram. Building on Keller’s work, we treat reporting inconsistencies as possible indicators of misreporting or governance failure [417, 418, 421, p. 18]. In the financial domain, consistency provides a materiality threshold that allows auditors to assess whether reported activities can be trusted [441]. Analogously, in platform governance, consistency enables evaluators to identify discrepancies between mechanisms—such as mismatches between reported account suspensions in transparency reports and moderation actions recorded in the SOR (See Section 5.4.2). Therefore, in this article, we use consistency as a method for identifying anomalies, traceability gaps, and audit-relevant discrepancies across DSA mechanisms. Such discrepancies do not, by themselves, establish legal breach. Especially in the early implementation period, inconsistencies may also reflect operational constraints and changing reporting requirements rather than immediate legal breach. As Keller notes, platforms may lack “standardized tracking and queuing tools” and may therefore modestly miscount some categories, while later technical specifications can also “reshape what platforms actually do in their content moderation” [419]. As Keller

puts it, “We should insist on precise data in some cases, and tolerate more imprecision in others” [420].

Besides, we assess how different forms of consistency can be used to evaluate compliance with the DSA by analyzing three of its key transparency and oversight mechanisms: (1) biannual transparency reports (number 1-4), (2) statements of reason (SOR) published in the Transparency Database, (3) the first systemic risk assessment and (4) independent audit reports. Focusing on Instagram as a case study, we ask:

- **RQ1:** How can different forms of consistency be assessed across DSA reporting instruments?
- **RQ2:** What challenges and opportunities arise in applying consistency checks to a VLOP’s DSA compliance?
- **RQ3:** How can a consistency-based approach support effective oversight and auditing under the DSA?

Our contributions are threefold. First, we propose a framework for multi-level consistency assessment of DSA compliance,¹ informed by audit theory, legal scholarship, and empirical research. Second, we apply this framework to Instagram’s transparency reports, SOR data, and risk assessments. Third, we identify practical implications for researchers, NGOs, regulators, and auditors. The rest of the paper is structured as follows. Section 5.2 presents our methodology, including the typology of inconsistency, the interdisciplinary methodological framing, and the case study selection. Section 5.3 reviews related work and legal foundations. Section 5.4 applies the three-level consistency framework to Instagram’s DSA reporting. Section 5.5 discusses key challenges in evaluating consistency across the relevant DSA mechanisms. Section 5.6 outlines recommendations for improving consistency-based compliance evaluation. Section 5.7 reflects on the broader regulatory relevance of consistency under the DSA. Section 5.8 discusses implications and limitations. Finally, Section 5.9 concludes.

Our findings inform ongoing discussions about regulatory implementation, compliance evaluation, and auditability under the DSA. By bridging granular decision-level data (from SORs) with aggregate reporting (from transparency reports) and risk-based assessments (from systemic risk evaluations), we provide a structured approach to understanding how consistency can be used as a tool for empirical compliance assessment and regulatory scrutiny.

5.2. METHODOLOGY

This study develops and applies a consistency-based methodology to evaluate how control mechanisms introduced by the Digital Services Act (DSA) can be used to assess platform compliance.

Focusing on Instagram — a very large online platform (VLOP) with high regulatory obligations according to Article 33 DSA — we analyze how these four key mechanisms (defined in Section 5.1) operate in practice for Instagram. Together, these mechanisms form

¹Throughout the analysis, cross-mechanism traceability is examined as a focused form of external consistency. This allows us to preserve the three-level framework while still analysing whether different DSA mechanisms can be rendered mutually intelligible for purposes of compliance evaluation.

a layered system of regulatory oversight that enables both qualitative and quantitative compliance evaluations.

To assess how these instruments function as tools of control, we propose and apply a multi-level consistency framework, distinguishing between:

- **Internal consistency** — alignment within a single report or mechanism;
- **External consistency** — coherence across multiple reporting instruments (e.g., between SORs, transparency reports, and risk assessments);
- **Historical consistency** — comparability across reporting periods.

We therefore retain three levels of consistency in this article: internal, external, and historical consistency. Cross-mechanism traceability is not treated as a separate fourth level, but as a specific analytic application of external consistency. In other words, when we examine whether systemic risk claims, moderation outputs, and audit findings can be meaningfully related across transparency reports, SOR data, systemic risk assessments, and independent audits, we are operationalizing external consistency across a particularly complex set of interlocking DSA mechanisms.

Our methodological approach combines:

1. A doctrinal legal analysis to identify the obligations, scope, and interrelations of the relevant DSA mechanisms, based on primary legal texts [225, 230].
2. A compliance evaluation of Instagram's reporting practices, drawing from its publicly submitted DSA transparency reports (Rounds 1–4 [522, 524–526]), its SOR entries from the DSA Transparency Database (capturing over 32 million moderation decisions in May 2025 [146]), and the first available systemic risk assessment and independent audit reports [523, 767].
3. A comparative consistency check to triangulate data across these mechanisms, enabling us to assess the plausibility, coherence, and completeness of Instagram's reporting under the DSA framework.

The relevance of consistency directly maps onto our research questions as introduced earlier (see Section 5.2). Together, these questions reframe DSA reporting not as discrete compliance actions, but as a system of interlocking mechanisms whose integrity hinges on their internal and mutual coherence.

Whereas previous studies have documented inconsistencies in single reporting streams, this paper uniquely triangulates four DSA mechanisms using a multi-level consistency framework, and demonstrates cross-mechanism gaps in a case study.

By synthesizing findings from all four mechanisms, our methodology enables a more holistic evaluation of whether Instagram's reporting aligns with the DSA's high-level control logic. It also highlights where reporting practices fall short of delivering accountability. In doing so, this paper contributes both an analytical lens and a practical toolset for regulators, auditors, and researchers tasked with interpreting the increasingly complex landscape of platform compliance under the DSA.

5.2.1. TYPOLOGY OF INCONSISTENCY AND PLAUSIBLE EXPLANATIONS

To move beyond the descriptive identification of discrepancies, we distinguish analytically between different applications of the three consistency levels and their possible compliance relevance (see Section 5.2). Internal inconsistencies concern the integrity of a single artefact, such as subtotal–total mismatches, undefined residual categories, or platform-specific reporting distortions created by cross-platform aggregation (see Section 5.4.1). External inconsistencies concern reconciliation across artefacts, such as divergences between transparency reports and SOR data, weak alignment between systemic risk claims and downstream moderation outputs, or broader failures of cross-mechanism traceability (see Section 5.4.2). Within this category, we treat cross-mechanism traceability as a specific application of external consistency that asks whether claims made across transparency reports, SOR data, systemic risk assessments, and independent audits can be meaningfully related to one another for purposes of compliance evaluation. Historical inconsistencies concern temporal instability, including unexplained category shifts, changes in reporting logic, or measurement drift across reporting rounds (see Section 5.4.3). This distinction adds an interpretive layer to the analysis: it clarifies not only that inconsistencies exist, but also what kind of evaluative problem they represent and what sort of follow-up inquiry they require. Internal inconsistencies are most likely to point toward reporting irregularities or category-design problems and therefore call for reconciliation logic, source tables, and workflow documentation. External inconsistencies are more heterogeneous: some are best explained by infrastructural friction or interpretive divergence, while others may indicate more substantive governance weaknesses or traceability deficits, requiring scrutiny of logging pipelines, submission rules, variable definitions, and cross-mechanism mappings. Historical inconsistencies are most plausibly associated with category drift, changing reporting logic, or shifts in internal measurement practices over time. In the empirical sections below (see Sections 5.4.1–5.4.3), we use this typology not merely to label discrepancies, but to guide interpretation by asking which explanatory class is most plausible in each case and what follow-up evidence would be needed to assess it more confidently. Table 5.2 operationalizes this typology for the Instagram case study by showing how concrete discrepancies can be read as different evaluative problems and by indicating the kinds of documentation or audit evidence needed for follow-up.

5.2.2. INTERDISCIPLINARY METHODOLOGICAL FRAMING

Our approach is interdisciplinary in a specific sense. First, we use doctrinal legal analysis to identify the scope, obligations, and interrelations of the relevant DSA mechanisms [294, 698]. Second, we conduct an empirical compliance evaluation that translates those obligations into observable variables and reporting indicators across publicly available platform artefacts (see Section 5.4). Third, we apply consistency-based auditability assessment to examine whether these artefacts can be rendered mutually intelligible across mechanisms and over time. In this respect, our approach is closer to a platform observability exercise than to a simple comparison of reports: it treats DSA artefacts not as neutral outputs, but as sociotechnical products of internal classification systems, logging infrastructures, reporting templates, and institutional design choices [420, 461]. This orientation also follows Kemper and Kolkman's broader point that disclosure alone does

not secure transparency; what matters is whether the information made available can be meaningfully understood and scrutinized by the relevant audience [422, p. 2086]. For DSA research, this means that public reporting artefacts must be assessed not only for their formal availability, but also for whether they support meaningful cross-mechanism interpretation and scrutiny.

This methodology is designed to identify anomalies, traceability deficits, and areas for audit-relevant follow-up testing. This is why the framework should be understood as a first-order auditability method rather than a definitive legality assessment. It can identify anomalies, traceability deficits, and areas requiring follow-up investigation, but it cannot on its own resolve all questions of substantive compliance. In this sense, our findings also support the broader caution that “Digital transparency — whether it is enacted through technological solutions or more classical administrative and organizational means — will not on its own provide an easy solution to the challenges posed by the growing role of platforms in political and public life” [305, p. 20]. This limitation also supports Kemper and Kolkman’s broader claim that transparency has practical value only where disclosed information can be taken up by audiences capable of critical and informed evaluation [422, p. 2092]. It cannot, on its own, establish final legal liability or resolve all questions of substantive compliance breach. Its value lies instead in showing where a platform’s reporting environment becomes difficult to reconcile, where explanations are missing, and where more targeted legal, technical, or organizational investigation is needed. Framed in this way, consistency functions as an evaluative lens for auditability rather than as a substitute for doctrinal adjudication [380, 420, 461].

5.2.3. CASE STUDY SELECTION: INSTAGRAM

We selected Instagram as a case study due to its significance as a VLOP with an estimated 272.0 million monthly active users in the EU [526], its ongoing regulatory scrutiny, and the public availability of its submissions across the three examined mechanisms.

Instagram is used here as an analytically rich case for methodological demonstration rather than as a statistically representative sample of platform behavior. The purpose is not to generalize inconsistency rates across platforms, but to show how a structured consistency framework can surface legal, infrastructural, and evaluative tensions across multiple DSA mechanisms. This is consistent with Leerssen’s broader observability perspective, which emphasizes enabling “different actors to develop their own observation capacities, adapting their analytical methods to the questions they want to ask” [461, p. 10]. Future work should extend this framework comparatively across providers and sectors in order to test which inconsistencies are platform-specific and which reflect broader features of the DSA reporting environment.

The European Commission has launched formal proceedings into Instagram’s compliance with the DSA, including concerns over political content visibility, deceptive advertising, and takedown mechanisms [233] — making it a timely and relevant case for identifying potential inconsistencies and accountability gaps. Similarly, in May 2025, the Commission concluded that TikTok had breached its obligations under the DSA regarding advertising transparency, making the first steps toward DSA enforcement and underscoring the broader challenges of ensuring compliance across major platforms [237].

For the SOR data, we analyzed moderation decisions submitted during the periods covered by Transparency Reports 2-4, starting from the operational launch of the SOR database on 26 September 2023 [234]. In total, Instagram reported over 58 million moderation actions during these periods [22]. For systemic risk assessments, we examined Instagram's first published systemic risk report in line with Article 34 DSA, focusing on its structure, self-reported mitigation strategies (Art. 35), and correspondence with transparency and SOR moderation data. The independent audit reports conducted by third party auditors like KPMG, FTI Consulting, Holistic AI, or EY are regulated under Article 37 DSA and act as a control mechanism for the systemic risk assessment reports the platforms issue. Similar as for Article 34 reports we analyse the first issued round of audit reports for Instagram [767].

5.3. RELATED WORK

Mechanisms for platform accountability predate the Digital Services Act (DSA), with early legislative frameworks such as Germany's Network Enforcement Act (NetzDG) and Austria's Communications Platform Law (KoPl-G) mandating transparency reports from online platforms [57, 284]. These laws required semi-annual disclosures of content moderation activities, but were widely criticized for producing non-standardized, opaque reporting structures and category inconsistencies [353, 747, 782]. For instance, Facebook was fined €9 million under NetzDG for inadequate reporting [617].

Alongside these legal obligations, voluntary transparency efforts such as Wikimedia's 2017 report [103], the Santa Clara Principles (2018), and the OECD Voluntary Transparency Framework (2020) aimed to enhance platform accountability, yet lacked enforceability and methodological coherence [28, 568, 589]. These early initiatives demonstrated both the potential and limitations of transparency mechanisms in platform governance [721].

The DSA builds on this background, introducing an enforceable, layered regime of control mechanisms. These include enhanced transparency reports (Articles 15, 24, 42) [546, pp. 177, 298, 490], Statement of Reasons (SOR) databases (Articles 17, 24(5)) [546, pp. 199, 298], systemic risk assessments (Article 34) [546, pp. 392], and independent audits (Article 37) [546, pp. 392], all further harmonized by the Delegated Regulation (DR) under Articles 24 and 88 DSA [224, 225, 230]. The DSA's legal requirements operationalize compliance monitoring, ensure traceability of decisions, and provide regulators with audit-capable documentation.

The DSA introduces tiered reporting obligations. Article 15 sets basic transparency requirements for intermediary services, mandating machine-readable, publicly accessible reports. Article 24 extends these requirements to platforms, and Article 42 imposes comprehensive biannual reporting for very large online platforms (VLOPs) and search engines (VLOSEs). Transparency reports must detail notice-and-action mechanisms (Article 16), the use of automated tools (Article 15(1)(b)), human moderation (Article 15(1)(e)), and provide a breakdown of moderation actions by legal or contractual grounds (Article 15(1)(c)). The Statement of Reasons (Article 17) database further requires platforms to publish justifications for each moderation action, including the legal or contractual basis.

Platforms' contractual governance is also addressed, with Article 14 anticipating a Terms and Conditions (TaC) Database for centralizing platform rules and enhancing transparency [22]. Article 34 introduces systemic risk assessments, requiring VLOPs/VLOSEs

to identify, analyze, and mitigate systemic risks, e.g. impacting fundamental rights, imposing negative effects on the electoral process, or disseminating illegal content. These assessments are subject to independent audits under Article 37 and further detailed in the Delegated Regulation (DR), which standardizes reporting and auditing, risks, and sets standards for methodologies [225].

Scholars have conceptualized traceability, auditability, and verifiability as key pillars of trustworthy algorithmic systems. For example, Kroll et al. highlight procedural regularity as foundational for meaningful oversight [443], while Binns et al. warn that transparency without consistency risks masking deeper systemic opacity [78]. Thus, consistency is not merely a technical goal but a structural precondition for effective compliance monitoring. Besides, Panigutti et al. propose a DSA-tailored risk management framework for algorithmic systems [584]. A further complication is that not every visibility-affecting intervention belongs in the same disclosure channel: some measures appear as content moderation outputs, while others are better understood as recommender-system or curation measures under Article 27. These algorithmic systems are at the core of platform design and through the regulation of the DSA ideally foster user-centered design, which is further explored in a speculative design study of short video platforms by Fabbri et al. [261]. External consistency analysis must therefore ask not only whether a measure is reported, but also whether it is being sought in the correct doctrinal location.

Recent empirical studies of DSA implementation illustrate both advances and persistent challenges. Trujillo et al. analysed 195 million SOR entries (September 2023–January 2024), revealing a 93% automated detection rate for Instagram's moderation and highlighting ambiguities in how automated tools were documented [710]. Drolsbach and Pröllochs studied Instagram's early DSA months, identifying platform-specific moderation differences [205]. Kaushal et al. critiqued vague justification references to platforms' terms and conditions (TaC) [413]. Terzis has further advanced the field by empirically investigating the practices and limitations of algorithmic audits under the DSA, revealing the need for more robust audit methodologies and independent oversight [693]. Also, Groesch et al. highlight the shortcomings of the Transparency Database in their work and stress that the database's objectives ("ensuring transparency, enabling scrutiny over content-moderation decisions, and monitoring the spread of illegal content") are not met [318, p. 1]. Aspromote et al. explore how Large Language Models can be leveraged in using the SOR database [54]. Dergacheva et al. on the other hand focus on the analysis of content moderation decisions logged in the SOR database of social media platforms within 24 hours [189]. These findings underscore both the analytical richness of new DSA data and the urgent need for interpretive, consistency-based tools.

5.3.1. THE DSA'S TRANSPARENCY AND DUE-DILIGENCE ARCHITECTURE

From a doctrinal perspective, the DSA should not be understood as replacing transparency with some wholly new accountability model. Rather, it multiplies transparency-related artefacts and embeds them in a wider due-diligence and risk-management framework [546, pp. 392–406]. As Husovec argues, the DSA is best seen as “a digital due process regulation bundled with risk-management tools,” one that moves the regulatory focus toward the design and governance of systems that generate risks rather than only isolated content decisions [377, p. 1]. This matters for the present study because it clarifies

why the value of layered-consistency testing lies not in any single report, but in the relationships between disclosure, risk articulation, institutional oversight, and auditability across the broader compliance environment. At the same time, both legal and technical scholarship caution against assuming that more transparency automatically yields clearer accountability. Keller stresses the need for “humility” about transparency mandates, noting that legal disclosure requirements can reshape internal platform practices, create significant compliance burdens, and drive de facto standardization in what platforms track and report [420]. Kemper and Kolkman make a closely related point in the technical accountability literature, arguing that “it is possible for organizations to share all available documentation, procedures, and code, yet this will not constitute transparency if that information is not understood by the relevant audience” [422, p. 2086]. They further conclude that “the value of transparency fundamentally depends on enlisting and maintaining critical and informed audiences” [422, p. 2092]. These insights are central for interpreting DSA artefacts, because mismatches may arise not only from poor compliance, but also from the way legal obligations are translated into technical and organizational reporting systems.

In sum, the DSA and its delegated acts establish a multi-layered framework for regulatory control in which consistency can serve as an important condition for effective oversight. Discrepancies across mechanisms—internal, external, or historical—may indicate weaknesses, frictions, or unresolved interpretive issues in platform governance and therefore warrant further regulatory scrutiny.

5.4. THREE LEVELS OF CONSISTENCY: A NEW METHODOLOGY TO EVALUATE THE COMPLIANCE OF DSA REPORTING MECHANISMS

In this section, we answer RQ1: *What methodology can be used to evaluate different forms of consistency of the means of transparency within the scope of the DSA for a VLOP?* Therefore, we propose three levels of consistency, as illustrated in Figure 5.1, for evaluating transparency reporting and the information provided in the DSA Transparency Database for testing compliance under the DSA:

- **Level 1:** Internal consistency: This ensures internal consistency in reporting, controlling for inconsistencies within the report under investigation by, for example, checking sums within a table.
- **Level 2:** External consistency: This tests reporting for consistency across two or more reporting mechanisms by, for example, comparing the total number of suspensions of the service from a transparency report with the correlating number of SORs with the same reason.
- **Level 3:** Historical consistency: This ensures consistency throughout reporting time frames. Information from previous reports can be compared with new reporting data or information. One example of a historical comparison is comparing the average reported number of service recipients between year X1 and year X2.

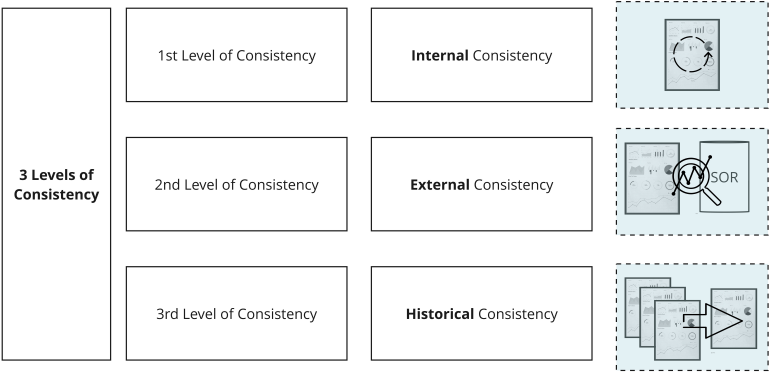


Figure 5.1: Three levels of consistency for evaluating transparency reporting and the information provided in the DSA Transparency Database (SOR).

5.4.1. INTERNAL CONSISTENCY

We assess internal consistency by comparing reported totals with subtotals in Instagram’s transparency reports. In several cases (e.g., Reports 1–4), the sum of disaggregated content removal entries does not match the reported totals, with deviations of 71,400,270 (Report 1), 1,370,980 (Report 2), 1,253,173 (Report 3), and 147,746 (Report 4) instances, respectively. This suggests ambiguity in the ‘Total (including other violations)’ category, which lacks clear legal or contractual definition by only providing a lump sum.

Such aggregation practices appear difficult to reconcile with Annex II of the Delegated Regulation, which mandates that totals must represent the sum of disaggregated categories and that double-counting should be omitted [227, p. 7]. Furthermore, inconsistencies arise from how Instagram presents cross-platform figures. For example, the ‘Number of business content removal measures’ aggregates Instagram and Facebook data, while adjacent tables refer exclusively to Instagram (see Table 15.1.c.(3) [526]). This blending obfuscates accountability and raises concerns about platform-specific traceability of platform-specific questions as outlined in Article 15(1)(c) of the DSA as “meaningful and comprehensible information” for content moderation.

In terms of illegal content reporting under Article 16, broad categories such as ‘Other Illegal Content’ remain vague and fail to align with the 17 defined categories in Annex II of the DR [227, p. 7-11]. Instagram’s failure to distinguish, for instance, terrorist content or hate speech within these submissions reduces legal clarity and conflicts with Article 16(6) obligations to provide granular reporting highlighted by the need for “decisions in respect of the information to which the notices relate” (see Table 15.1.b.(1) in Instagram’s reports [526]).

Additionally, Instagram’s claim that “all Article 16 notices are processed using manual review” [526, p. 9] lacks specificity. It remains unclear whether this applies to all content flagged as illegal under the DSA or only subsets. Instagram’s report demotion and removal statistics further reflect inconsistency in granularity: while some categories are broken down by guideline, others remain vaguely grouped, complicating internal consistency assessments (see Table 15.1.c.(2) [522, 524–526]).

Taken together, the inconsistencies identified in this subsection are most plausibly interpreted as reporting defects and category-design problems. In particular, subtotal-total mismatches, residual lump-sum categories, and vague illegal-content groupings suggest weaknesses in reconciliation logic and disclosure structure within the reports themselves, even if they do not, on their own, establish why those weaknesses arose.

5.4.2. EXTERNAL CONSISTENCY

External consistency refers to the alignment across the four key reporting and compliance mechanisms introduced by the Digital Services Act (DSA): (1) Transparency Reports, (2) the Statement of Reasons (SOR) Database, (3) Systemic Risk Assessment (SRA) Reports, and (4) Independent Audits. Each of these tools serves a different but interrelated role in assessing a platform's efforts to moderate content, mitigate systemic risks, and fulfil legal obligations under the DSA.

Comparing Instagram's Transparency Reports 2 and 3 with the SOR database reveals discrepancies. For example, Transparency Report 3 records nearly 18 million account terminations, while the SOR database lists only 12.6 million for the same period—a difference of over 30% [518]. Similar gaps appear in Report 2. These findings point to problems with data integration and category definitions across reporting mechanisms.

A further source of potential deviation may stem from the legal architecture of the DSA itself. Not all mismatches between transparency reports and SOR data necessarily indicate simple reporting failure. For example, some categories of spam or deceptive high-volume commercial content may be treated differently under the SOR regime than in aggregate transparency reporting, which can complicate one-to-one comparison across mechanisms. This possibility reinforces the need to interpret cross-mechanism discrepancies cautiously and in light of the specific legal design of each reporting obligation. Instagram's 2024 Systemic Risk Assessment (SRA) acknowledges the reporting fragmentation, noting that “reporting systems are not always fully aligned across internal metrics and DSA templates” and that some datasets submitted to the SOR are generated post hoc from logging systems rather than sourced directly from moderation workflows [523, p. 33]. This introduces both definitional and temporal inconsistencies that hinder auditability and transparency.

These limitations are echoed in the 2024 Independent Audit conducted by Ernst & Young. While the audit found that Meta's moderation infrastructure and mitigation protocols were largely adequate, it flagged Instagram's “asynchronous logging systems” as a key challenge to validating SOR completeness [767, p. 18]. It further noted that Instagram “does not yet have a fully standardised process for reconciling discrepancies across transparency reports and SOR entries,” limiting the reliability of cross-mechanism verification.

Compounding these issues is the inconsistent use of terminology and categories between mechanisms. For example, SOR entries often reference ‘account restriction’, while the transparency reports variably refer to ‘account terminations’ or ‘service suspensions’. Platforms like TikTok also report account restrictions without clear disaggregation [705], further complicating cross-platform comparisons. Although the DR attempts to harmonize language by linking transparency template fields with SOR keywords [227, p. 8-11], practical implementation remains uneven.

Moreover, the report identifies difficulties in harmonizing moderation outcomes across EU member states, stating that “risk indicators for political content over-enforcement rely on experimental classifiers that are not uniformly calibrated.” This lack of standardization limits the interoperability of data across transparency, SOR, and systemic risk reports and impairs external consistency evaluations.

Although the Delegated Regulation (DR) aims to reduce such inconsistencies by harmonizing reporting templates and aligning transparency fields with SOR keywords [227, p. 8–11], neither the SRA nor the audit confirms systematic use of these mappings. This raises concerns about the enforceability and comparability of DSA-mandated disclosures.

Beyond numerical inconsistencies, we also identify divergences in how the systemic risk mitigation measures are described and evaluated across the SRA and audit. For instance, the SRA highlights investments in safety-by-design and third-party partnerships as central mitigation strategies [523, p. 6–8], whereas the audit focuses more narrowly on governance and control frameworks, noting limited formalization and documentation of risk-tracking metrics across business units [767, p. 10–11]. Furthermore, while the SRA positions automation as a cornerstone of risk mitigation (especially regarding harmful content and child safety), the audit raises concerns about the lack of transparency and interpretability in automated decision-making systems [767, p. 19]. These strategic misalignments reflect broader tensions between risk framing, organizational accountability, and evidence-based compliance across mechanisms.

Besides, notable inconsistencies emerge in how illegal content and flagging mechanisms are represented across compliance documents. Instagram’s transparency reports provide limited granularity for illegal content removals, often aggregating categories such as “Other Illegal Content” without disclosing specific legal grounds or harmonized classification as required under Annex II of the Delegated Regulation (DR) and will likely change after the DR enters into force after August 2025 [227, p. 7–11]. As a result, external evaluations under Article 15(1)(c) and 16 become more difficult.

These reporting inconsistencies are mirrored in Instagram’s 2024 Systemic Risk Assessment (SRA), which outlines systemic risks such as disinformation, gender-based violence, and civic discourse disruption, yet fails to align these categories with enforcement statistics from the SOR or transparency reports [523, p. 10–15]. While the SRA emphasizes partnerships and safety-by-design, it does not quantify the efficacy of flagging mechanisms in mitigating specific risk vectors or link them to actionable thresholds for illegal content moderation.

Taken together, the observed inconsistencies across transparency reports, SOR entries, the SRA, and the independent audit point to a fragmented and only partially reconcilable compliance landscape. The current implementation of external consistency mechanisms currently limits the reliability of external evaluation of whether Instagram—or any VLOP—is fulfilling its DSA obligations. For external consistency to function as a meaningful compliance metric, platforms must not only synchronize reporting timelines and harmonize terminology but also establish interoperable technical pipelines that allow consistent, verifiable data exchange across all four DSA accountability layers.

From the perspective of the typology introduced in Section 2.1, the external inconsistencies discussed here are analytically more heterogeneous. Some are best understood as infrastructural frictions, especially where post hoc logging systems and asynchronous

submission pipelines complicate reconciliation. Others suggest interpretive divergence, particularly where similar moderation outcomes are labelled differently across mechanisms. A narrower subset may indicate more substantive governance weaknesses where no clear reconciliation process appears to exist across transparency reports, SOR entries, systemic risk assessments, and audits.

5.4.3. HISTORICAL CONSISTENCY

Historical consistency is examined by comparing trends across Instagram's DSA transparency reports (1-4). The total volume of content removals dropped by more than 50% between Report 1 and Report 2, by 67% between Reports 2 and 3, and by almost 45% between Reports 3 and 4 [522, 524–526]. These sharp fluctuations are not explained in the reports, making it difficult to interpret trends or assess data stability over time.

For example, Report 2 introduced a new content category — 'Spam' — that was absent in Report 1. While it accounted for the majority of removals in Report 2, its volume fell by over 80% in Report 3 (see Table 5.3). Categories like 'Firearms' (over 300%) and 'Violent and Graphic Content' (over 900%) exhibited increases between Reports 1 and 2 (see Table 5.7), which may relate to geopolitical events such as the Gaza conflict beginning in October 2023 [507]. However, the absence of explanatory notes for such surges diminishes interpretability.

Similarly, the 'Child Endangerment – Child Nudity and Physical Abuse' category saw a sixfold increase in removals between Reports 2 and 3, but a concurrent 24.3% drop in automation use for the same category. This suggests increased human oversight but is not elaborated upon in the report narratives. While in Report 2 and 3 Meta states that it has around 40,000 moderators employed, it also states that it has "5,400 language agnostic reviewers globally" in Report 2 but only 2,500 in Report 3. Meta's last Transparency Report however indicates 4,600 language agnostic reviewers for their service in Report 4.

Within the typology used in this article, the historical inconsistencies identified here are most plausibly associated with category drift, changing reporting logic, and shifts in internal measurement or documentation practices over time. Without explanatory notes from the platform, sharp fluctuations remain difficult to classify more precisely, but the typology helps distinguish temporal instability from more straightforward numerical reporting error.

In sum, Instagram's reporting across the three consistency dimensions — internal, external, and historical — reflects challenges in both implementation and conceptual alignment with the DSA's compliance logic. These inconsistencies undermine transparency, hinder empirical oversight, and highlight the need for harmonized, standardized reporting practices as outlined in the Delegated Regulations.

5.5. CHALLENGES IN EVALUATING CONSISTENCY ACROSS THE SOR AND TRANSPARENCY REPORTS

This section addresses RQ2 by examining the structural and semantic misalignments that undermine consistency checks across the DSA's four core mechanisms: transparency reports, SOR entries, systemic risk assessments, and independent audits.

Building on the typology introduced in Section 2.1, the challenges discussed below

can be read as recurring sources of reporting defect, interpretive divergence, infrastructural friction, and possible governance weakness. This helps move the analysis from discrepancy detection to a more structured account of why certain consistency problems matter for auditability and compliance evaluation.

5.5.1. STRUCTURAL AND SEMANTIC MISALIGNMENTS

A primary challenge lies in the structural divergence between transparency reports and SOR entries, SRAs and audit reports. For example while transparency reports aim to provide aggregated, platform-level statistics (e.g., account terminations, content take-downs), SOR entries document individual moderation decisions. However, platforms often use divergent terminology and aggregation logics that make reconciliation difficult. For example, Instagram refers to "account terminations" in its transparency reports, while using "account restrictions" in SOR entries. The lack of harmonized taxonomies across mechanisms—despite efforts outlined in the Delegated Regulation's keyword mapping annex [227, p. 8–11]—prevents automated cross-checking and undermines consistency. As Keller argues, late-arriving database API specifications created practical difficulties of interpretation and implementation for platforms subject to the DSA [419]. Additionally, transparency reports often differ significantly in how platforms categorize 'Terms of Service' violations, which complicates comparisons across platforms. For example, as shown in Tables, 5.3. Instagram's categories do not align directly with the standardized categories in the SOR database. This misalignment persists across platforms [710].

5.5.2. INTERNAL INCONSISTENCIES IN QUANTITATIVE REPORTING

Transparency reports frequently include category totals that do not align with underlying subtotals and appear difficult to reconcile with Annex II's rules on double-counting and categorical separation 5.3. For instance, Instagram's "Total (Including Other Violations)" category combines disparate content types without clear attribution to specific contractual or legal grounds. Similarly, automated versus manual decision shares are inconsistently labelled and defined, complicating interpretability of compliance with Article 15(1)(c) DSA. Additionally, it makes historic comparisons less efficient if categories are renamed, merged or split up (see Table 5.3 and [526]).

5.5.3. LACK OF ALIGNMENT ON ILLEGAL CONTENT AND FLAGGING MECHANISMS

These challenges are reminiscent of issues seen in Germany's NetzDG and Austria's Communication Platform Law [57]. As Wagner et al. highlighted, differences in flagging structures for legal and contractual violations skewed transparency reports under NetzDG [747]. Facebook faced penalties for non-compliance with these standards, which underscores the importance of harmonising flagging and reporting mechanisms for more meaningful transparency mechanisms [31, 617]. Similarly, the design of reporting mechanisms in the DSA (Art. 16) could face similar challenges influencing the transparency reporting quality and the SOR data [668]. While Article 16 of the DSA mandates clear reporting of notices concerning illegal content, Instagram's transparency reports group a large volume of removals under generic labels such as "Other Illegal Content", without specifying whether these fall under national laws or contractual provisions [520]. The

SRA, meanwhile, focuses on high-level risks such as hate speech or child safety but does not quantify their treatment through user flagging or automated tools [523, p. 16–18]. The audit corroborates these issues, noting that flagging processes lack documentation of how reports are classified, triaged, and attributed to legal grounds [767, p. 22]. These gaps impair traceability and impede meaningful enforcement under Articles 16, 17, and 22.

5.5.4. MISALIGNMENT OF CATEGORIES ACROSS MECHANISMS

Instagram’s reporting allows for four categories of illegal content in its transparency reports (see Table 5.4), while its ‘Terms of Service’ categories include thirteen distinct violations and an ‘Other Violations’ category. This disparity complicates alignment with the SOR database, which groups ‘Illegal and Harmful Speech’ and other categories inconsistently across mechanisms and jurisdictions. For instance, only six SOR categories matched ‘Illegal and Harmful Speech’ in flagged contents, compared with the 1,090 entries for the same category in Instagram’s Transparency Report 2.

Table 5.9 illustrates the difficulty of aligning systemic risks, mitigation actions, and reporting categories across Instagram’s DSA compliance mechanisms. Many risk areas identified in the Systemic Risk Assessment (SRA) are mapped inconsistently to Statement of Reasons (SOR) and Transparency Report categories. For instance, “disinformation” is linked to several different SOR and transparency categories, while some risks, such as electoral process interference, lack direct matches. This highlights significant gaps in cross-mechanism traceability and the need for harmonized, standardized classification across reporting instruments.

5.5.5. DISJUNCTIONS BETWEEN SRAS AND TRANSPARENCY REPORTING

Systemic Risk Assessment Reports (SRA) and transparency reports are conceptually complementary: the former identify systemic risks (Art. 34) and mitigation strategies (Art. 35), while the latter document enforcement outcomes. However, we find no formal linkage between identified risk vectors (e.g., electoral disinformation, gender-based harms) and moderation outcomes reported by Instagram. The SRA identifies strategic mitigation pillars such as content classifiers and third-party partnerships [523, p. 11–14], yet these tools are not reflected in any disaggregated way in transparency reports. The absence of impact metrics renders it difficult to assess the effectiveness of mitigation and risks obfuscating accountability. Instagram’s systemic risk assessment further illustrates this disjunction by listing multiple mitigation actions—such as “user education on deceptive content” and “enhanced fact-checking workflows”—without quantifying their impact on enforcement statistics or linking them to transparency report categories or SOR reasons. This disconnect reveals structural challenges in aligning risk mitigation strategies with operational transparency, weakening evaluability under the DSA.

For instance, Meta outlines initiatives to downrank “borderline content” that may propagate disinformation without violating platform rules, like Community Guidelines [323]. Yet, such interventions are neither reflected in moderation volumes (e.g. SOR) nor in risk quantification metrics (e.g. discussed in Table 15.1.c.(2) [526] under content demotion measures explicitly), illustrating the absence of traceability between systemic mitigation and platform enforcement outputs [523, p. 8–9].

A further interpretive limitation is that not all information relevant to ranking, demo-

tion, or visibility reduction necessarily belongs within the transparency report–SOR–SRA–audit bundle examined here. Some disclosures concerning recommender systems and ranking logic may instead fall, at least partly, under Article 27 DSA, which requires online platforms to set out in their terms and conditions the main parameters of recommender systems and any options available to recipients to modify or influence those parameters. As Leerssen has argued, forms of visibility reduction such as demotion or so-called “shadow banning” sit at the boundary between content moderation and recommender-system governance, making their legal and analytical classification contested [460, 461]. This means that the absence of a clear trace from systemic risk mitigation claims about downranking or borderline content to transparency-report or SOR categories should not automatically be read as a pure reporting failure. It may also reflect the fact that the relevant transparency obligations are distributed across different DSA instruments, including Article 27 disclosures, which are not systematically analysed in this article.

5.5.6. AMBIGUITIES IN ADVERTISING CONTENT

Transparency reports often fail to distinguish between user-generated and advertising contents limiting insight into the DSA’s new rules on advertising [725, 726, 760]. For instance, Instagram reports combined figures for ‘business content measures’ across Facebook and Instagram [525, p. 13], making it unclear whether content removal data in Tables 5.3 and 5.7 include advertising violations. By contrast, TikTok explicitly differentiates between ‘advertising’ and ‘user’ policies in its Transparency Report 2 [704]. Standardizing these distinctions could improve cross-platform comparisons and enable deeper insight into content moderation practices. This also illustrates a broader limitation of consistency-based evaluation in advertising-related contexts: meaningful assessment often requires additional legal and technical expertise to distinguish user-generated, commercial, sponsored, and platform-delivered promotional content and adequately display targeting criteria [735].

5.5.7. LACK OF STANDARDIZED MEASURES OF AUTOMATED ACCURACY

Platforms report accuracy of automated moderation differently, further complicating evaluations of consistency and measuring compliance. Instagram reports an ‘automation overturn rate’ of 8.4%, whereas TikTok provides ‘specific accuracy’ (98.6%) and ‘error rate’ (1.4%) metrics. Without standardized definitions for accuracy measures, cross-platform comparisons lack reliability and consistency. The typology also has practical value in this context as a follow-up protocol. E.g. internal inconsistencies should trigger requests for reconciliation tables, templates, datasets, and reporting workflow documentation. External inconsistencies should prompt scrutiny of variable definitions, logging pipelines across mechanisms, submission rules, and cross-mechanism mappings and documentation. Historical inconsistencies should lead evaluators to leverage on template version histories, explanatory documentation of changes in processes and systems, and internal notes on measurement changes. Where weak linkage exists between DSA-reporting mechanism claims and downstream outputs, regulators and auditors should request explicit explanation, additional information, and documentation of how measures were tested, monitored, and validated over time.

We have identified and evaluated seven key factors – structural and semantic misalign-

ments, internal inconsistencies, ambiguous categorization of illegal content, disjunctions between systemic risk assessments and transparency reporting, advertising content ambiguities, inconsistent automation reporting standards, and misaligned categorization across mechanisms – because these issues directly affect the reliability, comparability, and auditability of platform compliance data under the DSA. These factors were prioritized due to their significant impact on regulatory coherence and meaningful oversight, as well as their prevalence in existing compliance reporting practices. An overarching theme across these challenges is the critical need for consistency, both conceptually and operationally. The issues identified frequently amplify each other, undermining the mechanisms introduced by the DSA and complicating effective regulatory interventions. Conversely, addressing these interconnected issues in a structured way can enhance trust, facilitate regulatory accountability, and ultimately reinforce platform governance. In the following section, we present targeted recommendations and practical solutions aimed at systematically improving consistency across these mechanisms, thus bridging the identified challenges to a more robust and empirically verifiable compliance framework.

5.6. RECOMMENDATIONS TO EVALUATE CONSISTENCY ACROSS DSA COMPLIANCE MECHANISMS

This section addresses the third research question: *How can a consistency-based approach support effective oversight and auditing under the DSA?* Building on the three-level consistency framework presented in Section 5.4, we propose opportunities to enhance the regulatory effectiveness of DSA reporting mechanisms through greater alignment and mutual reinforcement across tools.

Recommendations for greater consistency should not be read as requiring identical metrics or identical questions across all DSA mechanisms. Transparency reports, SORs, systemic risk assessments, and audits serve partially distinct legal and institutional functions. The aim is therefore not uniformity as such, but sufficient traceability and reconciliation to enable meaningful scrutiny. Consistency in this context should therefore be understood as commensurability rather than sameness [377, 420, 461]. This caution is not merely methodological but institutional. One reason the DSA distributes transparency, risk assessment, and audit functions across multiple instruments is that they allow different actors to examine platform governance from partially distinct vantage points. Some divergence between these mechanisms may therefore be desirable, provided that it is documented, intelligible, and auditable. Moreover, stronger harmonization always entails tradeoffs: if platforms and auditors are pushed too far toward common templates and replicated metrics, compliance resources may shift from substantive inquiry toward reporting conformity and box ticking exercises. A consistency-based approach should therefore aim to improve traceability and comparability without collapsing the plural functions of transparency reports, SORs, systemic risk assessments, and independent audits into a single reporting logic. This is especially important in relation to systemic risks under Articles 34 and 35 DSA, where the goal of consistency should not be to force one-to-one translation of broad lawful-but-harmful risk categories into identical content-level reporting fields, but rather to ensure that platforms explain how strategic risk claims, mitigation choices, and downstream reporting outputs relate to one another.

Current Gaps in Compliance Evaluations European Commission investigations into platforms such as Instagram, TikTok, X, and AliExpress have not yet disclosed specific infringements under Articles 15, 17, 24, or 42 of the DSA. However, our comparative analysis in Section 5.4 reveals systemic inconsistencies across the transparency reports, the SOR database, Instagram’s systemic risk assessment (SRA), and its 2024 independent audit. These gaps suggest the need for more structured evaluation approaches that improve traceability and comparability across mechanisms without erasing their distinct institutional roles.

Integrating Multiple Consistency Checks

The three levels of consistency—internal, external, and historical—can be extended to enable richer evaluations across multiple DSA tools. In particular, we recommend:

1. *Cross-Referencing SORs and SRAs*: platforms should be expected to explicitly reference consistency findings from the SOR database in their systemic risk assessments. For example, discrepancies in moderation reasons or decision volumes could be flagged as a systemic governance risk. Instagram’s 2024 SRA partially implements this but lacks specific quantitative reconciliation. Although Instagram’s risk report concedes that “transparency reporting discrepancies may reflect systemic governance risks,” it stops short of integrating these observations into a structured consistency analysis or remediation workflow [523, p. 15]. Future systemic risk assessments should be required to flag anomalies in DSA reporting—such as divergent moderation volumes or absent removal justifications—as traceable risks under Articles 34 and 35 DSA.

2. *Aligning Audit and Reporting Metrics*: External audits should validate the quantitative consistency of transparency reports and SORs as part of their control assessment. While EY’s audit of Instagram mentions SOR data validation, it stops short of conducting full reconciliation checks. Future audits should be standardized to include consistency assessments that also disclose information of the auditing time frame, scope, data, and methodology used in greater detail, e.g. disclosing information about sampling criteria [661].

3. *Creating a Unified Compliance Matrix*: A structured matrix that links transparency report metrics, SOR entries, SRA findings, and audit conclusions would improve traceability and support empirical accountability reviews. Such a matrix could be integrated into Annex templates provided by the Delegated Regulation (DR) or as stand-alone section in SRAs and audits.

Standardization Priorities

Terminology and Classification Alignment: Shared definitions for moderation outcomes (e.g., account restriction vs. termination) must be enforced across all DSA mechanisms. The DR’s keyword mappings ([227, p. 8–11]) should become mandatory and used across reporting mechanisms. Where visibility reduction or downranking is concerned, future consistency assessments should also consider how Article 27 recommender-system disclosures relate to the moderation and risk-mitigation artefacts discussed here.

Synchronized Reporting Timeframes: The DR already introduces biannual reporting cycles (Article 2), but full harmonization of report publication dates is necessary to allow meaningful cross-mechanism analysis. To prohibit for example that moderation spikes linked to elections are covered by the Transparency Report of one platform, but not of another.

Automation Reporting Standards: As automated systems remain central to moderation, we echo the DR’s call for detailed classifier reporting, including accuracy, precision, and recall ([227, p.25]). SRAs and audits should explicitly assess the performance of such systems using standardized indicators. Additionally, it is essential to show where accuracy is achieved. This includes information about content types (reels, live content, video, comments, images, generative Artificial Intelligence content, etc.) and platform’s policies (terrorist content, disinformation, hate speech, etc.).

Meta’s own systemic risk assessment echoes this need, stating that “without harmonized metrics for automation accuracy and over-removal risk, comparative evaluations across time or region are not feasible” [523, p. 10]. This admission underlines the importance of embedding standardized metadata schemas and accuracy indicators—such as precision, recall, and classifier confidence thresholds—across all reporting instruments as foreseen in the Delegated Regulation [227, p. 25].

By operationalization consistency across multiple layers of the DSA’s compliance regime, we offer a roadmap for more reliable, evidence-based assessments of platform governance. The ability to triangulate insights from different DSA tools supports both proactive risk mitigation and reactive enforcement.

On the other hand, consistency should not be equated with full metric isomorphism across DSA instruments. Transparency reports, SORs, systemic risk assessments, and independent audits serve distinct legal and institutional functions, and some degree of divergence between them may be normatively explainable. The aim of consistency-based testing is instead to secure commensurability: clear definitions, documented mappings, and auditable explanations for why artefacts differ and how we can draw on them together to assess DSA compliance. Pushed too far however, standardization risks reducing methodological pluralism, context dependency, and pushes for a ‘one-size-fits-all approach’.

5.7. LEVERAGING CONSISTENCY FOR REGULATORY EVALUATION

The central message of our article is simple but crucial: *consistency matters* in platform governance under the DSA. Our proposed consistency-based methodology enables multi-layered insights that go beyond surface-level reporting metrics. Evaluating different forms of consistency across four key DSA mechanisms—transparency reports, the Statement of Reasons (SOR) database, systemic risk assessment (SRA) reports, and independent audits—provides a more holistic understanding of how platforms comply with their obligations.

5.8. DISCUSSION

In this article, we analysed the transparency and accountability mechanisms introduced by the Digital Services Act (DSA), focusing on Instagram as a case study. By combining legal analysis with a methodological framework for evaluating reporting consistency, we offer a novel perspective on how platforms disclose moderation activities and systemic risks under the DSA.

The typology developed in this paper adds an analytical layer beyond discrepancy detection by distinguishing which inconsistencies are more likely to reflect numerical

reporting error, category drift, reporting architecture, or deeper governance weaknesses. In that sense, the framework contributes a structured vocabulary for interpreting inconsistencies within DSA compliance contexts.

Addressing the complexities of DSA implementation and compliance requires close collaboration between computer science and law. As this article demonstrates, robust evaluation of platform reporting mechanisms is only possible by combining technical expertise in data analysis, information systems, and algorithmic transparency with legal knowledge of regulatory requirements, fundamental rights, and enforcement processes. Regulators, auditors, and researchers each bring vital perspectives: legal scholars interpret obligations and rights, computer scientists operationalize empirical consistency checks, and auditors synthesize evidence to inform oversight. Interdisciplinary and cross-sectoral engagement is essential, as the task of making DSA mechanisms effective is a collective endeavor—one that depends on ongoing dialogue, shared learning, and the integration of diverse skills and backgrounds to navigate this evolving regulatory landscape.

A further limitation concerns the relationship between broad systemic risk categories and content-specific moderation outputs. Not every inconsistency between systemic risk framing and downstream reporting should be read as evidence that platforms failed to translate risk assessment into identical content-level enforcement categories [401]. As Husovec shows, Article 34 DSA casts a wide net at the level of risk identification, whereas Article 35 mitigation is narrower and constrained by fundamental rights, legality, and the prohibition of general monitoring [379]. In particular, where lawful but harmful expression is concerned, the Commission cannot simply invent new binding content rules through risk mitigation practice. This means that while cross-mechanism inconsistencies may reveal traceability deficits, they do not necessarily imply that every systemic risk should map onto a content-specific moderation category in transparency reports or SOR outputs.

The presence of discrepancies does not permit a one-step inference to legal breach; rather, it requires distinguishing between at least four explanatory pathways: reporting error, interpretive divergence, infrastructural friction, and governance failure. Some inconsistencies may indeed point to weak reconciliation practices or inadequate reporting discipline. Others, however, may arise because the early DSA implementation period was marked by evolving templates, shifting SOR categories, and technical specifications that themselves reshaped what platforms recorded and how they submitted it [419]. A central contribution of consistency analysis is therefore not to collapse these explanations into a single conclusion, but to structure the inquiry into which kind of inconsistency is being observed, what evidence would be required to distinguish among competing explanations, and what kind of compliance concern follows from each.

Summary of Contributions.

We introduced a structured framework for assessing three levels of consistency: internal, external, and historical, with cross-mechanism traceability analyzed as a specific application of external consistency. These mechanisms were evaluated through three primary lenses: legal compliance, practical use case analysis, reporting obstacles, and opportunities for harmonization and improved compliance.

Our legal-doctrinal analysis clarified how Articles 15, 16, 17, 24, 34, 37, and 42 of the DSA, together with the Delegated Regulation, define transparency obligations. Our

methodological approach demonstrated how internal inconsistencies in Instagram's reports (e.g., 'Other Violations' lumping), external misalignment with SOR data, and lack of historical coherence (e.g., sudden category changes like 'Spam') compromise the value of transparency in practice.

Challenges and Limitations. Despite the promise of transparency tools under the DSA, several issues remain unresolved. Disparities in reporting formats, vague or undefined categories, and asynchronous reporting cycles undermine comparability across platforms and even across mechanisms within the same platform. As shown in Section 5.4.2, Meta's SRA and audit report both acknowledge that SOR submissions are often derived post-hoc, raising concerns about the accuracy and origin of reported data.

As demonstrated in Section 5.4, audit and SRA reports are not currently standardized, hindering effective cross-mechanism validation. These issues point to an urgent need for more integrative and standardized governance instruments.

Opportunities for Enhancing Compliance Evaluation. As discussed in Section 5.6, combining insights from transparency reports, SORs, SRAs, and audit reports offers a more holistic approach to evaluating DSA compliance. We argue that the proposed consistency layers enable a more rigorous and triangulated evaluation strategy (if reporting and auditing is standardized in a meaningful way). Aligning definitions, synchronizing timelines, and mandating shared classification schemes could turn these disparate mechanisms into a coherent reporting ecosystem.

This approach also strengthens the evidence base for regulatory interventions. For example, discrepancies in account restriction figures across mechanisms could be flagged for audit, while sudden shifts in moderation categories across time may suggest procedural inconsistencies warranting further investigation.

Interlocking Potential of DSA Mechanisms. The integration of SRAs and independent audits into the DSA ecosystem is not merely a compliance requirement—it is a strategic opportunity. These instruments can act as checks on platform self-reporting and provide regulators with contextual insights that go beyond raw numbers. When triangulated with SOR data and transparency reports, they enable a systems-level evaluation of risk, and procedural consistency.

In future iterations of the DR templates, Guidelines, and best practices policymakers could consider requiring platforms to explicitly reference and reconcile findings across these mechanisms. Doing so would elevate the status of consistency not just as a desirable feature, but as a formal legal expectation.

Balancing Standardization and Innovation. While the need for shared metrics and harmonized terminology is clear, regulators must avoid over-constraining platform innovation—especially with regard to platform-specific Terms and Conditions (TaC). As our analysis shows, standardization should focus on ensuring traceability and comparability, not imposing a one-size-fits-all model.

In sum, we argue that consistency is not just a metric for reporting quality but a proxy for the DSA's regulatory effectiveness. By leveraging all four transparency mechanisms in tandem, we unlock a more actionable and robust understanding of platform compliance.

5.8.1. LIMITATIONS

Consistency checks are only one layer of compliance evaluation and cannot exhaust the complexity of DSA enforcement in practice. In several domains, meaningful assessment requires additional sector-specific legal and technical expertise beyond the DSA itself. This is particularly true for areas such as advertising, recommender system design [556], and the categorization of illegal content, where substantive legality is often shaped by adjacent bodies of EU and national law. In such contexts, consistency analysis is best understood as a tool for surfacing auditability problems and traceability gaps, not as a substitute for the doctrinal and empirical work required to assess the legality of specific forms of content or platform conduct [380, 460, 725].

This is also relevant for ranking and visibility-reduction practices: some aspects of demotion, recommender-system design, and parameter disclosure may be governed partly through Article 27 DSA rather than solely through the transparency, SOR, systemic risk assessment, and audit artefacts analysed here. Accordingly, the present study cannot offer a complete evaluation of all transparency obligations relevant to ranking-related mitigation practices. A further limitation is that this study relies on publicly available compliance artefacts and does not draw on provider-side interviews, internal reconciliation documentation, or access to Meta's underlying reporting workflows. The explanatory classes developed in this article should therefore be read as structured interpretive hypotheses grounded in public-facing materials, rather than as definitive causal findings about why a given discrepancy occurred. This distinction is particularly important in an early implementation period, where mismatches may reflect not only reporting defects, but also evolving internal infrastructures, unresolved interpretive disputes, or transitional compliance practices.

Next to those points, the study has several limitations stemming from the novelty of the DSA framework and the evolving nature of its implementation. First, the misalignment of reporting time frames restricts full comparability across mechanisms. While Instagram's second and third transparency reports (1 October 2023–31 March 2024 and 1 April 2024–30 September 2024, respectively) overlap with the launch of the DSA Transparency Database (September 2023), the first report does not. As a result, historical consistency across all three reports cannot be triangulated with SOR data for the initial period [22, 148].

More generally, the framework proposed here is best understood as a first-order auditability method. Its primary strength lies in testing the internal integrity of reporting artefacts, the external traceability of claims across mechanisms, and the historical stability of categories and metrics over time. It is less well suited to determining the substantive legality of particular moderation outcomes or business practices in sector-specific domains, where compliance may depend on additional legal standards, contextual interpretation, and technical evidence beyond the DSA reporting architecture itself. In that sense, consistency analysis is most valuable as a diagnostic layer that identifies where more targeted doctrinal, technical, or organisational inquiry is warranted and audit testing might focus on in particular.

Second, the overall utility of transparency mechanisms remains constrained by definitional ambiguities. Core concepts—such as “account restriction,” “violation categories,” and “automation overturn rate”—are not standardized across transparency reports, SORs, systemic risk assessment (SRA) reports, and audit statements. This lack of shared termi-

nology limits the ability to conduct fully integrated evaluations across mechanisms.

Third, both the SRA and audit reports analyzed in this study follow non-standardized formats and diverge significantly in the level of detail provided. For example, while the audit document by Ernst & Young acknowledges the asynchronous nature of SOR logging, it does not quantify the potential reporting gap or specify corrective measures. Similarly, Instagram's SRA report describes internal compliance frameworks but does not provide sufficient granularity for meaningful external validation. These gaps limit the strength of cross-mechanism consistency checks.

Fourth, the Delegated Regulation (DR) under Articles 24 and 88 of the DSA was only recently adopted and its harmonized reporting templates will not become mandatory until 1 July 2025. Therefore, this study captures a transitional reporting phase. As future reports begin to reflect DR-aligned structures—including standardized templates, keyword mappings, and clearer definitions—comparability and reliability of compliance evaluations are expected to improve. Our findings provide an early diagnostic of transparency under the DSA. They do not represent a definitive compliance audit.

5.9. CONCLUSION

Transparency mechanisms under the Digital Services Act (DSA) mark a significant regulatory advance in improving accountability in platform governance. Yet, as this study has shown, their effectiveness depends not merely on disclosure, but on *consistency*—across reporting structures, categories, time frames, and platforms. We argue that transparency becomes meaningful only when reporting mechanisms such as the Transparency Reports, Statements of Reason (SORs), Systemic Risk Assessment (SRA) reports, and Independent Audits are not only publicly available but also mutually verifiable.

Consistency should therefore be understood as an auditability criterion and a mechanism for surfacing anomalies across DSA artefacts. Its value lies not in replacing substantive legal assessment, but in identifying where more targeted doctrinal, empirical, or technical investigation is required. Read in this way, consistency does not function as a shortcut to breach determination; it functions as a structured method for making platform reporting environments more interrogable to researchers, auditors, regulators, and civil society actors [420, 461].

Our findings demonstrate that evaluating consistency across three levels (internal, external, and historical) can reveal critical discrepancies and structural gaps in reporting. These inconsistencies highlight both the challenges of implementing the DSA in its early stages and the opportunities to develop a more robust compliance ecosystem. We identified ambiguities in reporting illegal content, mismatches in moderation terminology, misaligned time frames, and unclear accuracy metrics, all of which hinder comparative evaluation and undermine regulatory enforcement.

By analyzing Instagram as a case study, we proposed a novel consistency-based framework for assessing DSA compliance. This framework combines legal analysis with empirical validation across complementary reporting mechanisms. In doing so, it offers a pathway for future research, regulatory development, and audit practice.

Looking ahead, the full potential of the DSA regime will depend on the adoption of harmonized standards, such as those outlined in the Delegated Regulation (DR). Aligning terminology, reporting cycles, and evaluation criteria will enable regulators, researchers,

and platforms to unlock the regulation's transformative capacity, not merely as a formal obligation but as a substantive mechanism for accountability, comparability, and democratic oversight in digital environments.

Table 5.1: Applying the typology of inconsistency to Instagram's DSA reporting

Type	Level	Most plausible explanation class	Why	Follow-up
Subtotal–total mismatches in transparency reports	Internal	Reporting defect / category design problem	Reported totals do not reconcile with disaggregated subtotals, and residual categories such as “Total (including other violations)” remain undefined	Reconciliation logic, source tables, residual-category definitions, reporting workflow documentation
Mismatch between account terminations in TRs and SOR entries	External	Infrastructural friction / interpretive divergence	Reported terminations do not align with SOR-recorded actions; asynchronous logging systems and divergent labels (e.g. account termination/restriction) complicate one-to-one comparison	Internal logging pipeline, submission rules, variable definitions, cross-mechanism keyword mappings
Cross-platform aggregation of content measures	Internal	Reporting design problem; possible governance weakness	Instagram and Facebook figures are combined in otherwise platform-specific reporting, weakening platform-specific traceability	Reporting rationale, source datasets, internal aggregation rules, platform-level disaggregation logic
Sharp category shifts across reporting periods (e.g., spam)	Historical	Category drift / measurement change	Categories appear, disappear, or shift sharply in prominence across reporting rounds without explanatory notes, making temporal comparison difficult	Template version history, category mapping notes, internal documentation of measurement changes
Weak linkage between SRA risks and TR/SOR outputs	External	Possible governance weakness / traceability deficit	Risk claims and mitigation strategies in the SRA are not systematically connected to measurable moderation outputs in transparency reports or SOR data	Internal risk-to-metric mapping, mitigation testing evidence, audit testing documentation, cross-mechanism traceability matrix

Table 5.3: Instagram Content Removals by Policy Category Across DSA Transparency Reports 1–4 (Total and Automated Removals)

Policy Category	Report 1		Report 2		Report 3		Report 4	
	Total	Automated	Total	Automated	Total	Automated	Total	Automated
Adult Nudity and Sexual Activity	950,816	852,471	1,150,038	1,057,008	1,351,522	1,213,764	898,264	778,404
Adult Sexual Solicitation / Explicit Language	–	–	–	–	–	–	97,855	44,416
Bullying and Harassment	837,741	722,502	1,297,196	1,131,702	1,176,634	1,015,909	480,721	329,539
Child Endangerment / Nudity	47,154	14,077	22,582	7,832	133,229	5,927	84,166	22,760
Child Sexual Exploitation	60,364	16,886	145,984	107,142	110,226	82,605	–	–
Cybersecurity	–	–	–	–	–	–	3,406	3,327
Dangerous Orgs – Hate Orgs	46,338	23,433	70,430	41,212	45,839	26,225	–	–
Dangerous Orgs – Terrorism	26,038	21,571	166,098	151,789	76,224	66,356	–	–
Dangerous Individuals and Orgs	–	–	–	–	–	–	100,862	79,908
Fraud and Deception	–	–	–	–	–	–	115,444	32,603
Hate Speech	1,521,669	1,352,204	1,509,400	1,390,887	1,450,002	1,331,715	716,246	639,852
Restricted Goods – Drugs	49,478	44,405	45,810	40,893	32,877	22,290	–	–
Restricted Goods – Firearms	7,094	5,398	33,806	32,150	11,571	10,858	–	–
Restricted Goods and Services (Merged)	–	–	–	–	–	–	145,910	24,538
Spam	–	–	29,628,165	29,622,615	5,200,240	5,076,296	3,242,761	3,236,558
Suicide, Self-Injury	153,051	121,530	92,854	78,994	106,042	71,835	–	–
Suicide, Self-Injury and Eating Disorders	–	–	–	–	–	–	51,638	35,805
Third-Party Intellectual Property Infringement	–	–	–	–	–	–	145,848	41,725
Violence and Incitement	1,188,216	979,568	1,400,781	1,297,277	1,033,996	949,394	556,201	443,079
Violent and Graphic Content	10,184	8,020	105,287	102,390	155,372	146,248	28,833	24,993
Total (incl. other violations)	76,298,413	75,113,462	37,039,411	35,724,613	12,136,947	10,535,490	6,760,857	5,800,826

Table 5.4: Instagram Notices Submitted Through Article 16 Mechanisms in DSA Transparency Reports 1–4 (See Table 15.1.b.(1) in the reports)

Type	Notices Submitted				Content Actioned			
	Report 1	Report 2	Report 3	Report 4	Removed (R1)	Removed (R2)	Removed (R3)	Removed (R4)
Intellectual Property (IP)	231,334	103,244	89,859	43,895	74,336	37,262	29,657	14,729
Defamation	64,966	111,252	62,929	35,753	12,138	20,580	11,953	8,612
Privacy	7,626	29,878	32,267	22,224	1,305	6,134	6,078	3,216
Other Illegal Content	47,477	151,633	137,228	61,888	11,065	28,962	28,396	12,418
Total	351,403	396,007	322,283	163,760	98,844	92,938	76,084	38,975

Type	Content Restricted			
	Report 1	Report 2	Report 3	Report 4
Intellectual Property (IP)	0	0	0	0
Defamation	80	294	117	41
Privacy	4	15	11	4
Other Illegal Content	310	2,199	1,774	655
Total	394	2,508	1,902	700

Table 5.5: Instagram account-termination metrics in DSA Transparency Reports 1–4 and the Statement of Reasons database (data accessed 3 July 2025).

Report period	Reported terminations	Recorded in SoR	Difference
Transparency Report 1	9,506,546	109,919	9,396,627
Transparency Report 2	15,360,549	9,041,495	6,319,054
Transparency Report 3	17,991,379	12,560,641	5,430,738
Transparency Report 4	18,462,723	23,721,798	–5,259,075

Table 5.7: Percentage Change in Instagram's Policy Violations and Automated Removals Across Transparency Reports

Policy Violation	Removed 1→2 (%)	Removed 2→3 (%)	Removed 3→4 (%)	Auto 1→2 (%)	Auto 2→3 (%)	Auto 3→4 (%)
Adult Nudity and Sexual Activity	+21.0	+17.5	-33.5	+24.0	+14.9	-35.8
Adult Sexual Solicitation & Sexually Explicit Language	–	–	–	–	–	–
Bullying & Harassment	+54.9	-9.3	-59.1	+56.6	-10.2	-67.6
Child Endangerment - Child Nudity and Physical Abuse	-52.1	+490.1	–	-44.4	-24.3	–
Child Endangerment - Child Sexual Exploitation	+141.8	-24.5	–	+534.8	-22.9	–
Child Sexual Exploitation, Abuse, and Nudity	–	–	–	–	–	–
Dangerous Organisation - Hate Orgs	+52.0	-34.9	–	+75.9	-36.3	–
Dangerous Organisation - Terrorism	+538.0	-54.1	–	+603.7	-56.3	–
Dangerous Individuals and Orgs	–	–	–	–	–	–
Fraud and Deception	–	–	–	–	–	–
Hate Speech	-0.8	-3.9	-50.6	+2.9	-4.3	-52.0
(Restricted Goods and Services) Drugs	-7.4	-28.2	–	-7.9	-45.5	–
(Restricted Goods and Services) Firearms	+376.7	-65.8	–	+495.7	-66.2	–
Restricted Goods and Services	–	–	–	–	–	–
Suicide and Self-Injury	-39.3	+14.2	–	-34.9	-8.1	–
Suicide, Self-Injury, and Eating Disorders	–	–	–	–	–	–
Spam	–	–	-37.6	–	–	-36.2
Third-Party IP Infringement	–	–	–	–	–	–
Cybersecurity	–	–	–	–	–	–
Violent and Graphic Content	+934.1	+47.6	-81.4	+1176.2	+42.9	-82.9
Violence and Incitement	+17.9	-26.2	-46.2	+32.5	-26.8	-53.3
Total (including other violations)	-51.4	-67.2	-44.3	-52.4	-70.5	-44.9

Table 5.8: Yearly Overview of Organic Content Measures Including 'Other Violations' [Table 15.1.c.(1) in Instagram's Transparency Reports]

Report No.	Description	Total Numbers	Change from Previous Year	Difference
1. Report	First record of total violations	76,298,413	–	–
2. Report	Second record of total violations	37,039,411	Decrease	-39,259,002 (-51.44%)
3. Report	Third record of total violations	12,136,947	Decrease	-24,902,464 (-67.25%)
4. Report	Fourth record of total violations	6,760,857	Decrease	-5,376,090 (-44.30%)

Table 5.9: Cross-Mechanism Traceability of Instagram's Risk Mitigation and Reporting under the DSA

Systemic Risk (SRA)	Reported Mitigation (SRA)	Matching SOR Category	Matching Transparency Report Category
Disinformation and Civic Discourse Harm	AI classifiers for deceptive content; partnerships with fact-checkers; downranking	"Scams and Fraud", "Public Security Risk", "Platform Service Abuse"	Fraud and Deception; Spam; Restricted Goods
Gender-based Violence	Safety-by-design measures; safety centre updates; user reporting channels	"Non-Consensual Behavior", "Violence", "Pornography or Sexualised Content"	Hate Speech; Bullying and Harassment; Adult Sexual Solicitation
Child Safety	Dedicated child safety classifiers; increased human oversight; third-party audits	"Child Protection", "Nudity", "Pornography or Sexualised Content"	Child Sexual Exploitation, Abuse, and Nudity; Suicide and Self-Injury
Hate Speech and Extremism	Risk forecasting tools; external policy partnerships; automated classifier updates	"Harmful or Illegal Speech", "Public Security Risk"	Hate Speech; Dangerous Individuals and Orgs
Electoral Process Interference	Internal policy reviews; political ad transparency updates; partnership escalation protocols	[Not traceable in current SOR tags]	[No electoral or political content category]
Mental Health Harm (e.g. Eating Disorders)	Updated content guidelines; enhanced classifier precision	"Self-Harm", "Violence", "Pornography"	Suicide, Self-Injury, and Eating Disorders

6

DIGITAL SERVICES ACT'S SYSTEMIC RISK ASSESSMENT, INDEPENDENT AUDITS AND AUDITING METHODOLOGIES

"The scholars are everywhere believers, But never succeed in being weavers. He who would study organic existence, First drives out the soul with rigid persistence; Then the parts in his hand he may hold and class, But the spiritual link is lost, alas!"
"Wer will was Lebendigs erkennen und beschreiben, Sucht erst den Geist heraus zu treiben, Dann hat er die Teile in seiner Hand, Fehlt leider! nur das geistige Band."

Mephistopheles, Faust, Part I, IV. Studierzimmer/The study, Johann Wolfgang von Goethe,

A central part of the European Union's Digital Services Act (DSA) is the requirement for internal and external audits of online platforms. A key component of these audits is the analysis of platform systemic risk assessments, which cover wide-ranging and multifaceted topics such as the dissemination of illegal content, negative effects on human rights, the electoral process, and gender-based violence. The Delegated Regulation on how to conduct such audits under the DSA further specifies what is expected from platforms and auditors to ensure compliance with the law. This article discusses the pros and cons of different quantitative and qualitative methods that could be used to audit the varying systemic risks of online platforms and search engines under Europe's DSA regime. Following an evaluation of key methods, we propose an empirical mixed-method approach based on risk-specific sampling techniques combined with (legal) content analysis as a reliable auditing method for the DSA. We argue that audits should be based on the evaluation of content

samples representing spot checks for systemic risks. To develop this mixed-method approach, we first analyse the European provisions relevant to auditable content samples and their selection. Then, we explore sampling techniques that can be meaningfully applied in internal and external DSA audits and assess methodological techniques for testing systemic risk categories. Furthermore, we examine how the representativeness of a sample can be understood in different disciplines. Finally, we analyse the first systemic risk assessment reports of platforms with a view to understanding how the tests and methodologies (such as sampling techniques) used in these assessments evaluate systemic risks. By proposing a novel empirical mixed-method approach for conducting DSA audits, this article addresses the complex challenges of evidence-based risk assessment in content moderation. Furthermore, we contribute to the rapidly growing field of DSA compliance assessments, emphasizing the specificities of audits and systemic risks. In doing so, we recognize that there is no “one-size-fits-all” solution for DSA audits, but rather a use case-specific and platform-tailored approach is needed.

6.0.1. INTRODUCTION

The Digital Services Act (DSA) [224] represents a cornerstone of Europe's efforts to regulate online expression and intermediary services and aims to harmonize rules for user-generated and synthetic content on online platforms. The DSA aims to increase transparency and protect users from illegal content online. It aspires to achieve the so-called "Brussels effect", potentially influencing global regulatory standards [565]. It entered into force on February 17, 2024. The DSA applies to *intermediary services* operating within the European Union (EU) according to Art. 2(1). These services offer network infrastructure, such as hosting platforms (for example, cloud services) or online platforms (such as app stores or online marketplaces). Importantly, these services also include Very Large Online Platforms (VLOPs) and Very Large Search Engines (VLOSEs) [224], which face the most stringent requirements. They are defined as platforms with over 45 million average monthly users (or 10 % of the EU population) in the European Union (Art. 33 DSA). The European Commission published (and continuously updates) a list of designated VLOPs and VLOSEs, including platforms such as Amazon, LinkedIn, Wikimedia, and Facebook [141].

Therefore, the Regulation adopts a layered approach to obligations based on platform size and function, considering an increased risk for society due to content that is accessible to a large number of users. The DSA includes various new obligations for users, businesses, and platforms. Key provisions include the adaptation of Terms and Conditions (TaC) (Art. 14), transparency obligations (Arts. 15, 24, 42), the design of notice and take-down mechanisms (Art. 16), platforms' content moderation decision reporting (Art. 17), Trusted Flagger involvement (Art. 22), recommendation system disclosures (Art. 27), researcher access for independent evaluations (Art. 40), ex ante risk assessments (Arts. 34, 35), or third-party audits (Art. 37). These external audits are done by auditing firms like KPMG, EY, or Deloitte [362]. Recent enforcement actions underscore the DSA's impact. For example, on December 18, 2023, the European Commission initiated formal proceedings against X (formerly Twitter) for noncompliance, citing risks related to illegal content dissemination (Art. 34(1)(a)) and threats to civic discourse and electoral processes (Art. 34(1)(c)) [146]. The Commission also opened proceedings against TikTok on election risks [232] and against Temu for the sale of illegal products, their addictive platform design, failed disclosure of recommendation system, and failed compliance of researcher data access [231]. This paper focuses on sampling methodologies under the DSA, addressing two key research questions: (RQ1) What sampling methods are most relevant for assessing compliance under the DSA for internal risk assessments and independent audits?, and (RQ2) How do systemic risk assessment reports and audit reports from VLOPs and VLOSEs address the use of sampling methodologies, and what are the implications of their variability for the standardization of audit practices? (RQ3) The following sections provide an in-depth exploration of these questions.

This chapter is based on the preprint: Sekwenz, M.-T., Gsenger, R., Dahlgren, S., & Wagner, B. (2025). "Doing Audits Right? The Role of Sampling and Legal Content Analysis in Systemic Risk Assessments and Independent Audits in the Digital Services Act." doi:10.48550/arXiv.2505.03601. Supplementary material appears in Appendix B.

6.0.2. BACKGROUND

Various research endeavours have already looked into audits in the online space in general and the auditing requirements of the DSA specifically. Research has focused on auditing algorithms [383, 692] and their challenges. More recently, AI audits have emerged as a discussion point, as increasingly, approaches are proposed to audit AI systems [262, 451] to increase fairness and accountability [113]. Furthermore, some research elaborates on increasing auditability, for instance, in extended reality (XR) applications [310, 563]. These studies have highlighted issues with accessing systems, highlighting that black-box access is not enough [117]. AI can be used to perform audits [474]; however, these audits can be biased [321]. European Regulations are not the first to establish auditing as a mechanism of transparency. Moreover, the DSA, the Digital Markets Acts [240], and the AI Act [241] include auditing provisions. Effective auditing under the DSA requires a robust framework that ensures auditor independence and competence. Kaushal et al. propose guidelines for audit quality assurance, including performance standards and quality control procedures, which are essential for credible assessments [413]. They provide an empirical analysis of the DSA's transparency database, highlighting the legal obligations and the role of automated systems in ensuring compliance. Terzis et al. highlight that although the auditing mechanisms are not new, they have a larger scope than other legally required audits. However, they only apply to a smaller number of audited entities (namely, to VLOPs and VLOSEs) [692]. Moreover, to align sampling with specific risks, such as the dissemination of illegal content (Art. 34(1)(a) DSA), auditors must ensure that samples represent the population accurately, particularly in ex post content moderation scenarios. Studies such as Mubarak et al.'s work on disinformation emphasize the value of comparing deleted and undeleted content samples [544]. Various challenges in auditing remain, and as such, "auditing the auditors", finding best practices, and standards are still necessary [171]. Additionally, no public participation or involvement of data subjects is intended, even though data access for vetted researchers, as foreseen by Art. 40 DSA, is a first step towards some sort of public involvement [113]. Finally, the involvement of traditional auditors in algorithmic audits introduces complexities, such as potential conflicts of interest and the need for specialized expertise, which must be addressed to ensure effective oversight [692]. Per Art 34 and 37 DSA, audits assess content moderation quality, addressing risks such as illegal content the assessment of content moderation quality in detail to understand if, for example, the risk of disseminating illegal content is adequately addressed. This process, however, is criticized for being opaque and arbitrary [295]. As a result of the volume and diversity of online content [199], challenges of varied risk types [34, 80] and dynamic time constraints arise. The Delegated Regulation (DR) accompanying the DSA provides additional guidance for conducting audits, emphasizing the need for transparent frameworks and methodologies [226]. It clarifies the need for frameworks and methodologies used in the process of assessing risks and auditing, including a justification for why specific methods were chosen. These methodologies should include tests to assess compliance with the DSA (Art. 1(c) DR). Sampling methods, explicitly mentioned in Art. 10 (DR), are integral to these evaluations and widely discussed in the academic literature [77, 97, 178, 595]. The following sections will provide an overview of sampling under the DSA, as well as proposing mixed methods solutions to deal with legal requirements of the law regarding methodological details for systemic risk

assessments and independent audits.

6.0.3. AUDITING VERY LARGE ONLINE PLATFORMS AND VERY LARGE SEARCH ENGINES IN THE DIGITAL SERVICES ACT

The DSA establishes a rigorous framework for auditing VLOPs and VLOSEs; however, in several aspects, the framework is not unambiguous [692]. These auditing processes have two central components: internal systemic risk assessments (Art. 34 DSA) and external independent audits (Art. 37 DSA). These processes culminate in the generation of key deliverables, including risk assessment reports prepared by platforms (Art. 42(4)(a) DSA; Rec. 100) and audit reports authored by independent auditors (Art. 37(4) DSA; Art. 6 DR). If an audit report identifies deficiencies, platforms must implement corrective measures, document the process in an audit implementation report, and undergo re-evaluation (Art. 37(6) DSA). Systemic risk assessments are mandated annually under Art. 34(1) DSA and address four critical categories: (a) dissemination of illegal content, (b) adverse impacts on fundamental rights, (c) negative effects on electoral processes, threats to civic discourse and public security, and (d) harms related to gender-based violence, public health, and individual well-being. Platforms must evaluate their algorithmic systems (Art. 34(2)(a)), content moderation mechanisms (Art. 34(2)(b)), enforcement of Terms and Conditions (Art. 34(2)(c)), advertising systems (Art. 34(2)(d)), and data-related practices (Art. 34(2)(e)) to address these risks effectively. This layered approach reflects broader discussions in the fairness literature in the FAccT community about designing contextually aware and compliant systems. For example, Laufer et al. highlight the need for interdisciplinary collaboration to effectively address challenges like systemic risks, a principle that underpins the DSA's framework for compliance and auditing [455]. Platforms must also implement mitigation measures for identified risks (Art. 35 DSA). Independent audits, as outlined in Art. 37 DSA, are tasked with verifying compliance across a comprehensive range of requirements outlined in Arts. 10–48 of the DSA. These requirements reflect a critical shift from merely enforcing compliance to addressing the broader socio-technical dynamics that influence systemic risks, as Rateike et al. explored in their discussion on designing long-term fair policies in dynamic systems [611]. The audit process is integral to the DSA's enforcement strategy, ensuring that platforms not only comply with existing regulations but also actively address systemic risks. A “positive” audit outcome reflects full compliance, while a “negative” result necessitates immediate corrective actions and further evaluations. By embedding these auditing mechanisms, the DSA establishes a robust accountability framework to enhance the safety and transparency of digital platforms operating within the European Union [224]. This framework aligns with calls for accountability mechanisms in socio-technical systems that incorporate human oversight and actionable feedback loops [78].

LEGAL REQUIREMENTS FOR SAMPLING IN THE CONTEXT OF SYSTEMIC RISK ASSESSMENTS AND INDEPENDENT AUDITS

The DSA empowers the European Commission (EC) to provide additional rules through delegated regulations (DRs) (Art. 87), facilitating clarification and specificity in compliance processes, for example, the DR for researchers' access to platform data [228]. The EC's power has already been exercised to outline rules for conducting independent

audits (Art. 37(7) DSA)[226]. Auditing organizations must follow an audit procedure defined in Rec. 16 DR, which includes techniques such as data collection, methodological application, substantive analytical procedures, and other actions to gather and analyse evidence, excluding the issuance of opinions or reports. These methodologies should reflect the measures VLOPs and VLOSEs implement to mitigate risks (Art. 14 DR). Additionally, auditors are required to document and report any reasonable doubts arising from their assessment of internal controls and analytical procedures (Art. 13(2)(b) DR). The DR emphasizes the importance of robust methodologies, including qualitative and quantitative elements, in analysing systemic risks. For instance, Art. 3 DR delineates the scope of audits, requiring auditors to describe specific elements audited and the methodologies applied (Art. 37(4)(d) DSA; Art. 10 DR). Methodologies must link sampling to the minimization of detection risks (Art. 12(1) DR) and ensure alignment with the specific risks under investigation, such as the dissemination of illegal content (Art. 34(1)(a) DSA).

Sampling as a Methodology in Audits Sampling is explicitly recognized as a critical audit methodology to strengthen evidence (Art. 2(6) DR). Sampling methodologies are central to identifying risks that might otherwise remain obscured, particularly in algorithmically moderated environments where decisions are opaque [282]. Effective sampling ensures that a quantitative sample is representative of the content moderation quality of a platform, thereby enabling qualitative assessments of risks as outlined in Art. 34 DSA. Sampling methodologies must aim to minimize detection risks (Art. 12(1) DR; Art. 2 and Art. 3(2) (c) DR) and justify sample size and technique in audit and risk assessment reports (Art. 12(3) DR). The methodologies must account for materiality thresholds (Art. 10(2)(a) DR) and detail the analytical tests and procedures employed. Moreover, Art. 10(2)(b) DR demands the description of tests and other analytical procedures an auditing organization intends to use during their evaluation. According to Rec. 17 DR, a test is “an audit methodology consisting in measurements, experiments or other checks, including checks of algorithmic systems, through which the auditing organization assesses the audited provider’s compliance with the audited obligation or commitment”.

Regional and Contextual Specificities Art. 13 DR highlights the need for contextualized risk assessments, including regional and linguistic factors, particularly when analysing civic discourse and electoral debates (Art. 34(1)(c) DSA). For example, auditors must account for regional disinformation narratives, electoral laws, and political controversies [444, 499]. Art. 35(2) DSA further emphasizes the importance of presenting systemic risks by Member State and at the Union level. By considering these factors, auditors ensure that sampling methodologies are representative and relevant to the contexts being studied, aligning with Art. 13 DR’s mandate to address “probability and severity of risks”.

Methodological Requirements and Assurance Auditors must describe how information sources were selected and assumptions tested with impacted groups (Art. 13(1)(a)(v) DR). Methodologies must address algorithmic systems, content moderation practices, Terms and Conditions enforcement, advertising, and data practices (Art. 34(2) DSA). Art. 10(2)(a) DR mandates that methodologies make sense of data both quantitatively and qualitatively, with substantive analytical procedures used to assess compliance risks (Rec. 18; Art. 10(2)(b) DR). Auditors must provide a “reasonable level of assurance” in their evaluation of a platform’s controls to mitigate identified risks (Art. 13(2) DR). Sampling

methodologies, while critical, must be transparently documented in both risk assessment and audit reports to ensure methodological rigor and alignment with the DSA's objectives. By emphasizing rigorous sampling and contextual specificity, the DSA and DR ensure that systemic risk assessments and independent audits are robust, representative, and capable of addressing the diverse challenges posed by digital platforms.

6.0.4. MIXED METHODS – COMBINING STATISTICAL CONTENT SAMPLING TECHNIQUES AND CONTENT ANALYSIS FOR SYSTEMIC RISK ASSESSMENTS AND INDEPENDENT AUDITS

This section addresses research question two: What sampling methods would be most relevant for internal risk assessments and independent audits? According to Webb and Wang, the aim of sampling is the representativeness of the sample itself that presents the “[. . .] incidences selected accurately [and] portray[s] the population” [750, 98]. In an ideal world, those analysing the sample would be working with a perfect ground truth sample that also includes information, for example, about the appropriate time frame for data collection or geographical dimension of data collection and context. However, such an ideal ground truth sample is usually not attainable [771]. This is not only because of the imperfect access to data from a third-party auditing perspective but also because platforms themselves with assumed ideal knowledge can have problems attaining a “perfect” ground truth sample. This is not only due to the subjectivity of the decision of what constitutes a violation of the law or TaC [445] but also because of historic data bias, geolocation blocking, or protection from legal liability [281].

In the area of social media research, different methodological approaches exist on how to collect and analyse data. Data analysis within the context of systemic risk assessments and independent audits leads, according to Piña-García et al., to a knowledge gap regarding the data needed, the quantity of it, the questions and queries to ask, or the protection or quality of privacy [595]. Sampling techniques used within the context of the DSA, however, can have several ways to create a sample that meaningfully represents the population and must take the individual systems, content, and platform design into account. According to Rougier, risk assessments can “be made with a varying degree of confidence, reflecting both the innate difficulty of the task, and the amount of resources available” [628, 1081].

Online platforms samples can be drawn before publishing the content online (ex ante sample) or after the platform's internal moderation process is in place (ex post sample). While the first content sample includes a realistic picture of the violations and content uploaded to platforms, the second content sample provides a snapshot of what users will see after content moderation has taken place. Furthermore, ex post content moderation samples do not include already filtered-out content. This content class is either deleted due to violations of the TaC or the law, or can be geo-blocked in certain areas. Both sampling time frames can provide meaningful information for assessing systemic risks and DSA compliance of platforms. While an ex ante sample might indicate what systemic risks are prevalent, trending, or novel, ex post content samples might provide further insight into internal content moderation quality, challenges, and areas for improvement. Ex ante sampling can also help identify potential algorithmic vulnerabilities before they result in significant harm, an approach advocated for in fairness auditing [606].

Sampling which uses random chance is called probability sampling, in which repeated sampling using the same method results in different samples. Simple random sampling is the most intuitive method of random sampling: each member of a population has an equal chance to be included in the sample. Nonprobability sampling methods are deterministic in nature and always result in the same sample. The most common non-probability sampling method is convenience sampling, where a researcher takes members of the population that are easiest to access without any use of random chance. Simple random sampling can be used to create an auditable sample, for instance, to evaluate the prevalence of illegal content and TaC violations [544]. This method of sampling does not take “auxiliary information on the population” into account when selecting a sample [222, 12]. Such a sampling method could be used for both *ex ante* and *ex post* samples to understand the overall content uploaded on platforms [238, 541, 544, 594]. This could contribute to understanding what violations are common on a platform. However, there are also downsides to random sampling techniques, for example, if unequal inclusion probabilities harm the representativeness of a sample [309].

Ex post content moderation samples are more common in most research domains because of the data access and data collection possibilities. Researcher access depends on a platform's willingness to provide an API and data access. To illustrate the use of a social media platform's API, X (formerly Twitter) is a useful example, as social science research focused predominantly on the social media platform X, before free access to its API was terminated in 2023 [594]. Researchers usually choose a corpus that best fits to answer their research question [41], for instance, by choosing specific hashtags that refer to the topics they are researching [500]. This approach, however, is open to biases and might not represent every aspect of a topic or conversation. Moreover, data returned by the API might be subject to inherent biases that are not detectable for external researchers [605].

When accessing the API of X, only publicly available information (i.e. Tweets and metadata) can be accessed; a banned account, for instance, would not be included in a dataset [594]. However, the sampling strategy influences the quality of the data. To choose a representative sample, Rafail [605] proposes first defining a population pertinent to the investigation. That might be an “unbounded population”, which is best used with random sampling and includes every user and tweet. More often used in research are nonprobability sampling strategies that focus on “semi-bounded populations”, for which either the users or the content is chosen according to clear criteria pertaining to the research question. The users are chosen by consulting sources outside of the platform. Content restrictions are included by using hashtags. Finally, “bounded populations” focus on a sample of a very restricted user group, which focuses on a group of accounts. *Ex post* content moderation samples are limited as the sample is only collected as publicly accessible in that moment.

However, Pfeiffer et al. [594] have shown that a large number of posts are removed either by platforms or users. For risk audits in the context of the DSA (Art .35(c)), deleted posts are important for platforms to show compliance with the provisions of risk mitigation by means of adapting content moderation systems. In contrast to simple random sampling, the probability proportional to size sampling method takes information about the population of a sample into account to create an auxiliary variable that determines

some aspect in relation to size and the population being studied. For creating representative samples with this technique, more accurate information about a population is needed to correctly estimate inclusion probability. In a scenario with perfect knowledge about platform violations, for example, the reported numbers of orders received according to Art. 9 and 10 DSA about illegal content in a specific Member State could be known, [297] which might facilitate the estimating of good measurements for probability sampling to create a representative sample of auditable platform content. Following this logic, the prevalence of orders that refer to violations of the law could be used as a criterion to interpret samples' underrepresentation or overrepresentation of problematic legal content. However, it is highly likely that even the platforms themselves do not have perfect knowledge about themselves in that sense, for example, in cases of terrorist content and deletion obligations which link platforms' liability starting only after they receive removal orders per Art. 3 [239] as reflected by the non-existence of a general monitoring obligation of platforms for their uploaded content according to Art. 8 DSA.

When a population has substantially different subpopulations called strata, sampling within each stratum independently may have statistical and practical advantages. Stratified sampling allows different sampling strategies within strata. Stratified sampling in comparison to the other sampling techniques presented makes use of information about a population, but also considers the specificities of a sample by dividing the population into separate strata [222, 12]. According to Bilon and Clemente, this approach could start from "the date and time the post is created, content type, message, link, and post story" for which "content type" is the simplest category to base the strata on [77, 52]. These divided groups are called strata and are "non-overlapping subpopulations" [222, 12]. The creation of such groups, also called strata, can be based on additional information about the data, such as content type, and can increase precision [77]. Therefore, stratified sampling allows different sampling needs for subpopulations and can be adjusted, for example, to specific systemic risks or content types. In the case of systemic risks such as the groups most impacted in line with Art. 13 DR, regional aspects of content according to Art. 13(2) DR or correlating languages of Member States might be useful categories for defining strata. Content types, on the other hand, could include the selection of images, video content, or comments to be included in a sample that addressed a specific risk, for example, gender-based violence. Grafström and Schelin highlight the limitations of stratified sampling as a method. According to them, the classes would get too numerous or small to be able to meaningfully interpret the results [309]. Additionally, the sampling time is highlighted in the context of content samples on Facebook according to Bilon and Clemente for posts "as elementary unit" under days, weeks, and simple random sampling (with and without replacement) [77, 44].

ON THE ALIGNMENT OF SAMPLING TECHNIQUES AND SYSTEMIC RISKS

Because systemic risks are different in nature, aligning them to sampling methods is a critical part of preparing audit methodologies. One example of sampling criteria involves testing compliance with Art. 23 DSA, which sets rules for handling users who repeatedly exhibit malicious behaviour on platforms. Sampling can focus on user accounts, including those not classified as repeatedly violating Art. 23 DSA and those that have exhibited such behaviour [772]. User IDs or account handles can serve as legitimate sampling criteria to

construct an auditable sample representing, for example, the risk of accounts spreading misinformation.

The challenge of Member State-specific risk Another important aspect of sampling involves Member State-specific risk domains, which requires sampling based on geographic location (e.g. an uploader's location or a recipient's location if no VPN is used), content targeting specific Member States, or content language [444]. Content evaluation can have many layers: for instance, a TikTok post might include German background text, spoken French, English and French hashtags, and an English description, as reflected in event- and language-based sampling practices [705]. Comments can vary in language, and emojis might hold different meanings depending on the language context [35]. These linguistic complexities affect keyword selection, abbreviations, and acronyms, which are critical for creating meaningful samples for compliance testing under the DSA.

The challenge of keyword/hashtag selection Hashtags, while sharing challenges with keywords, provide additional layers of content metadata and can inform event-based testing [705] and moderation decisions [34]. Similarly, sampling based on content similarities can enhance audits in line with content hashing initiatives [64, 285, 306, 459, 757].

Sampling based on meaningful time frames is another challenge (see Table 6.1). For instance, auditing systemic risks affecting electoral processes (Art. 34(1)(d) DSA) requires ensuring data collection coincides with election periods in Member States [705]. Missing relevant time frames could result in underrepresentation of violations in a sample and increase detection risks, undermining an audit's accuracy and relevance.

The challenge of capturing context Content specificity must also inform sampling decisions in line with Arts. 12 and 13 of the DR. Content samples may include main posts, threads, sub-threads, comments, likes, or warning labels, as discussed in Section 6.0.7. Auditors may require additional contextual information, such as chronological timelines, to draw meaningful conclusions for specific risk categories. For example, a seemingly neutral post like “yes, let's do that” might reveal a violation when linked to prior antisemitic statements. Conversely, counter-speech, which may appear violative in isolation, could only be understood as compliant in context [742]. Chronological and contextual analysis is thus essential for assessing compliance meaningfully within digital conversations.

The challenge of sampling interval selection The feasibility of analysing sample sizes also varies with content type. For example, YouTube's diverse content – ranging from Shorts (brief videos) to long-form videos and live streams – creates challenges in determining appropriate sample sizes [382]. Feasibility considerations must be justified under Art. 12(3) DR.

The challenge of violation location Additionally, the positioning of violations within content varies across platforms. On X (formerly Twitter), violations could appear in the main post, images, text within images, hashtags, or comments. These variations, coupled with platform-specific designs [e.g. Pinterest's pins], algorithms, and features, add complexity to the creation of representative sampling methodologies. Lovato et al., for example, compare the concept of compliant user consent across platforms [485]. Auditors must consider these factors to develop auditable and representative samples and explain their rationale. These examples underscore the complexity of selecting and applying sampling methods for compliance testing under the DSA. We propose using

(legal) content analysis as a substantive analytical procedure for auditing VLOPs' and VLOSEs' compliance (e.g., [299–303, 677]). This approach can provide valuable insights into systemic risks identified in Art. 34 DSA [444].

LEGAL CONTENT ANALYSIS

Legal content analysis adapts methods and traditions from disciplines such as political science and sociology [61]. The emphasis on legal content analysis aligns with Bogiatzis-Gibbons' advocacy for reclaiming democratic oversight in AI systems for the purpose of achieving oversight of content moderation systems. This approach can guide the development of codebooks and coding practices that align with democratic values and legislative requirements under the DSA [86]. Through this approach, empirical legal scholars aim to bring the rigor of social science to the study of law [99, 1]. Brook posits that content analysis bridges traditional legal methodologies, such as doctrinal research, with quantitative empirical elements [99, 3]. Hall and Wright highlight its value in increasing objectivity within legal research [335]. Content analysis involves emergent coding, in which codes arise from data itself (e.g. analysing reasons for violations in a sample), or a priori coding, which uses predefined criteria such as a codebook [633]. Hall and Wright further discuss the iterative process of creating and refining codebooks to categorize data [335], with testing and adjustments ensuring clarity and usability [326].

Bar-Ziv outlines three stages of content analysis: (1) Data making, (2) Inferring, and (3) Narrating [61, 475]. Salehijam, however, emphasizes formulating a research question as the initial step, followed by data collection, coding, reasoning, and, finally, presenting findings to the target audience [633]. Hall and Wright argue that content analysis enables legal scholars to derive qualitative insights and, subsequently, quantify these for deeper legal analysis, as needed for DSA audits [335]. Systematic application of content analysis is often referred to as Systematic Content Analysis (SCA), a replicable technique for analysing texts, including case law and legislation [633, 35].

While effective, content analysis has limitations. Brook cautions against the lack of comparability across fragmented legal domains [99, 1], an issue relevant to European Member State law and the DSA. Bar-Ziv highlights the problem of missing data, such as cases resolved outside courts [61]. For DSA-related audits, analogous issues arise when deleted content or risks are excluded from samples, potentially skewing metrics and underrepresenting risks. Ensuring robustness in content analysis involves coding samples by multiple individuals to compare subjective decisions, with inter-coder reliability measured using metrics like Krippendorff's alpha or Cohen's kappa [369].

For systemic risk assessments and external audits, we propose a priori coding with risk-specific codebooks tailored to the risks being assessed. For example, codebooks should include legal provisions linked to deletion or geo-blocking obligations under Art.34(1) (a) DSA for risks such as disseminating illegal content. Non-legal risks, like negative effects on electoral processes, require bespoke coding schemes [444]. After developing risk-specific codebooks, content analysis should involve double-coding each sample by at least two coders. Measuring inter-coder reliability ensures transparency and robustness in subjective decisions during the auditing process. We further propose a "Moderation Decision Supervisory Team", comprising experts on the analysed risks, to resolve disagreements between coding teams. For instance, if legal coding team 1 does not

classify a video as violating the TaC, but legal coding team 2 does, the supervisory team would make the final determination. This structure enhances the quality and reliability of audits by enabling external auditors to generate both qualitative and quantitative insights into the content moderation practices of VLOPs and VLOSEs, aligning with Art. 10 DR.

6.0.5. ENSURING REPRESENTATIVENESS OF CONTENT SAMPLES

In general, representativeness ensures that the sample reflects the population characteristics on all the variables of interest [309, 278]. According to Art. 12(2) DR, both the sample size and the methodology used to assess it must be representative. However, as discussed in earlier sections, sampling methods and required sample sizes vary based on platform, content type, risk assessed, content distribution, assessment period, and the availability of historical or real-time data. These diverse requirements underscore the complexity of creating representative samples for auditing. Notably, the selection of sample size and methodology often depends on the subjective judgment of auditors [628, 1084–1085] and may be influenced by differing views on “how to do audits right”.

As highlighted by Pfeffer et al., the quality of sampling and methodology can decline if data are not collected continuously, leading to temporal bias in the sample. Their findings indicate that data availability decreases by approximately 2% per month within the first 18 months after content publication [594, 6]. Similarly, the choice of API for data sourcing (e.g. Streaming API vs Academic Research API) can result in varying data samples [37].

This decay in data availability starts almost immediately. For example, in one instance, “1 out of 30 Tweets collected 24 hours after posting was unavailable on the Academic API 10 seconds after posting” [594][49:7]. Additional challenges include representativeness issues arising from distributed data centres, differing geo-blocking policies, and accessibility conditions. Under such circumstances, achieving a highly representative sample, let alone a perfect ground truth sample, becomes increasingly difficult. Pfeffer et al. also highlight the challenges posed by deplatforming user accounts. For example, when U.S. President Trump’s Twitter account was deplatformed, metadata associated with the account (e.g. retweets) was also rendered unavailable for analysis [716]. Such scenarios align with risks identified under Art. 34 DSA, including malicious user behaviour and the frequent provision of manifestly illegal content per Art. 23 DSA. While larger sample sizes may mitigate some sampling method challenges via the law of large numbers, the methodology itself significantly influences sample quality and, consequently, representativeness [496]. For example, Siersdorfer et al. conducted an in-depth study of commenting and rating behaviour using a sample of over six million comments on 67,000 YouTube videos, utilizing metadata such as titles, tags, categories, descriptions, upload dates, and user engagement statistics [673]. Moreover, qualitative analysis can meaningfully enhance representativeness [182].

Given these challenges, creating a representative sample under the DSA requires detailed documentation and robust arguments tailored to individual audits. There is no universal definition of sample representativeness; instead, auditors and platforms must perform context- and risk-dependent analyses that are both visible and transparent [259, 266]. This aligns with Art. 12(3) DR’s requirement to justify the selection of sample size and methodology, as well as Art. 13 DR’s requirement to describe the sources of information for audit evidence (Art. 2(6) DR). By providing thorough documentation on

risk estimation, distribution, error, and population, auditors can foster a better understanding of subjective decisions, increase the robustness of their studies, and improve the reproducibility of the auditing process. Such explanations also contribute to defining the materiality threshold in accordance with Art. 37(4) DSA(d) and Art. 10(2)(a) DR.

6.0.6. CONDUCTING INDEPENDENT AUDITS AND SYSTEMIC RISK ASSESSMENTS

According to the DR, the goal of sampling is to minimize the detection risk of the audit (Art. 2(3) DR and Art. 12(1) DR). Detection risk is defined in Art. 2(8) DR as “the risk that the auditing organisation does not detect a misstatement that is relevant for the assessment of the audited provider’s compliance with an audited obligation or commitment”. Thus, auditing risk refers to instances when external auditors fail to identify mistakes, omissions, or misunderstandings during their audit. The DR also defines two additional risk categories: inherent risk and control risk. The role of auditors and the element of subjectivity in deciding in audit methodologies and certainty levels however is critiqued under the formulation of the DSA [692].

Inherent risk refers to the vulnerabilities stemming from “the nature, the design, the activity and the use of the audited service, as well as the context in which it is operated, and the risk of non-compliance related to the nature of the audited obligation or commitment” (Art. 2(1) DR). This highlights the importance of individual designs and services offered by a VLOP or VLOSE. For example, the presence of malicious users may vary based on whether the platform features a marketplace or live-streaming functionality. Control risk, on the other hand, pertains to the possibility of the audit being misdirected or unable to identify risks accurately, which cannot be resolved through internal controls (Art. 2(11) DR). For instance, external auditors might fail to identify the risk of gender-based violence affecting groups like LGBTQ+ users. Under Art. 37(2) DSA, VLOPs and VLOSEs must provide “cooperation and assistance necessary to enable [the auditors] to conduct those audits in an effective, efficient and timely manner, including by giving them access to all relevant data and premises and by answering oral or written questions”. If an auditing organization lacks access to necessary data and cannot adequately analyse, identify, and assess risks (Art. 13 DR), it must note these shortcomings in the audit report (Art. 37(5) DSA). However, such data inaccessibility can increase detection risk and undermine the robustness of the analysis. This definition underscores the critical role of context in risk classification and evaluation. Context is also essential for conducting systemic risk assessments and external audits. Transparency reports (Art. 15, 24, and 42 DSA) and Statements of Reason (Art. 17 DSA) benefit from contextualization and can serve as additional evidence in audits. Art. 13 DR mandates that audits consider the probability and severity of risks, aligning these factors with the appropriateness of the assessment and the minimization of detection risk. Additionally, the probability and severity of risks must be assessed for the groups most impacted. For example, the LGBTQ+ community might face a heightened risk of gender-based violence (Art. 34(1)(d) DSA), or female politicians may be targeted in disinformation campaigns before elections [444, 458].

These considerations add complexity to the auditing process, which may also be constrained by available resources. Regional and linguistic aspects of detection and control risks must also be addressed, documented, and justified within audit reports. Member

State-specific samples should be collected and analysed in their respective languages to account for these regional and linguistic differences. Additionally, risks in these samples must be tested on groups most impacted, with documentation explaining who these groups are and why they were chosen. Such transparency and rigor are essential for maintaining the credibility of the audit process. To minimize detection risk, multiple samples should be drawn within representative time frames, such as the period preceding an election. These samples should cover all Member States and their languages, addressing regional and linguistic aspects comprehensively. The auditors' decisions must be documented and justified in the audit report. If an audit lacks sufficient rigor and a reasonable level of assurance regarding systemic risks under the DSA, the credibility of the audit report and process may suffer. Transparent methodologies, including clear documentation of sampling choices, are essential for maintaining the integrity of audits [606, p. 429]. This could lead to a decrease in the reasonable level of assurance and negatively impact detection and control risks.

6.0.7. COMPARISON OF SYSTEMIC RISK ASSESSMENTS AND AUDIT REPORTS

In this section we compare the first systemic risk assessment reports of VLOPs and VLOSEs and the first audit reports, analysing how these reports have addressed sampling as a methodology to assess systemic risks and to minimize detection risk. We analyzed 19 reports from 2023, observing significant variations in the depth of information provided, methodological rigor, and transparency across platforms and auditing firms. The analysis focuses on the strengths and weaknesses of the sampling methodologies, the consistency of data across sources, and the implications for evaluating compliance the DSA and the DR. In the following sections we compare first systemic risk assessment reports, and second audit reports, with a focus on the following: sampling as a method used, the time frame, or the content used in sampling.

SYSTEMIC RISK ASSESSMENT REPORTS

Within systemic risk assessment reports, platforms vary in the level of detail they provide regarding methodologies (see 6.1). The best practice example is Snapchat, which defines confidence intervals, a $\pm 5\%$ margin of error, and employs the KPI "Policy Violating Prevalence" (PVP) rate, encompassing Terms and Conditions (TaC) violations and illegal content [677, p. 5, p. 78, p. 216.]. Their inspections are based on daily random sampling, aligning with proactive risk mitigation strategies, such as monitoring Public Stories. YouTube, by contrast, uses the "Violative Video Rate" (VVR) as a random sampling metric to identify gaps in automated systems. This metric estimates the percent of undetected violations on the platform, evaluated by humans using "statistically significant samples" [303, p. 100]. However, this approach lacks clarity regarding the sample size, the method used to calculate it, and details about the sampled content (e.g. comments, YouTube Shorts, or videos, including min./max./average length). Despite these gaps, the mixed-methods approach in the VVR aligns with the methodology proposed in this article. LinkedIn employs "positive samples" (assumed to mean false positives) of accounts and content on a weekly or monthly basis to evaluate the precision of automated content moderation using "regular sampling" but does not specify sample size or confidence intervals [477, p. 19, p. 23, p. 28.]. Similarly, AliExpress conducts weekly quality checks on

product listings, focusing on multilingual content in high-risk product categories, both internally and with external professional service providers [35, p. 24, p. 31.].

Table 6.1: Platform Sampling Methods and Analysis in Relation to Systemic Risks

Platform	Sampling Method	Method of Analysis	Frequency	Focus Area	Source
AliExpress	Not clear	Multilingual sampling for illegal listings	Weekly	Illegal listings compliance	pp. 24–31
Amazon Inc.	Proactive sampling	Manual and automated compliance controls	Ongoing	Seller verification, safety	pp. 6–18
Apple	Not clear	Manual triage integrated with automated systems	Annual	App Store risks	pp. 66–81
Booking.com	Not clear	Scenario-based systemic risk assessment	5 months	Reviews and discrimination	pp. 1–2
Google Platforms	Statistical sampling	Random and flagged content review	Periodic	Content policy compliance	p. 100
LinkedIn	Positive sampling	QA audits and prevalence exercises	Weekly, Monthly	Abuse trends, algorithm QA	pp. 19, 28
Facebook	Not clear	Governance-based feedback loops	Continuous	Moderation governance	p. 59
Instagram	Not clear	Governance measures and user feedback	Continuous	Harmful content moderation	p. 59
Microsoft Bing	Not clear	QA sampling and user activity evaluation	Annual	Search and ad safety	pp. 3–22
Pinterest	Not clear	Moderation error trend analysis	Ongoing	Moderation error reduction	p. 7
Snapchat	Random sampling	Testing for policy violation prevalence rates	Daily	Detection blind spots	pp. 5–216
TikTok	Targeted sampling	Context-specific reviews for election content	Event-driven	Election misinformation	p. 33
Wikimedia	Not clear	Human rights-based approach	Periodic	Risk register	Online resource
X (Twitter)	Not clear	Crisis management and systemic risk profiling	Annual	Crisis management, risks	pp. 6–10
Zalando	Not clear	Partner business compliance reviews	Annual	Partner content compliance	pp. 1–30

Pinterest incorporates “calibrated sampling” for “Root Cause Analysis” and “Error Trend Analysis” as part of its quality assurance program [596, p. 7p. 7]. Meta, on the other hand, provides limited insight into sampling methodologies, mentioning only “continuous testing” without elaborating further [519, p. 59 Meta]. TikTok employs “scenario-specific sampling” to detect election-related risks, such as misinformation, focusing on top-rated videos and high-priority events [705, p. 33]. Regarding sampling frequency, platforms demonstrate varied approaches: continuous (Snapchat, Meta), periodic (YouTube, Pinterest), event-driven (TikTok), or specified intervals (LinkedIn, AliExpress)

AUDIT REPORTS

Audit reports conducted by auditing firms provide critical insights into methodologies, as required by Arts. 12 and 13 of the DR as illustrated in Table 6.4. The sampling methods across the referenced audit reports exhibit notable variations in their complexity, scalability, and applicability to different compliance contexts. A notable example is Snapchat’s audit report, which clearly specifies a materiality threshold of 95% and an actual or projected error margin of 5% [257, p. 63]. Similarly, all audit reports conducted by EY include

Table 6.2: Comparative Analysis of Platform Strengths and Weaknesses

Platform	Strengths	Weaknesses
AliExpress	Multilingual checks; high-risk targeting	Limited transparency on emerging risks
Amazon Inc.	Automation and manual reviews; scalable accuracy	Unclear handling of false positives
Apple	DSA-aligned; stakeholder consultation	Low intervention thresholds
Booking.com	Tailored transaction-specific approach	Narrow scope beyond reviews
Google Platforms	Wide sampling and broad coverage	Minimal systemic risk insights
LinkedIn	QA audits; abuse pattern refinement	Lacks detailed long-term risk plans
Facebook	Appeals-driven feedback loops	Lacks broader systemic context
Instagram	Feedback integration	Weak systemic abuse data focus
Microsoft Bing	Combines ranking with QA	Limited external transparency
Pinterest	Trend analysis with consistent QA	Minimal focus on systemic risks
Snapchat	Statistically robust (95% CI)	Gaps in emerging risk handling
TikTok	Context-specific targeting	Lacks general risk evaluation
Wikimedia	Human rights-centered	Lacks sampling/selection criteria details
X (Twitter)	Focus on hate speech/misinformation	No granular systemic data
Zalando	Reviews of recommender systems	No external validation

dedicated sections on sampling methods [249–258]. These reports detail sampling approaches, such as the selection of sample sizes: “e.g., a sample of 25 when the population is greater than 250 occurrences or 10% of the population size, with a minimum sample of 5 when the population is less than 250 occurrences” [250, p. 11].

EY’s reports consistently employ “attribute sampling to determine if the sample selected has the desired attribute (e.g. correct or incorrect, present or absent, valid or not valid),” forming a “qualitative statistical [random] sample” [255, p. 13]. These methods evaluate whether specific attributes (e.g., presence, correctness, or validity) are consistently upheld within the sample population. The reports also highlight that no other sampling approaches are applicable when the populations lack quantitative dimensions, such as monetary balances in financial audits [255, p. 13]. In cases of exceptions within the sample selection, EY addresses the need to extend sample sizes to create substantive evidence, noting, “e.g., for sample sizes of 25, we tested at least 15 additional items or 40 in total” [249–258]. However, its application is limited in highly dynamic systems where attributes may evolve over time, potentially leading to outdated evaluations. Furthermore, the reports lack details on what those attributes are and which audit criterion or systemic risk category they link.

In contrast, *KPMG* audits do not provide the same level of detail regarding sampling methodologies in their reports [185, p. 8]. *Deloitte*, on the other hand, employs both statistical and non-statistical sampling methods and do not provide enough detail on understanding detailed decisions on samples, content types, sampling frequency, time frames, or confidence intervals.. However, reports audited by *KPMG* (e.g., TikTok, AliExpress, and Amazon Inc.) predominantly employ random sampling with increasing selection size linked to risk levels associated with tested controls. *Deloitte*, (e.g., LinkedIn and Zalando) on the other hand, employs both statistical and non-statistical sampling methods (leveraging auditor judgment to define tolerable error rates and sample sizes) [185, p. 8]. However, *Deloitte* reports only indicate the use of mathematical models for determining sample sizes, without specifying which models were used or their appli-

cation to auditing criteria, risk categories, or content types. Nevertheless, it introduces subjectivity, which may impact reproducibility and comparability across audits. When auditors decide not to use statistical sampling is also not indicated in the audit reports, which makes detailed evaluation inaccessible.

Certain reports offer more granular details about sampling practices. For example, Zalando's audit report describes the testing of a sample of 16 notices, with a benchmark time frame for reply set at 30 days [184, p 37]. Such variations in reporting practices underscore the need for standardization in audit methodologies to enhance transparency and comparability across platforms.

6.0.8. CONCLUSION AND FURTHER RESEARCH

This article examined the strengths and weaknesses of quantitative and qualitative analytical methodologies in systemic risk assessments and audits under the DSA and DR, emphasizing the critical role of sampling as a methodological foundation. By analyzing 19 reports from VLOPs, VLOSEs, and independent auditors, we highlighted significant disparities in the transparency, robustness, and applicability of sampling techniques. These findings underscore the need for standardized yet adaptable approaches to address the complexities of socio-technical systems.

Our analysis identified systemic risk assessment reports as generally offering platform-specific metrics tailored to unique operational risks. For example, Snapchat's "Policy Violating Prevalence" and YouTube's "Violative Video Rate" illustrate how statistically robust sampling can illuminate systemic risks and guide proactive mitigation strategies. However, inconsistent documentation of sample sizes, criteria, and methodologies limits the replicability and transparency of many reports. Conversely, audit reports often focus on compliance and scalability, providing detailed sampling thresholds, while other firms lack comparable methodological transparency. This variability highlights a gap in aligning systemic risk assessment and audit methodologies.

We proposed a mixed-method approach that integrates quantitative sampling with qualitative legal content analysis, balancing statistical rigor and contextual depth. Sampling methodologies should be tailored to the platform's characteristics, risks, and context, ensuring representativeness and relevance. Qualitative legal content analysis complements these methods by contextualizing violations and aligning findings with systemic risk categories outlined in the DSA.

The mixed-method approach offers actionable insights for both platforms and regulators. First, it enables auditors to identify and address gaps in content moderation systems, linking systemic risks to measurable KPIs. Second, it provides platforms with a framework for adapting sampling methodologies to dynamic risks, such as those related to misinformation, election integrity, or gender-based violence. Lastly, the approach ensures that audits not only verify compliance but also contribute to the continuous improvement of platform governance.

By bridging methodological gaps and addressing variability in current practices, this article contributes to the growing field of DSA compliance research. Our proposed mixed-method framework offers a pathway for advancing fairness, accountability, and transparency in the governance of socio-technical systems.

Table 6.4: Platform Sampling Methods and Analysis in Relation to Audit Reports

Category	Auditor	Sampling Method	Context Information	Source
AliExpress	KPMG	Random sampling with risk-adjusted selection size.	Random selection considers illegal exceptions for >5% of the selection, linking population size to significance levels.	AliExpress DSA Assurance Report, KPMG, Sept. 2024, pages 12, 35, 75, 98-99, 108.
Amazon Inc.	KPMG	Random sampling with risk-adjusted selection size.	Population size linked to significance levels.	Amazon DSA Audit Report 2024, pages 34, 95, 98-100.
Apple	EY	Attribute sampling techniques.	Sample size based on population size, with a focus on valid/invalid attributes.	Apple DSA Audit Report 2024, pages 16-17, 19, 30-31, 38-49, 51-62, 89-96, 100, 115, 125.
Booking.com	Deloitte	Statistical and nonstatistical methods.	Sample sizes determined by population size, control failure risk, and tolerable deviation rate.	Booking.com DSA Audit Report 2024, pages 14, 82, 90, 95, 105, 108, 113, 137, 145.
Google Platforms	EY	Attribute sampling techniques.	Sample size ensures appropriate attribute representation.	Google DSA Audit Reports 2024, pages 13-14, 27, 38-43, 45, 47, 49, 51, 55-57, 66, 75, 78, 80, 96-97.
LinkedIn	Deloitte	Statistical and nonstatistical methods.	Sample sizes based on population, control risk, and tolerable deviation rate.	LinkedIn DSA Report 2024, pages 7, 17-18, 30, 32, 38, 41, 43-45, 49, 51, 53-54, 56-57, 58, 61-62, 64, 72-73, 75, 84, 86, 94-95, 100-101, 103, 106, 135, 137.
Facebook	EY	Attribute sampling techniques.	Minimum sample size established for tested attributes.	Facebook DSA Audit Report 2024, pages 11-16, 18, 25-34, 36-37, 45-52, 56, 59-61, 80.
Instagram	EY	Attribute sampling techniques.	Same methodology as Facebook, targeting attributes.	Instagram DSA Report 2024, pages 11-16, 18, 25-35, 37-38, 46-51, 53, 57, 60-62, 79, 81.
Microsoft Bing	Deloitte	Statistical and nonstatistical methods.	Sample sizes based on population and control risks.	Microsoft Bing DSA Audit Report 2024, pages 8, 32, 107.
Pinterest	EY	Sampling aligned with regulations, considering risk evaluations.	Challenges included limits in algorithm assessments.	Pinterest-2024-DSA-EY-Independent-Audit-Report, pages 6, 12, 19, 25.
Snapchat	EY	Statistical sampling focusing on transparency metrics.	Reviewed transparency in ads, moderation, and personalization systems.	Snapchat DSA Report 2024, pages 12, 66, 80, 104.
TikTok	KPMG	Random sampling with size linked to control risk.	Population-linked significance testing.	TikTok DSA Audit Report 2024, pages 9-12, 21, 25, 28, 36, 38, 41-43, 49-50, 52, 59, 62, 68, 78, 86, 91-92, 96, 99-100, 105.
Wikimedia	Holistic AI	Non-statistical methods.	No statistical data provided.	Wikipedia DSA Audit Report 2024, pages 104, 107.
X (Twitter)	FTI Consulting	Random sampling.	Randomized communication tests for clarity and consistency.	TIUC DSA Audit Report 2024, pages 55, 65, 100, 113, 139-141, 153, 199-200, 207, 261, 319.
Zalando	Deloitte	Random sampling.	Cross-verification using legal product databases.	Zalando DSA Audit Report 2024, pages 20, 22, 34, 36-39, 41-42, 47, 49, 51-52, 69-70, 82, 85, 87-88, 96, 119, 121, 160, 164, 167, 198.

7

LEGAL DOCTRINAL ASPECTS FOR CONTENT ANALYSIS – DEVELOPING A LEGAL CODE BOOK FOR ADVERTISING AND ELECTORAL CONTENT

*If I have to search the whole legal code,
If I have to read through the whole statute book,
With a quibble or a paraphrase I'll find some obstacle. /
Se tutto il codice dovessi volgere,
se tutto l'indice dovessi leggere,
con un equivoco,
con un sinonimo qualche garbuglio si troverà.*

Dr. Bartolo, Act I, Scene III, Le Nozze di Figaro, Composer: Wolfgang Amadeus Mozart,
Libretto: Lorenzo Da Ponte

The rise of digital platforms has transformed political campaigning, introducing complex regulatory challenges concerning transparency, content moderation, and the targeting of political advertisements. This paper presents a taxonomy for analysing political content in the EU's digital electoral landscape, aligned with the Digital Services Act (DSA), the Regulation on the Transparency and Targeting of Political Advertising (TTPA), and the European Commission's Guidelines on the Mitigation of Systemic Risks for Electoral Processes (G-E-DSA). Using a legal-doctrinal methodology, we construct a codebook for systematic content

analysis of user-generated and political-advertising content. The framework addresses systemic-risk identification, advertising-repository transparency, and sponsor-disclosure requirements. We apply the taxonomy through sample-based annotation of content on Very Large Online Platforms (VLOPs), evaluating platform adherence to transparency and electoral-integrity obligations. By bridging legal analysis and empirical methodology, the study strengthens compliance evaluation and supports an interdisciplinary understanding of online risks. It provides a framework for platforms, advertisers, regulators, and auditors navigating the EU's evolving digital political landscape.

7.1. INTRODUCTION

After the 2016 US elections and the Brexit referendum, political consensus formed in the EU that novel rules for the context of elections are needed to increase online protection [336, 355]. The digital dimensions of political debate and their potential negative consequences can have manifestations offline, which was illustrated by the storm on the US Capitol in 2021 or the attempt to replicate such behavior in Germany [559, 601]. Governmental processes aim to protect voters, politicians, and political debates, including the obligation to decide how much regulation is needed for disinformation, digital campaigning, or content moderation. The European Union's key regulations and soft law provisions that regulate the digital side of elections include:

- The Digital Services Act (DSA) [224]
- The Regulation on Transparency and Targeting of Political Advertising (TTPA) [24]
- The Commission's Guidelines on Mitigation of Systemic Risks for Electoral Processes under the DSA (G-E-DSA) [23]

This article introduces key provisions relevant to the political context of online advertising and provides a codebook for content analysis in line with the DSA, TTPA, and G-E-DSA, aiming to support platforms' systemic risk assessments and external independent audits under Art. 34 DSA.

These regulations target different aspects and domains of political speech. The DSA mainly sets rules for content moderation and platform compliance with human rights and laws, demands regular risk assessments and audits, mitigation measures for risks, and creates obligations for some platforms to build ad repositories.

The TTPA that includes norms on how campaigns and actors must be verified, which information must be included about spending and targeting, and ad repositories for platforms and for a new aggregated EU repository [745]. Therefore, understanding the interplay of these new regulatory measures is important for understanding the compliance obligations for platforms for user-generated or AI-generated political ad content. A recent example of the enforcement of novel regulations can be the opening of the proceedings against the Very Large Online Platforms (VLOPs according to Art 33 DSA) X (formerly known as Twitter) and TikTok [232]. X is investigated for potentially violating several provisions of the DSA, such as the risk on civic discourse and electoral processes (Art. 34(1)(c) DSA) or Art. 39 DSA laying down rules for online advertising transparency. Such violations can have severe consequences and could lead to fines of significant parts of the platforms' annual revenue [224].

Therefore, platforms, advertisers, political candidates and campaigners, as well as regulators and civil society have to follow these regulations which complement each other and provide additional details for specific use cases like generative AI.

In this article, we aim to introduce the key provisions relevant for the political context of online advertising to contribute to the ongoing debate about compliance with these regulations. Moreover, we provide a codebook for content analysis in line with the DSA, the TTPA, and the G-E-DSA. The codebook can be used to empirically test representative content samples of user-generated or ad-content to analyse systemic risks of Art 34 (1)

DSA (mainly focusing on negative consequences on the electoral process according to Art 34 (1) (c) DSA).

Systemic risks that need to be assessed under the DSA can occur in various domains relevant for elections. For instance, risks can be situated between legal and harmful speech, mis- or disinformation, malicious flagging, Artificial Intelligence (AI) and deep-fakes, targeting, or campaigns against candidates or specific groups. Disinformation is defined in the Code of Practice of Disinformation (CDP) as “verifiably false or misleading information [that is] created, presented and disseminated for economic gain or to intentionally deceive the public [and] may cause public harm [intended as] threats to democratic political and policymaking processes as well as public goods” and is an example of (potentially) harmful speech that not necessarily crosses the threshold of illegality [143, p. 1]. Following the fine line between illegality and harmful speech, however, is posing questions to the protection of fundamental rights, such as freedom of speech or the right to free and fair elections [248]. Furthermore, if concepts like mis- or disinformation are not tangible and clearly definable [405] the automatic detection of such violations can be hard to operationalise for automatic content moderation techniques like classifiers, or the use of Large Language Models (LLMs) [104, 736].

To address the complex interplay of regulations and palette of risks to ensure free and fair elections of the digital age we will bring together three regulatory sources governing the election in the European Union (EU). By combining the three sources we aim to provide a detailed overview of the demands for politicians, campaigners, advertisers, platforms, regulators, Trusted Flaggers, public authorities, interdisciplinary researchers (e.g. computer or social scientists), journalists, and other overseeing bodies in the EU. In this text we build on prior work for testing the systemic risk on the electoral process in the context of the German Federal elections in 2021 [444]. For this article we will adapt our code book defining systemic risks under Art 34(1) (c) DSA for the notions needed to test the TTP, and the most recent G–E–DSA. To our knowledge, the selection of these legal sources adequately portrays the relevant regulatory landscape for the digital aspects of the electoral discourse and digital campaigning. This article aims to answer the following research questions:

- RQ1: Which taxonomy can measure compliance with the DSA, TTPA, and G–E–DSA when assessing systemic risks to electoral processes in political content?
- RQ2: What content annotation categories are necessary to be included in the content analysis process in line with the DSA, the TTP, and the G–E–DSA?

By answering these research questions, we will provide a consistent overview of applicable regulations in the context of elections in the digital realm and a method to comply with these regulations. This collection will contribute to understanding the complex interplay of the different regulatory sources and demands, helping advertisers, political campaigners, platforms, users, election overseeing bodies, regulators, and researchers better understand compliance at the intersection of the three regulatory sources and contribute with methodological tools to the scientific debate dealing with compliance measurement and automation of test and audits.

7.2. POLITICAL ADVERTISING AND REGULATION

Political campaigns and the creation of political ads play an important role in elections. Politicians create comparatively cheap advertising strategies (in contrast to TV advertising) and fine-tune their message, visual, or appearance to target a specific voter audience, e.g., through gender filtering [363]. This separation between the political messages provided to the voters for different voter segments can have the consequence that not all voters see the same message of a political party [67]. This fragmentation, however, can create diverging political realities, e.g. through the distribution of generative-targeted AI content [398].

Risks to the electoral process might stem from the message conveyed by the ad content and the ad delivery process that decides who the target audience should be. In the ad delivery process, users might be already selected as a target audience who should be nudged towards changing their political opinion or forming a voting decision [33]. Targeting in the ad display phase is also linked to the financial side of advertising. The price of ads usually includes an auctioning system defining different prices for users [33]. For instance, users with different political leanings or affiliations can influence the price of the message presented in the ad (higher prices for presenting ads that do not match interest categories). Additionally, presenting the ad in different time frames can be relevant to prices (e.g., presenting the ad closer to the election date might have consequences on the pricing of the ad), or the targeting for different user groups can be a significant price indicator (of young mothers [581]). Nevertheless, financial data show that political players spend more than ever on digital campaigns [670]. How exactly such auctions work is not a transparent process and is only supported by aggregated numbers in repositories or transparency reports [745].

Attempts of new European regulations to make financial aspects of advertising more transparent include the indication of information in the user interface about high-level targeting criteria (for example, through the “why am I seeing this ad function”) or through the mandatory ad repositories in the aggregated form of spending [224, 232].

By creating publicly accessible political ad repositories, more information about the electoral campaigns, the messages that politicians frame for specific audiences, visuals used, clips created, sponsors endorsed, or overall spending should create a better overview of the digital side of the election landscape. How meaningful transparency mechanisms are, however, is also very much related to the design of the tools used [747] and can be shifted to create different, more desirable visibilities [268]. Creating ad repositories for political content is now mandatory under the DSA and the TTPA on a platform level and for VLOPs and Very Large Search Engines (VLOSEs) on a European level [224, 232, Art. 13 TTPA, Art 39 DSA]. However, the completeness and accuracy of such repositories are criticised [508]. On the other hand, such inaccuracies could pose a potential systemic risk for platforms and their systemic risk assessments and, risk mitigation obligations and transparency reporting [224, Art. 35].

However, it is very complex to define ‘political content’ on platforms (for example, what constitutes political advertising in cases of influencer content [509] and how to label or filter it automatically. Several platforms, such as TikTok, have claimed not to allow political content and failed [579].

How can such claims be meaningfully and empirically tested in the light of new regula-

tions [663]? In empirical tests, false positives and negatives can shed light on the question of what constitutes good content moderation practices for identifying political ads. False positives can be illustrated as cases in which a piece of content is misidentified as political advertising. For instance, the ad of a building company is classified as political content. False negatives, therefore, are cases that do not classify political content as political. For example, a political party post advertising an event is not labeled as a political ad. The automation of these processes is an intersection with legal and contractual requirements like the Terms and Conditions (TaC) that has to be modeled and implemented by computer scientists to automate the processes needed for regulatory compliance. If ‘political content and advertising’ on a VLOP like TikTok are prohibited in their TaC [146], it is still relevant to ask how many false positives and false negatives are acceptable under new regulations and how to measure them [226].

Besides, new risks may emerge for safeguarding political campaigns and voters online due to the increased use of generative AI. These risks are in the scope of the EU’s AI Act [23]. Such risks could stem from the use of image generation in political contexts [732]. In the context of the campaigns for the US presidential election 2024, generative AI was used to generate the voice of the democratic candidate Joe Biden, to convince voters to ‘save’ their vote for the election in November and not vote in the primary elections [602]. On the other hand, deepfakes are another manifestation of the negative consequences of the digital sphere, or voice-generated content [510]. Moreover, these often have a gendered component. For example, Kübler et al. found that female politicians are targeted more often than male colleagues in disinformation-promoting content [444]. Furthermore, different content forms of disinformation, such as deepfakes of audio content, can add complexity to the regulation and enforcement of the created rules; for example, how will the advertising repository for an audio-based platform be searchable and explorable in a meaningful way to spot risks early on? However, not only digital risks may influence the political parquette before elections. For example, the platform Cameo was used as a vehicle in a disinformation campaign in the context of the ongoing Russian attack on Ukraine, instrumentalising American actors such as Elijah Wood to create individualised videos to create the impression of President Volodymyr Zelensky having a substance abuse problem [140].

7.3. INTRODUCING THE LEGAL PROVISIONS RELEVANT FOR THE CONTEXT OF ELECTIONS AND CAMPAIGNING IN THE DSA, TTPA, AND G–E–DSA

7.3.1. AD TRANSPARENCY AND SYSTEMIC RISK IN EUROPE’S DIGITAL SERVICES ACT

Under the DSA, VLOPs and VLOSEs defined in Art. 33(1), must test periodically the systemic risks defined in Art 34 DSA and mitigate the defined risks according to Art 35 DSA. Those platforms are measured against the threshold of reaching 45 million average monthly users within the EU. Their analysis of platform risks will be recorded

in the audit report in line with Art 42(4) (a) DSA and Rec. 100. These claims made in the audit report on the other hand will be tested by external independent auditors according to Art 37 DSA. The mandatory risk assessments should cover five areas of audit according to Art. 34 (2) (a-e) DSA: the design of the algorithmic systems in use (including recommendation algorithms), the content moderation systems, the terms and conditions (including enforcement of the rules), systems for advertisement (including the creation of the ad, the audience selection tools and ad-delivery techniques), and data practices of VLOPs and VLOSEs. Advertising systems create a particular risk if run by VLOPs and VLOSEs according to Recital 95 DSA “for example in relation to illegal advertisements or manipulative techniques and disinformation with a real and foreseeable negative impact on public health, public security, civil discourse, political participation and equality”.

In Art 34 four risk categories are defined laying down the testing frame for the internal systemic risk assessments which have to be “identified, analyzed and assessed”. To the context of elections Art 34(1) (c) DSA is most relevant for the context of elections defining a systemic risk as “any actual or foreseeable negative effects on civic discourse and electoral processes, and public security”. Additionally, to the problem domain of political content in the context of elections the exercise of fundamental rights as stated in Art 34 (1) (b) DSA might be applicable [248]. Specifically, the right to freedom of speech (Art 11 [248]), the right to Freedom of assembly and association (Art 12 [248]), the Right to vote and stand as a candidate at the elections to the European Parliament (Art 39 [248]) and the parallel protection for municipal elections in Art 40.

Regarding advertising, the DSA furthermore defines minimal general rules for advertising online in Art 26 DSA. This provision demands the identifiable and consistent indication that the content is labelled as advertising content (Art 26 (1) (a) DSA), for whom the advertising is presented (b), who paid for the ad (c), and the main parameters used in the process to identify the audience of the ad (information about aggregated targeting criteria) and the possibility to adjust those according to literal d. According to Art 26 (3) in conjunction with Art 4 (4) and Art 9 (1) GDPR [238] shall be prohibited. This may include targeting based on “[...] racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person’s sex life or sexual orientation”. Recital 95 DSA additionally stresses the need for information about targeting criteria and the relationship to vulnerable groups and situations. A demand echoed in the Delegated Regulation for testing systemic risks on these marginalized groups [226].

Additionally, Art 39 DSA is creating novel rules to increase ad transparency through advertising repositories. Content presented in the repository has to indicate at least the ad itself (Art 39 (2) (a) DSA), for whom the ad is presented (b), who has paid for the ad (c), the duration of presentation (d), information about the parameters of targeting (inclusion and exclusion of groups to the desired audience) of groups (e), commercial communication (f), and the total number of views broken down by Member State (g).

If the advertising content however would violate the terms and conditions (Art 14 DSA) of the platform or be regarded as illegal according to Art 3 (h) DSA it should not be included in the repository but should be indicated within the transparency mechanism referred to as Statement of Reason according to Art 17 DSA or within the reporting of

received orders from public authorities according to Art 9 DSA. This information covers the reporting on content moderation decisions for platforms at a meta-level. According to the regulation codes of conduct should complement the compliance measures used in the domain of online advertising and the systemic risks that could be faced by platforms (Art 45 DSA). Such codes should provide key performance indicators as measurements of compliance according to Art 45 para 3 DSA. Systemic risks are furthermore creating the potential invitation to create new codes of conduct according to Art 45 para 2 DSA. Should a platform fail to address the systemic risks adequately the Commission, the Board (cite) and the relevant signatories (e.g. Who Targets Me for the Code of Practice on Disinformation [143]) will be consulted for repairing the failed attempt (Art 45 (4) DSA).

More specifically, Art 46 DSA defines voluntary and tailored rules for online advertising to increase the need for transparency highlighting Art 26 and 39 DSA. Additionally, Art 46 (2) (c) DSA includes information on the monetization of data from signatories. This need for standardization is also addressed in Recital 107 DSA and the relation to the technical implementation of such compliance demands.

7.3.2. REGULATION ON TRANSPARENCY AND TARGETING OF POLITICAL ADVERTISING

The Regulation on Transparency and Targeting of Political Advertising (TTPA) sets out crucial provisions for safeguarding electoral integrity and transparency in political advertising within the digital domain. The TTPA mandates that political advertisements be clearly identifiable and that platforms provide transparency on ad targeting, financing, and sponsorship. It addresses systemic risks by defining guidelines that ensure voters have access to reliable, unbiased information and can easily identify the sources and funding behind political messaging [23].

Under the TTPA, all political ads must display the sponsor's name and include sufficient information to identify the individual or entity paying for the advertisement, ensuring transparency in sponsorship (Art 3(10) TTPA, Art 8(1)(b) TTPA). Platforms are required to maintain publicly accessible, searchable repositories of political advertisements, detailing the sponsor, the duration of each ad, the aggregated targeting criteria, and information on who viewed the ad (Art 13 TTPA, Art 39 DSA). This repository allows both the public and auditors to access political ad data, providing oversight and enhancing transparency around targeted political campaigns [23].

The TTPA further requires VLOPs and VLOSEs to clearly label political ads, distinguishing them from other content. Influencers promoting political content must also declare their sponsorships, with platforms providing functionalities for these declarations (see also Guideline Sec. 3.2.1(f)). This measure ensures that audiences can differentiate organic content from paid political advertising and understand the affiliations behind the content [23].

An essential component of the TTPA is transparency in targeting criteria. Platforms must disclose the criteria used for ad targeting, allowing users to understand why certain ads are shown to them, especially where ad targeting includes demographic or interest-based segmentation (Art 3(11) TTPA, Art 8(1)(f) TTPA). Platforms must also uphold ethical targeting practices by providing a “why am I seeing this ad” feature that offers users a straightforward explanation of why they were targeted for specific political messages,

further reinforcing transparency in ad delivery [23].

The TTPA also addresses emerging risks associated with synthetic media, mandating that generative AI content, especially deepfakes, be appropriately labeled. Such labeling helps prevent the spread of misinformation and ensures that voters are not misled by synthetic representations of political figures or events (see also Guideline Sec. 3.3). Media literacy initiatives are also recommended, equipping users with knowledge on how to critically evaluate digital content, particularly AI-generated material. These initiatives support the democratic process by promoting informed decision-making among voters [23].

Through the TTPA's comprehensive approach, the EU aims to create a transparent, accountable environment for digital political advertising. By defining clear guidelines for sponsorship disclosure, ad targeting transparency, influencer accountability, and synthetic media labeling, the TTPA contributes significantly to maintaining the integrity of digital electoral processes.

7.3.3. THE COMMISSION'S GUIDELINES ON THE MITIGATION OF SYSTEMIC RISKS FOR ELECTORAL PROCESSES PURSUANT TO THE DSA

The European Commission's Guidelines on Mitigation of Systemic Risks for Electoral Processes under the Digital Services Act (G-E-DSA) aim to address digital risks that may compromise the integrity of electoral processes, particularly risks posed by VLOPs and VLOSEs. The guidelines outline best practices and recommended measures that VLOPs and VLOSEs are encouraged to adopt to manage risks related to disinformation, generative AI, and manipulative advertising. These measures are intended to protect public discourse, support transparency, and uphold electoral integrity.

In accordance with Article 35(3) of the DSA, VLOPs and VLOSEs should integrate these guidelines into their risk mitigation frameworks. The guidelines emphasize the importance of transparency in political advertising, requiring that ads be clearly labeled and that platforms maintain accessible ad repositories. These repositories should detail essential information, including the identity of the sponsor, targeting criteria, and the reach of each ad. This transparency enables users to understand the source and intent of the content they encounter, while helping regulators and auditors monitor the digital landscape during election periods [23].

Generative AI represents a particularly complex risk, and the guidelines encourage platforms to adopt advanced measures for identifying and labeling AI-generated content. Using watermarks, metadata, and other technical tools, platforms can ensure that AI-generated content is authenticated, which is critical in political contexts to prevent the spread of misleading information that could manipulate voter perceptions. To further minimize these risks, platforms are advised to proactively inform users about potential inaccuracies in AI-generated content and to recommend verification through reliable sources. Additionally, the guidelines suggest platforms conduct regular testing, including red-teaming exercises with both internal teams and external experts, to identify and control unintended effects of generative AI content before public release [23].

Political advertising transparency is a cornerstone of the G-E-DSA, aligning closely with the TTPA. The guidelines mandate that VLOPs and VLOSEs label political ads in a way that is clear, visible, and identifiable in real-time. For influencer-based political content,

platforms must provide functionalities that allow influencers to transparently declare sponsorships, ensuring that users can distinguish between organic and paid political messaging [23].

In addition to labeling, platforms are encouraged to respect 'silence periods', which are times when political advertising is restricted under Member State laws to prevent undue influence on voters immediately before elections. The guidelines also advocate for media literacy initiatives that educate users on recognizing manipulated or AI-generated content, especially when it involves political or election-related information. These initiatives are seen as crucial steps to empower users with the skills needed to critically assess the content they encounter online [23].

By implementing these guidelines, platforms can more effectively identify and manage systemic risks associated with digital electoral processes. Through measures like transparency in political advertising, monitoring of AI-generated content, enforcement of silence periods, and media literacy initiatives, the G-E-DSA contributes to creating a safer digital environment that respects democratic principles and empowers users in an increasingly complex information landscape.

7

7.4. METHODOLOGY

This study employs a legal doctrinal research approach to develop a comprehensive taxonomy for assessing compliance with the EU's Digital Services Act (DSA), the Transparency and Targeting of Political Advertising Regulation (TTPA) and the Commission's Guidelines on the Mitigation of Systemic Risks for Electoral Processes (G-E-DSA). Building on prior work [742] and aligning with regulatory obligations, we systematically structure a taxonomy that addresses key compliance requirements, annotates risk categories, and establishes content analysis codes specific to electoral integrity and transparency in political advertising.

We aim to contribute to regulatory compliance in the context of the EU by highlighting key provisions in the three legal sources and showing how they complete each other. Furthermore, we want to contribute to the systemic risk assessment and external audit process, according to Art. 34-35 and Art. 37 DSA by proposing a codebook specifying content annotation classes, especially targeting the systemic risks addressed in Art. 34(1) (c) DSA. These content annotation classes are based on relevant literature and legal norms and build on prior work contextualizing the negative effects on the electoral process as proposed in Art. 34 (1) (c) DSA [444]. The codebook defines the following:

1. Code names that can be applied to individual pieces of content (e.g., E-1),
2. sources that influenced the creation of a code or legal provisions (e.g., [405], or Art. 39 (2) (a) DSA),
3. A qualification that determines what the code should entail (e.g., hoxes).
4. A definition providing a better overview of what is meant by the code and how it should be applied and understood.
5. Examples to enable a better-aligned application of the codes and increase inter-coder reliability.

We propose content analysis as a qualitative methodology for conducting systemic risk assessments under the DSA in line with Art. 34 DSA and testing systemic risk assessment reports as external independent auditors on a sample basis, according to Art. 37 DSA and the Delegated Regulation [226].

The development of this taxonomy involves a doctrinal analysis of the regulatory texts, synthesizing legislative language, definitions, and thematic requirements across DSA, TTPA, and G-E-DSA. By cross-referencing legal provisions with existing frameworks on disinformation, political advertising, and content moderation, we derive annotation categories that capture the core regulatory obligations. These classes are relevant to future research for computer scientists to automate risk analysis techniques, as well as for legal and social science to have a quality control mechanism to conduct systemic risk assessments. Each category is explicitly tied to legal provisions, with clear definitions, qualifications, and examples to guide the analysis. For instance, in assessing transparency requirements for political ads, our taxonomy incorporates TTPA's mandates for sponsor identification, targeting criteria disclosure, and ad labelling in real-time, reflecting the regulatory emphasis on ensuring public awareness of the origins and targeting mechanisms behind political content.

Content is annotated based on this taxonomy, which includes distinct classifications for transparency, systemic risks, and ad repository completeness. For example, categories are defined according to transparency requirements (e.g., clear sponsorship labelling, real-time targeting disclosure) and systemic risks as outlined in Art. 34 DSA, explicitly focusing on risks to civic discourse, electoral processes, and fundamental rights. Annotations are refined to account for the G-E-DSA's emphasis on managing synthetic and AI-generated content, labelling requirements for synthetic media, and targeted approaches to disinformation in electoral contexts.

This structured taxonomy allows for the application of content analysis techniques on a diverse sample of political content, drawn from prominent VLOPs identified as central to electoral advertising. Using stratified sampling, we select advertisements and user-generated content from two VLOPs, representing a broad cross-section of digital content types, including text, images, and audio. The content sample is annotated in line with the taxonomy, with coders assessing compliance with transparency and systemic risk obligations by applying each annotation category. We measure the consistency and accuracy of our coding process through Krippendorff's Alpha, which allows us to refine our taxonomy through pre-tests and ensures reliability in annotation.

Our taxonomy-based approach culminates in a comprehensive report that evaluates compliance across three main regulatory domains: transparency, systemic risk, and completeness of the political ad repository. The structured taxonomy standardises our content analysis methodology and supports future comparative research by defining clear categories for the complex regulatory landscape of digital political advertising. By mapping content categories directly to legal provisions, our framework bridges the gap between doctrinal interpretations and empirical content analysis, providing a consistent tool to assess systemic risk mitigation practices, compliance with the DSA and TTPA requirements, and potential impacts on electoral integrity.

This methodology demonstrates a novel use of doctrinal research to operationalise

regulatory compliance in digital electoral contexts, emphasising the need for empirical methods that align closely with legal taxonomies. Our findings will contribute to the ongoing regulatory discourse, offering platforms, advertisers, and policymakers guidance in managing transparency and systemic risks, ultimately strengthening protections for democratic processes in the digital age.

7.5. RESULTS: ANNOTATING, REPORTING, AND TRANSPARENCY OBLIGATIONS IN POLITICAL CONTENT AND ADVERTISING

In the context of EU regulatory frameworks such as the DSA, TTPA, and G-E–DSA, platforms face increasing obligations to systematically address and report on risks associated with political content, advertising, and misinformation. These frameworks require platforms, especially VLOPs and VLOSEs, to engage in rigorous monitoring and annotation of content to identify, assess, and mitigate systemic risks, particularly those affecting electoral processes.

To facilitate these requirements, the proposed taxonomy is a structured methodology for classifying and annotating political content based on risk factors identified in the DSA, TTPA, and G–E—DSA. Categories within this taxonomy include specific annotations for disinformation, procedural electoral threats, misrepresentation in political advertisements, and the ethical implications of generative AI use in political contexts. The taxonomy supports compliance by enabling platforms to label and classify content, ensuring transparency in advertising practices, and adherence to reporting requirements for political ads and synthetic media.

Furthermore, the taxonomy provides a foundational structure for reporting obligations. Article 39 of the DSA mandates the creation of ad repositories with detailed information on the nature, targeting, and sponsorship of political advertisements. Our framework defines categories for annotation, facilitates content analysis across various media forms, including text, image, video, and audio, and enables platforms to achieve regulatory compliance while preserving transparency in political discourse.

Through systematic annotation, this framework aligns with EU obligations on transparency, enabling a more accountable, transparent, and equitable political advertising ecosystem. By using the taxonomy, platforms and regulators can more effectively monitor and assess compliance with transparency requirements, ultimately fostering a secure and resilient digital electoral environment.

We developed nine new annotation categories, specifically focusing on the need for these three regulations to be laid out for annotating, especially political advertising content. By selecting these new categories Ro1-9, we open our coding framework to encounter new regulatory and compliance challenges for testing systemic risks in the context of 'negative consequences for the electoral process'. The coding guideline developed in prior work can be used as an annotation framework to use the adapted taxonomy of this text [742].

We propose using a cluster sampling method in elections or a random sampling method to create a representative dataset of relevant political (advertising) content for annotation [663]. Using content annotation and sampling together, we believe that empirical legal studies on new regulatory challenges can be tackled interdisciplinary

Ro-1: Gender-based violence	“Violence against women” means gender-based violence directed against women or girls, disproportionately affecting them, including acts causing physical, sexual, psychological or economic harm, threats, coercion, or arbitrary deprivation of liberty, in public or private life.	Example: Online content targeting a female candidate before an election, falsely claiming she lacks an academic degree [444].
Ro-2: Sponsorship and identification of sponsors	Political ads must display the sponsor’s name and details to identify who paid.	Example: A political ad without sponsor information.
Ro-3: Targeting Criteria Disclosure	Platforms must disclose targeting criteria for political ads.	Example: An ad targeting women in a zip code without disclosing criteria.
Ro-4: Ad Repository Requirements	Platforms must maintain a public repository of political ads with sponsor and targeting details and transparency about algorithms.	Example: A platform not archiving political ads.
Ro-5: Political Ad Labelling	Political ads must be clearly labeled in real-time.	Example: Political ads without labels, making them appear organic.
Ro-6: Influencer Political Content	Influencers must declare political content; platforms must label it.	Example: Influencer posts endorsing candidates without disclosure.
Ro-7: Labelling of Synthetic Media	Synthetic media in political ads must be labeled.	Example: AI-generated candidate videos without clear labels.
Ro-8: Silence Period Enforcement	Platforms must respect national silence periods before elections.	Example: A political ad shown during the silence period.
Ro-9: Monitoring and Fact-checking	Platforms must label fact-checked disinformation.	Example: Misinformation without a fact-check label.

Table 7.2: Overview of newly added categories considering negative consequences in the electoral process. For detailed sources, see Annex C.0.1.

while still being flexible enough to incorporate new categories for emerging risks and online harms. By using legal doctrinal methodologies to develop new coding frameworks for specific content domains, we contribute to the ongoing debate on evaluating and measuring compliance within the intersection of several legal domains.

7.6. CONCLUSION

This paper introduces a comprehensive taxonomy for political content and advertising under the EU’s digital electoral framework, encompassing the DSA, TTPA, and G-E–DSA. By developing a structured approach for annotating, reporting, and monitoring political content, this taxonomy provides a foundational tool for VLOPs, VLOSEs, and regulators to assess systemic risks impacting civic discourse and electoral integrity.

The taxonomy proposed here facilitates a nuanced analysis of political content, identifying risk factors and enhancing transparency in advertising practices. It also serves as a basis for improving content moderation strategies, contributing to a more secure and informed digital public sphere. By systematically addressing the regulatory demands, this approach not only aids platforms in meeting their compliance obligations but also strengthens public trust in the integrity of the electoral process within digital environments.

This approach does present some limitations. Within this text the definition of online platforms is used to refer to digital service providers not necessarily in the same understanding as the DSA, since the other regulations analyzed in this text do not share the same differentiation into hosting providers, online platforms, VLOPs or VLOSEs. We will, however, refer to the definitions outlined in the DSA in the correlating sections.

Overall, this work advances the field of digital regulatory compliance by providing a detailed framework for risk assessment, transparency, and ethical practices in political content and advertising. Future work will focus on refining these annotation categories in response to emerging risks and on developing empirical methodologies to validate and operationalize the taxonomy in various electoral contexts.

8

SCALING UP DSA EVALUATIONS – LEVERAGING LARGE LANGUAGE MODELS FOR AUDITING ONLINE PLATFORMS UNDER THE DIGITAL SERVICES ACT

"[...] Entfalte du die alten Pergamente, Nach Vorschrift sammle Lebens Elemente Und füge sie mit Vorsicht eins ans andre. Das Was bedenke, mehr bedenke Wie. Indessen ich ein Stückchen Welt durchwandre, Entdeck' ich wohl das Tüpfchen auf das i. Dann ist der große Zweck erreicht; Solch einen Lohn verdient ein solches Streben: Gold, Ehre, Ruhm, gesundes langes Leben, Und Wissenschaft und Tugend – auch vielleicht."

" [...] Unfold all your ancient parchments, Then, by rote, collect life's elements, And place them together with due care, Consider What, more deeply consider How. Meanwhile round the world, a bit, I'll fare, And find the last dot on the 'i', for now. Then the great work will see its final stage: Great effort will merit great reward, you'll see: Gold, honour, fame, a long and ripe old age, And science too – and virtue, possibly. Farewell!"

Homunculus, Faust, Part II, Act II Scene II: A Laboratory/ Hochgewölbtes enges gotisches Zimmer, Johann Wolfgang von Goethe

This paper investigates the feasibility and potential role of using Large Language Models (LLMs) to support systemic risk audits under the European Union's Digital Services Act (DSA). It examines how automated tools can enhance the work of DSA auditors and other

ecosystem actors by enabling scalable, explainable, and legally grounded content analysis. An interdisciplinary expert workshop with twelve participants from legal, technical, and social science backgrounds explored prompting strategies for LLM-assisted auditing. Thematic analysis of the sessions identified key challenges and design considerations, including prompt engineering, model interpretability, legal alignment, and user empowerment. Findings highlight the potential of LLMs to improve annotation workflows and expand audit scale, while underscoring the continued importance of human oversight, iterative testing, and cross-disciplinary collaboration. This study offers practical insights for integrating AI tools into auditing processes and contributes to emerging methodologies for operationalizing systemic risk evaluations under the DSA.

8.1. INTRODUCTION

Recent regulatory frameworks, such as the European Union's Digital Services Act (DSA) [224], highlight the increasing importance of managing the work process of auditing digital platforms at scale. Specifically, Article 34 of the DSA identifies four systemic risk categories and the obligation to assess online content through independent audits which are conducted by third-party auditing firms like KPMG or E&Y. Furthermore, the work to audit digital platforms under the DSA also concerns regulators in enforcement and researchers or NGOs as providers of evidence in the compliance process. Traditional audit and content moderation evaluation methods often rely heavily on human review processes, which can be labor-intensive and challenging to scale effectively. [169].

Advances in Large Language Models (LLMs) offer novel and expanded possibilities to address these challenges by automating the auditing process [36, 129, 337] and make it more scalable and cost-effective [127]. Yet, humans remain crucial in auditing content moderation quality and DSA compliance, particularly for complex [127, 207, 337] and interpretative tasks where LLMs may require significant fine-tuning and validation [76, 325]. This paper investigates the potential application of LLMs for auditing online content through automated classification and annotation techniques. To better understand the potential of these technologies in supporting the work of third-party auditors, researchers, and regulators, an interdisciplinary expert workshop with participants from law, social sciences/economics, and computer science was held at *TU Delft* in 2025.

Through collaborative discussions and interactive group work, participants drew on their diverse disciplinary perspectives to explore and critically assess model choices, parameter settings, and prompting techniques—highlighting how cross-domain expertise is essential for successfully navigating the technical, legal, economic, and societal complexities of auditing online platforms. This paper synthesizes the interdisciplinary discussions and evaluations, contributing to the ongoing debate about the potential and responsible integration of AI-based approaches into regulatory auditing practices to enable effective platform governance and meaningfully support the work of auditors.

Our results offer novel insights into the potential for LLM-supported DSA audits through structured workflows and iterative prompt design strategies in interdisciplinary teams to enhance accuracy, increase the audit scope and offer explanation and transparency of annotation decisions.

8.2. RELATED WORK

Recent studies recognize LLMs' utility in automating basic annotations [555, 687] and content moderation [272, 435], yet their limits in legal interpretation, historical content biases, cultural nuance, and hallucination resistance persist [49, 208, 370, 612, 778].

To address the inherent limitations of LLMs, structured workflows that combine human reviews and LLM-driven annotations have been proposed to improve reliability [543, 610, 749, 780]. Several authors suggest rigorous methodologies for annotation of content, proposing systematic prompt optimization, and fine-tuning to enhance annotation accuracy and reliability [127, 717, 718]. LLMs' ability to perform structured deductive coding methods using codebooks for streamlined annotation can differ in efficiency, and it depends on the content type (e.g., image or video) and structure of the material (e.g.,

comment to posts) to be analyzed [129, 207].

Certain LLM application scenarios demand precise alignment with legal standards — for example, auditing approaches under the DSA [54, 337, 548].

While LLMs can effectively assist human annotators, their standalone annotation capabilities for complex tasks remain inadequate. Therefore, cautious integration and validation strategies are warranted [76, 292, 325]. Our study expands this research by integrating interdisciplinary expert perspectives to evaluate the potential and feasibility of using LLMs for auditing political advertisements under the DSA's systemic risk framework.

8.3. METHODOLOGY

An expert workshop (N = 12) involving participants with interdisciplinary backgrounds and experience in collaborative, cross-disciplinary work was conducted at TU Delft in February 2025 (see Table D.1 in the Appendix for an overview of participants' expertise).

The workshop comprised four main elements: (i) an introduction to the DSA and a methodology of LLM-supported content annotation for DSA audits; (ii) interdisciplinary breakout sessions for each team; (iii) a hands-on session refining LLM-supported audits in three teams (details below); and (iv) discussion and feedback.

Experts were divided into three interdisciplinary groups based on their primary and secondary topic preferences, expressed through the collaborative tool *Menti* [515]. Each group focused on a distinct key aspect: 1) The **Legal Team**: Examined how LLM annotations could support DSA compliance, evidence collection, and enforcement, emphasizing annotation's potential for systemic risk evaluation. 2) The **Social Science Team**: Focused on user perspective challenges, emphasizing accountability and transparency necessary for user trust in automated audits. 3) The **Tech Team**: Discussed the general potential of LLMs based on inherent weaknesses and identified strategies for improving LLM accuracy, emphasizing prompt engineering techniques to enhance overall reliability and measurement.

The results of the breakout and hands-on sessions were documented on a Miro board [535] and subsequently analysed qualitatively (see Fig. D.1 in the Appendix for documentation of the boards).

The workshop findings were analyzed using a structured qualitative content analysis [92, 93] to systematically derive insights from visual collaborative outputs.

First, we identified key themes of prominent textual elements (headings, bullet points, post-its, and highlighted texts), using color codes to distinguish the teams: Legal (white), Social Science (green), Tech (yellow) (see Fig. D.3 in the Appendix). The thematic analysis of the Miro board outputs was conducted by the lead author, drawing on Braun & Clarke's reflexive thematic analysis framework [92]. The process began with a close reading of the visual data—including post-its, bullet points, and headings from the breakout and hands-on sessions. Initial codes were inductively derived from these visual elements, guided by the workshop structure (Legal, Social Science, Tech teams) and their respective color-coding (Legal – white, Social Science – green, Tech – yellow). The lead author and workshop organizer, whose background is in platform governance and regulation, discussed iteratively the coding process with a co-author and co-organizer of the workshop, specializing in platform governance with a communication science background. These discussions helped to challenge assumptions, refine thematic categories, and ensure

broader interdisciplinary validity.

In the next step, the extracted insights were systematically organized into thematic categories, reflecting regulatory compliance, annotation strategies, methodological challenges, and recommendations [93]. The visual clustering on the Miro board supported the transparency of this process, allowing for traceability from raw data to final themes. This process resulted in a set of higher-order themes reflecting shared concerns and recommendations across the interdisciplinary teams. An example coding trace is provided in the Appendix D.2 to illustrate the development from raw data to thematic categories. In the Results section (Section 8.4), we synthesize and interpret the identified themes and insights into a coherent narrative across teams, contextualizing them within the existing literature on qualitative methods and annotation processes for legal compliance auditing.

8.4. RESULTS

8.4.1. BREAKOUT SESSIONS

The breakout sessions produced six thematic clusters: Technological Aspects, Organizational Aspects, Methodology and Quality Control, User Aspects, Platform-Related Aspects, and Regulation (see Fig. D.3 in the appendix).

Technological Aspects. Participants emphasized the value of structured prompts and efforts in explainability and control. Open-source models like LLaMA [521] and DeepSeek [338] were preferred “to ensure reproducibility.” Setting the temperature parameter to zero was seen as key to reduce randomness and ensure consistency. Accessing LLMs via Application Programming Interfaces (APIs) makes it possible to specify parameters and manage training/test data splits, contributing to greater methodological stability. Automatic prompt optimization using auxiliary models (e.g., Claude [43]) was considered helpful for improving prompt quality. Additionally, standardized formats (e.g. JSON) were suggested to facilitate processing and structure output.

Organizational Aspects. LLMs were seen as useful for updating codebooks and detecting rare categories. They can support the audit process as dialogic counterparts during the coding process, fostering reflection and documentation. One group proposed to “use the LLM to update the codebook itself,” while another highlighted the potential to “annotate rare codes with a human to verify.” However, assumptions about LLM neutrality were questioned—“no filtering in LLMs, or is there?”—underscoring the need for continued expert oversight.

Methodology and Quality Control. Participants called for repetitive testing and mandatory explanations in outputs. Concerns about automation and platform opacity were raised requiring outputs in automated processes to be reviewed by humans. Participants highlighted the need for model explanations to indicate how specific formulations or examples have impacted the output and enable effective prompt refinement. Further, making use of models’ reasoning capacities may effectively reduce hallucinations [696]. Prompts should be tested repeatedly with varied phrasings to assess consistency. ‘Few-shot prompting’ and the use of an ‘NA’ option were debated due to potential downsides: the former can lead to overfitting on examples, while the latter may cause models to overuse ‘NA’ when facing even slight uncertainty.

User Aspects. Participants debated the limits of platform user empowerment through DSA-introduced rights—both as content creators and as voluntary contributors to audits via content reporting. They cautioned against excessive responsabilization, questioning, for example, “Why would a user have to explain that?” They emphasized that “being a skilled auditor is something separate,” highlighting the distinction between professional auditors and everyday users. As such, they stressed that user-facing tools must include clear explanations and set appropriate expectations, to avoid losing valuable insights that users can offer within the audit scope.

Platform-Related Aspects. Discussions focused on inconsistent enforcement and lack of transparency. One participant asked whether platforms are “even enforcing” their own rules. Participants agreed that LLMs could help reveal enforcement gaps, if embedded in accountable processes.

Regulations. Finally, participants discussed the challenge of translating DSA requirements into LLM prompts. Structured legal definitions were seen as helpful, but overly technical language risked confusing the model. The consensus favored simplified legal guidance co-developed with domain experts.

8

8.4.2. HANDS-ON SESSIONS: EVALUATING PROMPT STRUCTURES ACROSS TEAMS

In the hands-on sessions, interdisciplinary teams collaboratively designed and tested prompt structures using multiple LLMs across diverse content formats documented in the Miro board (see Fig. D.4 in the Appendix). Contributions clustered around: *Structure*, *Prompts*, *Testing*, *Legal Input*, and *Limitation of Scope*.

Structure. Participants emphasized the value of hierarchical prompts, where contextual framing, role-setting, and constrained output logic were clearly layered. JSON formatting and codebook references helped clarify classification schemes. As one team noted, inserting the image “as a comparison with text” helped validate shared context and reduce ambiguity in multimodal classification. However, others noted that model performance generally improved when images and especially videos were translated into descriptive text.

Prompts. Prompt construction varied across teams but typically included a mix of structuring the classification process for LLM use and setting a concrete context for categories. Common examples included: “Please extract the Codenames and Definitions of the (XXX) Categories from the Codebook.” Others preferred multi-step prompts, beginning with legal or domain-specific framing and ending with specific instructions (e.g., “Given the Codebook, identify a category from the image.”). Participants warned against overfitting to specific examples, emphasizing the need to generalize without reducing clarity.

Testing. Testing strategies involved small-scale experiments on representative instances. Incremental steps such as “test e.g., D1 first, then D2” were used to control for failure modes and isolate confounds. Testing across instances was seen as key, as missing factors may distort results. Prompts that explicitly asked for explanations (“demand explanations”) were found to improve transparency and help identify weak reasoning.

Legal Input. Legal grounding was widely seen as essential for improving annotation relevance. Participants found that “legal definitions from the law work well,” but flagged risks when using rare or overly complex legal terms. A preference emerged for using

legal commentaries as contextual examples and allowing for “open definitions of a legal concept or several.” Teams proposed that legal clarity in prompts supported both model accuracy and downstream audit explainability.

Limitation of Scope. Across all teams, prompt scoping was treated as a design strategy. Limiting prompt complexity by splitting up codebooks or restricting the prompt to a few categories increased classification precision. One participant noted that the “limitation of the knowledge base works well,” supporting the idea that reduced scope leads to fewer hallucinations and better alignment. Participants advised against combining multiple tasks in a single prompt, favoring instead modularization and multi-prompt strategies.

8.5. DISCUSSION: CONTRIBUTIONS TO PROFESSIONAL PRACTICE AND PLATFORM OVERSIGHT

This study explored how LLMs can support systematic risk audits under the DSA through an interdisciplinary expert workshop. This approach emphasizes that the task is not merely a technical challenge, but a multidisciplinary design problem that requires coordination across legal interpretation, model behavior, and user-facing transparency.

The use of the visual Miro board supported rapid iteration, interdisciplinary comparison, and collective reflection. The session revealed that prompt engineering is a cross-cutting design task—spanning technical, legal, and methodological concerns essential for responsible LLM use in audits. This study offers a multi-dimensional contribution to the operationalization of systemic risk audits under the DSA by exploring how prompt engineering and LLM-supported annotation workflows can support auditors and other practitioners in their auditing work. Prompt design presents a multidisciplinary challenge that requires balancing legal clarity, model stability, and user transparency. The iterative experimentation across teams demonstrated that carefully constructed, focused, and legally grounded prompts produce more reliable and auditable outputs to support the work of DSA auditors—an essential requirement for applying LLMs to systemic risk audits under the DSA.

The findings demonstrate how structured prompts—especially those using hierarchical formats, legal references, and scope-limited inputs—can have scaling-up potential of audit scope. Participants emphasized the importance of rationale generation (“Let the LLM explain why it made a particular decision”) as quality control mechanisms. These approaches increase reliability and interpretability—two main challenges for LLM use in auditing and crucial for systemic risk assessments.

The hands-on sessions provided meaningful insights for future work developing replicable frameworks for using LLMs in annotation contexts. The use of iterative testing, modular prompt decomposition, and multimodal scaffolding reveals how prompt design choices affect both model performance and human-AI collaboration. Importantly, participants highlighted the need for interdisciplinary co-design across social, economic, policy, legal, and technical domains to avoid misalignment between regulatory intent and model interpretation. This suggests new avenues for empirical studies on measurements such as intercoder reliability in hybrid (i.e., human + LLM) annotation designs and setups.

The study demonstrates how LLM-supported workflows can be designed to scale content analysis for enforcement and evidence collection while also creating documentation

for other stakeholders at scale.

Structured prompts—with embedded legal logic, scoped queries, and demand for justifications—may enable regulators in testing audits to automate large-scale parts of the work process.

Participants highlighted how prompt chains and modular codebook design help manage complexity, allowing for targeted testing across diverse content formats and DSA risk categories. These methods allow practitioners not only to scale but also to systematically identify edge cases and iterate on category definitions. The findings suggest that AI-assisted auditing can support ongoing oversight and comparison across platforms, and support practitioner’s work on the empirical evaluation of platform compliance over time.

Overall, this study highlights that employing LLMs to augment auditing is a design challenge at the intersection of domains that opens novel opportunities to audit at scale in the age of AI-based content generation. Therefore, designing human and LLM-based auditing approaches for the DSA requires methodological rigor, legal clarity, and context specific technical finetuning to ensure that audits remain interpretable, enforceable, and meeting regulatory demands.

8

8.6. CONCLUSION

This paper explored and assessed the feasibility and potential of employing LLMs for supporting the work process of auditing content, specifically targeting systemic risk evaluation under the DSA. Through an interdisciplinary expert workshop and hands-on prompt engineering sessions, we offered novel insights into both the potential and limitations of integrating LLMs into regulatory auditing practices. Our findings highlight that while LLMs provide significant opportunities for scalable and efficient annotation, their practical application requires careful calibration, structured prompting, and rigorous legal and methodological validation. The interdisciplinary approach underscored critical dimensions including legal precision, user-centric transparency, and technological robustness necessary for responsible deployment. This research represents an essential first step toward operationalizing AI-driven auditing within regulatory frameworks, emphasizing the value of and need for collaborative, multidisciplinary and multi-stakeholder expertise and exchange. Future work should focus on building and exploring a broader range of datasets, refining annotation designs and methodologies, and developing iterative co-design processes involving experts from multiple domains to ensure that LLM-supported audits reliably enhance regulatory enforceability.

III

CONDUCTING EMPIRICAL EVALUATION OF DSA COMPLIANCE

9

EVALUATING THE SYSTEMIC RISK OF DISSEMINATING ILLEGAL CONTENT ON ONLINE PLATFORMS UNDER THE DIGITAL SERVICES ACT

"[...] In Italy, six hundred and forty; in Germany, two hundred and thirty-one; a hundred in France; in Turkey ninety-one; but in Spain already one thousand and three!"

"In Italia seicento e quaranta, in Alemagna duecento e trentuna, cento in Francia, in Turchia novantuna, ma in Ispagna son già mille e tre!"

Leporello, "Madamina, il catalogo è questo", Act I, *Don Giovanni*; music by Wolfgang Amadeus Mozart, libretto by Lorenzo da Ponte

Content moderation is a vital condition that online platforms must facilitate, according to the law, to create suitable online environments for their users. By the law, we mean national or European laws that require the removal of content by online platforms, such as EU Regulation 2021/784, which addresses the dissemination of terrorist content online. Content moderation required by these national or European laws, summarised here as 'the law', is different from the moderation of pieces of content that is not directly required by law but instead is conducted voluntarily by the platforms. New regulatory requests create an additional layer of complexity of legal grounds for the moderation of content and are relevant to platforms' daily decisions. The decisions made are either grounded in reasons stemming from different sources of law, such as international or national provisions, or

can be based on contractual grounds, such as the platform's Terms of Service and Community Standards. However, how to empirically measure these essential aspects of content moderation remains unclear. Therefore, we ask the following research question: How do online platforms interpret the law when they moderate online content? To understand this complex interplay and empirically test the quality of a platform's content moderation claims, this article develops a methodology that facilitates empirical evidence of the individual decisions taken per piece of content while highlighting the subjective element of content classification by human moderators. We then apply this methodology to a single empirical case, an anonymous medium-sized German platform that provided us access to their content moderation decisions. With more knowledge of how platforms interpret the law, we can better understand the complex nature of content moderation, its regulation and compliance practices, and to what degree legal moderation might differ from moderation due to contractual reasons in dimensions such as the need for context, information, and time. Our results show considerable divergence between the platform's interpretation of the law and ours. We believe that a significant number of platform legal interpretations are incorrect due to divergent interpretations of the law and that platforms are removing legal content that they falsely believe to be illegal ('overblocking') while simultaneously not moderating illegal content ('underblocking'). In conclusion, we provide recommendations for content moderation system design that takes (legal) human content moderation into account and creates new methodological ways to test its quality and effect on speech in online platforms.

9.1. INTRODUCTION

Understanding how online platforms interpret legal requirements to take down online content remains challenging for scholars and media regulators.[744] Although platforms publish transparency reports and other legal requirements stemming from international, national, or contractual sources for transparency and compliance of content moderation, the vast majority of decisions regarding online content moderation remain extraordinarily difficult to study in depth.[590, 739, 782]

At the same time, there is an increasingly broad societal debate about the governance of social media platforms,[215, 267, 304, 333, 352] of which the governance of online content moderation is a key aspect.[201, 202, 354, 453, 731] However, this debate takes place in an empirical vacuum. For example, online platforms make statements about the extent to which they moderate content, such as Meta's claim about Facebook that the "prevalence of hate speech from July 2020 to September 2020 was 0.10–0.11%",[516] without backing up this statement with verifiable empirical data.

As such, it should be unsurprising that platforms have been heavily criticised for lacking transparency, accountability, and procedural safeguards in their content moderation decisions.[28, 348, 622, 737, 761] At the same time, critics argue that Meta's Facebook blocks either too much[622] or too little[358] content. Considering these claims of overblocking and underblocking, it is argued that online platforms are doing content moderation wrong at all levels.[122, 286]

Perhaps the first and most crucial question is this: *How do online platforms interpret the law when they moderate online content?*

By the law, we mean national or European laws that require the removal of content by online platforms, for example, EU Regulation 2021/784, which addresses the dissemination of terrorist content online. Content moderation required by these national or European laws, summarised here as "the law", stands in contrast to the moderation of pieces of content that is not directly required by law but instead is conducted voluntarily by the platforms. While all online platforms claim to follow the law, whether it is of international, national, or contractual origin, when they moderate online content,[517, 715, 768] the law is not a static object.[84, 722] Differences in how platforms and their critics interpret legal rules for content removal could explain why claims of overblocking and underblocking content are so common.[55, 181, 360]

Content moderation can be based either on the law or on the platform's Terms of Service (ToS). Often, an overlap between the two grounds is possible. Therefore, the relationship between legally required content removal and content removal based on a platform's ToS is blurry for various reasons. To a certain degree, all platforms have integrated legal norms into their ToS; some legal frameworks, such as the EU's Audiovisual Media Services Directive (AVMSD), even explicitly require platforms to change their ToS.[13, 246] Similarly, the Digital Services Act (DSA) has created new rules regarding ToS, which are setting the stage for future research in the content moderation domain. According to Art. 14, the Act demands that ToS "include information on any restrictions that [providers of intermediary services] impose in relation to the use of their service".[246] Since at the time this research project was conducted, the DSA was not in place, we will not focus on it here. Some platforms, such as Meta's Facebook,[739] have explicitly tried to keep their ToS separate from country-specific national legislation, for example,

the German Network Enforcement Act (NetzDG)[15] and the Austrian Communication Platform Act (KoPL-G).[57] There has also been a frequent concern that platforms might implement takedown legislation overbroadly to limit the likelihood of legal liability from incorrectly interpreting the law.[118, 135, 782]

Finally, many countries have criticised large international platforms, such as Twitter (now named X), for their lack of willingness to follow ‘local’ laws.[617, 758] This claimed lack of willingness of large international platforms to implement legal frameworks effectively has been a critical driver of regulatory debates around the German NetzDG and Austria’s KoPL-G.

This article attempts to answer this fundamental question: *How do online platforms interpret the law when they moderate online content?* In doing this, we aim to provide a much-needed empirical basis for the existing debate on content moderation online. After clarifying the scope and key definitions, we discuss our case selection. We decided to study content moderation in Germany, as we believe this is one of the most interesting country use cases. Germany has complained the most loudly about the lack of international platforms’ adherence to the law[211, 617] and was among the first European Member States to craft corresponding regulations.[7, 15, 57]

To study content moderation in Germany, we first explain why content should be legally removed in the specific national context and how ToS typically allow for content removal. We then present our methodology and its results when applied to our dataset. We analysed the content moderation data of a medium-sized German online platform with approximately 720,000 posts of user-generated content per month. This online platform uses both paid moderation staff and community moderators to conduct its content moderation.

Based on this analysis, we attempt to make empirical claims about the relationship between ToS and legal content moderation. Additionally, we analyse how online platforms interpret the law in content moderation. This opens up discussions about legal content moderation decisions. In conclusion, we suggest that content moderators receive legal training and are provided with enough time in their work to make informed decisions. Furthermore, we see the need for new standards for legal interpretation in the content moderation domain. We also suggest more transparency about the content moderation process on online platforms. We identify the need for better insight into the reasons for making moderation decisions on platforms. At the same time, we understand the complexity of the underlying problem and the difficulty of interpreting the law in content moderation. We believe that a more misinterpretation-aware understanding of content moderation and problem-aware design solutions are central to effectively addressing the challenge. Finally, we emphasise the need for easy and reliable researcher access to validate and control online platforms’ content moderation systems and processes.

9.2. DEFINITIONS, SCOPE, AND LIMITATIONS

To answer our research question: *How do online platforms interpret the law when they moderate online content?*, it is first necessary to develop a research design that considers how to study the relevant legal provisions and ToS.

The research question relates to how platforms interpret the law, which shapes online environments.[685] In this context, we study “the law as a behavioural system”[685] and,

specifically, the role of the law in inducing behaviours by those trying to enforce legal rules. Notably, by law or legal provision, we differentiate between what is required by legal provisions that require the removal of content and what is required by contractual law, such as ToS, which is defined as all the platform rules, codes, and norms used to moderate content on an online platform that are not explicitly mandated by the law in a specific jurisdiction. An example of the contractual relationships between users and platforms is an online platform's Community Standards (CS), which are included in the ToS. The CS describe undesired behaviour within the online platform and the consequences for violating such rules, for example, de-platforming, applying warning labels, etc.

9.3. LEGAL GROUNDS FOR GERMAN PLATFORMS TO MODERATE ONLINE CONTENT

This section provides an overview of the norms relevant to the German context of content moderation, including the perspectives of human rights and European, international, and national provisions.

9.3.1. OVERVIEW OF HUMAN RIGHTS, EUROPEAN PROVISIONS, AND INTERNATIONAL PROVISIONS IN THE SCOPE OF CONTENT MODERATION IN GERMANY

Platforms located in Germany must follow German law, and Germany has its own national provisions that apply to online platforms. Furthermore, Germany is a Member State of the EU. The EU protects human rights as a primary law in its Charter of Fundamental Rights, which is applicable in the context of EU law. Moreover, the European Convention on Human Rights (ECHR) protects cases that are not limited to application within the context of EU law. Additionally, the national constitutional norms addressing freedom of information and expression can be found in Art. 4 and Art. 5 of German Basic Law.

The Audiovisual Media Service Directive (AVMSD)[13, 16] harmonises the audiovisual media landscape by proposing rules that are within the scope of traditional forms of media, which include broadcast TV and on-demand services. In its definition of “services”, the AVMSD includes linear and nonlinear services. The AVMSD addresses the rules for video-sharing platform services (VSPs) in Art. 28a and 28b.

The obligations for VSPs include taking “appropriate measures” to protect minors from content that might “impair their physical, mental or moral development”, such as content that contains incitement to violence or “hatred directed against a group of persons or a member of a group based on any of the grounds referred to in Article 21 of the Charter”, or content that constitutes a criminal offence “under Union law, namely public provocation to commit a terrorist offence [...] offences concerning child pornography [...] and offences concerning racism and xenophobia”. According to Art. 28b(3) of the AVMSD, whether a measure of a VSP is classified as “appropriate” or not shall recognise the content in question. Furthermore, the factor of “appropriateness” weighs the potential harm of the content, the characteristics of the protected group, and the rights and legitimate interests of the case. Those interests may include the VSP, the creators of the content, and the general public.

The AVMSD prohibits the inclusion of ex-ante moderation measures (such as upload

filters). However, it foresees specific rules for the protection of minors under Art. 6a and Art. 28b(1) lit a to avoid the risk posed to minors, for example, through Child Sexual Abuse Material (CSAM). While we acknowledge the risk that CSAM content poses to online platforms, our analysis does not specifically focus on CSAM content on online platforms due to the specific use case and legal uncertainty about the proposed regulation. The “appropriative measures” to be taken by VSPs are further described in Art. 28b(3) AVMSD as actions involving ToS, notice-action mechanisms, information about the flagging process, age verification, parental control systems, content-rating functionality, or media literacy measures and tools.

Additionally, international norms in the domain of online platforms are primarily binding regulatory instruments for states. They might also be relevant to platforms in specific contexts (e.g., consumer protection) connected to the services provided.

Platforms operating in Germany take into account several provisions from the corpus of international law for the regulation of speech. This corpus can include sources such as the Security Council resolution 1566, the General Assembly’s Declaration on Measures to Eliminate International Terrorism, the 1999 International Convention for the Suppression of the Financing of Terrorism, the International Covenant on Civil and Political Rights (ICCPR), the Declaration of the Rights of the Child, the United Nations Convention against Transnational Organized Crime and the Protocols, Article III of the Convention on the Prevention and Punishment of the Crime of Genocide, the Rome Statute of the International Criminal Court, with the corresponding Articles 25(3) and 6 focusing on the prevention of the incitement to or promotion of genocide.

9.3.2. NATIONAL PROVISIONS IN GERMANY RELEVANT TO CONTENT MODERATION

Within national law, the German NetzDG has regulated aspects of content moderation since 2017.[15] The NetzDG provides a list of norms that shall be qualified under its definition of “unlawful content”, according to Art. 1(3). Furthermore, the national regulation includes a biannual reporting obligation that provides insight into the complaint management of unlawful content from platforms. Additionally, NetzDG’s Art. 2 requires platforms to report on their mitigation efforts around “criminally punishable activity”, information on the mechanisms intended to flag unlawful content, and details about how the decisions in the process of setting content moderation actions (e.g., to delete the content in question or to block access to it) are made. Also included in the required information are the quantity of unlawful content, the details about the flagging of users (complaint body or “normal” platform user), and the reason for flagging the content in the first place. In addition, the NetzDG asks for detailed information regarding the human content moderation teams and their organisations. Since the recent revision of the law (NetzDGÄndG), the concrete link between Art. 3d, the AVMSD, and the NetzDGÄndG has become even clearer and stresses the importance of holistic compliance and auditing approaches across legal norms.[20]

Information about how the human staff is educated, their linguistic expertise, and how these teams are supported must also be provided. German law demands information about the external parties consulted in the decision-making process. Such a report must include information about the time of action the platform requires after being informed

about a violation. In addition, the flagging user and the user who uploaded or generated the content must be informed of the decision taken. Art. 3 para. 2 of NetzDG defines the specifics of the system design for the notice and action mechanism. Paragraph 2 gives platforms a general time span of 24 hours to act upon unlawful content after they receive the notice, unless the platforms have agreed upon longer reaction time spans with competent law enforcement. Art. 4 of the NetzDG includes provisions on fines of up to 5 million euros.

9.3.3. TERMS OF SERVICE OF ONLINE PLATFORMS IN GERMANY

Platforms seem to “prefer” violations of the ToS over illegal content violations.[739] ToS violations are, per se, of a contractual nature (e.g., breach of contract) and, therefore, not necessarily illegal. A ToS violation is also easier to enforce due to platforms’ greater discretion in enforcing the ToS rules they have drafted themselves. Enforcing Community Standards (CS), however, must be automated to handle the sheer amount and velocity of user-generated content. Each platform creates ToS and CS individually, shaping the notion of allowed and inappropriate behaviour. Several advantages and insights into the content moderation process shaped by the CS are only inside the platforms; such advantages might include a rich, internally labelled training dataset (internal ground truth dataset) that, for example, can be used for algorithms and machine learning (ML) purposes, the possibility of changes in the interpretation of CS, the wording or the consequences of a violation of the CS that build an understanding, the examples or training data sources of AI tools to enforce content moderation, and so on. Therefore, more advantages result in a focus on moderation-harmonisation and standardisation. Because of the economically favourable concept of enforcing ToS violations, large online platforms have prioritised reporting mechanisms through design features, such as dark patterns. Dark patterns refer to design choices that should nudge the user to a specific desired behaviour.[98]

One example of such a design nudge is to allow easy reporting of ToS violations for users and cumbersome reporting features for reporting illegal content online.

Furthermore, platforms can customize rules and reasons for moderating content within their ToS for specific reasons. Such detailed rules of online behaviour are often referred to in the appendices of the existing CS (for instance, including COVID-19 moderation reasons). To better understand the ToS rules in Germany, we decided to take a close look at Meta’s (then Facebook’s) Community Standards. As one of the largest online social networks and the largest social network in Germany, it seemed like an obvious choice to better understand ToS and online content moderation. When looking closely at Meta’s CS, we identified 24 categories specifying behaviour that constitutes different categories for moderation. Additionally, we added the CS of a smaller German platform to our categories.

9.4. RESEARCH DESIGN AND CODING METHODOLOGY

The empirical analysis of the content moderation data relied on an anonymised German-language dataset containing all posts by online forum users from 1 January 2019 to 30 April 2019, regardless of whether they had been deleted by moderators or flagged by users.

The dataset covered 2,887,228 user-generated posts.

All information about the users who made the posts was removed in a previous step towards the anonymisation of the dataset. However, there was always the theoretical possibility of re-identifying users using their expressions and specific content, as their posts with usernames attached were publicly available online. As such, the anonymisation process did not provide full anonymity for the users, as it was possible to reverse it. As a research team, we agreed that we would not take any steps to identify users to ensure their anonymity.

Importantly, the second dataset also included the reasons for the deletion of the content. This allowed us to interpret the reasons for the platform's deletion of content. However, we only looked in detail at whether the platform deleted the posts for 'legal' reasons, i.e., that they believed they were compelled to remove these posts under German law, or whether they removed the posts for 'other' reasons, which we categorised as platform ToS violations.

Additionally, because of the ToS and the high degree of contractual freedom in creating the CS, the granularity of such provisions varies from platform to platform. Those variations might address the clauses as such (e.g., including specific rules for moderating content addressing COVID-19 in specific ways), the detail of the definition per category (for example, how far the description of "nudity" could be drawn), the level of granularity of the explanation for a definition (for instance, how could "nudity" be understood in the context of the platform), the number of examples given (e.g., what emojis could be moderated in the context of unsolicited online behaviour), or the act followed by the violation of such rules (such as the deletion of the content, visibility reduction, warning labels, strikes to the user's account, etc.). Because creating and designing ToS is complex and costly, large online platforms, such as Meta/Facebook CS, provide a rich example of more "sophisticated" rules upon which to base an analysis. In our study, we used five leading categories in the code book provided in Annex I: Violence and criminal behaviour, Safety, Objectionable Content, Integrity and Authenticity, and Respecting Intellectual Property. We included five subcategories for the class referring to Violence and Criminal Behaviour, six for Safety, five for Objectionable Content, seven for Integrity and Authenticity, and one for Respecting Intellectual Property, totalling 24 distinguishable CS.

To code the data, we first generated two samples of the dataset from the online platform, with 800 randomly selected items of all content deleted by the forum moderators (0001) and 800 randomly selected items of all content posted on the platform (0002), out of a total dataset of 2,887,228 items. For the sample of 800 randomly selected items of all content, we expected a response distribution of approximately 0.5%, giving us a margin of error of 0.32%, with a confidence level of 80%.

For the sample of 800 randomly selected items of deleted content, we expected a response distribution of approximately 5%, giving us a margin of error of 0.99%, with a confidence level of 80%.

While we appreciate that using a confidence level of 80% is lower than would typically be used and that the margin of error with a sample of this size is quite large, the goal of this study was not to achieve a perfect quantification of problematic content but instead to establish broadly accurate estimates for how platforms interpret the law. The margin of error is thus believed to be acceptable, as long as it is clearly communicated to the reader

and the details are made transparent at every stage of the research process.

9.4.1. CODING GERMAN LAW

Once the data were reduced to a representative sample, they were coded based on a previously developed taxonomy of international and German law,^[702] including categories for Meta's/Facebook's CS, which, to our knowledge, represent one of the most detailed CS available. This was combined with the CS from a representative of a small/medium-sized platform from a Member State, as mentioned above in Section 2 and which is provided in Annex I. In addition, to address the need for content not falling into a contractual or legal category, the research team created two new categories: 4-1- and 4-2- under the class "otherwise sensitive content", which addressed the possibilities of the annotators classifying presumed problematic content without a legal reason (4-1) and to form legal analogies (4-2) that cover judicial gaps or very intense content. Notably, additional data that we received from the platform about how the moderators had categorised the data was not shared with the coders at any time during the coding process. The coders only knew that they had two datasets to code but not that one of them consisted of deleted content. They were unaware of the platform's categorisation of the data points or even that all the content in one sample had been deleted and that the other constituted a sample of all platform content. They were only provided with two representative samples of content from a medium-sized German platform and asked to code this content with the codebook at hand. This step was taken to ensure that the coders were not biased by the platform's pre-existing decisions about content moderation, as they were not aware of them.

The core coding team consisted of two advanced law students who, at the time, had studied law for at least three years. They were supported by three senior colleagues who comprised the Moderation Decision Supervisory Team, two of whom had an extensive legal background, and one of whom had received a German habilitation in law and was recently appointed as a law professor in Austria. The two advanced law students coded the dataset independently before discussing their results in a meeting with the Moderation Decision Supervisory Team to make a final determination of all the diverging decisions of the two legal coders. This process ensured the sophisticated legal qualification of all the moderation decisions within both samples.

At the beginning of the coding process, a test run with a smaller representative sample of 90 posts from each dataset was conducted to ensure that the taxonomy could be applied effectively to the dataset. This test run led to some clarifications of the taxonomy, as well as a wider debate among the team about how different legal categories could be meaningfully interpreted. The initial analysis produced an intercoder reliability rate on the question of the legality of the content of 83.15% for Test_Dataset_0001 (all deleted content) and 96.55% for Test_Dataset_0002 (all content), which we hoped to improve in the main analysis.

Importantly, the coders also had access to contextual information about the post, should it be required. This contextual information consisted of other posts in the same thread of conversation, with two levels of granularity in the context (main thread and sub-thread). In the case of the German platform we were studying, to our knowledge, the content moderators employed by the platform also had access to contextual information

in the main thread. Only for some pieces of content was the context unavailable. Such cases of unavailable context were documented by the coders. The coders also recorded the number of situations in which they required additional contextual information to decide whether the context was needed for the main thread or the corresponding sub-thread.

In addition, the coding team kept track of the time needed to make the coding decision for each comment and inserted the coding decision into the template used. Coding time was tracked separately for the additional time needed to interpret additional contextual information without separating the level of necessary context (inclusive time measurement for the main thread and sub-thread). This time tracking helped provide a deeper understanding of the time it takes to code legal information and how much additional time was needed for decisions that demanded a deeper context. The time measured for both coders to decide without requiring additional context was between five seconds and 53 minutes in total. For decisions that consulted an additional layer of context, the time span was between 22 seconds and 43 minutes. One hundred and eighty-three decisions required partial consultation of the context, and 141 decisions needed detailed consultation of the context. Information about the context, however, was only available in 162 cases, leaving 21 cases without context.

The coders for this study also had the possibility of discussing inappropriate, disturbing, or traumatising content with the Moderation Decision Supervisory Team when needed. Content moderators for large online platforms are often not employed by the platforms but instead are often contractors. Content moderators who are contractors often have different working conditions in which to make their content moderation decisions compared to their colleagues employed by the platforms. Former contractor content moderators for Meta's Facebook called Cognizant, for example, had to decide on a piece of content in about 30 seconds. Since the type of content (short text, long text, text with emojis to accentuate the meaning, text with emojis to replace words, pictures, pictures and text, videos, audio, 3D environments, etc.) is also related to its complexity and, therefore, the time needed to make a moderation decision, no meaningful legal interpretation was generally possible following the level of detail used for this analysis. Strict "productivity" goals, such as the number of moderated pieces per day or per hour or accuracy rates of about 95%, also increase the pressure on moderators when making their decisions.

Finally, an essential part of the coding framework was a common "heuristic of standards of proof", which had to be applied when making the coding decisions and was designed as a benchmark for the coding team to have the same "degree of strictness" in reaching the coding decisions. The Moderation Decision Supervisory Team discussed this extensively within the coding team and agreed on coding the content at a final stage based on a heuristic of standards of proof that adheres to a variety of different legal standards differentiating between international law, criminal law, and civil law because "since the Normative Taxonomy of Expressions deals with legal circumstances rooted in international and domestic (German and Austrian) law, we had to apply different standards of proof".^[702]

Therefore, the heuristics of standards of proof mimic the different law categories used in the taxonomy. For international law categories, the heuristics of standards of proof of the European Court of Human Rights (ECtHR), the International Court of Justice (ICJ),

and the International Criminal Court (ICC) were considered. The common element for the three courts is the shared definition of standard of proof and the quality of “beyond reasonable doubt”. Additionally, the ECtHR demands that “such proof may follow from the coexistence of sufficiently strong, clear and concordant inferences or similar un-rebutted presumptions of fact”. Furthermore, the ICJ includes “preponderance of evidence”, a heuristic in which the parties of a case can agree on the standards to use for the case.

For classifying heuristics of standards of proof for German criminal law, we decided to use three different standards of proof following the jurisprudence: the standard of “initial suspicion”, which is not only a presumption according to §§ 152 II and 160 Code of Criminal Procedure (StPO), the standard of proof for convictions according to § 261 Code of Criminal Procedure (StPO), and the standard of “sufficient suspicion” according to §§ 170 I and 203 Code of Criminal Procedure (StPO). Because the standard of “initial suspicion” would be too broad for classification, and the second standard of proof for convictions “cannot be operationalised since an assessment is only possible after full evidence has been considered (in oral proceedings before the court)”, we chose the third standard of “sufficient suspicion”. The criminal law heuristics, therefore, can be understood as follows: “A sufficient suspicion exists if, after a preliminary assessment of the facts resulting from the entire file content and the results of the evidence, a conviction of the accused is more likely than an acquittal, and there is, therefore, a predominant probability of a conviction. Due to the vagueness of the legal term, the prosecution has a margin of judgment when filing a lawsuit. In doing so, she makes the prognostic decision as to whether, after a preliminary assessment of the factual and legal situation at the end of the main hearing, she would probably come to an application for conviction. To examine the sufficient suspicion of the crime, both the legal and the actual aspects must be assessed. The matter must, therefore, be punishable from a legal point of view and can be proven with the means available.”

For the third category of deciding on the heuristics of proof, which covers private law claims, we decided to take “credibility” as the standard to use according to § 294 Code of Civil Procedure (ZPO), which focuses on overwhelming credibility: “A fact is made credible if it can be regarded as true with an overwhelming probability”. We prefer this definition and understanding of standard of proof (§ 294 ZPO) compared to § 286 ZPO as an alternative understanding, which would require full proof of evidence, a definition we believe is not appropriate for digital content.

After both members of the coding team had finished working on the dataset, the coding decisions of both coders were compared to determine the intercoder reliability. In the final dataset, the intercoder reliability was 92.50% for Dataset_0001 (all deleted content) and 97.00% for Dataset_0002 (all content), leading to a combined intercoder reliability for both samples of 94.75%. Following Ali et al., Cohen's kappa is used to measure intercoder agreement for classifying objects into categories for two raters with an expected above-chance base agreement proportion in their rating. We argue that Cohen's kappa is a reliable metric for measuring intercoder reliability, that the result is well within a reliable margin of error for coding results, and, thus, that our coding process based on the taxonomy we developed is robust enough to provide reliable results.

The legal coders noted their degree of certainty about their decision...

9.5. RESULTS

The empirical findings of this study show considerable divergence between the platform's interpretation of the law and our own legal analysis. Our comparison of the platform's moderation decisions and the results of our legal coding highlights both the subjectivity and the practical constraints of content moderation in practice.

9.5.1. EMPIRICAL COMPARISON: PLATFORM DECISIONS VERSUS LEGAL CODING

Based on our coding, a significant proportion of content deleted by the platform for supposed legal reasons was not, in fact, illegal according to our team's legal evaluation. Conversely, there were also instances in which content that was left up by the platform should have been removed to comply with the law.

We identify two key types of systematic errors:

- **Overblocking:** Content that is legal is deleted by the platform, based on an incorrect or overly cautious interpretation of legal requirements.
- **Underblocking:** Content that should be removed under the law is left up, due to either misclassification or a lack of legal knowledge on the part of the platform moderators.

Overall, the data suggest a high rate of misclassification, with a notable tendency toward overblocking. This indicates that platforms, possibly out of an abundance of caution to avoid legal liability, err on the side of removing content that may, in fact, be protected by law.

9.5.2. SUBJECTIVITY AND CONTEXT IN MODERATION DECISIONS

Our analysis further demonstrates that content moderation decisions are highly context-dependent. Coders often required substantial contextual information—such as preceding and following posts, thread context, or user history—to make accurate determinations regarding legality.

The coding process revealed that:

- Approximately 20% of decisions required context from the main thread or sub-thread to be assessed correctly.
- The need for context added significantly to the time required for each moderation decision.

9.5.3. TIME AND ACCURACY CONSTRAINTS

A key challenge identified in our study is the trade-off between the speed and the quality of content moderation. The time measurements recorded by our coders suggest that accurate legal classification of content frequently exceeds the time typically available to content moderators in professional settings, especially those working under strict productivity requirements.

For example:

- Decisions without context were typically made within a few seconds to a minute.
- Decisions requiring additional context or legal research could take up to 40 minutes.

These findings highlight the practical limitations of expecting consistently high legal accuracy from platform moderators, especially when productivity quotas and limited training are factors.

9.6. DISCUSSION

Our results indicate that platforms frequently misinterpret legal requirements, resulting in both the overblocking of legal content and the underblocking of illegal content. The empirical findings underscore the challenge of aligning platform moderation with nuanced legal standards, especially given the speed and scale required for day-to-day operation.

One central finding is the subjective nature of content moderation. The empirical process highlighted the considerable degree of interpretation involved in legal classification, as well as the impact of limited context, moderator experience, and time constraints. The legal coders' disagreement rates, while ultimately resolved through supervision, demonstrate the complexity and ambiguity of many moderation decisions.

Moreover, platform moderators tend to err on the side of caution, prioritizing risk avoidance. This has important consequences for online freedom of expression, as legal but controversial or borderline content may be disproportionately suppressed.

9.6.1. IMPLICATIONS FOR PLATFORM GOVERNANCE

Our study highlights the need for increased transparency and accountability in content moderation processes. In particular, we recommend:

- **Legal training:** Platform moderators should receive comprehensive legal education relevant to their jurisdiction and content moderation duties.
- **Adequate time and support:** Moderators should be given sufficient time and contextual resources to make well-informed decisions.
- **Clear standards:** Platforms should develop more transparent and consistent standards for moderation, grounded in law but clearly distinguished from contractual ToS-based decisions.
- **Transparency:** Public reporting on moderation outcomes, criteria, and disagreements between legal interpretations should be standard practice.
- **Access for researchers:** Independent researchers should have access to anonymized platform data and moderation processes, enabling external validation of moderation practices.

9.6.2. POLICY RECOMMENDATIONS

Policymakers and regulators should:

- Establish clearer guidelines for distinguishing legal obligations from platform-specific contractual restrictions.
- Encourage or require platforms to implement robust notice-and-action mechanisms that are both user-friendly and legally compliant.
- Mandate external audits or reviews of content moderation systems, focusing on legal accuracy and the proportionality of moderation decisions.
- Foster collaboration between platforms, legal scholars, and civil society to improve interpretative standards and reduce the risks of both overblocking and underblocking.

Finally, the study underlines the importance of building systems that acknowledge the inherent uncertainty and subjectivity in legal interpretation. Rather than aiming for perfect legal certainty in every individual decision, platforms and regulators should design systems that can be transparently reviewed, contested, and continuously improved.

10

EVALUATING THE DESIGN OF REPORTING MECHANISMS FOR ILLEGAL CONTENT IN PRACTICE: USER PATHWAYS, INTERFACES, AND COMPLIANCE

“It is true that architecture depends on facts, but its real field of activity is in the realm of significance.”

Ludwig Mies van der Rohe, address at the Illinois Institute of Technology, 1950

Online platforms face persistent problems involving harmful and illegal content. Flagging can empower users to report content they believe violates the law or platform rules. The European Union’s Digital Services Act (DSA) seeks to harmonize aspects of this reporting process. This paper examines how reporting mechanisms support user action through semi-structured interviews (N = 12) with regulatory authorities and professional reporting experts, using walkthrough case studies of Facebook and TikTok interfaces. We identify a tension between empowering users through additional reporting options and burdening them with complex interfaces and procedures: users may need to understand legal concepts, participate in a formal process, and distinguish illegality from violations of terms of service. Design choices, including the number and length of reporting steps, also shape expectations about procedural transparency. We conclude with design insights for supporting users and other stakeholders in the reporting process.

This chapter is based on: Sekwenz, M.-T., Wagner, B., & Parkin, S. (2025). “It Is Unfair, and It Would Be Unwise to Expect the User to Know the Law!”—Evaluating Reporting Mechanisms under the Digital Services Act. In *FAccT* '25, pp. 532–546. ACM. doi:[10.1145/3715275.3732036](https://doi.org/10.1145/3715275.3732036). Research instruments and supporting tables appear in Appendix F.

10.1. INTRODUCTION

Offensive content regularly appears on social media platforms (SMPs), undermining the experience and safety of users. Removing harmful/illegal content is a central task of content moderation on SMPs [317]. As content moderation does not catch all violations, users can report through flagging items or content, also referred to as a notice action, or reporting mechanism (RM) [439]. Content reporting helps platforms catch violations missed by automation and reflects user experiences. For example, in February 2025, a technical failure on Instagram led to a spike in reports after users were exposed to graphic content, prompting a public apology from Meta [108].

Regulatory efforts around the world are moving to reconcile activity on these platforms with legal oversight, as some content goes beyond breaking a platform's Terms of Service (ToS) and into being recognizably against the law. In the European Union, the most important regulation in this domain is the Digital Services Act [224]. It is regulating the process of content moderation, albeit not setting material rules for how the content itself has to be moderated [694]. This raises questions around how the additional user rights afforded through the DSA are empowered, and users appropriately protected from harmful, illegal content while navigating the law amid existing platform-specific ToS.

This study seeks to situate the challenges faced by users and various stakeholders within the framework of the DSA reporting mechanisms (RMs), drawing on interviews with legal and regulatory experts and a heuristic walk-through of reporting interface designs, based on real-world implementations on SMPs.

Our overarching research question is: How do usability and legal compliance interact in reporting mechanisms for illegal content under Article 16 DSA?

We address this question through three sub-questions:

1. How do experts perceive the ways in which different users' flagging behaviours affect the usability of DSA reporting mechanisms on social media platforms?
2. How do experts interpret differences in the usability of DSA reporting mechanisms during heuristic walkthroughs of selected use cases?
3. How do experts explain the effects of interface design on DSA reporting within the wider content-moderation process?

The goal of the DSA is to “set out harmonized rules for a safe, predictable, and trusted online environment” (Article 1 DSA), and it applies to platforms that offer their ‘services’ to users in the EU. Similar to Section 230 of the U.S. Communications Decency Act [720], which limits liability for problematic speech on SMPs, the DSA imposes no obligation to ‘generally monitor’ content (Article 8 DSA). DSA flagging covers illegal content reporting mechanisms as outlined in Article 16(1) DSA. These must be ‘easy to access’ and ‘user friendly’ as key obligations under EU law (Recital 50, Recital 52, Article 16 (1) DSA).

This represents a new convergence of legal-interface design. The remit to bring legally-recognized reporting structures direct to users, and at the same time ensure easy-to-use processes, brings potential for a range of tensions in reporting interfaces and associated processes. For instance, users often do not understand the categories or processes involved in reporting content and follow a “hedonic mindset” [540, p. 9]. Flagging is a multi-actor process and “a complex interplay between users and platforms,

humans and algorithms, [implementing] social norms and regulatory structures of social media.” [175, p. 412].

We argue that the way SMP interfaces are designed and the affordances these interfaces provide are increasingly driven by legal frameworks. As such, we have developed a methodology (Section 10.4) that takes these legal frameworks into account, providing a systematic perspective on how both law [139] and technology [765] shape human behavior. To this end, we conduct a study with regulators and other expert users rather than the average users of the RMs; these participants have knowledge of the hard bounds of legal requirements. This is reflected in our adaptations to the heuristic walkthrough approach.

Our findings (Section 10.5) contribute to the developing discourse around *legal design*, and methods to examine social, legal and technological dimensions in tandem. Our expert participants noted that the DSA provides new possibilities for empowering users and flaggers in the reporting process, which on the other hand, also burdens users due to the need for legal understanding of filing illegal content reports, compounded by a need to distinguish illegal content from content against an SMP’s ToS.

Our results also illustrate ambiguities in the role of TFs and how non-expert users can reasonably engage in flagging processes. We found tensions between DSA obligations of creating ‘easy to access’ and ‘user friendly’ RMs and differences in design choices on SMPs, which can further complicate the reporting process. Besides, interface designs may impact the reporting process, user experience, and what can be expected of the user. Additionally, experts also highlighted that SMP moderation activity remains opaque even to them, highlighting the burden on non-expert users.

Despite the promise of the Digital Services Act to create safer online spaces through user-friendly reporting mechanisms, the complexity of legal requirements and opaque platform practices could undo efforts to encourage user reporting, leaving open questions as to whether reporting systems can effectively empower non-expert users. In discussing our findings (Section 10.7), we contribute recommendations for practitioners including how to embody legally-recognized user rights within (usable) interface designs, feedback loops for users reporting content and increased process information, and better cooperation between actors in the complex content moderation ecosystem.

10.2. BACKGROUND

Where there has been prior research on ‘flagging’ of content, this has for example been in a USA context [175, 671] where a flag is a report of content, as a mechanism of governance. Increasingly under emerging legislation around the world, flagging can encompass the reporting of content under either Terms of Service (ToS) or as illegal content (which is then subject to legal process ‘outside of’ the platform on which it is reported). Addressing legal/regulatory compliance and design aspects in tandem, specifically for (user) flagging, is a notable research gap (where prior work has examined platform-driven reporting structures [551, 776]), and is especially relevant for Europe’s Digital Services Act (DSA) [224] which creates new rules for reporting systems in Article 16 for online platforms operating in the European Union according to Article 2. The DSA imposes stringent obligations on SMPs for example on the design of illegal content RMs (Article 16(2) DSA). To ensure compliance, national Digital Service Coordinators (DSCs) and, ultimately, the European Commission oversee enforcement, with fines of up to 6% of annual global

turnover for violations (Articles 49–52, 56, and 72 DSA).

Other jurisdictions like the UK [569], India's [534], or California's [112] echo the DSA's aim of regulating SMPs. In the EU Germany [284], France [275], and Austria [617], had laws prior to the DSA.

Unlike platform ToS, which platforms set unilaterally, the DSA establishes 'must-moderate' criteria which can overlap with and complement a platform's ToS regarding problematic content such as deepfakes [270], antisemitic material [491], or harm against animals [531]. Therefore, RMs are legal instruments of the "notice and takedown policy" [375, p2].

Few prior EU regulations demand similarly transparent and 'user-friendly' mechanisms to report content [239, 243, 244]. The DSA specifically requires that Reporting Mechanisms (RMs) be "clearly identifiable, located close to the information in question" and at least as accessible as mechanisms for reporting ToS violations (Recital 50, Article 14 DSA). Users must be able to flag multiple pieces of content in a single notice (Recital 53 DSA), and those reporting must be informed and offered means of redress (Recital 58 and Article 16(5) DSA). Additionally, platforms must process notifications promptly, objectively, and transparently, disclosing any automation involved (Recital 61 and Article 16(6) DSA).

The DSA also introduces Trusted Flaggers (TFs), who receive priority treatment for their notices. Accredited by DSCs (see Section 10.6 Recital 61, Article 22 DSA), TFs are expected to demonstrate expertise and unbiased judgment, further professionalizing the moderation process. These Trusted Flaggers would be active alongside Professional Reporters (representing users), and normal users. Article 16(1) DSA regulates 'illegal' content reporting. Because 'illegal' content reporting requires understanding of the legal domain (difference of ToS and law, legal jargon, etc.) and is a "barrier for [normal] user engagement" [671, p. 5] DSA-RMs are an expert system.

10.3. RELATED WORK

In studying flagging and reporting behaviors in online gaming environments, Kou and Gui [439] note that flagging content is not simply a problem-solving exercise, as content can differ in both 'flaggability' and toxicity. Jhaver et al. study end-user perceptions toward social media content moderation tools [393], finding that participants were reluctant to set content filters, and expected platforms to moderate offensive content on their behalf.

Wei et al leveraged insight from experts to evaluate online safety advice for users in the context of online hate [753], finding that muting content is regarded as a 'best practice' but that it should also be reported to SMPs and civil society organizations (mirroring the social role of flagging mentioned by our participants, Section 10.6.3).

Related research [395] surveying hundreds of US users found that participants managed only their own feeds using content-hiding features, as a means to still allow the free speech of others. Zhang et al. [776] raise questions about the transparency of the reporting process and how content moderation effort is distributed between users and an SMP, where our work identifies many moving parts between stakeholders (all while the SMPs themselves remain as something of a mystery to other stakeholders). Where Gomez et al. [296] raise concerns about the arbitrary nature of content moderation decisions (especially where AI is involved) and the impact on human rights, our findings

(Section 10.6.2) indicate that reporting entities acting on behalf of users equally have little insight into decisions around content removal (thereby compounding harms to affected users). Similarly, Naab et al. [551] explore reporting as a visible bystander intervention, prompting a need for users and platform moderators to have a dialogue.

Previous work on design and compliance with the NetzDG for Reporting Mechanisms (RMs) focused on how Facebook and Twitter nudge users to use Terms of Service (ToS) rather than legal reporting [747], by assessing necessary clicks to report under the law or ToS. Our findings signal that general users are not equipped with the necessary legal knowledge to navigate reporting mechanisms effectively, shifting the burden on users to get educated in how to use the RMs (potentially while upset by content), where a knowledge asymmetry has been noted in prior work on how platform consent functions [484].

The growing body of work around deceptive (or ‘dark’) patterns in website interfaces and processes is also relevant to our study (e.g., [501, 502], see Section 10.6.3). Gray et al. [313] highlight the need for having legal knowledge in the design domain, seeing it not solely as a limitation but also as empowerment, and actively fostering the integration of legal and UX design [310]. Our findings explored options for integrating flagging into SMP interfaces, but also highlight a platform’s power over that design as the ‘host’, and the interplay between different user groups and their particular needs (e.g. users and Trusted Flaggers). We also characterize challenges in how the navigation of legal-process interfaces relies on prior understanding of the law, in order to be able to competently exercise one’s choices.

10.4. METHODOLOGY

Here we describe our approach to answering our Research Questions, by way of an expert interview study. The need to study legal aspects of RM is evident in complying to Article 16 DSA, which regulates the design of RMs.

It has been noted that “legal and procedural complexities often overshadow usability concerns, resulting in reporting interfaces that are challenging for average users” [671, p. 5]. Hence, we interviewed practitioners with knowledge of the legal complexities of the DSA. Here, we bring together the perspectives of Professional Reporters who address legal aspects of user needs daily, experts specializing in challenges related to DSA-reporting, and regulators with expertise in enforcement and evidence within the DSA-regime. Based on their knowledge of the DSA, and prior experience interacting with users in actively reporting content, this serves to anticipate challenges for users who are not familiar with the legally-recognized reporting process, nor how it sits alongside other reporting paths such as ToS.

10.4.1. STUDY DESIGN

The interview sessions were semi-structured and are based on main Recitals and Articles governing reporting mechanisms in the DSA, which were hinging on the overarching research questions. We did link the relevant provisions in the interview protocol to each question in the Appendix F.0.1. Each interview opened with questions to understand the work of the participants (with interview questions included in Appendix F.0.1), their

experience, and their connection to users; this includes whether there is a particular group or kind of user that they represent. The interview would then move into understanding reporting/flagging under the DSA, and expectations on different user groups, relating this to the existing connection the participant has with users who actively report content.

The goal was to discuss how “professionalized” reporters would navigate the interfaces, and how this relates to their expectations for the users they represent in their work, who through the DSA would have similar expectations imposed upon them.

After opening questions, the interview would move to a ‘legal design’ heuristic walk-through exercise, as detailed in Section 10.4.2. After the heuristic walkthrough and discussion of live reporting options, participants were prompted to consider how interfaces should or could be re-engineered, to better encourage and best support reporting by users. The walkthrough mapping would then serve as discussion points for legal requirements and design options.

The interview protocol (as in the Appendix E.0.1) differed slightly for Regulators (‘Regs’), and for User Representatives (‘UReps’). The interview protocol for Regs was focusing on oversight activities of the DSA, expected platform compliance, and expectations of TFs and transparency through it. P-UReps on the other hand were providing practical details on flagging, operational challenges, user concerns and expectations.

10.4.2. LEGAL DESIGN HEURISTIC WALKTHROUGH

Each interview includes a walkthrough exercise. Our walkthrough approach is based on a heuristic walkthrough [652], which in brief, first examines a user task, then moves to open examination of the interface. Here, the user task is reporting a piece of content on a social media platform. We explored usage through structured examples (such as reporting in different countries, and ToS vs. illegal content). We have adapted the heuristic walkthrough method as, to our knowledge, no visualization method already exists to represent legal concepts in a task flow across interfaces, for use in interface assessment research.

The core adaptation is that we represent legal requirements, representing individual interface screens as steps in a process, alongside salient legal-process information appearing in each step. This is demonstrated in showcasing details of information users have to provide as indicated in Figure 10.1. Similar to Gray et al.’s examination of dark patterns [313], we include information pertinent to the ‘task flow’ of a law-related interface. In this case the structure of the steps in the interface process, and specifically the information that a user would have to understand on each screen, and conversely, prompts for the user to provide information. These represent the complexity and amount of text that must be parsed or provided per reporting step. We provide a comparison for how illegal reporting is implemented on Facebook’s RM in all different EU member states (constituting in four design versions: F (France), DE (Germany), AT (Austria), and all other member states and a comparison for hiding content on Facebook.

The adapted heuristic walkthrough used a Miro board which was populated with high-level visualizations of wireframes (from the 5.21.2024) of the reporting interfaces and processes for both Facebook and TikTok, as case study examples of design implementations in the scope of the DSA. We focused on elements such as process length or showing all text per process step, or visualizing UI solutions per member state, as

illustrated in Figures 4-15 in the repository.¹ This is in contrast to, for example, Gray et al.'s use of wireframes to categorize interface elements according to interface design patterns [312].

The visualizations of reporting details are illustrated in the Figure 10.1 (and can be found in the repository). We used colors (e.g., blue for ToS and red for the law) annotation labels (describing main RM categories). The visualizations were explained with an overview by the interviewer, in each session. Participants were then invited to think aloud [53, 562] and provide immediate thoughts about the wireframes. Afterwards participants were able to discuss freely aspects relevant to them without any order imposed on them akin to the second, open phase of a heuristic walkthrough. The interviewer or the participants controlled the Miro board. If asked to the interviewer steered participants but shared the screen via Teams. Participants were provided a link to freely explore the Miro board themselves.

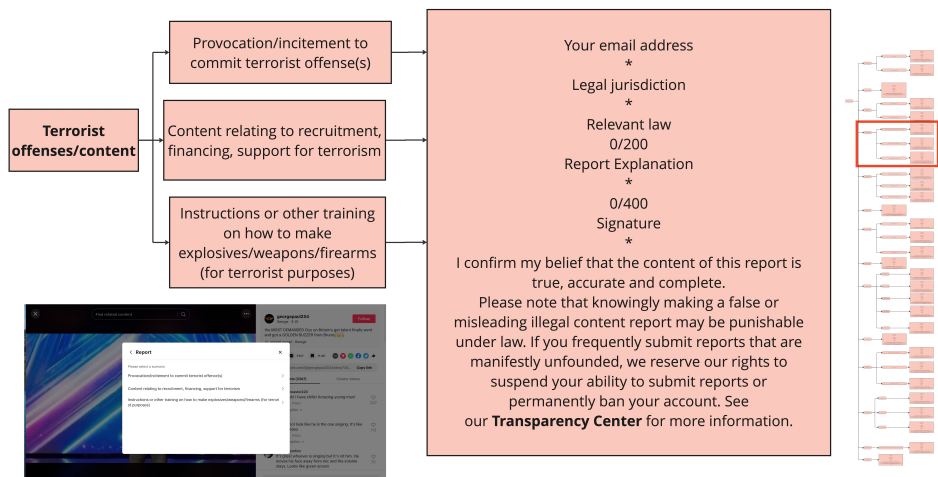


Figure 10.1: Selection of Wireframe Element for 'Terrorist offense/content' under Illegal Content Reporting on TikTok. Including screenshots of the original interface design, a wireframe element laying out the entire text in that reporting step in a bird's eye perspective, and an overview of where this reporting step is located in the entirety of illegal content reporting options on TikTok.

10.4.3. RECRUITMENT

Similar to Wei et al.'s study of online expectations to reduce online hate content [753], we interviewed subject matter experts. Elias and Bezerianos echo the value of experts who "provide insights into usability challenges that may be inaccessible to regular users" [218, p. 4].

Expert participants will have an awareness of the immutable legal imperatives/constructs which an interface must include (much like how in dark patterns research, a consent process must allow for both 'Accept' and 'Decline', for instance [501], and cannot be

¹<https://doi.org/10.4121/16efb52d-b94d-4180-a5f6-89d93028a8c2>

Int	ID	Group	Role	Yrs	Dist.
1	P-Reg1	Reg.	Data analyst	4	A
1	P-Reg2	Reg.	Legal officer	3.5	A
2	P-URep1	URep.	Policy advisor	3	B
3	P-URep2	URep.	Lawyer	25	A
3	P-URep3	URep.	Lawyer	5	A
4	P-URep4	URep.	Legal advisor	3	A
5	P-URep5	URep.	Project mgr., re- porting DB	12	C
6	P-Reg3	Reg.	Legal officer	12	A
6	P-Reg4	Reg.	Legal officer	12	A
7	P-URep6	URep.	Dep. Dir., fact- checking org.	4	B
8	P-URep7	URep.	Exec. dir., NGO	7	C
9	P-Reg5	Reg.	Dir. of regulatory ops	10	A
10	P-Auth1	PubAuth	Legal officer	10	D
11	P-URep8	URep.	Senior policy ana- lyst	5	B
12	P-URep9	URep.	Lawyer	1	A

Table 10.1: Interview Participants: Group, Role, Years of Experience, and Distance to Users. Distance Codes: A = Direct interaction, B = Policy representation, C = Builds tools for users, D = Regulator/public authority.

removed even were it deemed to improve usability). Further, we build on the reasoning of Habib & Cranor [329] when assessing the usability of privacy choice mechanisms, that recruiting experts can inform an estimate of the burden on users to use a policy/legal interface, including the difficulty for a user to enact specific choices and the alignment of user-facing interface designs with regulatory and legal-expertise needs. This is especially the case with the DSA, as experts have knowledge of later ‘offline’ steps outside the interface.

Prospective participants were approached through professional contacts, either directly or via an existing contact using snowball sampling to enrich the selection. Interviews were conducted online via Microsoft Teams, and lasted approximately one hour each. Interviews were audio- and video-recorded to inform the analysis of the legal-heuristic walkthrough. Reporting expertise covers illegal content, disinformation, hate speech, human rights aspects, political advertising and copyright issues. Each Reg. and P-Auth. interview (see Table 10.1) relates to one country within the EU; to preserve anonymity we do not declare which country. The ‘UREps’ organizations are all active within an EU country, the UK or the US. Participants were not remunerated for their participation.

10.4.4. RESEARCH ETHICS

The study was reviewed and approved by our institution's Human Research Ethics Committee (HREC). We gathered consent from participants after they reviewed the informed consent form. Participants could halt the interview at any time, with discussions up to that point being discarded. An overview of the question set was shared with each participants requesting it ahead of the interview. Transcripts were produced from interview audio, then the original recordings deleted and transcripts used for subsequent analysis. There were three interviews (1, 3, and 6) where two participants from the same organization participated. This was suggested and arranged by the contacted participant, and was not initiated by the researchers.

10.4.5. DATA ANALYSIS

All interviews were conducted in English. There was one interviewer for all of the interviews.

We applied Thematic Analysis (TA) [133] to analyze the interviews, specifically a codebook-based approach [94] (codebook in Appendix F2). We employed TA to characterize interactions between stakeholders in the complex content reporting and moderation ecosystem, and because identifiable elements of the DSA and online reporting environment are in a state of transition, with stakeholders shifting their views and expectations around them. The lead author, who conducted the interviews, was one coder (Coder1). A second coder, Coder2, is a co-author. Coder1 has a background mainly in law and design, computer science, and journalism; Coder2 a computer science and (security and privacy) usability background.

All transcripts were coded separately by both coders, in small batches of transcripts; codebook meetings occurred after every 2-3 transcripts. After each such batch, both coders would compare and discuss their codebooks. Codebooks were combined by Coder1 after each round of discussion, with both coders then using the reconciled codebook in the next round. Although there was not great consensus between coders, this is acceptable for this kind of TA [94]; crucially, neither coder disagreed with the coding of the other. For instance, Coder1 was looking for explicit mention of "user friendly" discussion (relative to the DSA), and Coder2 was interpreting discussion of users' primary tasks (browsing social media) against reporting – as a secondary task – as a critical consideration for user-friendliness.

There were also many codes which recorded one stakeholder's role in a process or situation, and codes mirroring the reciprocal process from another stakeholder's position (e.g., "uncertainty in reporting" and "legal uncertainty"). Both coders chose to carry this approach through the whole coding process – to the extent that a passage of text can have multiple codes and classifications – to facilitate coding of a complex, multi-stakeholder set of processes.

10.4.6. LIMITATIONS

We did not talk to users directly. Since the DSA regulates how flagging for "illegal" content has to be organized, focusing on experts that can explain in detail their – and users' – challenges focuses more clearly on our research questions. As a study design decision, we chose instead to use NGOs and flaggers as a proxy for the motivated user to

understand their challenges with legal aspects of flagging, similar to [752] and [329]. The most meaningful study of users would be to meet them at the exact moment they are offended by a piece of content and motivated to report it, following the process until after a moderation decision was made, potentially leading to the case reaching the internal complaint-handling system of the platform (Article 20 DSA), or an out-of-court dispute settlement body (Article 21 DSA). Given the positive response we had from participants, we envisage this as future work with these stakeholders, when users liaise with them for guidance.

Our wire-frame visualizations were designed to highlight particular legal design challenges, representing a potential bias in the issues that we aimed to discuss, but nonetheless enriched the discussion with participants; many participants noted the wireframe visualizations as being eye-opening or very useful in their own right as a knowledge artifact – future work will engage more closely with regulators and reporting entities, to explore how these wireframe visualizations can be used further as conversation devices for policymakers.

10.5. FINDINGS

We structure our findings according to seven themes that emerged from the analysis, as below, with each theme discussed in a subsection. We note which part of the interview protocol they relate to (as in the ‘Semi-structured Interview Guidelines’ in the Appendix E.0.1): Content Reporting – A Multi-Actor Process [1.1]; Professional reporting [1.2]; Content moderation and moderation decisions [1.2], Differentiating illegal content from ToS [1.3]; Differences in SMP (via the heuristic walkthrough in SMPs) [1.4]; Design of RMs [1.5]; and; The process of change with DSA [1.5]. Each subsection is opened with an overview summary.

10.5.1. CONTENT REPORTING – A MULTI-ACTOR PROCESS

DSA flagging as a form of user empowerment is dependent on a user’s motivation to report harmful content while engaging within a legal process, and a need for legal knowledge to navigate Reporting Mechanisms (RMs). Participants did not expect that users should know the law. Regulators and user representatives need to navigate how best to interact with users, to ensure that their legal experience can reach users, for them to be able to report content effectively.

Connection to users The 12 participants work on digital rights in member state and global contexts (P-URep1, P-URep8), directly enforce user rights through flagging (P-URep2, P-URep3, P-URep4, P-URep9), make data on flagging available to researchers (P-URep5), make users/journalists “aware about platform functionalities” (P-URep7), ensure accurate information environments online for users (P-URep6), or are public authorities enabling flagging for EU member state authorities (P-Auth1), Regulators from three EU member states (P-Reg1-5).

On Social Media Platforms (SMPs), users are seen to focus on “just scroll, [...] rather than [...] report stuff” (P-URep7). Reporting is described as a secondary user task on a platform, where the form of content can influence users’ ability and motivation to report

it, as “the casual user is gonna be like, ‘I don’t have time for that’ unless they’re already like, really motivated” (P-URep5).

Regarding motivations, participants represent users in a range of contexts, such as “consumer protection”, users affected or witnessing “digital violence” (P-URep9), like “dehumanizing insults”, “defamation”, “threats”, or “unconsensual or deepfake images”, to help users to “file complaints” (P-URep9). Yet other participants support users with in-house legal expertise (P-URep2, P-URep3, P-URep4, P-URep9). The consultation of legal experts is also brought up in the design of reporting mechanisms themselves, where “pages are advising the user even to first consult a lawyer before they fill out this form” (P-URep9).

10.5.2. LEGAL REPRESENTATION FOR USERS

RMs for illegal content, require some degree of legal understanding in their design. According to P-URep9, users must provide an “explanation to why they think the content is illegal” as stated in Article 16(2)(c) DSA. Such statements must be given in *bona fide* in accuracy and completeness as lined out in literal d of the same provision.

P-URep4, on the other hand, does not see Social Media Platforms (SMPs) as having high legal expectations from users in the reporting process. Expectations on users to know the law are evident in design decisions as on TikTok, “stating the jurisdiction” (P-URep9, P-URep8), which is seen as a legal version of asking “[in] what country do you have your permanent residency”? P-Reg1 notes: “[If] users [...] were not able to report the content or [...] it was not handled [...] according to the DSA, [...] users have the right to complain about alleged violations” to regulators. Where Digital Service Coordinators (DSCs) step in to provide help to users in enforcing their rights “in their struggle to find it, or in their despair because they don’t get an answer” DSCs provide help to users and are even “contacted by lawyers because the user didn’t get...very far with this complaint.” (P-Reg1). In many cases however information to “judge the case” is not provided by users (P-Reg1).

10.6. CONTENT REPORTING

Professional reporting emphasizes the importance of supporting users in the DSA flagging process, such as reporting illegal content through Trusted Flaggers (TFs) acting on behalf of users and utilizing dedicated reporting tools. Empowering and educating users are considered crucial for fostering their active engagement in the reporting process.

Upholding representation of users TFs are prioritized in the content moderation process (P-Reg5, P-Auth1), though how exactly prioritization is facilitated and against whom “is up to the platforms” (P-URep5). We note here that TFs report according to Article 22 of the DSA, representing a specific expertise and competence in the DSA flagging process. However, queuing of flags is opaque according to P-Reg2, “if a post gets 1000 notices of being illegal or not compliant with the [Terms of Service] that must get more priority than a post with like one notice, right?”.

The TF role is still new and during our interviews only a few were accredited by DSCs. Several participants of this study were in the process of being accredited. Some

SMPs have existing ‘pre-DSA’ flagging programs for similar fast-lane treatment (P-URep7). TFs have the “big problem [of] resources” (P-Auth1), echoed by P-URep7, since TFs are not allocated any specific funding with their accreditation (P-Reg3, P-Reg4, P-Auth1). Resources spent can be in conflict of outcomes – “maybe at the end of the day [...] you’ve removed, [...] 10 videos with 100 views each and you know, 10s of thousands of Euros and of people’s time has been spent doing this” (P-URep7). This raises questions about how much end-users can rely on TFs to fill the ‘gap’ in their ability.

Reporting that supports the user To file a report, users also must provide contact information (name and email address) according to Article 16(2) c DSA, “but I could imagine that that will prevent maybe some people who want to stay anonymous from reporting” (P-URep8). Requesting contact information may curb user reporting (P-URep3), where “anonymity” in the process is an important feature (P-URep5), yet it may be that “the lawyer of the perpetrator says [...] You have to come to court” (P-URep3). P-URep7 on the other hand stresses the need of “educating and helping users behave better” in the reporting process and to prevent problematic content in the first place.

The need for building tools Many User Representatives have built their own tools for users to flag content, to reduce the burden (P-URep2, P-URep3, P-URep4, P-URep5, P-Auth1). Intended users vary though, spanning platform users (P-URep2, P-URep3, P-URep4), researchers and journalists, or “small and medium enterprises” (P-URep5), to authorities and regulators (P-Auth1). These aim to help to, for example, “send notifications or removal orders to hosting service providers” (P-Auth1). The need to build these user support tools often comes from the need for “evidence” (P-URep2, P-URep3), addressing “unawareness” about dealing with illegal content (P-Auth1), the necessity to create a “public record of content removal requests” (P-URep5), or because “users needed a way to help report harmful or illegal content” (P-URep4). Such tools may be intended to collect less information than SMPs’ existing reporting processes (P-URep2, P-URep4), or “[help the user] report things [...] even if they aren’t sure of the legal definitions” (P-URep9). To expect users to be familiar with the law is seen as “wildly ambitious” (P-URep5).

10.6.1. DIFFERENTIATING ILLEGAL CONTENT FROM TERMS OF SERVICE (ToS)

Users cannot be expected to have legal expertise to use DSA Reporting Mechanisms (RMs), and may struggle to differentiate illegal content from Terms of Service (ToS) violations. Perpetuating in differences in member state law that create complexity and define the scope together with European Union law of “illegal content” under Art 3(h) DSA. There can be some content where the (il)legality is not clear-cut. If a user successfully reports content as “illegal” the platform’s liability changes since the user’s notice creates “actual knowledge” according to Articles 6 and 16(3) DSA removing the platform’s liability exception according to Article 8 DSA. Platforms are perceived to have a preference of ToS over legal flagging, according to P-Reg1, P-URep8, P-URep9, where the latter comments that “the platforms [...] more often rely on their Terms of Service because they’re private entities, at the end of the day it’s their terms of service ” (P-URep9). Professionals are

expected to use illegal content RMs, while regular users rely on ToS without understanding the differences.

As expressed by P-Reg3: “it’s [...] so tough also for us lawyers to decide on this if it’s against any criminal law”, where there is ‘gray content’ as a tension in content moderation (P-URep4), as it is “always so hard to know the fine line between illegal and legal” (P-URep2). Differentiating illegal content from ToS violations, P-URep8 expects RMs to use “language that does not require from me to be a lawyer in order to understand what is not allowed.” P-URep1 regards a website’s DSA reporting path as intended for TFs, while users might use reporting under both – the ToS and the law – to flag as described in Section 10.2.

Reporting against ToS was seen as “more straight forward” (P-URep8) and “easier” (P-Reg1) than illegal content reporting. The “underlining of legal consequences” (P-URep8), and mention of “punishable under law” (P-URep9), during reporting could dissuade user reporting and lead to under-reporting of content as illegal.

Put together with user expectations and the design of the reporting process (Section 10.5.1) and user motivations, there is a paradox that by legally recognizing a user right to report content, the process has potentially become too complex and daunting for users who are *unexpectedly* offended and unfamiliar with how to report offensive content.

10.6.2. CONTENT MODERATION AND MODERATION DECISION

Non-transparent moderation complicates understanding of Reporting Mechanisms (RMs) for different users and regulators alike. However, benchmarking across platforms can enhance comparability and DSA compliance.

Regarding certainty and visibility in content moderation, the “decision lies with the companies” (P-Auth1), where there is uncertainty about the involvement of humans and AI in reviewing flags (P-URep1, P-URep2, P-URep8, P-URep9). This is characterized as “you report in hope rather than expectation” by P-URep7.

How SMPs deal with moderation is opaque: “What policy kicked in? How is this being moderated? If it’s being even moderated?” (P-URep8). SMPs have different options like “Disable access to content or remove content entirely” (P-Reg1), as well as “geo-blocking” (P-Auth1), “demotion” or “labels” (P-URep6). Content moderation action is not guaranteed, where perhaps “the flag won’t be acted upon...for various reasons, e.g., depending on the [...] culture. What’s acceptable in Brazil may not be acceptable in Pakistan” (P-Reg5), constituting a further aspect of uncertainty in content moderation.

Design of reporting mechanisms (walkthrough exercise) Through the walkthrough exercise, our participants noted that there are numerous overlaps between DSA and Terms of Service (ToS) categories, complicating illegal content reporting, even for content where the legality is unclear.

ToS/law reporting path Terms of Service flagging can be “more detailed” (P-URep8) and create an overlap with the legal reporting categories. The difficulty for users’ understanding between contractual and legal violations becomes evident in the numbers of the DSA transparency reports and the statement of reasons database (P-URep5, P-URep8), that show lower numbers of illegal content report flags compared to ToS equivalents. The online statement of reasons database captures all content moderation decisions under

Article 17 the DSA, but “is not really useful at the moment” (P-Rep3) [148]. Kaushal et al. highlight a similar concern, that despite promising accountability, the implementation of DSA transparency mechanisms risks being “a facade of compliance without substantive insights” [414, p. 8], further stressing that defensibility for legal compliance is often prioritized over user-friendly design – we find similar concerns here among stakeholders tasked with supporting reporting.

It was regarded that users have to “probably click through several pathways” P-URep8 in order to report illegal content compared to ToS violations (P-URep9) (See also Figure F.1). This represents a variation on the ‘excessive number of legals’ concern noted by Rossi et al. [626], where immediately a user would need to decide between a platform-internal report and going down a ‘legals’ route and entering into a legal process. Further, where Rossi et al. note ‘difficult comparability’ in assessing legal information, we find for flagging under the DSA, participants note this between legal categories which overlap with each other, as well as legal categories overlapping with ToS categories.

Differences in member state design Notably, before encountering Facebook’s illegal-content reporting mechanism during the walkthrough, no participant, including the regulators, was aware that the interfaces differed across national contexts such as France, Italy, and Germany (see Fig. 16 in the repository, *Selection of Wireframe Elements for Differences in Illegal Content Reporting in France, Austria, and Germany on Facebook*).

10.6.3. DESIGN OF REPORTING MECHANISM

This section addresses the design of reporting mechanisms regarding their legal demands and effects on usability.

The act of reporting: motivation, emotion, and duty Users might be in upsetting or “shocking” situations like “Oh my God, there is such thing [as this] on internet (P-URep4). So, you have to do something about it.” E.g., for cases of “terrorist” or “far-right” content according to P-URep4. P-Reg1 also mentions the influence of encountering problematic content like “seeing hate speech” online, and the potentially deterring influence it might have on a user’s flagging behavior. The personal involvement of the user was identified as a core motivation for reporting, as “It really depends if you’re personally affected”, exemplified in cases of “cybermobbing” (P-URep2). P-URep2 made a general observation, that the reporting of content on SMPs by users and reporting entities is “outsourcing the problem” of finding problematic content. Furthermore, navigation on SMPs should not be an essential step in reporting according to P-URep5, as the user “shouldn’t have to go to someplace else in the website and [do] work”.

Easy to access interface design The “three dots” as a metaphor for the UI opening up the reporting button were mentioned regularly by participants, in relation to a need for interfaces to be easy to access (P-Reg1, P-URep1, P-URep9), though the subsequent reporting steps are seen as more “complicated” (P-URep1). Reporting access should be “next to every unit of content” that invites the user “do you want to flag this?” (P-URep5). Also, the explanations of ToS categories, and finding “information easily on the platform,

preferably at one place”, describe qualities of better reporting mechanisms in the opinion of P-URep8.

Reporting labyrinth – a form of dark patterns? There is an immediate tension in having the reporting mechanism on – and under the control of – a platform. Users were regarded as likely to find the notion of reporting features useful, but not easy to use upon entering the reporting process itself.

P-Auth1 stresses the need to report “swiftly and [in only a] few steps”, yet the design may falter if a user has to “click through many things until you actually arrive to what you’re looking for” (P-URep8). Such reporting mazes are described by P-URep8 as “dark patterns” regulated in Article 25 DSA, and mentioned in the context of, e.g., privacy policies, in [311, 314]. Dark patterns were also stressed in the context of RMs under the German NetzDG, as influencing other forms of reporting such as numbers in transparency reports [747]. Crucially, the length of the reporting process might dissuade users from reporting, akin to “I guess I’m not going to flag it after all, if I have to click through, you know, one to seven different screens. [...] I just wanted to, you know, put an asterisk on this horrible video” (P-URep5).

Support of different user needs Participants suggested that platforms direct users to specific representative entities, akin to “Hey, did you know that there are these collective redress organizations that can take your case and here is the contact” (P-URep8). Users often seek support by user representatives because of the lack of response of online SMPs (P-URep4), or a need for the engagement of “real persons”, “If it’s a really frightening situation” (P-URep2).

Hiding content as the user-friendly default The option to “mute” content or accounts according to “personal taste” (P-Auth1) is seen as “another option” according to several participants (P-URep6, P-Reg95, P-Auth1), e.g., P-URep6 explains “I don’t wanna deal with it, so I just made you invisible” and P-URep7 qualifies hiding as “perfectly reasonable behavior”. However, this kind of personal content management “does not help at all against this kind of content being spread online [...] and everyone else still can see it” (P-URep8), where “closing your eyes... it’s not the solution” (P-URep6). This is akin to how flagging has been regarded in the literature up to now (e.g., [393]), driven in part by respect for freedom of expression; this then raises further tensions for the future, as to whether a user will be considering whether they are singularly offended, or reporting and expressing offense ‘on behalf of’ a community of like-minded users.

10.6.4. STAKEHOLDER COOPERATION

The DSA harmonizes content moderation across the EU. Communication is important for enforcement. Clarity is needed on what defines an ‘easy to access’ and ‘user-friendly’ Reporting Mechanism (RM). Tensions may arise with other user rights, such as GDPR. In order to make compliance work, “cooperation” and “coordination” are key (P-URep4, P-Reg5, P-Auth1). Regulators have to work together (e.g. in the board Article 61 DSA) and also with Trusted Flaggers (TFs) (P-Reg5, P-Auth1), otherwise there is a risk of fragmented handling (P-Reg5). For example, there may be rare “geo-blocking” in contrast to “removal

orders” enforced in “the whole of Europe” (P-Auth1), and global tensions for example with the US 1st Amendment [46]; the latter is already notable in the different ways flagging is discussed here compared to prior literature. The need for cooperation extends to reporting interfaces, for instance in how to design reporting mechanisms under the Terms of Service (ToS), which according to P-URep8 is also not “subject to any reporting obligations”.

10.7. DISCUSSION

Here we reflect on our findings, and the implications for users and the support provided by other stakeholders. Our findings emphasize the urgent need for platforms to balance compliance with the DSA and intuitive design. We found that it is necessary to define channels for support and feedback that connect users with Professional Reporting entities, in order to reduce reporting complexity and burden (where even our practitioner and regulator participants remarked on several pitfalls). These needs are especially critical given how reporting mechanisms are situated on the very same platforms they oversee, where complex reporting processes compete with Terms of Service and content-hiding alternatives as a way for the non-expert user to act on offensive material.

10.7.1. USERS AND THE LAW: EMPOWERMENT AND BURDEN

There is the immediate contradiction that by recognizing content reporting as a user right under Article 16 DSA, signaling offense has potentially become a more daunting and unclear process (generally and from the user perspective) (Section 10.6.3). For instance, does the user need to go to court at some point in the future? This is compounded by users potentially being emotionally-charged at the time they experience offensive content on a platform. This tension is stressed by the current platform shifts, exemplified by Meta, demanding more user interaction in content moderation through less institutionalized content moderation, and more reliance on ‘community note’ systems [213, 332, 407, 643, 701]. This could lead to two very different digital realities in the US and Europe, for instance leaving content online in the US while having to remove it in the member states.

Some of our participants – both user representatives and regulators – believed that platform users should liaise with user representative or NGO organizations, to make a legal report of offensive content (Sections 10.5.1 and 10.6). Yet, how users reporting would become aware of these organizations is not clear, as social media platforms do not signal or redirect to specialized organizations (e.g., for copyright or personal harassment). Additionally, such organizations could face resourcing demands and financial shortcomings that have no provision within the DSA, where financial incentives from SMPs and dependencies with these can impact the reporting of content [694]; given that reporting may involve making decisions on divisive topics (e.g. transgender content [695]), power dynamics play a role [260]. It remains to be seen whether easy-to-use reporting interfaces will compete with existing user options to dismiss content from their feeds (Section 10.6.1). In a sense, the DSA appears to assume a direct correlation of the legally-determined “severity” of offensive content with the motivation and energy it would activate in the offended user.

The DSA is in its early stages, but how to keep users motivated once they report

content is unclear, especially if they are expected to make a distinction between offensive and illegal content. One focal point is feedback – the UK Action Fraud service allowed reporting of crimes, but focused on passing information to police forces rather than victim support [170]. Making a comparison to phishing reporting and differentiating between a malicious or benign email [450], users who only receive feedback for reporting a malicious email may begin to doubt whether they are spotting bad content properly and reduce their efforts. We recommend that future effort examines how to support a user who is brought into the (legal) reporting process, and avoid erosion of their motivation by burdensome or uncertain processes.

One direction for future engagement with users would be to fabricate reporting processes within an experimental setting, based on real-world implementations, and lead users through the process (similar to how we distilled real-world implementations here). This could identify ‘pain-points’ where users encounter problems (given the expectation for legal knowledge and comprehension), where support could be targeted (either in terms of simplifying the process in a way that removes an expectation of legal knowledge, or where expert flagger assistance can best be positioned to compensate for the difficulties users have (to reduce the need for Capability [527] to complete the reporting task). Our approach of using a wireframe of the reporting process, and a legal design walkthrough, can be repurposed as a user journey, to track where any pain-points occur within the reporting process (for instance, identifying issues with comprehension or unclear choices). Given the current positioning of reporting infrastructure on the platforms themselves, there are also issues of Motivation relative to the perceived effort of engaging with the reporting process; specifically, a user may weigh up their sense of offense with a piece of content against their perception of the immediately available (and usable) reporting mechanisms.

A user may engage directly and fully with the legal reporting process, or *satisfice* [674] and accept the first available option that seems satisfactory for getting the content either out of their feed or into a system where it will be scrutinized by the platform (rather than tirelessly seeking out the optimal solution, in this case, reporting within the legal framework of the DSA); they might do this by choosing to ‘hide’ content from their feed, or reporting it against a platform’s own Terms of Service. The aforementioned ‘pain-points’ are then not only about usability issues with the interface and decisions, but also the ‘usability’ of the legal process embedded within the interface and its impact on user motivation [527].

Given the current positioning of reporting infrastructure on the platforms themselves, a potential future user study would be useful to understand how users would weigh up content and their perception of available and usable reporting mechanisms (notably, to hide content, report against Terms of Service, or engage directly with legal reporting).

There are user concerns the DSA will not solve – malicious use of flagging features becomes an increased cost for all stakeholders. Flagging of divisive content (e.g., views on immigration) seem to have been put to one side, deferring instead to the character of the law and there being offensive content that is somewhat “universally” agreed as being illegal. Crawford & Gillespie [175] note how a flag arbitrates “the negotiation around contentious public issues” (Section 10.6.4). However, the findings of Eyert et al. [260] raise the possibility that societal influence and power dynamics may be what ultimately

resolves any such deadlock.

10.7.2. USER ACTION IN THE PRESENCE OF PLATFORM OPACITY

Even with the DSA in place, the content moderation processes within Social Media Platforms (SMPs) remain opaque to external actors (Section 10.6.2). At present, the SMPs alone have the capacity to see what is happening within their own online environments. This leaves questions as to what “good enough” effort by SMP looks like, as the DSA arguably results in monitoring for signs of non-compliance, as opposed to monitoring for proof of utmost compliance (i.e., removal of everything that should be removed) (Section 10.6.2).

We can articulate our findings around DSA reporting mechanisms, in terms of behavior adoption factors via the COM-B behavior adoption model [527]. To consider reporting entities and Trusted Flaggers (TFs) as (professional) users themselves, our participants perceived a lack of transparency despite the additional controls (impacting capability) (Section 10.6), and expected the same among general users. Participants’ legal expertise helped them to understand the scope of specific categories, but did not remove the need to navigate ‘gray areas’ of content, or decide if content was clearly legal or illegal. Legal aspects were seen to dampen user motivation despite being positioned to empower users. The opportunity to report was seen as being hindered by the placement of reporting mechanisms away from content, but also in part because more convenient options to ‘hide’ content supplanted reporting features in platform interfaces.

As TF roles expand, another recommendation would be to examine the issue where some participants thought users would consult Professional Reporters (PRs), or that users and flaggers would act independent of one another, or that PR reports would be triangulated with or would verify user reports, once seen together by platform moderators. As the DSA is being established in practice, there is a need to explore how communication and visibility of flagging pathways can be best signaled to users, or otherwise how the interplay between users and TFs can be coordinated – this would offset the perception of a few of our participants, that content moderation is being outsourced to users. Building on the proposal for user studies in Section 10.7.1, a user study could include a range of conditions, each with different signposting for reporting assistance within a fabricated interface. These could include a user being relied on to interpret the process themselves – as we found is presently the case, unless they are already aware of flagger/reporter entities they can consult – or designed to include a range of signposting options, acting to cue the user to contact and involve a flagger entity, delegate the process to a flagger, etc.

One future direction is to synthesize reporting processes based on real-world implementations, and lead users through the process in a user study. Our mapping approach can be used to track and identify ‘pain-points’ where a user encounters problems (especially given the expectation for legal knowledge and comprehension). This would be where support could be targeted, either in terms of simplifying the process more toward a layperson, or where expert the assistance of trusted flaggers or other reporting entities can best intervene to compensate for the difficulties users have.

10.8. CONCLUSION

In conclusion the DSA is promising EU-harmonization of the regulation of reporting mechanisms (RMs) of Social Media Platforms (SMPs) while also creating new user rights and PR-roles like Trusted Flaggers. Our findings show the entwined and complex relationship between the different actors of the reporting process. RM are a tool for the user but not always designed in a way that supports the user who therefore, has to seek help at user representative organizations to guide them through the legal complexity that filing a report under the DSA can bring. How to better design RMs that adhere to imprecise DSA obligations, like being 'user friendly', while also following user needs is even for experts in the field a challenging task. Further work is needed in the field of legal-design addressing RMs and their regulation. Issues like transparency, UI complexity, and user information, can foster alignment of SMP with the DSA, aiming for a safer online environment for a stakeholder-multitude.

11

AUTOMATING DIGITAL SERVICE ACT USER SUPPORT: LARGE LANGUAGE MODELS AS USER SUPPORT FOR FILING ILLEGAL CONTENT REPORTS

"I could not help laughing at the ease with which he explained his process of deduction. "When I hear you give your reasons," I remarked, "the thing always appears to me to be so ridiculously simple that I could easily do it myself, though at each successive instance of your reasoning I am baffled until you explain your process. And yet I believe that my eyes are as good as yours."

Dr. Watson, A scandal in Bohemia, Chapter I, Arthur Conan Doyle

Reporting illegal content online is a tedious but important task that can help ensure individual safety and foster healthy digital communities. This paper investigates how Large Language Models (LLMs) can support users in reporting illegal online content under the European Digital Services Act (DSA). We designed an empirical study with German users combining interactive reporting scenarios and a representative survey to examine how LLMs affect comprehension of legal and procedural language. Two deployment conditions were tested: (1) LLMs as assistants explaining jargon and guiding reporting choices, and (2) LLMs as proactive agents pre-classifying content for user confirmation. User outcomes were shaped by AI literacy, trust, and legal confidence. We discuss implications for user

empowerment, legal design, and regulatory compliance, highlighting both the potential and risks of integrating LLMs into reporting interfaces.

This chapter is based on: Sekwenz, M.-T., Biswas, S., Hermann-Gsenger, R., & Gadiraju, U. (2026). "AI at the Front Lines of Platform Governance: Using LLMs to Support Illegal Content Reporting under the Digital Services Act." In *FAccT '26*, pp. 2848–2884. ACM. doi:10.1145/3805689.3812301. Study materials and supplementary analyses appear in Appendix G.

11.1. INTRODUCTION AND BACKGROUND

Online platforms have become core societal infrastructure [597] and *de facto* public spaces shaping political discourse, cultural participation, and access to information [144]. As a result, governments increasingly regulate how platforms address harms such as the dissemination of illegal content [134]. The European Digital Services Act (DSA) harmonizes content governance rules across the European Union [224], aiming to create a “safe, predictable and trustworthy online environment” (Rec. 3 DSA).

Article 16 mandates that platforms provide user-facing mechanisms for reporting illegal content. What is ‘illegal’ is defined in Article 3 (h) DSA.¹ The requirements of Article 16 DSA are relevant to a wide range of platforms including *hosting intermediary services* (Art. 3(g) DSA), *online platforms* (Art. 3(i) DSA), *online search engines* (Art.3 (j) DSA), and *Very Large Online Platforms* (VLOPs) and *Very Large Search Engines* (VLOSEs) according to Article 33 DSA.² Reports must include (i) a sufficiently substantiated explanation of the reasons (lit. a), (ii) a URL, (iii) identifying information of the reporting user, and (iv) a good-faith statement. While other DSA provisions impose systemic risk-management obligations for VLOPs and VLOSEs (Arts. 34–35, 37), Article 16 (which is applicable to hosting intermediary services, online platforms, search engines, VLOPs/VLOSEs), explicitly foregrounds the role of end users in initiating enforcement processes. Non-compliance can result in fines of up to 6% of global annual turnover for providers of VLOPs and VLOSEs (Art. 74). Recent enforcement actions against major platforms [163], including a 120€ million fine against the platform X [163], underscore the regulatory importance of DSA compliance.

The DSA further expands platform responsibility to content that is blurring the boundaries of ‘illegality’ in cases of disinformation³ which is at times referred to as “lawful but awful” content [128, 415]. Regulating speech, therefore, entails risks of over-removal and requires trust in regulators, platform participation, and avenues for user challenge and involvement. However, accurately distinguishing illegal content from harmful but lawful expression remains difficult even for experts, let alone ordinary users [490]. Reporting

¹The definition of ‘illegal content’ states: “any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law.” Recital 12 further clarifies: “‘illegal content’ should broadly reflect the existing rules in the offline environment. In particular, the concept of ‘illegal content’ should be defined broadly to cover information relating to illegal content, products, services and activities. In particular, that concept should be understood to refer to information, irrespective of its form, that under the applicable law is either itself illegal, such as illegal hate speech or terrorist content and unlawful discriminatory content, or that the applicable rules render illegal in view of the fact that it relates to illegal activities.”

²The VLOPs and VLOSEs have additional reporting and assessment duties compared to other platforms within the scope of the DSA. Such platforms have e.g., additional transparency reporting duties (Art. 42 DSA), or the obligation to conduct *Systemic Risk Assessments* (Art. 34 DSA) and be externally evaluated through *Independent Audits* (Art. 37 DSA). To qualify as a VLOP/VLOSE platforms must reach a threshold linked to “number of average monthly active recipients of the service in the Union equal to or higher than 45 million” according to Article 33 (1) DSA.

³See for example Art. 34(1)(c) which covers “any actual or foreseeable negative effects on civic discourse and electoral processes.” This is further discussed in Recital 84 “[...] When assessing the systemic risks identified in this Regulation, those providers should also focus on the information that is not illegal, but contributes to the systemic risks identified in this Regulation. Such providers should therefore pay particular attention on how their services are used to disseminate or amplify misleading or deceptive content, including disinformation.”

requires mapping ambiguous posts to specific legal provisions under uncertainty. Errors at this stage can lead to under-reporting, over-reporting, or misrouting, with downstream consequences for both users and platforms. While platforms have no general monitoring duty (Art. 8), intermediary services (Rec. 20) and hosting providers (Rec. 22), however, do not have a *carte blanche* for liability questions mitigation as “other actors [...] should also help to avoid the spread of illegal content online, in accordance with the applicable law” (Rec. 27) and “it is important that all providers of hosting services, regardless of their size, put in place easily accessible and user-friendly notice and action mechanisms that facilitate the notification of specific items of information that the notifying party considers to be illegal content” (Rec. 50). VLOPs and VLOSEs, furthermore, are obliged to mitigate the dissemination of illegal content (Arts. 34–35, Rec. 80 and 87). Alternative approaches, such as community-based moderation, face challenges of scale [203, 290, 561], time constraints [38], increasingly convincing AI-generated material [429], linguistic diversity [446, 513], evolving semiotics, and heterogeneous content types [373, 397, 774].

In response, platforms have increasingly explored AI-based support for moderation and reporting workflows [498]. Yet, AI assistance can amplify systemic risks rather than mitigate them, and human collaboration with decision-support systems is error-prone [83, 219]. A dominant approach to address these challenges has combined decision support with explainable AI (XAI) [70, 344, 645], aiming to foster appropriate reliance through post-hoc rationales [277, 646]. This paradigm assumes relatively reliable recommendations and treats explanations as justificatory aids. In legal classification tasks, however, users must deliberate among competing interpretations *before* committing to a category, and a single suggested label may anchor decisions even when incorrect [564]. A growing body of work shows that appropriate reliance remains difficult to achieve—even with XAI—and that explanations can increase overreliance via illusory explanatory depth, particularly in low-friction or conversational settings or when explanations are hard to verify [216, 269, 343, 345, 347, 412, 447, 511, 635, 644].

To address these limitations, Miller [530] proposed *evaluative AI*: decision support that provides evidence for and against multiple options rather than prescribing a recommendation. This hypothesis-driven approach surfaces structured arguments, statutory elements, and evidentiary considerations without prescribing an outcome, thereby supporting user deliberation and mitigating over- and under-reliance. Evaluative AI is particularly suited to legal reporting contexts, where responsibility rests with the user and overconfident automation may undermine agency and legal reasoning.

In this work, we examine how different architectures of AI assistance influence provision-level accuracy, the quality of users’ substantiated explanations, and robustness under erroneous AI output in illegal-content reporting. We do not frame AI assistance as a mechanism for automated compliance or enforcement. Instead, we focus on user–AI interaction at the reporting interface and its implications for users’ ability to meet the procedural requirements of Article 16 DSA relevant to all intermediary services. References to VLOP/VLOSE’s systemic-risk provisions are limited to downstream effects of misclassification at the reporting stage, rather than claims of direct compliance or risk-mitigation effects (Arts. 34–35).

To that extent, our study is guided by the following overarching research question—*“How do different AI assistance architectures in illegal-content reporting interfaces shape*

user decision quality, explanation quality, and robustness to erroneous AI output in Article 16 DSA reporting tasks?” Based on this question and informed by existing literature, we derive six hypotheses that henceforth structure our empirical analysis. As described below, H1–H3 examine provision-level accuracy under correct and erroneous AI assistance. H4 examines whether assistance architecture improves the quality of users’ own substantiated explanations. H5 tests whether different AI error conditions differentially degrade performance and whether **EVALAI** attenuates these losses relative to **XAI**. H6 examines downstream risk-relevant outcomes of misclassification at the reporting stage, namely over-removal and misrouting.

- **H1:** Under AI error conditions, participants using **EVALAI** will achieve higher provision-level accuracy than participants using **XAI**, indicating lower overreliance on incorrect AI outputs.
- **H2:** Under AI error conditions, participants using **EVALAI** will select legally closer incorrect provisions than participants using **XAI** (smaller misclassification distance), but this reduction will be attenuated for Out-of-Scope errors relative to Near-Miss and Overbreadth errors.
- **H3:** When AI assistance is correct, **EVALAI** and **XAI** will yield comparable provision-level accuracy, but decisions under **XAI** will be faster than under **EVALAI**.
- **H4:** Participants using **EVALAI** will produce higher-quality substantiated explanations than those in **BASLINE** and **XAI**, as reflected in stronger legal element coverage and reasoning depth.
- **H5:** AI error conditions (Near-Miss, Overbreadth, Out-of-Scope) will differ in the magnitude of their impact on reporting performance, and **EVALAI** will differentially attenuate these error-specific performance drops relative to **XAI**.
- **H6:** Relative to **XAI**, **EVALAI** will better align with downstream DSA-relevant reporting goals by reducing over-removal under Overbreadth errors and reducing misrouting under Out-of-Scope errors.

Hypotheses We investigate these hypotheses through a controlled experiment ($N = 450$) using a custom reporting interface modeled on existing social-media workflows. Participants evaluated a simulated feed containing illegal content grounded in German criminal law,⁴ selected mandatory legal categories, and submitted reports under one of three assistance conditions. The primary outcomes in our study capture provision-level legal accuracy; while secondary outcomes assess explanation quality, decision time, and trust-related measures.

⁴We selected this context as a member-state example where illegal-content reporting was regulated prior to the DSA [747].

This paper contributes to FAccT scholarship on content moderation and platform governance in three ways:

1. introducing an experimental paradigm for AI-assisted content reporting under regulatory constraints;
2. providing empirical evidence on how assistance conditions affect accuracy, reasoning quality, decision time, and trust under correct and erroneous AI assistance; and
3. deriving design and policy implications for DSA-aligned reporting interfaces.

11.2. RELATED WORK

Prior research examines how generative AI tools support human tasks across domains such as design [308, 672], finance [593], creative writing [82], and content moderation [580]. Across these settings, effectiveness depends not only on model quality, but also on how assistance is presented, how easily outputs can be verified, and whether the interface supports appropriate reliance rather than passive acceptance [188, 346, 683].

Content Moderation and Reporting. User reporting plays a central role in content moderation ecosystems [754], yet the scale and complexity of online platforms exceed what purely human moderation can sustain [498]. As a result, platforms rely on automated moderation systems that classify user-generated content and trigger governance outcomes [306]. Hybrid approaches that combine automation with user input show promise, but also raise concerns about overreliance on AI [347, 371].

Content moderation is often described as operating across two phases: *ex ante* moderation before publication, typically relying on automated filtering, and *ex post* moderation after publication, where user reporting plays a key role [290]. Reporting interfaces, however, are often opaque and misaligned with user expectations [438] and mental models [777]. Users are reluctant to engage deeply with reporting [394], though they value nuanced options, proactive platform action, and personal moderation settings [395]. User characteristics further shape flagging behavior [550]. While accessible platform help centers can aid policy understanding [504], most users lack the legal knowledge to file substantiated reports [668], echoing broader findings on the inaccessibility of rules like the terms of service [505, 566].

Legal Design Challenges. The DSA highlights legal design (Art. 25) and bans features that “deceive or manipulate [...] or otherwise materially distort or impair” user decision-making [191, 357, 463] as exemplified in the Commission’s first DSA-based fine against X for the “use of the ‘blue checkmark’ for ‘verified accounts’ [which] deceives users.” [163]. The DSA also requires “user-friendly” reporting mechanisms (Art. 16). Yet, which design solutions meet these standards remains unclear [668]. As regulation reshapes design practice, interdisciplinary approaches become necessary [310]. Poorly implemented legal UX can take the form of *dark patterns*, raising usability, legal, and ethical concerns [311] under which reporting mechanisms as *Interface Interference* can be subsumed, driven by *Language Inaccessibility* and *Complex Language* [314].

AI-supported reporting. Large Language Models (LLMs) as a form of AI, increasingly shape moderation systems [295], including detecting harmful content across languages [446, 561] and generating “Statements of Reason” under the DSA [54]. LLMs are also

applied in legal practice and research to simplify statutes and provide advice [331, 448, 571], but they remain prone to hallucinations and misclassifications [371, 462], raising ethical concerns.

Salimzadeh et al. [636] show that complexity and uncertainty shape user performance in AI-supported tasks; in legal content reporting, this makes meaningful contestability especially important because interfaces must enable users to understand, challenge, and revise automated reasoning [136, 428, 769].

A central challenge in AI-supported reporting is not only whether AI assistance is present, but how it is structured and how susceptible users become to inevitable output errors. Conventional explainability often follows a *recommend-and-justify* logic, foregrounding a single predicted option and rationale. Such designs may reduce friction, but can also anchor users to erroneous outputs when the reasoning is difficult to verify [105, 327, 343, 391, 432, 437, 512, 729]. By contrast, more deliberative interfaces may reduce harmful overreliance, though often at a cost in effort and usability [105, 530, 729]. Taken together, this literature suggests that assistance architectures foregrounding a single recommendation may be especially vulnerable to harmful reliance under AI error, whereas evaluative architectures that present evidence for and against alternatives may better align with human decision processes and support correction and recovery. This expectation informs *H1* and the premise of *H2*, which test whether Evaluative AI improves provision-level accuracy and reduces misclassification distance relative to Conventional XAI when the AI advice is wrong.

Recent work proposes more argumentative and contestable forms of AI support [273, 530, 769, 770]. User judgments are also sensitive to how alternatives are presented under model error, highlighting the importance of choice structure and choice independence [124, 220]. Together, these lines of work point toward *evaluative* forms of AI assistance that support deliberation across alternatives rather than merely justifying one recommendation. This informs *H3*, which examines whether the two AI-assisted conditions remain comparable in accuracy when the AI advice is correct, while differing in decision time.

Because evaluative interfaces expose users to structured supporting and opposing considerations, they may support richer user justifications as well as better selections [273, 530], motivating *H4*. Prior work on automation bias, error tolerance, and contestability also suggests that plausible, overbroad, and legally irrelevant errors may shape reliance differently and carry distinct downstream consequences [105, 220]. This motivates *H5*, which tests whether AI error conditions differ in their performance impact and whether Evaluative AI attenuates those penalties relative to Conventional XAI. It also motivates *H6*, which examines whether these differences extend to downstream risk-relevant outcomes such as disproportionate intervention under Overbreadth errors and routing failures under Out-of-Scope errors (described in detail in the following section).

Together, these strands of work motivate our comparison of manual reporting, conventional XAI, and evaluative AI under systematically controlled legal error conditions in a user-facing illegal-content reporting task.

11.3. EXPERIMENTAL DESIGN AND PROCEDURE

We conducted a controlled experiment with a between-subjects factor (*assistance condition*) and a micro within-subject manipulation (*correct AI assistance* vs. *AI error condition*) applied only in the AI-assisted arms. Participants were blind to their condition assignment.

Assistance conditions Participants were randomly assigned to one of three conditions:

- **BASELINE:** Manual reporting without AI assistance (control).
- **EVALAI:** Evaluative AI, presenting structured pros and cons for multiple candidate legal provisions, inspired by Miller [530].
- **XAI:** Conventional XAI, presenting a single suggested legal provision with a brief rationale.

Each participant evaluated a 12-post feed 11.1: 10 benign filler posts (identical across participants) and two illegal posts drawn from four legal case categories (§ 86a symbols [L_6], § 111 incitement [L_10], § 130 hate speech [L_15], drug sales [L_27]; Appendix G.9.3). The two illegal posts were A/B variants of the same case. In AI-assisted conditions, one illegal post included correct AI assistance and the other included an injected AI error condition. Error order was counterbalanced.

AI assistance architectures Figure 11.1 illustrates the reporting interface in the two AI-assisted conditions. In all conditions, participants saw the same candidate legal provisions. What varied across conditions was the assistance architecture. In **XAI**, the interface pre-highlighted one AI-suggested provision and displayed a short rationale (*recommend-and-justify*). In **EVALAI**, no provision was preselected; instead, participants could inspect structured supporting and opposing arguments for candidate provisions (*deliberation-oriented*). In **BASELINE**, no AI-generated reasoning was shown.

AI error conditions To capture distinct human–AI risk profiles, we introduced three randomized AI error conditions that meaningfully operationalize errors in our context:

- **Near-Miss (NM):** A semantically close but incorrect provision (e.g., § 111 instead of § 130), representing a setting in which the AI output is plausible and may increase automation bias risk [210, 542].
- **Overbreadth (OB):** A related but overly broad provision (e.g., § 166 instead of § 130), representing a setting that may encourage over-removal or overly expansive legal interpretation [287, 306].
- **Out-of-Scope (OS):** A legally irrelevant provision, representing a setting in which the AI output misroutes the report away from the appropriate legal frame [63].

These conditions are not intended to encode a universal severity ranking. Rather, they operationalize distinct reporting risks at the human–AI interface: plausible misguidance (Near-Miss), overbroad categorization (Overbreadth), and legally irrelevant routing (Out-of-Scope).

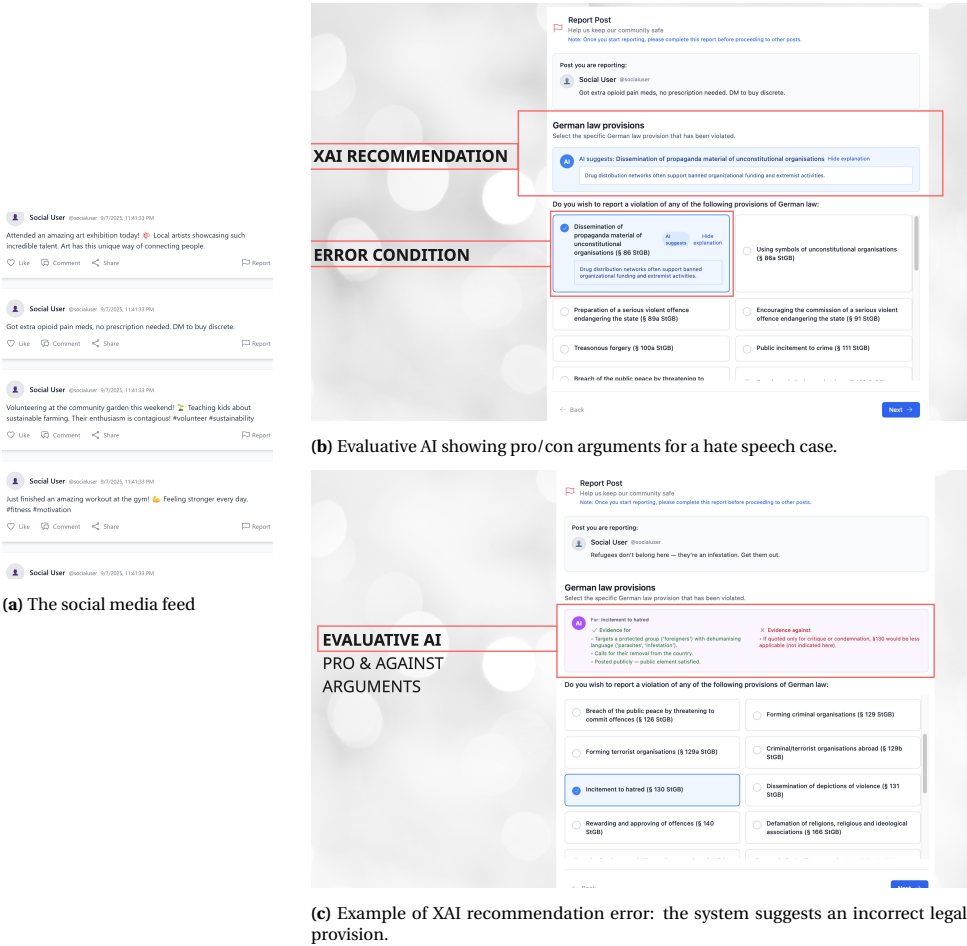


Figure 11.1: Components of the experimental interface. (a) the social media feed the users were presented

Condition manifestation In both AI-assisted conditions, the injected error pointed to the same incorrect legal provision for a given case; only the presentation architecture differed. In **XAI**, the incorrect provision was highlighted and an explanation was provided. In **EvalAI** the same incorrect provision received comparatively stronger supporting arguments, while competing categories were comparatively downplayed; outside the targeted mappings, categories were often represented with weaker, minimal, or placeholder pro/con content rather than fully elaborated arguments. In **BASELINE**, no AI assistance was provided and no error manipulation was applied. Representative mappings are listed in Appendix Table G.5.

Procedure The study consisted of a pre-survey, the reporting task, and a post-survey (Appendix G.4–G.5; Fig. 11.2). **Pre-survey.** Participants completed measures of need for

cognition [478], AI literacy (MAILS) [116], trust in automation [433], and legal knowledge, including attention checks (Appendix G.4.10–G.4.11). **Reporting task.** Participants reviewed the 12-post feed under their assigned condition. Reporting required selecting a legal category (with AI support, where applicable), completing an *Explanation of Reasons* (statutory element marking and textual evidence; Art. 16(2)(a) DSA), choosing a moderation action, and indicating proportionality. To preserve internal validity, progression was locked and cross-post comparison was prevented. **Post-survey.** Participants completed measures of perceived usefulness and ease of use (TAM) [179], sense of agency [689], and workload (NASA–TLX short) [340]. Participants exposed to AI errors were debriefed.

Cases We modeled the reporting interface on Facebook’s illegal-content workflow using German statutory categories (Fig. G.2). Legal cases were derived from German case law, translated into fictionalized scenarios using ChatGPT (Appendix G.9.2), and cross-checked by two authors. Four cases were selected to span easy and hard classification challenges [372, 634] (See, Appendix Table G.7).

Variables The independent variables were assistance condition (between-subjects), AI trial type (correct-assistance vs. error trial; within-subjects), and AI error condition (NM, OB, OS) for error trials only.

Primary outcomes were (a) categorization accuracy (binary match to ground truth) and (b) misclassification distance (0 = correct; 1 = near-miss; 2 = related-but-wrong; 3 = out-of-scope). *Secondary measures* included decision time, Explanation-of-Reasons coverage, and explanation quality (1–5 rubric; two independent coders; Appendix G.7). Additional measures captured moderation actions, false positives on benign posts, interaction logs, self-reported confidence, trust, workload, and demographic covariates.

Participants. We recruited Prolific participants located in Germany (fluent in English; approval rate $\geq 90\%$). Compensation was set to a fixed £1.60 base payment (10–15 min) in addition to a bonus up to £2.00 (i.e., a maximum of £3.60). Participants could earn bonus payments by correctly categorizing malicious posts and by providing high-quality explanations of reasons, assessed based on coverage of key legal elements and proportionality considerations. We recruited 489 participants in total, and after exclusions due to technical issues or false-positive reports on benign filler posts, $N=450$ were retained, with 150 participants per condition (EVALAI, XAI, and BASELINE; Appendix Table G.6). The final sample comprised 293 men, 152 women, and 5 non-binary participants, with mean age $M=31.32$ years ($SD=8.7$, range 18–71).

11.4. METHODOLOGY

We ran a controlled experiment following Ozanne et al. [580] using a custom web app (*SocialNet*) that mimics a social-media reporting workflow. We evaluate six hypotheses using models matched to outcome type: logistic models for binary outcomes, ordered logit models for misclassification distance, linear models for decision time and coder-mean explanation ratings, and paired difference-in-differences specifications for error-induced performance changes. Full model equations and particulars, robustness checks, and supplementary analyses are reported in the appendix. Our goal is to test whether differ-

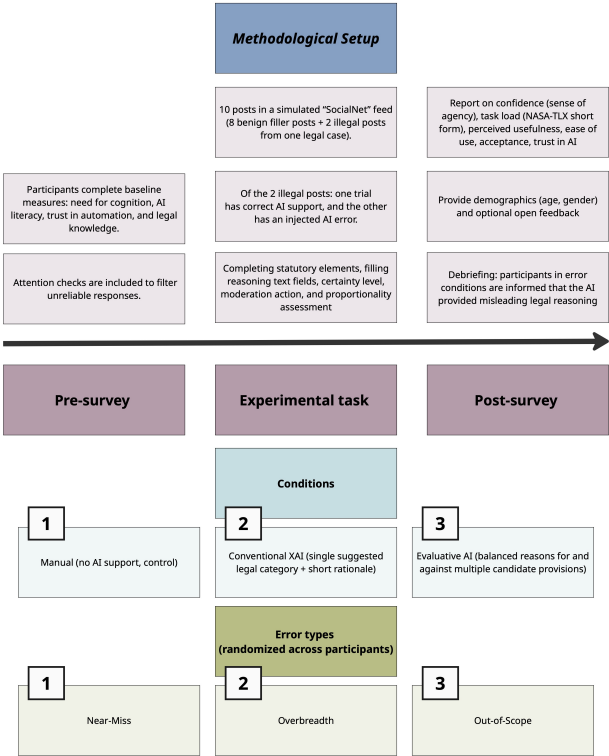


Figure 11.2: Overview of the experimental flow, assistance conditions, and randomized AI error conditions used in the study.

ent assistance conditions shape reporting performance and downstream risk-relevant outcomes in DSA-aligned reporting tasks (Table G.1).

H1: Provision accuracy under AI error. Among AI-assisted participants (Evaluative AI vs. Conventional XAI), we estimate the effect of Evaluative AI on provision-level accuracy on illegal-content error trials (one error trial per participant by design). The outcome of interest here is binary provision accuracy. We fit a logistic model with AI error condition and case fixed effects, reporting robust standard errors, odds ratios, 95% confidence intervals, and two-sided p -values. Robustness checks add pre-task covariates and exclude participants who reported benign filler posts.

H2: Misclassification distance under AI error. On illegal-content error trials in the AI-assisted conditions, H2a tests whether Evaluative AI reduces misclassification distance relative to Conventional XAI, and H2b tests whether this effect varies across AI error conditions, with particular attention to Out-of-Scope errors. The outcome is an ordinal misclassification distance score (0= correct, 3= farthest misclassification). We estimate proportional-odds ordered logit models with participant clustered robust standard errors. Robustness checks add pre-task covariates, exclude participants who reported benign filler posts, and compare conditions using non-parametric tests.

H3: No-error trials (correct AI assistance). We analyze the no-error trial in the two AI-assisted conditions only (one no-error trial per participant; $N=300$, 150 per condition). H3a tests provision-level accuracy using a binomial GLM with robust standard errors and a TOST equivalence test on the difference in proportions ($\delta = 0.05$). H3b tests decision time using OLS on log-transformed response times with robust standard errors, with a one-tailed Mann–Whitney U test as a non-parametric check. Robustness checks add case fixed effects where available.

H4: Explanation-of-Reasons quality. We analyze illegal-content reports with completed coder ratings across all three assistance conditions (BASELINE, XAI, EVALAI). Each report was independently rated by two trained annotators. We examine four coder-rated Explanation-of-Reasons dimensions: element coverage, proportionality reasoning, reasoning depth, and perceived overall quality. For each dimension, the dependent variable is the mean of the two coder ratings. We estimate linear models with (AI) assistance condition as a categorical predictor, participant-clustered robust standard errors, and pairwise contrasts across conditions. Participants could contribute up to two rated reports, which motivates clustering at the participant level. Inter-rater reliability was acceptable to high across dimensions (Appendix), supporting aggregation.

H5: AI error condition impact and attenuation. For AI-assisted participants only, we construct a paired difference-in-differences panel in which each participant contributes one illegal-content no-error trial and one illegal-content error trial. The participant's AI error condition (Near-Miss, Overbreadth, or Out-of-Scope) is assigned at the pair level. We examine two classes of outcomes: provision accuracy and the four coder-rated Explanation-of-Reasons dimensions. Provision accuracy is analyzed using a paired difference-in-differences logistic model with participant-clustered robust standard errors. Explanation outcomes are analyzed primarily using participant-level first-difference OLS models on coder-mean ratings, with ordered-logit sensitivity analyses reported in the appendix. H5a tests whether the error penalty differs by AI error condition; H5b tests whether Evaluative AI attenuates these error-specific performance drops relative to

Conventional XAI.

H6: DSA-relevant downstream risk outcomes. H6 evaluates whether, relative to Conventional XAI, Evaluative AI yields user reporting behavior more consistent with downstream risk-relevant outcomes in DSA-aligned reporting. Analyses are restricted to AI-assisted error trials and estimated separately by AI error condition. For Overbreadth errors, we model over-removal as a binary indicator of whether the participant selected a severe enforcement action. For Out-of-Scope errors, we model misrouting as a binary indicator of whether the participant failed to route the case as out-of-scope. Each outcome is analyzed with a binomial GLM with a logit link and participant-clustered standard errors. When variation is insufficient for model estimation, we report descriptive statistics only.

11.5. RESULTS

We first report provision-level outcomes under AI error (H1–H2), then comparisons under correct AI assistance (H3), followed by Explanation-of-Reasons outcomes (H4–H5) and downstream risk-relevant outcomes (H6).

H1: Provision accuracy under AI error Evaluative AI substantially improved provision-level accuracy under AI error. Participants in the Evaluative arm were correct on 74.0% of error trials ($n = 150$), compared to 46.0% in the Conventional XAI arm ($n = 150$), a difference of 28 percentage points. In a logit GLM with heteroskedasticity-robust (HC1) standard errors, including AI error condition and case fixed effects, Evaluative assistance was associated with an approximately ten-fold increase in odds of selecting the correct legal provision relative to Conventional XAI (OR = 10.96, 95% CI [4.35, 27.65], $p < .001$). Results were numerically stable across estimation choices: a covariate-adjusted model including pre-task measures (MAILS, NCS-6, legal knowledge, Propensity to Trust Automation) yielded a comparable effect (OR = 11.10, 95% CI [4.52, 27.20], $p < .001$). The Evaluative advantage was present across all error conditions (Fig. 11.3). Independent of assistance condition, Out-of-Scope errors were associated with higher baseline accuracy than Near-Miss errors (OR = 10.34, 95% CI [3.77, 28.41], $p < .001$), while Overbreadth errors showed a smaller but still positive shift relative to Near-Miss (OR = 2.71, 95% CI [1.09, 6.72], $p = .032$). Moderation analyses indicated that the Evaluative advantage was attenuated for Out-of-Scope errors relative to Near-Miss errors (Evaluative $\times$ Out-of-Scope: OR = 0.07, 95% CI [0.02, 0.27], $p < .001$), while no reliable attenuation was observed for Overbreadth errors (Evaluative $\times$ Overbreadth: OR = 0.40, 95% CI [0.11, 1.46], $p = .165$). Case fixed effects were not statistically significant overall; full model tables and balance checks are reported in the Appendix.

Conclusion. H1 is supported: when the AI makes an error, Evaluative AI substantially improves provision accuracy, increasing correctness by 28 percentage points and yielding an order-of-magnitude increase in the odds of selecting the correct legal provision relative to Conventional XAI for the reference AI error condition (Near-Miss).

H2: Misclassification distance under AI error Evaluative AI reduced misclassification distance relative to Conventional XAI on AI-error trials. In the primary ordered-logit specification (without case fixed effects), the Evaluative coefficient was negative and statistically significant ($\beta = -1.53$, $SE = 0.40$, $z = -3.88$, $p < .001$), indicating a shift in probability mass toward smaller (less severe) distances. Interpreted as a proportional-

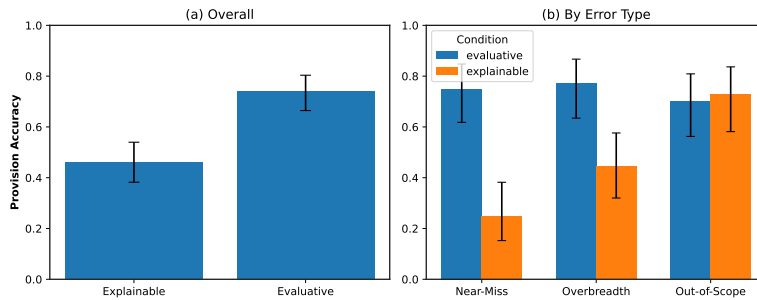


Figure 11.3: Provision-level accuracy under AI error regimes. (a) Overall accuracy on AI-error trials by assistance condition. (b) Accuracy by error condition (NM, OB, OS). Bars show proportions correct; error bars indicate 95% Wilson confidence intervals.

odds effect, this corresponds to $OR = 0.22$ with a 95% confidence interval of $[0.10, 0.47]$.⁵ Figure 11.4 visualizes the corresponding distributions: compared to Conventional XAI, Evaluative AI concentrates more probability mass on distances 0–1 and less on distances 2–3. Error condition moderated the effect. Relative to Near-Miss errors as the reference AI error condition, Out-of-Scope errors were associated with smaller distances overall ($\hat{\gamma}_{OS} = -1.34$, $p = .001$). The Evaluative advantage was significantly attenuated for Out-of-Scope errors (Evaluative $\times$ Out-of-Scope: $\hat{\phi}_{Eval \times OS} = 1.62$, $p = .007$). By contrast, there was no evidence in the interaction tests that Evaluative's improvement differed between Overbreadth and Near-Miss errors (Evaluative $\times$ Overbreadth: $\hat{\phi}_{Eval \times OB} = -0.07$, $p = .898$). Restricting the sample to Near-Miss and Overbreadth errors (H2b specification), the Evaluative effect remained strong at the Near-Miss baseline ($\hat{\beta} = -1.61$, $SE = 0.40$, $z = -4.02$, $p < .001$; $OR = 0.20$, 95% CI $[0.09, 0.44]$), while the Evaluative $\times$ Overbreadth interaction again was not statistically significant ($\hat{\phi}_{Eval \times OB} = -0.06$, $p = .919$), indicating comparable improvements for Near-Miss and Overbreadth errors. Results were robust across alternative specifications. Adding covariates (MAILS, NCS-6, legal knowledge, and Propensity to Trust Automation), including case fixed effects as a design-control robustness check, or excluding benign reporters did not alter conclusions. Non-parametric Mann–Whitney tests provided convergent evidence, showing significantly smaller misclassification distances under Evaluative AI overall ($U = 8128$, one-tailed $p < .001$), as well as within Near-Miss ($U = 734$, $p < .001$) and Overbreadth errors ($U = 841.5$, $p = .0003$).

Conclusion. H2a is supported: when the AI is wrong, Evaluative AI yields smaller misclassification distances than Conventional XAI. H2b receives partial support: the Evaluative benefit is attenuated for Out-of-Scope errors, while differences between Near-Miss and Overbreadth errors are not detectable in the interaction tests.

H3: No-error trials H3a: *Accuracy.* When the AI was correct, provision accuracy was high in both conditions and did not differ reliably between assistance forms. Descriptively, accuracy was 86.7% under Conventional XAI and 82.0% under Evaluative AI. In the binomial GLM with HC1 heteroskedasticity-robust standard errors (one observation per participant) and no case fixed effects, the Evaluative vs. Conventional XAI contrast

⁵Threshold (cutpoint) parameters are reported in the Appendix but are not substantively interpreted.

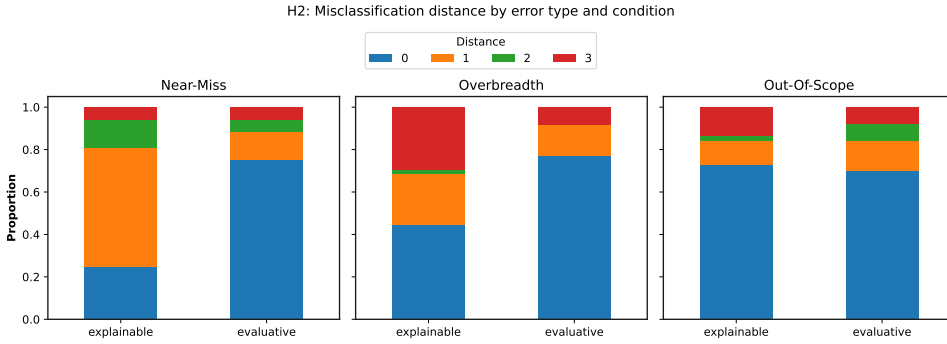


Figure 11.4: Misclassification distance under AI error by error condition and assistance condition.

was not statistically significant (OR = 0.65, 95% CI [0.35, 1.22], $p = .192$). Results were unchanged when adding case fixed effects as a design-control robustness check. A TOST equivalence test with margin $\delta = \pm 0.05$ did not support equivalence. While the upper bound test was satisfied ($p = .011$), the lower bound test failed ($p = .468$), indicating that equivalence could not be established under the specified margin. *H3b: Decision time.* Decision time analyses showed a consistent speed advantage for Conventional XAI. In the OLS model on log-transformed decision times with robust (HC1) standard errors, Evaluative AI was slower than Conventional XAI (coef = 0.13, $SE = 0.07$, $p = .064$), indicating a directionally consistent delay that did not reach conventional significance. Results were substantively unchanged when adding case fixed effects as a design-control robustness check. The non-parametric Mann-Whitney U test on raw decision times corroborated this directional effect: decisions were significantly faster under Conventional XAI than Evaluative AI ($U = 10012$, one-tailed $p = .0498$). Median decision times were 203,690 ms for Conventional XAI and 235,769 ms for Evaluative AI, corresponding to a substantial practical difference.

Conclusion. H3b is supported: when the AI assistance is correct, **XAI** enables faster decisions than **EVALAI**. H3a remains inconclusive: although the accuracy difference is not statistically significant, equivalence between Evaluative and Conventional XAI is not established under the prespecified margin.

H4: Explanation-of-Reasons (EoR) quality We tested whether Evaluative AI improves the quality of users' Explanation-of-Reasons (EoR) submissions relative to **BASILINE** reporting and Conventional XAI across four coder-rated dimensions: element coverage, proportionality reasoning, reasoning depth, and perceived overall quality. Across all four dimensions, we found no reliable differences between assistance conditions. In linear regression models on coder-mean ratings with participant-clustered standard errors, coefficients comparing Evaluative AI to Conventional XAI were small and statistically indistinguishable from zero (all $|t| < 0.70$, all $p > .48$). **BASILINE** reporting likewise did not significantly differ from Conventional XAI on any dimension. For element coverage, the largest estimated contrast was between Evaluative AI and **BASILINE** reporting ($\beta = 0.14$, $SE = 0.11$, $t = 1.24$, $p = .216$), but this difference was not statistically significant. No evidence of differences was observed between Evaluative AI and Conventional XAI

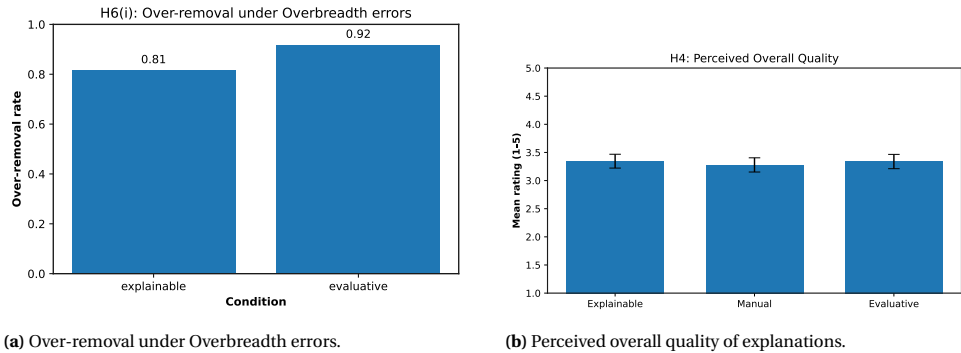


Figure 11.5: Downstream enforcement risk and explanation quality. **(a)** Over-removal rates under Overbreadth errors show no reduction under Evaluative AI. **(b)** Perceived overall explanation quality does not differ reliably by assistance condition. Together, the panels illustrate that gains in decision accuracy do not necessarily translate into improved justifications or proportional enforcement.

($\hat{\beta} = 0.007$, $p = .953$). Figure 11.5b illustrates the distribution of perceived overall quality ratings. Distributions for Evaluative AI and Conventional XAI largely overlap across rating levels, with no systematic shifts consistent with improved explanation quality.

Conclusion. H4 is not supported. Although Evaluative AI improves decision accuracy and error handling (H1–H2), it does not reliably improve coder-rated explanation quality relative to **BASILINE** reporting or **XAI**.

H5: AI error condition impact and attenuation Accuracy. In the paired difference-in-differences logistic model, AI error conditions substantially reduced provision accuracy relative to correct AI trials, and the magnitude of this error penalty varied across AI error conditions. Relative to Near-Miss as the reference AI error condition, the accuracy decline was significantly smaller for Out-of-Scope errors (error trial $\times$ OS > 0 , $p < .001$) and smaller for Overbreadth errors (error trial $\times$ OB > 0 , $p = .025$). Near-Miss errors produced the largest accuracy drops. Both Overbreadth and Out-of-Scope errors were significantly less damaging than Near-Miss errors, with Out-of-Scope errors showing the smallest declines. Differences between Overbreadth and Out-of-Scope were marginal. Wald contrasts confirmed that error penalties differed by AI error condition (OS vs. Near-Miss: $p < .001$; OS vs. Overbreadth: $p = .054$). **Attenuation in accuracy.** Evaluative AI significantly reduced the baseline error penalty relative to Conventional XAI (Evaluative AI $\times$ error trial > 0 , $p < .001$), indicating overall attenuation of error-induced accuracy losses. However, this attenuation was moderated by AI error condition. The three-way interaction for Out-of-Scope errors was negative and statistically significant (Evaluative AI $\times$ error trial $\times$ OS < 0 , $p < .001$), indicating that the additional attenuation provided by Evaluative AI was smaller for Out-of-Scope errors than for Near-Miss errors. No reliable moderation was observed for Overbreadth errors (Evaluative AI $\times$ error trial $\times$ OB, $p = .232$). Taken together, Evaluative AI most strongly buffered the accuracy penalty associated with Near-Miss errors, while incremental gains were reduced for AI error conditions that were already less harmful under Conventional XAI. *EoR quality (coder ratings; first-difference models).* We next examined whether AI error conditions affected explanation quality differently.

Using participant-level first-difference linear models on continuous coder-mean ratings, error trials were not associated with statistically reliable declines in EoR quality across any dimension. In particular, Out-of-Scope errors did not produce larger reductions than Near-Miss or Overbreadth errors in *element coverage*, *proportionality reasoning*, *reasoning depth*, or *perceived overall quality* (all $p > .39$). Accordingly, H5a is not supported for EoR quality. *Attenuation in EoR quality*. Evaluative AI did not significantly attenuate error-related changes in explanation quality on any EoR dimension. Interaction terms between Evaluative AI and Out-of-Scope errors were small and non-significant across all measures (all $p > .39$). Thus, H5b is not supported for EoR quality. *Robustness*. Sensitivity analyses using proportional-odds (ordered logit) models on discretized coder-mean ratings yielded substantively same conclusions, with no sign reversals or emergent significant effects (Appendix).

Conclusion. H5a and H5b are supported for accuracy but not for explanation quality. Evaluative AI selectively attenuates error-induced accuracy losses—most strongly for Near-Miss errors—while providing limited additional benefit for error conditions that are already less harmful under Conventional XAI. Across AI error conditions, however, Evaluative AI does not reliably improve the structure or completeness of user-provided explanations under AI error, indicating that its primary benefit lies in supporting correct decision outcomes rather than downstream explanation quality.

H6: DSA-relevant downstream risk outcomes *Overbreadth errors (H6(i))*. The Overbreadth panel comprised 102 error trials from 102 participants (54 Conventional XAI; 48 Evaluative AI). Severe enforcement actions were frequent in both conditions, with higher raw over-removal rates under Evaluative AI (91.7%) than Conventional XAI (81.5%). A binomial GLM with participant-clustered standard errors indicated that the Evaluative AI condition was associated with higher odds of selecting a removal-type action relative to Conventional XAI (log-odds $\beta = 0.92$, SE = 0.64; OR = 2.50, 95% CI [0.72, 8.68]), though this difference was not statistically significant ($p = 0.149$). Thus, we find no evidence that Evaluative AI reduces disproportionate enforcement under Overbreadth errors. *Out-of-Scope errors (H6(ii))*. The Out-of-Scope panel included 94 error trials from 94 participants (50 Evaluative AI; 44 Conventional XAI). Misrouting was near-universal: all Evaluative AI trials (100%) and nearly all Conventional XAI trials (95.5%) failed to select the out-of-scope routing option. Because the outcome exhibited insufficient variation, a regression model could not be meaningfully estimated. We therefore report descriptive statistics only and treat H6(ii) as not identifiable in this dataset.

Conclusion. Overall, H6(i) is not supported: Evaluative AI did not reduce over-removal under Overbreadth errors and showed descriptively higher, though non-significant, rates of severe enforcement. H6(ii) was not identifiable due to near-zero variance in out-of-scope routing behavior, indicating a measurement or interface limitation rather than a treatment effect. A summary of the hypotheses can be found in Appendix G.1.

11.6. DISCUSSION

This study examined how different AI assistance architectures shape user reporting behavior in content moderation workflows, focusing on accuracy under AI error, explanation quality, and implications for DSA-relevant reporting objectives. The results reveal a consistent pattern: Evaluative AI improves decision outcomes when the AI is wrong, but

does not reliably improve the quality of user-provided explanations. This distinction is consequential for governance-critical settings, where correct decisions and higher quality user provided justifications do not necessarily coincide.

Reducing overreliance under AI error. Under error conditions, Evaluative AI substantially increased provision accuracy (H1) and reduced misclassification distance (H2), indicating less harmful reliance on incorrect model suggestions. These effects are consistent with evaluative assistance functioning as a *deliberation scaffold*: by surfacing competing considerations rather than a single recommended category, the interface makes disagreement explicit and discourages one-shot acceptance of erroneous outputs [486, 537]. Even when users did not fully correct an AI error, evaluative support helped avoid the most severe misclassifications, keeping users' selections closer to legally relevant provisions. For moderation pipelines, this can improve the quality of information entering downstream processes under uncertainty [742].

Selective attenuation and targeted deployment. The paired DiD analyses show that Evaluative AI does not uniformly mitigate all error-related failures. Benefits were strongest for Near-Miss errors, attenuated for Out-of-Scope errors, and not detectably different from Near-Miss for Overbreadth errors. This pattern argues against a one-size-fits-all assistance policy [695]. Instead, the results suggest that evaluative support may be most useful in high-uncertainty or high-risk situations—for example, when model confidence is low, category entropy is high, or user choices conflict with AI suggestions [366]. Such targeted use could preserve robustness benefits while limiting added time and cognitive load, a key consideration given that evaluative interfaces slowed decisions when the AI was correct (H3b) and full human–AI review is not scalable in high-volume moderation contexts [449].

Why explanation quality did not improve. Despite accuracy gains, Evaluative AI did not reliably improve coder-rated Explanation-of-Reasons (EoR) quality (H4–H5). This dissociation suggests a mechanism mismatch: evaluative assistance primarily supports *choice revision* (selecting a better provision) rather than *argument construction* (articulating richer justifications). In our design, the reporting interface already required users to engage with statutory elements and evidence fields, which may have elevated baseline explanation quality across conditions and compressed variance. As prior work shows, providing AI explanations or guidance alone does not reliably improve reasoning quality or calibration [529]. Improving EoRs likely requires assistance that directly targets articulation—such as gap-highlighting, prompts for missing elements, or retrieval-backed evidence cues—without generating legal arguments on the user's behalf.

Implications for DSA-oriented design. The DSA requires user-friendly reporting mechanisms and sufficiently substantiated notices, but does not specify how interfaces should support users in meeting these standards [668]. Our findings indicate that Evaluative AI can improve the reliability of user reports under AI error by reducing overreliance and severe misclassification, thereby improving the *quality of information* entering moderation pipelines at the reporting stage. However, improved classification accuracy should not be conflated with legally sufficient explanations: Evaluative AI supports *better decisions*, but not necessarily *better justifications*. It is therefore best understood as a performance-enhancing design intervention under uncertainty, not as a mechanism that enforces regulatory compliance [647].

Downstream risks and platform costs. Downstream risk proxies (H6) highlight important limits. Under Overbreadth errors, over-removal rates were high in both conditions and descriptively higher under Evaluative AI, with no statistically reliable reduction in disproportionate enforcement. For Out-of-Scope errors, misrouting was near-universal, leaving insufficient variation to identify treatment effects. These results underscore that accuracy gains do not automatically translate into proportionate enforcement and suggest that interface defaults can dominate user choices [196]. More broadly, lowering reporting friction could increase the volume of actionable notices, raising moderation costs and governance burdens [119].

Limitations and future work. This study relied on a custom reporting interface with a curated set of legal provisions, which may have reduced real-world ambiguity and increased baseline performance relative to deployment settings. The structured explanation fields, while useful for measurement, may scaffold reasoning in ways that differ from deployed platform environments. Moreover, the experiment captures short-term behavior and cannot assess learning, strategic adaptation, or adversarial misuse over time. Future work should evaluate evaluative assistance in platform-integrated settings, link user reports to moderator outcomes, and explore adaptive or constrained writing support that better aligns gains in decision accuracy with meaningful improvements in explanation quality and downstream governance outcomes.

11.7. CONCLUSION

This study demonstrates that Evaluative AI improves user accuracy under AI error in illegal content reporting by reducing harmful overreliance relative to conventional XAI. Whereas conventional XAI interfaces encourage one-shot acceptance of incorrect suggestions and increase susceptibility to biases favorable to machine outputs, Evaluative AI supports more cautious and contestable decision-making under uncertainty. At the same time, we find no reliable improvements in the quality of user-provided explanations, indicating that gains in decision accuracy do not automatically translate into more substantiated or higher-quality reasoning.

These findings underscore that reporting interfaces are an important design space under the Digital Services Act and can improve performance under the right conditions. Evaluative AI is best understood as a deliberation scaffold that improves decision outcomes in error-prone settings, rather than as a mechanism that guarantees legally sufficient explanations or independently satisfies regulatory requirements. Important trade-offs remain: evaluative assistance increases deliberation time and may raise the volume of actionable reports, with implications for platform costs and downstream governance burden.

Future work should examine longitudinal effects on trust, learning, and strategic behavior; test generalizability across jurisdictions and platform contexts; and explore complementary support modalities—such as adaptive or constrained assistance—that directly target explanation articulation. Advancing such designs will be critical for balancing accuracy, user agency, and regulatory accountability in human-AI collaboration.

12

“THERE IS LITERALLY ZERO FUNDING”: UNDERSTANDING THE EMERGING ROLE OF TRUSTED FLAGGERS UNDER THE EU DIGITAL SERVICES ACT

“To me, going to the unknown is absolutely rule number one.”

Marina Abramović

The European Union's Digital Services Act (DSA) introduced regulatory mechanisms for addressing illegal and harmful content online. The recognition of Trusted Flaggers (TFs) is one such mechanism, accrediting entities with expertise, platform independence, and competence in identifying and reporting illegal content. With the DSA's TF role being roughly one year old, we interviewed representatives of seven such TF organizations to learn about their experiences of becoming a TF and how it impacts their interactions with online platforms and with individual users. We also conducted a workshop with TF representatives, primarily at their request, because they wanted to share experiences of their new role and learn from one another. We found that accreditation can be cumbersome, that TF resources remain unchanged despite increasing workloads, and that platform priorities often diverge from those of TFs. We conclude with recommendations for future research into understanding user representation within the DSA and the need for standardization measures tailored to the needs and resource constraints of TFs.

12.1. INTRODUCTION

Online platforms are a central component of today's interaction between individuals, whether it is to communicate [688], find information [427], shop [728], or just be entertained [763]. However, these online spaces pose different risks to users – one of which is illegal content online. To have a regulatory answer to this and many other risks¹ the EU has created the Digital Services Act (DSA) [224].

This novel regulation forms obligations for transparency, assessment, reporting, and design for online platforms like Instagram, TikTok, X, or Pornhub, as well as for search engines (See Art. 33 DSA [145]). Other countries are also pursuing comparable approaches to regulating online platforms and their processes for responding to harmful content, as in the UK [457], India [534], or US specific-rules in Florida [21] and Texas [25]. Breaching the DSA's obligations can have large financial consequences, as recently shown in the first fine issued against X of 120 million Euro [163].

The DSA includes transparency obligations (Arts. 15, 24, 42, 17, 39 DSA), systemic risk assessment and mitigation (Arts. 34-35 DSA), independent audit structures (Art. 37 DSA), and new user-facing redress mechanisms (Arts. 20-21). In particular, the DSA also creates a distinct Trusted Flagger (TF) role (Art. 22 DSA), to patrol the online space and identify illegal content online (Art. 3(h) DSA). The TF role is accredited by national DSA regulators², and must be recognized by large and small, from large social media platforms to small online stores. In the content moderation process TFs must be prioritized (Rec. 61).

With such a shift in the regulatory landscape, there will be a period of adjustment among organizations and individual reporters that pursue TF accreditation, and in the relationships between TFs and regulated platforms. It is important to understand any challenges which organizations and individuals face in attaining and using TF capabilities, as the enforcement of the DSA – and reduction of illegal and harmful content online – relies on these mechanisms functioning smoothly. The TF role, as a mechanism, is only a matter of months old at the time of writing [168], where it is critical to capture initial challenges and successes at this formative stage in the DSA [667], due to the novelty and legal complexity of the work, and capabilities beyond previous mechanisms such as Google's Priority Flagger program [298].

With this, our overarching Research Question (RQ) asks: *How do Trusted Flaggers enact the Article 22 DSA mechanism in practice, and what frictions and accountability implications arise from this for end-users representation?* This leads to distinct Sub Research Questions (SRQs) are: **SRQ1:** Are there challenges which Trusted Flaggers (TFs) have experienced in becoming and acting as a TF while the DSA is in its early stages, particularly in comparison to their activities before obtaining TF accreditation? **SRQ2:** What gaps emerge between the capabilities the DSA assigns to Trusted Flaggers (TFs)

This chapter is based on the preprint: Sekwenz, M.-T., Beadle, K., & Parkin, S. (2026). “There Is Literally Zero Funding: Understanding the Emerging Role of Trusted Flaggers under the EU Digital Services Act.” doi:10.48550/arXiv.2603.29874. Interview and workshop materials appear in Appendix H.

¹E.g. negative effects on the electoral process and public security, gender-based violence, child safety, and negative effects on fundamental rights.

²Digital Service Coordinators according to Art. 22(2)DSA.

and the resources, tooling, procedural support, and public visibility? **SRQ3:** How do Trusted Flaggers' (TF) experiences reveal differences in how platforms operationalize and prioritize DSA content reporting obligations, what are key challenges for TFs, and how are these practices reflected in DSA transparency and accountability artifacts (e.g., transparency reports and Statements of Reasons)?

This paper makes four contributions. First, it provides an early empirical account of how accredited Trusted Flaggers enact Article 22 DSA in practice – including the tools and activities involved – based on interviews with representatives of seven Trusted Flagger organizations and a validation workshop. Second, it identifies six recurrent implementation challenges that shape Trusted Flagger work: platform frictions, capacity constraints, coordination challenges, platform incentives, operational ambiguity, and mismatches in accountability artifacts. Third, it shows that the formal prioritization and institutional recognition granted to Trusted Flaggers do not, by themselves, resolve the practical burdens of reporting illegal content online, as the mechanism remains constrained by fragmented platform infrastructures, limited resources, and uneven visibility of their effort to platform users and regulators. Fourth, it contributes to HCI and platform governance research by demonstrating how the Trusted Flagger mechanism redistributes content moderation labor to platform-external organizations, while creating new questions of end-user representation, design of online reporting mechanisms, standardization, and accountability under the DSA.

The paper is structured as follows: Section 12.2 outlines the Trusted Flagger mechanism under the DSA and its relationship to notice-and-action and transparency obligations. Section 12.3 reviews related work. Section 12.4 describes our methods. Section 12.5 reports the findings. Section 12.6 discusses implications and recommendations for DSA implementation and accountability. Section 12.7 concludes the paper.

12.2. BACKGROUND

Unlike platform-internal - or 'commercial' - content moderators, who have been the focus of much prior research (e.g., [682]), Trusted Flaggers (TF) are platform-external, typically not-for-profit or civil society actors whose notices of allegedly illegal content, under the provisions of the DSA, should receive priority treatment by platforms.

This is codified in Article 22 DSA – a “fast-lane” for illegal content (Art. 3(h) DSA) – for notices (Art. 16 DSA) submitted through content reporting interfaces by a TF (Art. 22 DSA) [625]. TFs are positioned as an institutional response to illegal content that harms users (like CSAM content or disinformation). Figure 12.1 provides an overview of the Trusted Flagger framework under the DSA, distinguishing between the accreditation process and the tasks performed once an organization has been accredited. Through their mandate-bound flagging activity, the TF role operationalizes specialized expertise that reporting entities such as NGOs have in the identification and substantiation of presumed illegal content. The TF role then empowers existing experts and expertise, to negate the challenges in reporting content which non-experts would face, in terms of collecting evidence and understanding the law [667].

The TF role brings implications for illegal content reporting quality (as a form of external expert reporting) [684], and user representation (Art. 86) interact under a novel regulatory regime and create tensions and challenges for actors, pose new questions of

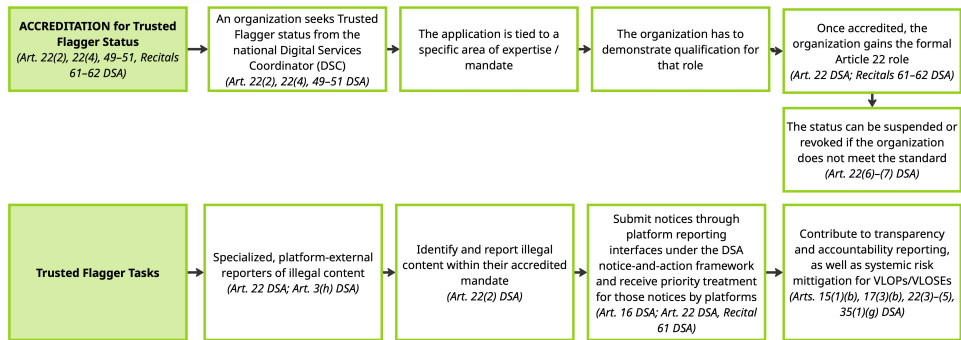


Figure 12.1: Overview of Trusted Flagger accreditation and tasks under the Digital Services Act.

Alt text: The figure is organized in two horizontal rows of green boxes. The first row presents the accreditation pathway for Trusted Flagger status under the DSA: an organization seeks status from the national Digital Services Coordinator, the application is linked to a specific area of expertise, the organization must demonstrate its qualification for the role, it then gains Article 22 status, and that status may later be suspended or revoked if the criteria are no longer met. The second row presents the core tasks of Trusted Flaggers: they act as specialized, platform-external reporters of illegal content, identify and report illegal content within their accredited mandate, submit notices through platform reporting interfaces under the DSA notice-and-action framework, and contribute to transparency and accountability mechanisms.

how best practices could be fostered, and how TFs can best support user groups and help platforms to identify already moderated illegal content online.

For the research community in human-computer interaction, TFs offer a concrete case of how regulation reshapes interactions between platforms, TFs, and users, through DSA-governed processes and supported through systems and interfaces which shape Europe-wide content moderation in practice. TF designation introduces procedural prioritization (Rec. 61–62) that structures the timing of platform decisions, an aspect with normative consequences for affected parties beyond efficiency considerations [684]. TFs also change the evidentiary and substantive quality of notices by adding external expert reporting. Finally, TFs foreground questions of representation and accountability (Art. 86), because particular organizations are empowered to escalate certain harms and illegality claims within platform processes. These interactions can produce new dynamics for the actors involved and point to open questions about how best practices should be developed and institutionalized across digital platforms.

12.3. RELATED WORK

Flagging as content moderation measure Content moderation can be conducted by humans and automated systems [288, 306]. Algorithmic solutions of content moderation mainly try to address the problem of scale [290]. Recent work shows that algorithmic moderation can produce inconsistent decisions on the same content, raising concerns

about fairness and discrimination [295]. Such degrees of fairness in content moderation can be perceived by users [111, 489, 690] as a form of procedural justice mechanism [409]. Under the DSA, platforms must enforce content moderation decisions in a way that is not arbitrary or discriminatory, and that supports diligent processes (e.g. Art. 35(1)(c) DSA). How to find out what is “not arbitrary” in practice within a (semi-) automated content moderation process however is still unclear, and therefore also how compliance should be assessed [693].

Content reporting (or, flagging [174]), as another part of the moderation process, is an approach for bringing problematic content to the attention of platforms through the use of dedicated UI features [671]; either by users (as a community-based approach [440]), or by experts like TFs (where such expert flagging has the potential to change the effectiveness of the content moderation process [280]). HCI research also shows that interface-level filtering and presentation decisions can shape how represented groups are perceived [637]. This is relevant for Trusted Flagger infrastructures, where platform-side reporting and prioritization mechanisms may similarly influence how harms, reporters, and affected communities are represented within moderation workflows. However, the design of such reporting systems is a complex legal UI design challenge [667]. Related HCI work on regulatory technologies similarly shows that practitioner needs are not satisfied by detection alone, but by tools and infrastructures that support legally meaningful assessment. As one interviewee in Rossi and Parkin’s CHI study put it, “what I’m interested in is something that violates the law” [627, p. 7], underscoring the gap between research-facing categories and enforcement-facing legal criteria. Flagging can be used to open up the content moderation queue, and in turn feed in to moderation decisions like deletion [550], or soft moderation measures like the application of warning labels [62, 390]. Flagging can be seen as a form of countermeasure to problematic content online, or ‘trust and safety’ issues [196, 539] (such as mitigating harmful content online [50]). It is then the human counterpart to algorithmic content moderation [190, 306, 592, 642]. Previous research shows that who reviews content can shape moderation outcomes, and that a wider mix of perspectives may reduce political bias; Trusted Flaggers offer a different model, based on formal expertise [697].

The design and procedural transparency of such flagging mechanisms, on the one hand, can influence the flagging behavior of end-users and influence their motivation to act [777]. More broadly, HCI scholarship on information navigation argues that complex socio-technical systems should be designed with “transparency about the inner workings” and “more opportunities for user control” [365]. Although developed in the context of news navigation, this insight is highly relevant to DSA reporting and Trusted Flagger processes, where users and intermediary organizations must navigate opaque platform procedures, shifting evidentiary requirements, and unclear prioritization logics. New regulations such as the UK Online Safety Act (Sec. 20) [569] or the DSA (Art. 16) create new obligations, for how such reporting interfaces must be designed [667]. Lanius et al. find that flagged content generally made participants view that content more negatively, and reduced self-reported willingness to engage with it [454]. On the other hand, content moderation measures can enable unintended harms [131], as they can be misused to silence specific groups [47, 487], i.e. content reporting can be used to silence marginalized or at risk groups [48], be perceived differently by various user groups like young users

[109], or users with different accessibility needs [366], or can create privacy concerns for users [473].

Trusted Flaggers in the DSA By introducing the new role of TFs the content moderation process on platforms changes too (Art. 22 DSA). The DSA embeds transparency and accountability requirements through annual public reporting (TF reports) and regulator-facing disclosures (Art. 22(3)). The DSA also links TF reporting activity to platform transparency reporting (Arts. 15(1)(b)) and statements of reasons (Art. 17(3)(b), 22(3)-(5)), as well as in systemic risk mitigation measures (Art. 35(1)(g)) [413, 658]. TF status (accreditation) is granted by national regulatory authorities of the DSA – the Digital Services Coordinators (DSCs) (Art. 22 (4), (7-8), 49-51) – for a designated area of expertise (within a specific mandate of illegal content). This status is conditioned on demonstrable competence, independence from platforms, and diligent, accurate, and objective flagging (Art. 22(2)). The status as a TF can be suspended and revoked by DSCs in cases where notices are imprecise, inaccurate, or inadequately substantiated (Arts. 16, 20, 22(6)-(7)) [44].

TFs in the big picture are one of the roles that positively contribute to the assessment (Art. 34(1)(a) DSA)³ and mitigation of systemic risks (Art. 35 DSA)⁴ which is mandatory for VLOPs and VLOSEs [212, 330].

A Legal Lens on Trusted Flaggers Legal scholarship has argued that TF arrangements function as *privileged* third-party access to platform notice-and-action, and that their legitimacy is contested because they promise expertise while raising concerns about opacity, accountability, and over-removal. Appelmann & Leerssen [44] conceptualize trusted flagging as a heterogeneous governance practice that is varying by legal construction, degree of privilege, and point of intervention in moderation, and map competing narratives that frame TFs as (i) a source of expertise and inclusion, (ii) an unaccountable co-opting of public/private power, or (iii) a performative signal of inclusion rather than a structural shift in control. Further, Van de Kerkhof’s constitutional analysis [723] highlights how formalizing TF roles can intensify freedom-of-expression and rule-of-law concerns. There are then tensions in how TFs would exercise reporting decisions over online platforms.

Prior work has examined TF arrangements as privileged, expertise-based reporting channels that platforms may prioritize, often through bilateral and opaque partnerships. Van de Kerkhof analyzes how Article 22 DSA formalizes this model while foregrounding persistent tensions around over-removal, transparency, and fundamental rights [724]. The author further highlights implementation risks that matter for socio-technical studies of notice-and-action (Art. 16 DSA), including the contestability of certification criteria (e.g., “objectivity”), cross-border legality frictions (e.g. the definition of ‘illegal’ content under Art. 3(h) DSA), and potential gaps in user-facing transparency when TF reporting is operationalized through platform workflows [724].

³This systemic risk category addresses the identification, analysis, and assessment of the ‘dissemination of illegal content’.

⁴Here the DSA specifically links in Art. 35(1)(g) to the relationship between TFs and VLOPs and VLOSEs and the mitigation measure of “initiating or adjusting cooperation with trusted flaggers”.

Especially differentiating ‘accurately’ [751] between legal and illegal content is challenging [742, 764], even for widely accepted topics like CSAM content national definitions differ [434]. However, under which legal reason such reports are filed matters for transparency reporting and documentation and assessment of systemic risks [364, 747].

Finally, Van de Kerkhof proposes harmonization and standardization measures (e.g., under Art. 22(8) and Art. 44(1)(c) DSA) and discusses resourcing options for TFs [724]. It is then important to understand how TFs experience and interpret their reporting successes and setbacks, and how the TF role affects the existing relationships they have with platforms and new relationships which are enabled by the TF role.

12.4. METHOD

Here we describe our multi-stage approach to understanding Trusted Flagger activities. The flow of research activities is illustrated in Figure 12.2.

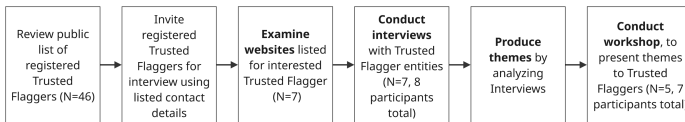


Figure 12.2: Steps in our engagement with Trusted Flagger entities.

Alt text: Flow diagram showing the six stages of the study. The researchers first reviewed the public list of 46 registered Trusted Flaggers, then invited listed organizations to participate in interviews. They examined the websites of seven participating Trusted Flagger organizations, conducted seven interviews with eight participants, analyzed the interviews to develop themes, and finally conducted a validation workshop with seven participants from five Trusted Flagger organizations.

12.4.1. SEMI-STRUCTURED INTERVIEWS

We conducted seven semi-structured interviews with eight participants, with one interview involving two representatives from the same Trusted Flagger (TF) organization. The interviews were designed to explore how accredited TFs experience their role in practice, while being semi-structured to leave space for participants to foreground issues they considered especially important. The full interview guide is provided in Appendix H.1.

During the interviews, we also showed participants screenshots from the Statement of Reasons (SOR) database, covering Trusted Flagger notices across platforms from 25 September 2023⁵ to 5 November 2025. We did this to elicit reflections on DSA accountability artifacts, and whether those artifacts aligned with the participants’ own reporting experience (Figures H.1 and H.2; [166]). Each interview lasted approximately one hour and was conducted via Microsoft Teams using its automatic transcription function.

12.4.2. RECRUITMENT

Participant recruitment was based on the public list of accredited Trusted Flaggers published under the DSA [168]. We sent an interview invitation to the listed contact email

⁵Launch of the SOR database.

address of each TF in the list on the 30th of October 2025, and waited for responses of TFs until the 26th of November. At the time of recruitment there were 46 registered Trusted Flaggers.

Because this is a small and publicly identifiable group of organizations carrying out resource-intensive work in the public interest, we adopted a deliberately restrained recruitment strategy. We contacted each listed TF once through the public contact address provided in the directory, and did not send reminder emails. This was both a methodological and an ethical decision: we did not want to place repeated pressure on a small population of publicly-listed organizations that represent user needs, simply to increase our sample size. Our study results evidenced this to be a reasonable position, as we found that TFs operate under substantial workload pressures (See Section 12.5.4). The participants who took part in the study responded quickly, typically within one or two days, with similar strong interest contributing to the opportunity to hold the validation workshop (Section 12.4.6).

12.4.3. PARTICIPANTS

Table 12.1 summarizes participants' roles, the general scope of their organizations, and whether they later attended the workshop (Section 12.4.6). The organizational scopes represented in our sample also aligned closely with both the official TF areas of expertise listed by the European Commission, and with the illegal-content category labels used in the DSA Transparency Database/API [1], which supported our use of Statement of Reasons screenshots as prompts that were substantively relevant to participants' accredited mandates (as illustrated in Appendix Figures H.1 and H.2). As such, our recruitment achieved coverage of all the recognized areas of TF expertise.

For organization 'P2', two representatives joined the same interview call separately (P2); we identify these as P2-1 and P2-2 (as in Table 12.1). As the TF directory is a public list, we knowingly define Organization Scope in a manner that is high-level but is indicative of participant roles. Our recruitment proved sufficient according to theme saturation borne out by the interview analysis (Section 12.4.5). Across the seven interviews, participants described notably similar interactions with platforms, regulators, and reporting processes, despite differing organizational mandates and substantive areas of expertise. After the first five interviews, no substantially new codes were identified in the coding process, with the remaining interviews reinforcing rather than expanding the thematic structure. This aligns with thematic analysis approaches which treat saturation as the point at which additional data no longer materially extends the interpretive themes [95]. The subsequent workshop, which was attended by representatives from most of the interviewed organizations (See Table 12.1), provided an additional opportunity to validate the identified themes (as shown later in Figure 12.4), and confirmed that the patterns identified from the interviews were recognized in practice.

12.4.4. PARTICIPATING ORGANIZATIONS – WEBSITE ANALYSIS

Contributing to the answer to RQ2, we also conducted an analysis of the participants' organization websites. The website analysis allowed us to investigate how organizations communicate their Trusted Flagger status to users and how Trusted Flaggers expect to interact with users. We first located these websites using the public list of accredited

Part.	Role / Position	Organization Scope	Workshop
P1	Lead TF team	Online information integrity and public-interest harms.	Yes
P2	P2-1: Legal counsel; P2-2: Leads communication/cooperation with platforms	Consumer- and product-related illegality.	P2-1: Yes; P2-2: No
P3	Scientific researcher	Identity- and belief-related harms and illegality.	No
P4	Legal advisor and content analyst	Interpersonal harm prevention.	Yes
P5	Cybercrime specialist	Online harm-related activity.	Yes
P6	Hotline manager; lead analyst team	Online safety reporting.	No
P7	Team lead TF	Consumer-related illegality.	Yes

Table 12.1: Overview of interview participants, including role/position, organizational scope, and workshop attendance.

TFs [168]. One author then proposed a set of categories upon which to interpret how components of the websites related to the research questions. These categories were then discussed and refined among all authors. The same author then noted websites information in those categories – this included languages used, accessibility considerations, working area matching expertise, funding sources, signpost status, published TF report, visible employee roles, and user contacting process, asks of user reports, platforms mentioned, privacy concerns, distinction between content types, and reflections. This allowed us to characterize how the various TF websites communicate what they do, who they support, and whether they mention that they are a TF. Insights gained from the website analysis were carried into the interviews (notably Section 12.5.2).

12.4.5. INTERVIEW ANALYSIS

Interview transcripts were analyzed using codebook-based thematic analysis (TA) [94]. This enabled us to identify the capacity, processes, and perceptions of Trusted Flaggers when they engage with platforms, regulators, and users. All authors began by familiarizing themselves with the data by reviewing the transcripts. Two authors independently conducted open coding on the first three transcripts. All authors then met to cluster codes around certain themes before continuing open coding of the remaining transcripts. The first coder had 35 codes and the second coder had 27 codes, yielding a total of 62 codes between them.

We did not measure inter-rater reliability given the iterative, inductive nature of our analysis (where the quality of codebook-based TA is not expected to rely on this as a measure [91, 506]). However, there was considerable overlap in the two sets of codes, supporting further discussion. Instead, the codebooks of the two coder-authors were used as reference artifacts during discussion, given the interdisciplinary nature of the analysis, and that the coders brought their own research backgrounds and experience relative to the study. While the two coders were aligned on the broad themes developed, the first coder focused on identifying the perceptions of TFs (especially through a lens of fairness), while the second coder aimed to identify the processes and interactions which TFs engage in. Both coders had experience conducting and analyzing interview-based

studies prior to this study. All authors met or communicated regularly (at least every few days) to discuss the coding process.

Six themes emerged from the interview analysis: **Platform frictions:** Platforms obstruct or fragment Trusted Flagger work; **Capacity constraints:** Capacity, resourcing, and training constraints shape what Trusted Flaggers can do; **Coordination challenges:** Regulatory coordination is uneven and sometimes difficult; **Platform incentives:** Platforms' incentives and response patterns shape outcomes; **Operational ambiguity:** Illegality is operationally difficult to determine; and **Accountability artefacts:** Transparency and accountability artefacts can misrepresent Trusted Flagger labour and roles. These themes are presented in Sections 12.5.3–12.5.8. One purpose of the subsequent workshop (Section 12.4.6) was to validate the six themes with participants; no participant disputed the presence of any theme.

12.4.6. VALIDATION WORKSHOP

We planned to conduct a workshop with participating organizations, contingent on sufficient interview uptake and interest. Several interviewees explicitly requested an opportunity to meet other Trusted Flagger organizations and learn about our broader results. Seven participants from five organizations represented in the interviews joined the online workshop (Organizations 1–2, 4–5, and 7; see Table 12.1). Organization P4 contributed more than one participant because its members were particularly interested in meeting other TF organizations.

The workshop was structured around an overview of the themes which emerged from the interview analysis (Section 12.4.5). We began the workshop by stepping through a set of slides (see Figure 12.4); each theme had one slide, which included 2–3 additional statements relating to each theme, for which we sought to understand (dis-)agreement, and which otherwise served as prompts for discussion (to validate the themes). Early presentation of overview themes was also an effort to make attendees feel comfortable voicing opinions on each theme afterwards. To this end, we encouraged attendees to speak up at any time with comments. After the presentation of themes, we encouraged open discussion, where we intended the workshop to also serve as a co-creation opportunity for Trusted Flaggers. To lower the barrier to participation, we supported both spoken interventions and written input during the workshop. The workshop lasted 90 minutes, as scheduled. Two of the authors were present, with one presenting the results, and the other present to monitor discussions. After this, remaining time would be given to participants to interact; we detail broad outcomes of this discussion in Section 12.5.9. Participants were invited to share their opinion on findings and were invited to correct or change any observation of the authors, by either speaking out loud or using a feedback tool during the workshop.

12.4.7. ETHICAL CONSIDERATIONS

The study went through institutional Human Research Ethics Committee (HREC) review and received approval. Consent for interviews and the workshop was obtained prior to the interviews. Data was stored according to a data management plan (DMP) that was also reviewed and approved before HREC review. The DMP included several data protection measures, e.g., data minimization. The interviews and online workshop were

recorded. Participants were informed in advance that other participating organizations would be attending the online workshop. Multiple authors were present on the online workshop call, to take written notes in case any participants objected to recording, but no participant joining the online workshop objected.

Participants did not receive any financial compensation for taking part in the study, but instead received advance view of the analysis. For participants, one distinct benefit from the study was that the planned workshop met a request from several TFs, and it became a meeting point to share experiences. Most of the organizations involved in interviews also then accepted the conditions of participation and joined the Teams-based workshop call. In preparation for the workshop, the summary of themes from the interview analysis (as in Appendix H.3) was reviewed by the authors several times, in an effort to limit any bias that may be introduced, so as not to sway the participants to make particular statements.

There is a risk that the participating organizations could be identified from the public list of registered Trusted Flaggers, given that – at the time of writing – it is a list of tens of organizations (rather than hundreds). The findings were written with this in mind, with consideration of how to convey meaningful detail while also being mindful of specific practices of any one organization.

12.4.8. LIMITATIONS

We recruited Trusted Flagger practitioners on a best-effort basis - we contacted the listed contact address in the directory of TFs, but only made one attempt, with an initial email invite. We did not send reminders, so as to limit any disturbance for these practitioners. Between sending the initial invites and the planning of the workshop, there was approximately four weeks, allowing TFs to respond. Although the number of participants is limited, the scale and intensity of their day-to-day reporting work was substantial. At the time of the recruitment there were 46 registered Trusted Flaggers. During the interviews (Section 12.5), participants described handling large batches of URLs, liaising with many different platforms, and navigating complex reporting infrastructures across jurisdictions and content types. Several participants were involved in reporting at significant scale, including engagement with hundreds of platforms or repeated submissions involving hundreds or thousands of URLs. For this reason, the interviews provide insight not only into individual perspectives, but into organizationally-embedded practices within a nascent but high-impact DSA reporting ecosystem.

One further limitation is that many participants reported only recently gaining Trusted Flagger status. These are early days for the DSA, so many activities have yet to become routine. However, our analysis already identified early challenges, where not all of these challenges will necessarily be resolved over time, but may require additional attention (e.g., platform opacity).

12.5. RESULTS

Here we detail the results of our interview and workshop analysis, arranged by code groups. Sections 12.5.1 and 12.5.2 provide foundational context for our participant cohort and their activities. Sections 12.5.3 to 12.5.8 describe the themes identified from the

interview analysis (Section 12.4.5). Closing the results section, Section 12.5.9 describes the outcomes of the online workshop we conducted with representatives of several participant organizations. A graphical overview can be found in Figure 12.3.

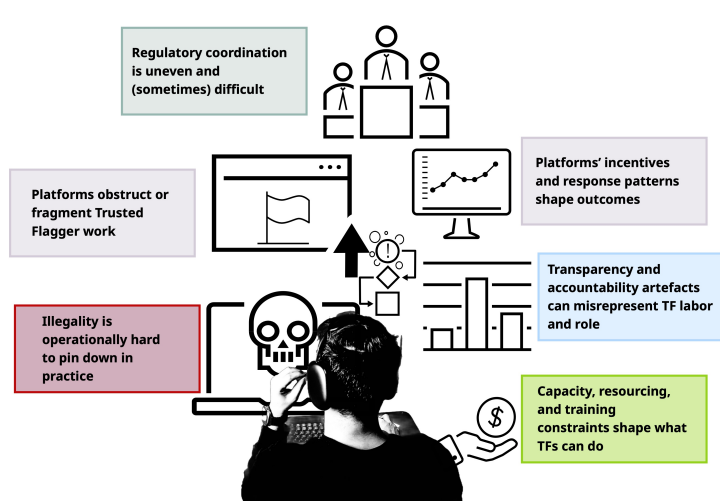


Figure 12.3: Graphical Overview of Key Findings

Alt text: The illustration summarizes six recurring themes in Trusted Flagger practice: platform frictions, coordination challenges, operational ambiguity around illegality, platform incentives, misrepresentation through accountability artifacts, and capacity constraints. These themes are arranged around a central image of a person working within a complex digital reporting and platform environment, emphasizing the socio-technical and organizational pressures that shape Trusted Flagger work.

12.5.1. TRUSTED FLAGGER ACTIVITIES

The process of becoming a Trusted Flagger (TF) could take several months (P1, P4). P7 attributed their 6–7 month process of accreditation work to there being a lot of questions and paperwork which “took for us more than 200 hours of multiple people working on it.” The process requires financial and organizational data, as well as proof of language and legal capacity, to be submitted to the national regulatory authority (the DSCs).

As P2-1 clarified, the ‘quite extensive’ application covered “expertise, independence, and our processes,” and P4 noted that “you have to have a legal background in house,” echoed by P7. Existing success rates and tooling capabilities were also seen as positive signals for accreditation (P2-1). P5 observed that processes differed across countries, with some member states having infrastructure in place sooner than others.

TFs must define their mandate, i.e., their area(s) of expertise and the categories of illegal content they are authorized to report. They must also evidence their “capability to identify” content as stated by P7. According to Art. 22(1)(a-c) DSA TFs must provide evidence of competence, impartiality, and objectivity within their area of expertise. The TF role therefore demands additional ‘capabilities’ (P1). TFs follow a strict scope where

they do not step outside their mandate even when contacted about content beyond it (P4), especially as “there’s already so much and we cannot do it all” (P5). P6 added that demonstrating platform independence requires evidence that their organization did not receive platform funding, or only negligible amounts.

Motivations to become a TF included the status the role brings and its ‘highly relevant’ value for their activities (P2-1), as well as interest in how the DSA unfolds – P7 spoke of their curiosity, and P5 described it as “a first kind of pilot phase.” P7 also aimed to build pressure on platforms to remove illegal content. TF status further signalled reliability (P2-1, P4) and encouraged platform engagement, since they have to acknowledge us as a reliable entity” (P2-1). P2-2 noted that some platforms proactively reached out with a dedicated reporting channel (as did P3), and P4 reported differences in report prioritization. P5 saw involvement in the DSA as an opportunity to influence regulation in ways that better reflect their organization’s cause.

TF activities as a new routine: Participants described TF entities as supporting both enforcement and recovery, for example by helping owners of ‘credible’ accounts (P1) that were taken down to restore them, addressing situations where reporting or take-down mechanisms were used against non-harmful entities. P1 also noted that the DSA’s scope is “much more narrower” than TFs’ day-to-day practice, because it focuses on illegal content, whereas in practice, TF work often extends to examining harmful or “highly problematic” content that does not meet the illegality threshold. Within their TF context, P3 reported following internal guidelines and, when appropriate, “I would report it to the platform via the specific process that they have.”

P1 described routine activities of “finding, vetting, checking, and documenting content, preserve[ing] the evidence,” and writing the report. This was mirrored by P2-2, who emphasized that reporting centers on “accuracy, compliance with all legal requirements and platform policies”⁶;

Documentation work would also capture “URLs, [...] infringement identifiers” and other supporting information (P2-2). P4 explained that they use tables, combined with a series of steps, to identify the relevant contact-point addresses within specific organizations and then route reports accordingly (in some cases also to law enforcement).

After submitting reports, TFs must monitor how platforms respond, which adds a further task that scales with the number of outstanding reports. P2-2 also noted that reports may be ‘rejected’ or that platforms may request “additional information.” P3 voiced that “in around half of the cases, they disagree with our judgment”, or geo-block content instead of EU-wide deletion of illegal content. P4 noted similar country-level blocking, but also commented that platforms have fast repose times and little disagreement with their classification noting that ‘community guidelines’ may be used to challenge a report (rather than legal norms). Since TFs must substantiate an annual report about their activities, they must also evidence their work through, e.g., “screenshots of the violation[s],” as outlined by P7. Different approaches to this documentation process appeared to develop at each TF.

Costs of TF activities: Where there are regular, routine activities of TF entities, par-

⁶Platform policies refer to platforms’ Terms and Conditions (TaC) according to Art. 3(u) as “[...] all clauses, irrespective of their name or form, which govern the contractual relationship between the provider of intermediary services and the recipients of the service.” and further regulated in Art. 14 DSA.

ticipants described accompanying costs. The experiences of participants resembled the longer end-to-end explanation provided by P1:

“you need to have the monitoring, you need to train the people. They need to find the specific violations, tailor all the documentations, collect the evidence, store the evidence. [...] then write report [...] All of this evidence put in the document, then register in the registrar where we need to measure when it’s reported and how much time does it take to take it down and then split the report to 20, 30 or 60 reports and then submit it, then track if you received confirmation e-mail and what is the timestamp of that confirmation e-mail. So you could put it in the Excel sheet so you could measure when it will be taken down. Then you will wait for the response and then when the response comes you will need to copy paste and check when that report was taken down so you could later write down the final report.” (P1)

These activities take effort, but already in these early stages of DSA roll-out, are chipping away at the resolve of TF members, who – where there are not additional resources to cover the costs – are acting somewhat as volunteers. P1 points to the hurdles of the reporting process that seem to be “designed to make your work miserable” and that every platform handles reporting interfaces and processes differently, which collectively impacts the quality of TF work in that “we [...] need to do [a] crazy amount of things, but you just are so under resourced that you just are not capable to do [that well].” Compounding this was that some participants could not reason why processes were as protracted or divergent as they seemed to be, where “it’s extremely unproductive the way the current system works, and it’s highly, highly problematic without a reason” (P1). Another view of this, as exemplified by P5, is that “We’d rather be cautious than making more numbers”. P3 noted that “there is no specific budget allocated for” DSA-related activities in their organization, such that it is “basically just an extra task that come on top of all of the other things” they do in their role. Aside from the sentiment toward reporting, we saw variety in how many platforms TFs engaged with – some restricted it to the VLOPs, focusing on social media content, whereas others reported engaging with hundreds of platforms year after year. Some participants also referred to ‘projects’ (e.g., P1, P4), as defining research or particular content they were focused on at a particular time (where P3 also referred to conducting ‘research’). P5 noted that their organization had been exploring ways to have platforms themselves allocate more resources to moderation of content (rather than it being them who would have to do so).

12.5.2. TRUSTED FLAGGER ACTIVITIES AND USERS

Many of our participant Trusted Flagger (TF) entities did not directly interact with members of the public in a reactive way. Rather, the TFs had a cause that drove their activities, which was seen as benefiting society, for instance P1, “Our job is to expose [...] illegal things and inform users.” A few participating TFs did plan for involving individual users, though seemingly as it was in the nature of their cause to do so (such as where the cause was around interpersonal harms or based on individual characteristics). This would then be a phenomena to monitor, as TF registrations grow over time. Similarly for P2-1, “we

not only fight illegal content on behalf of economic reasons for our members, but also a big social economic reason to protect consumers and governments”.

However, some TFs regard user reports as their starting point (such as P4, which would typically “receive [a] report from the public” or trusted organizations like NGOs). User reporting was more essential for P6 and P7, whose organizations were already dedicated to directly interacting with users before the DSA. These reports sometimes came from users who were “aware of our role as Trusted Flaggers” and included “thorough documentation and clear legal reasoning” (P7). However, the TF status is only one tool for these organizations as “90% of the reports we handle come from the online form” (P6).

P4 noted that they would sometimes be contacted by “someone I guess that tried everything they could before,” who they would then assist. Others, such as P1, do not routinely interact with individual users or citizens and do not see reporting for users as a key TF task.

TFs such as P3 and P5 are somewhere in-between, being contacted about content but responding on a best-effort basis, informed also by the amount of detail provided by the individual user (with it being “a matter of capacity”, P5). This adds shape to the developments noted in prior work [667], where the participants of the Sekwenz et al. study were not sure how TFs and users should be aligning with each other. As noted by P1, a TF may identify and report content faster than an individual user or citizen, resulting in lower reporting rates from those users (and less utility in providing dedicated reporting visibility).

(Lack of clarity on) Outward-facing visibility and role: When discussing TF visibility, such as the necessary online presence and website for a TF, it would appear to be something regulators and TFs still need to clarify (P1). As P1 put it, “So first they were saying that this is a mandatory thing, but then in the law it’s not mandatory thing. So it confused us. So in the end we didn’t put on our website [...] like there are other ways to report things on our website, but it’s not... not everything is about trusted flaggers.” highlighting that a single channel for reporting is preferable for users. P6 noted that “[TF status] is a very precious thing, but I don’t think it was very well communicated to the public.” Beyond any lack of clarity, another view among TFs was that it did not appear to change the nature of their interactions with users. P3 did though note that as a TF, they will be publishing public reports.

Responding to users who challenge reports: During the interviews, not all mention of users was about those individuals who shared the cause of the TF. In some cases, there were platform users who were the very same actors that TFs were acting against, i.e., those parties considered to be proliferating harmful material. This included actors seen as using DSA mechanisms to have their reported accounts reinstated, as noted by P1, where “we again spend our time,” as (in particular) ‘big’ accounts were “Reappearing after [...] we receive information that [...] they were taken down. So some of these big pages were taken down already three times and now the four times again they are being restored.”

Visibility via websites: Looking at how TFs communicate with users (Section 12.4.4), we found that there was no standard way organizations presented their TF status. Organizations used blog posts, press releases, and social media posts to communicate their TF status. The websites were also particularly helpful in understanding how the organizations’ working areas compared to their stated area of expertise under the DSA—with three

organizations selecting more areas of expertise than their websites communicated.

Some organizations included instructions on how users could contact them. The websites were largely accessible by people who spoke different languages as the participants' websites covered a total of 10 languages. All organizations were transparent of their funding sources, whether they came from government sources, donations, corporate memberships, or even the platforms themselves.

12.5.3. PLATFORM FRICTIONS

Platforms were seen as perpetually having content violations which needed effort to report and address, such that they would “communicate with the platforms on a daily basis” (P2-2). As noted by P1, “Most of the platforms violate DSA service, [...] or making things over complicated [...] But then you need to prove that”.

Views on the platforms' approach to DSA implementation were mixed, especially when it came to the differences in reporting pathways both within and between platforms. This was either seen as the reality, but also potentially intentional, that “[platforms] you know, they're doing everything to make this work complicated [because] we take out the revenue” (P1). This can result in a “frenemy situation” where TFs offer help and support but “they don't treat us as friends.” (P1). P3 highlighted an additional dynamic that tilted in favor of platforms, where “it also feels like it's a way to externalize the responsibility for hateful or illegal content that should belong to these social media platforms that are making huge profits from what users post.”

P4 remarked that where they had previously used templated e-mails to report content, they now needed to match reports to the different implementations of reporting channels that have emerged on platforms in response to the DSA. P7 found that with various reporting channels now existing, it “makes it difficult for us to especially choose the right reason we're reporting to get on the right path for reporting.” More practically, some platforms limited the amount of reports TFs could make, with P5 only being able to “report 10 URLs max at one time” after “they send you a link that gives you access for one hour [...] and if you were too slow, you have to ask again.” However, P1 saw issues of harmful content specificity with bigger platforms making “our work more complicated so they could keep as much of the revenue as long as possible. It's a very clear logic and very simple.”

P5 noted that some ‘smaller’ platforms fall outside the Trusted Flagger provision, resulting in “no harmonized flagging system”. Where a website did not mention the DSA, it could be “quite hard to find how to contact them.” P5 described relying on an email address that eventually stopped responding to requests to remove content. It is also difficult for external observers (e.g. TFs) to determine whether a platform is sufficiently small to fall outside the relevant DSA obligations, and DSCs may not have a complete overview. P5 therefore asked, “which platforms can we actually flag content to under our trusted flagger status”. These differences between TF experiences with smaller platforms and VLOPs resonate with earlier research suggesting that larger platforms will concentrate demand for audit activity [456].

What we are seeing here is that the machinations of larger platforms are drawing TF resources more than smaller platforms. Perhaps related are experiences such as those of P2, that platforms may ask for additional information for reports, and as P2-1 noted,

although the request may be clear there can still be a need to further explain TF's rationale and potentially even the need for stating the right of prioritization of TF reports. P2 and P3 noted roundabout discussions with platforms, that sometimes platforms re-request information and afterwards "ghosted" TFs (P3). P3 also speculated that the responses they receive from TF reports may be automated.

12.5.4. CAPACITY CONSTRAINTS

Participants were clear that they cannot act beyond the mandate they declare in their Trusted Flagger registration. The mandate collided with resourcing especially when considering irregular contact from individual users or other organizations, as "we are thinking how to support them. But this is again kind of problematic, you know, because it's to support them, it's cost for us" (P1).

Tooling: Many participants reported having developed, crafted, or obtained dedicated monitoring tools as a way to manage their work and streamline their efforts (e.g., P1, P4). This included AI-driven tools which automatically scan platforms for 'infringing content' (P2-1) and "automatically generate a list of potentially illegal content" (P6). These tools may be specific to their cause, raising questions as to how transferable those tools may be for others to learn from (as became a theme during the later workshop, Section 12.5.9). The monitoring capabilities were seen as a factor which differentiated the speed of TFs from that of citizen users (P1).

It was seen as necessary to bolster monitoring tools with subsequent data processing and classification capabilities. Crucially, this suite of capabilities then needed review by humans (P1, P5), or needed human involvement as part of engagement with other stakeholders (P2-1). Advances in tooling were then seen as a way to move beyond 'old school' reporting using tools such as databases (P5), where "we know things could be much faster".

Resourcing TF activities: Several participants gave shape to the limitations within which Trusted Flagger (TF) activity happens. For instance, "we don't have dedicated people who only work on this [...]" Like how can we dedicate people if we don't have resources on that?" (P1). As such, many employees would be involved in trusted flagging "on a rolling base, whoever has time" (P7) rather than being the task of a single employee. That TF activity may be relying on personal motivation was also evident, such as referring to it as doing it out of interest, or fun using up their "leftover" resources (P1). Similarly, P3 emphasized that TF activities were in addition to a range of other duties and roles they otherwise had within their organization, without additional budget. P5 framed human resources and expertise as "our biggest asset."

Discussion of resourcing also touched on legal resources, but also the spectre of liability for reports when faced with legal representation from platforms. P1 emphasized the 'vulnerability' of facing platforms in court cases. This was described in the context of providing evidence for DSA infringements where platforms in court "need to discredit our evidence", such that "we're making ourselves vulnerable, working for free" (P1). The scrutiny needed to protect against such scenarios is, in the case of P7, a mandatory "four eyes principle" for content evaluation. P4 remarked that "we have less power but more transparency" where "We don't have so much power to force them to do anything".

Skills came back into frame, as noted by P1: "You need to have the legal knowledge,

you need to have the technical knowledge, you need to collect the evidence.” Legal skills are further required after platforms fail to take down reported content and TFs “appeal, that’s when it comes to the legal team [of the platforms] themselves and not the moderator teams” (P4).

Some TF entities may have membership payments which cover their costs (perhaps as a non-profit organization), whereas others may be supported by volunteer efforts (much like a charitable organization). However, TF organizations struggle to fund flagging activities in addition to their existing activities. P6 went as far to say, “we don’t really have a budget right now. It’s just people are already paid for their service to [Organization 6].”

12.5.5. COORDINATION CHALLENGES

Becoming a Trusted Flagger (TF) was seen as a way to smooth the path, as necessary to “flag non-cooperative platforms” to a DSC or the European Commission (P2-1). P1 on the other hand saw it as “we’re doing institutions work” and therefore not only supporting content moderation processes of platforms, but also contributing to DSA enforcement, needing to coordinate across where they needed to coordinate to streamlining their arguments for three lenses: the DSCs, the platforms, and the local authorities. This coordination with stakeholders was further highlighted by other participants, referring to national law enforcement and judicial powers (e.g., P1, P4, P5). Similarly, P6 referred to their trusted flagging activities as happening “almost parallel” to law enforcement.

One example is this distance from investigations may be in how, as P5 put it, some content may be ‘kind of coded’, using terms which members of a community know the meaning of, but which by itself is not directly against the law (and which it was speculated, the platforms themselves could be checking before allowing on the platform). Similarly, P3 reported – when discussing how platforms differ in their reporting processes – that their DSC “said deal with it like you have to do your own work to see how things work”.

P2-1 noted being in regular communication with the DSC and Commission, regarding clarification of guidelines and procedures; this was echoed by P4. P4 cautioned that there can be cases where a platform is “not compliant or does not want to cooperate”, where P7 noted similar experiences and a need to approach the DSC, after they had “filed a number of complaints of Trusted Flagger cases with [a platform]” and they were repeatedly declined. P3 brought attention to the importance of coordinating with other reporting entities, wondering aloud how other TFs are managing even if they are dealing with different kinds of content and mandates. P4 also expressed a wish to meet other TFs facing similar challenges. P4 discussed that (aspiring) TFs are in need of templates for notifications to platforms, as well as details on annual TF reporting duties. We also saw that – perhaps due to the DSA still finding shape – TFs would also “write some recommendations” (P1) to be taken, ultimately, to the Commission. This included how DSA reporting will look in the end after years and decades of being in place, as P1 put it, “we will need to create everything and then you know, set a standard.”

12.5.6. PLATFORM INCENTIVES

Many participants noted that the effort required to report was compounded by platforms, big and small, having different reporting pathways. As remarked by P1, “with every platform, the process, how you submit the reports is different. There is no single form,”

but rather every platform designs reporting and TF mechanisms differently.

Some platforms reached out to TFs to establish a reporting channel, though this did not always happen smoothly, where there were anecdotes of how a platform may require a TF “to create an account, [yet] it asks for a phone number, but a personal phone number. But obviously as an organization [...] we cannot use our private ones because then we already have a [private] account” (P5). Technical implementations may even mean that “you need to report one URL at a time, which is hilarious” (P1). The cumbersome nature of reporting did serve to remind TFs that “it’s a zoo and it’s difficult for us” (P7) and “this field is still very, very, very new” (P1).

In contrast, P2-1 and P2-2 noted that platform policies may fill in gaps in EU regulation, where those policies provide additional mechanisms useful for the cause of the organization outside of its TF status (such as ‘defamatory comments’). However, they also stressed the need to “have a more uniform way of implementing the trusted flagger status” (P2-1).

The differences in reporting also left TFs with a sense that their reporting activity needed “tailoring” for violation to improve reaction times of platforms (P1). It was seen that the move to structured DSA reporting has resulted in more cumbersome, slower reporting pathways. This has created a situation where “some platforms apply unnecessary requirements to take down notices” (P2-1). As noted by P1, “with Meta we can only report 20 URLs at a time. Some of our reports will contain 800 URLs or 2000 URLs. So that means that we will need to instead of sending one report. [...] They are taken down bit by bit, 20 by 20,” whereas in the past, “we were submitting by emails and they were kind of reacting.”

12.5.7. OPERATIONAL AMBIGUITY

The type of content being reported often made it difficult for participants to determine legality. P2-1 noted experiencing “a very high success rate” with their reports, but that “IP infringing content is a little more black and white...rather than maybe some gray or political content or something like that.” On the other hand, P6 found that platforms were slow to delete “accounts that are specialized into linking [illegal] content that’s not on the platform which is external to the platform.” P7 explained that platform features pose different challenges, such as Reels on Instagram that ‘vanish’ and can not easily be reported or evidenced for TF activity. P5 often struggled to determine the legality of content where external documentation was required but only alluded to in the content saying, “if [users] make a vague statement referring to legality, such as I have the papers [that would dedifferentiate between a legal and illegal case], we also flag it,” such edge cases of certainty are still challenging and P5 is waiting for clarification on how their organization should handle such cases.

12.5.8. ACCOUNTABILITY ARTIFACTS

When presented with the reporting data captured by the Statement of Reason (SOR) database including all TF reports Participants variously had difficulty in relating the platform reporting data to their own reporting activities, where “It does not reflect reality” (P1), and “it just feels not very transparent” (P5). Compounding this, as P1 noted: “Every platform has different community standards” with overlaps of about 80-90%, these how-

ever are formulated in a always different and “very unfriendly way” demanding time of TFs.

Such observations of unsynchronised community standrads and the SOR input logic using it's own categorization terms mirror those found in the analyses of statement of reasons databases across several platforms [586].

Several participants noted difficulties in relating statement of reasons data with their own activity when presented with the screenshots illustrating TF activity (See Figures H.1 and H.2), for instance P3, commenting that “maybe even under 10 [reports] were in the database while we had done 10s of [reports],” leading them to wonder if the database is “not updated properly”.

Participants were left to come up with their own justifications for why reports did not show up in the public database. Participants hypothesized that the terms of service was a wide umbrella (though not the correct one) which covered illegal content implicitly (P6), but also significantly plays down the TF activity – who **only** report *illegal* content through reporting mechanisms. Meanwhile P7 argued, “the platform steers you into what is favorable for them...if we were certain we flagged illegal content under specific laws, the platforms came back to us saying, ‘thank you,’ but this doesn’t violate our terms of services.”

12.5.9. RESULTS - WORKSHOP EVENT

The workshop primarily functioned as (i) a check of our interview-derived themes among our five participants (Including Organization 1-2, 4-5, and 7, See Section 12.5, 12.4.6 and Table 12.1) and (ii) a peer-to-peer exchange, where participants had an opportunity to compare platform-specific reporting workflows, practices, and interactions. During the workshop discussion, participants corroborated the themes identified in our analysis, as illustrated in Figure 12.4. This led to discussion of practical day-to-day experiences. Participants collectively stressed that platform implementations can shape what becomes visible as TF action. Organization 2 described Meta’s TF portal constraint: “if we report to Meta through the trusted flagger channel, we’re only limited to 20 URLs”. Organization 1 similarly explained, “for example, last week reported a single campaign. And it contained 2 1/2 thousand Facebook pages URLs. So that meant that we submitted 130 reports and it’s very time consuming”. In response to URL limits per submission (see Sections 12.5.1, 12.5.3, and 12.5.6), participants compared technical workarounds such as adding additional URLs in free-text fields. There were then concerns that data registered in non-standard ways may not survive the journey through to platform confirmations or downstream transparency and accountability artifacts such as the SOR database (see Section 12.5.8).

The workshop also underscored that measuring “prioritization” as a main TF quality compared to other user reports is currently methodologically under-specified: participants debated whether take-down time should be computed from the moment of submission versus platform acknowledgment of ‘receiving’ TF reports, how to evidence submission timestamps, and how missing confirmations and partial/batched processing distort measurement. Some participants would illustrate their view by discussing their own tooling for reporting (see Section 12.5.4) that visualized weekly/monthly volume, takedown rates, and shifts in response times over time, illustrating both the feasibility

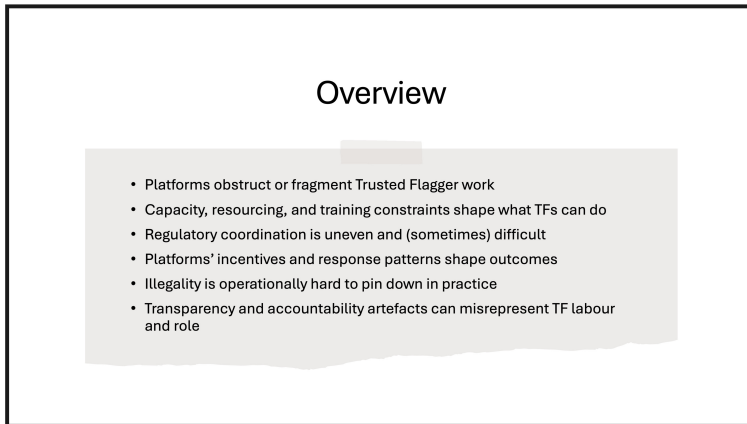


Figure 12.4: Overview-slide of the main challenges shaping Trusted Flagger work (as presented to workshop attendees), including platform obstruction, capacity constraints, uneven regulatory coordination, incentive-driven platform responses, difficulties in assessing illegality, and the limits of transparency and accountability artefacts.

Alt text: This slide summarizes the six cross-cutting themes emerging from the analysis of Trusted Flagger practice. Under the heading “Overview,” the slide lists the main findings in bullet-point form: platform frictions, capacity constraints, coordination difficulties, platform incentive structures, operational ambiguity in identifying illegality, and the limits of transparency and accountability artefacts in representing Trusted Flagger labour and role.

and the fragility of comparative metrics without shared standards. Additionally, human resources and capacity was discussed, as illustrated by organization 5 “we rely on our existing resources which are not enormous” and P1: “One person is able to submit maybe 20 reports per hour and then gets tired”. Prior research on collaborative disinformation detection shows that explanation design can shape human–AI verification workflows, providing a “tooling” solution for TFs [647], as embodied in this quote from P1 about the use of their in-house tool:

“So I submitted 884 reports, [...] reported URLs almost 15,000, take down rate 99 percent, 30 reports are still waiting for the decision. Average take down time. [...], so we are calculating how many reports we submit per week. So you can see our record is 1651 reports URLs per week and 89 reports per week. [...] These are the average take down times from the reports [...] So you see that in week 24 it was 332 hours.”

Another core topic of discussion was cross-member-state variability in accreditation and oversight practice (e.g., differing requirements for demonstrable legal expertise). There was a call for more systematic on-boarding by regulators – such as maintaining contact points and reporting-channel information for platforms – to reduce duplication of effort when TFs are newly designated (see Sections 12.5.1 and 12.5.5). Similarly. Finally, when asked about deeper cooperation with platforms on systemic risk mitigation

(Art. 35(1)(g)), participants reported little evidence of proactive platform engagement; there was discussion of practical coordination among TFs, for example by sharing templates, metrics, and “lessons learned”, as well as concerns shared (Organization 5: “[...] when we flag something to them and we say we are flagging this because it’s illegal and in addition it’s also breaches your policy, your community guidelines. They only reply that yeah, [...] We removed it because it breaches our community guidelines and not because it’s illegal.” (see Sections 12.5.3 and 12.5.6).

12.6. DISCUSSION

Here we discuss how our engagement with Trusted Flaggers (TFs) links to outstanding challenges in both the emergence of DSA infrastructure and user support more generally.

The opaque nature of large platforms and providers A common narrative among our participants was the large scale of content that they were seeking to manage. If the reporting pathways on one platform are cumbersome, it absorbs attention because Trusted Flaggers have limited resources – each platform then ‘benefits’ mutually when each other individual platform is difficult to report through, in terms of reducing the burden placed on them to remove content. From our interviews, it does not appear to be any clearer what is happening within the larger platforms and hosting providers, as seen in the discussion of the Statements of Reasons (SORs). One concern this creates is, if it is not clear whether reporting patterns are feeding back into algorithms, will it be clear how reporting is resulting in identifiable change on the platforms? Future work would consider how to bolster TF resolve, especially when considering that these organizations do not receive any additional funding for the potentially increased amount of reporting effort (and staffing) involved. Some participants spoke of taking on and training staff, but this is while there are also headlines of large platforms cutting their moderation teams; increases in reporting are outsourced but not resourced.

DSA resourcing and empowering user rights Our findings suggest that Trusted Flaggers are positioned as a kind of ‘advanced’ crowd-sourced reporter; they advance the causes they already supported and align therefore with their accreditation mandate (Art. 22 (2) DSA). The TF role is a mix of new capabilities to keep up with growing online content, and coordinated processes which – when working well – mean a fast-track reporting process with the platforms hosting content. These capabilities help TFs to ‘run to stand still’, but with no additional resources. Referring to the DSA and the remit to recognize user rights to report illegal content, the TF apparatus does not enable any sort of ‘reporter mobility’ where new causes emerge and are empowered by a TF role. Given the requirement for skills and time reported by our interviewees, some causes will gain TF representatives, and some will remain left behind as the upfront costs remain high and uncompensated. We expect the latter to include under-represented communities [341] as before the existence of TFs, given that there is no additional resourcing.

The Trusted Flagger role and signaling legitimacy of user concerns Of great concern is whether under-represented causes will remain under-represented with the DSA, but

now with a narrative that the DSA has empowered them (as if ‘job done’ and move to the next concern, meaning users may be forgotten, as a potential unintended harm [131] of the DSA). That TFs do not have additional resources makes a kind of sense within their remit and mandate, but requires clarity in communication as to what the DSA – and specifically, the TF role – is and is not for wider society. For instance, the DSA will seemingly not create any capacity to reach beyond a TF mandate to find marginalized user groups (i.e., those who experience “pervasive negative treatment or exclusion” [503]). The DSA will not suddenly, by its own existence, undo that ‘exclusion’ in the specific sense of bringing onboard targeted reporting skills and a centralizing community hub such as a TF organization. That intent would still need to grow on its own, and we did not see that the TF role ‘freed up’ energy, as registration and continued activity used up resources but in different ways, rather than making reporting more accessible.

There is essentially a cap on TF capability, throttled by resources (of which there is nothing additional coming for TFs), so individual capability must grow somehow, or outreach capabilities under the DSA must grow. This relates to comments by P4 during their interview, that “I know some organization even don’t apply because they know that there won’t be any financial compensation.”

Trusted Flaggers captured in DSA Transparency Mechanisms As described in Section 12.3, VLOPs and VLOSEs must assess systemic risks, including the dissemination of illegal content, and adopt mitigation measures such as adapting content-moderation processes (Art. 35(1)(c) DSA) or cooperating with TFs (Art. 35(1)(g) DSA) [100, 584]. Platform actions are also reflected in transparency mechanisms such as the Statement of Reasons database, which has been criticized for representing platform practices inaccurately or unevenly [206, 349, 711]. Future work should therefore examine interactions among TF reports, systemic-risk assessment reports, transparency reports, independent audit reports (Art. 37 DSA), and the Statement of Reasons database.

12.6.1. RECOMMENDATIONS

Here we provide initial recommendations based on the outcomes of our interviews and workshop with Trusted Flaggers.

Standardized reporting processes. Participants reported it being a drain to have to learn the ins and outs of sometimes very different reporting paths on different platforms. While this is not standardized, it also risks eating at TF resolve, if it continues to be unclear *why* the reporting paths are different. This could turn out to be platforms also learning how best to implement DSA requirements, such as Art. 22(8) for guidelines and Art. 44(1)(a-c) for standards for reporting, design of submitting notices for users and TFs, where these guidelines and standards do not exist at the moment. However, it was not clear at this stage, whether platforms were indifferent to reporting being difficult for TFs to do.

Such guidelines for example could also be provided by national regulators, as exemplified by Ireland’s DSC [549]. This need for standardization is also echoed in Art. 44(1)(c) DSA, providing the Commission with the regulatory power to set standards for “electronic submission of notices by trusted flaggers under Article 22, including through application programming interfaces.” Calls for standardized reporting and oversight channels align

with proposals for “legal compliance APIs” as a scalable enforcement interface [293].

Meeting places for reporter organizations. The workshop activity we conducted was explicitly requested by participants, and was well-attended. Trusted Flaggers would need such events to happen, if opportunities are needed to learn from each other and reduce up-skilling and reporting costs. Supporting TFs through national DSC-led events, cross-national through the Board (Art. 61), as well as better connection to the Commission, could show better coordination, sharing of best practices and more streamlined enforcement efforts for regulators on all levels. If there were no further effort by the wider enforcement apparatus, it may be that academic researchers can provide the meeting place (which would relate also to the need for objective measurement and audit, which is of academic interest too).

Do not lose users in keeping to mandates. Trusted Flagger participants spoke about their particular causes, and keeping to mandates defined around those causes. Further work should seek to understand the extent to which there are different kinds of TF, whether it be one more like a policing role keeping an eye on online content, or a support role to help users find support in a community of like-minded reporting entities. In turn, future work will explore whether the mechanisms of the DSA make individual reporting actions any easier for the unfamiliar and unskilled user. Such work will also leverage on the TF reports to further investigate how TFs document their work and interaction with platforms. Related to this is, if the mechanisms of the DSA support reduction of exclusion, but do not act to increase inclusion, then attention should be given to standardizing the reporting process; it may also be necessary to provide public guidance to users on what harmful content is, but also what is within their rights to do about it.

Ideally this would include knowing how to find an organization which can support them in reporting content. If regulatory capacity to support users impacted by harmful content does not increase, we may see other strategies emerge from the ground up, such as ‘counterspeech’ operations [132]. This, however, would vastly increase the scale of content that must be navigated online.

Act on the lack of knowledge about impacted users. Recent research on at-risk user groups in privacy and security asks what we know about such groups, but also what we do not know [755]. If the TF role stays within a set mandate, then it is perhaps on researchers and general advocacy activity, to proactively understand which users or user groups we *do not* know about, who want to participate in online discourse safely but do not (and do not know about reporting structures). From our analysis, any expectation of ‘proactive’ activity relied on the spirit of the represented TF organization. We highlight this simply as the amount of content appeared daunting for our TF cohort, and we can only expect that it would be overwhelming for an individual user. This raises a question, as to whether every user should expect to be able to find a TF for their needs (be it specific or general), or may already get a signal of sorts from the ecosystem that their cause is not important, if no TF can be found which represents it (then already undermining their willingness to report, constituting a compound unintended harm [131]). This phenomenon has been examined previously as the ‘likeness ideal’, and related harms such as ‘exclusionary

representativeness' [126]. Any 'emerging' cause will need to develop TF capabilities quickly, and already know that they will be able to enact the role when registering with the DSC/Commission. A further question is whether an outreach and awareness capability is needed to reach under-represented users, for instance in a similar model to spyware clinics [342] and privacy consultations for local communities [614]; a question would be how to move this clinical outreach model online and have it connect users to TFs via the same platforms the content is on (as is the current model).

12.7. CONCLUSION

In conclusion, our findings show that Trusted Flagger status increase the speed and likelihood of platform action, but the new role comes with the need for standardized processes, tooling, and costs, is constrained by fragmented reporting channels, and limited feedback from regulators and platforms. These operational frictions, coupled with insufficient activity capturing of TF's reporting in transparency artifacts (SOR) only insufficiently reflect lived practice, and create an accountability gap for evaluating TF work, platform responses, and its implications for user rights. Strengthening these TF mechanism therefore requires tailored standards for reporting illegal content, better platform cooperation, and a strategy that responds to the lack of additional resources to support the range of TF activities.

13

CLICK, BRANCH, AUDIT: A DECISION-TREE TOOLKIT FOR ASSESSING FUNDAMENTAL RIGHTS IMPACTS UNDER THE DSA OF VERY LARGE PLATFORMS & SEARCH ENGINES

The EU Digital Services Act (DSA) introduces an accountability architecture for Very Large Online Platforms and Search Engines (VLOPs/VLOSEs), requiring annual systemic-risk assessments and mitigation measures (Arts. 34–35) as well as independent audits (Art. 37). A central obligation is to identify and assess systemic risks that entail actual or foreseeable negative effects on the exercise of fundamental rights (Art. 34(1)(b) DSA). However, common practice-oriented tools for translating these legal duties into concrete audit procedures and tests remain limited. This paper addresses that gap through a decision-tree risk-assessment toolkit. It guides assessors through (i) contextual information, (ii) subsystem identification, (iii) stakeholder identification, (iv) risk identification and assessment, and (v) traceable reporting aligned with the DSA's Delegated Regulation on audits. To avoid reifying static “vulnerable groups”, we conceptualize vulnerability as situational and relational and use it to make distributional impacts and materiality judgements explicit. We develop the toolkit through design-science methods and evaluate it with experts in a focus-group setting. The contribution is an interdisciplinary and practically usable pathway for evidencing systemic risks under the DSA.

13.1. INTRODUCTION

Very Large Online Platforms (VLOPs) such as Meta, X, TikTok, Google, Amazon and others [145] have become quasi-governmental actors. VLOPs not only wield enormous economic power, but also heavily influence fundamental rights [68]. The European Union (EU) adopted the Digital Services Act (DSA) [224] with the aim of creating safer digital spaces in which fundamental rights are protected.¹ The DSA imposes additional obligations on designated VLOPs and Very Large Online Search Engines (VLOSEs), which meet or exceed the threshold of 45 million average monthly active recipients in the EU (Art. 33 DSA) [224]. VLOPs have become central infrastructures for communication with friends and family [58], trade [598], information retrieval [3], and political [426] and civic participation in Europe [408]. The EU has positioned itself as a regulatory pioneer in explicitly requiring providers to assess and mitigate systemic risks [316], including negative impacts on the exercise of fundamental rights [528]. Yet, translating these obligations into evaluative practice remains methodologically under-specified—particularly when it comes to capturing platform-driven, foreseeable harm as experienced by recipients of the service.

Design decisions of these platforms directly impact individuals [533, 632, 714], societal norms and democratic processes [554]. On an individual level the design of VLOPs can introduce significant and varied risks to users,² such as reduced self-satisfaction and mental and physical well-being due to harmful algorithms [533, 632] and even an increase in suicide rates [714]. The greater the power, "... the greater responsibility to act fairly and protect the vulnerable party" including digital aspects of user's expression [492]. However, many VLOPs' business models commodify personal data [81], creating tensions with the EU Charter [176] and regimes such as the GDPR [4, 493].

VLOPs must identify, assess, and mitigate systemic risks (Arts. 34–35 DSA), including actual or foreseeable negative effects on the exercise of fundamental rights (Art. 34(1)(b) DSA) [224]. Under Art. 37 DSA, independent auditors must evaluate compliance with these obligations. Developing the expertise needed for complex, multi-actor audits is particularly difficult while the DSA remains a comparatively new framework with limited case law and continuing interpretive uncertainty.³ Failure to refrain from deceptive design (Art. 25(1) DSA)—misleading 'verified account' signals—and shortcomings in the advertising repository and researcher data access (Arts. 39 and 40(12) DSA) have already led to a €120 million fine for X (formerly Twitter) [163].

Susceptibility to harm may place any user or other system stakeholder in a vulnerable position. Relevant indicators include perceived pressure toward prolonged use [172], time spent [479], restlessness [381], negative feelings when unable to use social media [712],

Supplementary evaluation protocols, decision-tree materials, and the walkthrough documentation for this chapter appear in Appendix I.

¹See Art. 1(1) DSA: "[...] setting out harmonised rules for a safe, predictable and trusted online environment that facilitates innovation and in which fundamental rights enshrined in the Charter, including the principle of consumer protection, are effectively protected."

²See e.g. the definition for systemic risks Art. 34(1)(d) DSA "[...] serious negative consequences to the person's physical and mental well-being."

³The DSA has applied in full since February 2024.

difficulty limiting use [620], sleep deprivation [39], disruption of social life [781], distorted perceptions of use [588], and fear of missing out [567]. These can signal vulnerability associated with social-media use [351]. Harms can also result from malicious conduct directed at other users, including doxing [678], malicious reporting [779], and trolling [773]. This article focuses instead on actual or foreseeable negative effects on the exercise of fundamental rights, particularly for system stakeholders in vulnerable positions.

Existing work has not yet sufficiently operationalized how actual or foreseeable fundamental rights impacts, particularly for users in vulnerable positions, can be systematically identified, evidenced, and assessed in socio-technical evaluations (like DSA systemic risk assessments and independent audits) beyond a primary focus on user-introduced harms.

To address this gap without reifying static “vulnerable groups”, we conceptualize vulnerability as situational and dynamic. The toolkit captures (i) contextual information, (ii) subsystem identification, (iii) stakeholder identification, (iv) risk identification and assessment, and (v) reporting on negative effects on fundamental rights.⁴ It can help providers and auditors develop a shared, evidence-based understanding of compliance with the Delegated Regulation’s audit requirements [226, 661].

This paper answers the core research question: *How can a structured risk-assessment methodology, operationalized as a decision-tree toolkit, improve the identification, assessment, and documentation of adverse effects on fundamental rights under Art. 34(1)(b) DSA for VLOPs and VLOSEs?*

To answer the main research question, we developed the following sub-research questions:

Our approach bridges legal interpretation and socio-technical evaluation, offering a methodological pathway for assessing structural harms, vulnerable positions [557, 775] as central concerns in FAccT research. Based on interdisciplinary work in computer science [136], including algorithmic auditing and accountability [137, 692, 730], human–computer interaction (HCI) [279], and legal compliance this toolkit enables users to identify [88], analyze, and assess the systemic risk of negative effects on fundamental rights. Building on related work that distinguishes between *risk-uncovering* and *risk-measuring* studies in algorithmic auditing [584], we structure our assessment pipeline accordingly.

We draw inspiration from existing toolkits for algorithmic fairness and equity [187, 442], AI Act [339], and fundamental rights impact assessment [87, 88]. The paper’s toolkit is organized in two phases. First, we conduct an exploratory risk-uncovering phase to surface gauge systemic risks. Second, within the risk-measuring phase, we operationalizes these signals into assessable indicators to support a defensible assessment of the severity, prevalence, and distribution of effects, including differential impacts on users in vulnerable positions.

13.2. BACKGROUND AND RELATED WORK

Recent scholarship highlights how VLOPs present distinct risks to democratic processes, vulnerable groups, and fundamental rights [444, 639, 692, 742]. Prior work has examined

⁴This operationalization follows the DSA’s accountability architecture: VLOPs and VLOSEs must conduct systemic-risk assessments (Arts. 34–35) and undergo independent audits (Art. 37).

the scalability challenges of content moderation [290], the limitations of transparency reporting [288], and the potential of algorithmic audits for ensuring accountability [692]. While audit practices for AI systems are increasingly formalized [197, 497, 608], the literature on DSA compliance remains underdeveloped. In particular, methodological frameworks for assessing *systemic risks* under Arts. 34 and 35 DSA are missing.

13.2.1. SYSTEMIC RISK ASSESSMENT UNDER THE DSA

The DSA creates a novel auditing structure, primarily shaped through the introduction of two auditing circles – an internal (VLOP/VLOSE’s systemic risk assessment) under Art. 34, and an external (auditor’s independent audits) under Art. 37 DSA. Together, internal and external audits create a continuous accountability loop, but without shared standards their credibility depends on consistently translating audit tests into transparent, evidence-based methods for substantiating rights-relevant risks and mitigations. Auditing under the DSA is resource-intensive and difficult to scale, which makes scalable approaches essential for meaningful oversight [662, 692].

These two (internal and external) auditing obligations are further regulated in the DR [226], which provides details on methodologies, tests, evidence and risk for platforms and auditors [661]. Both audit obligations have annual reports as a final output (Arts. 34(1) and 37 (4) DSA, Art. 6 DR) and the two auditing instruments assess risk identification and mitigation, including on fundamental rights.

Operationalization of such methodologies in line with *audit procedures*⁵ and selected *tests*⁶ is the next step necessary for meaningful compliance and assessment of such under the regulation.

The DSA grounds Art. 34(1)(b) obligations in the Charter of Fundamental Rights (the Charter) [224, 248]. According to Recital 94 systemic risk assessments should follow a “case-by-case basis” analyzing “in-depth” (See also Recital 80). It leaves open how to conduct risk assessments, which rights to prioritize, and which indicators can credibly operationalize harm in socio-technical systems (Recital 81). Fundamental rights may be directly linked to the scope of the DSA.⁷ For VLOPs, these rights imply obligations to protect users’ ability to communicate, access information, and participate without undue interference or harm, as operationalized in Art. 34(1)(b), shifting EU digital governance from illegal content to the rights impacts of design, recommender systems, and moderation at scale.

An additional complexity stems from the different test structures associated with fundamental rights. Following Malgieri and Santos [494], fundamental-rights risk assessment should distinguish *violations* from *interferences*: while violations are typically treated as

⁵Art. 2(16) DR which “means any technique applied by the auditing organisation in the performance of the audit, including data collection, the choice and application of methodologies, such as tests and substantive analytical procedures, and any other action taken to collect and analyse information to collect audit evidence and formulate audit conclusions, not including the issuing of an audit opinion or of the audit report.”

⁶Art. 2(17) DR defining “[...] an audit methodology consisting in measurements, experiments or other checks, including algorithmic systems, through which the auditing organisation assesses the audited provider’s compliance with the audited obligation or commitment.”

⁷Examples include political rights—such as freedom of thought, freedom of assembly and association, and the right to vote—as well as rights related to copyright and business, such as the right to property, freedom of the arts and sciences, and the freedom to conduct a business. Table 13.7 presents these examples in a structured form.

binary, interferences lie on a continuum across socio-technical interaction. The severity of such interferences depends on contextual factors, including social norms (e.g. community standards of platforms), legislative interpretations (definition of illegal content), and how individuals or groups perceive the impact. As established by the Court of Justice of the European Union (CJEU) in several cases⁸ the “seriousness” of an interference is variable and measurable. This positions severity as a central concept for proportionality analysis under Art. 52 Charter and for statutory obligations that require private entities to assess necessity and proportionality, such as Art. 35(7)(b) GDPR. Accordingly, what must be measured in practice is not the violation itself, often a legal conclusion, but the *degree of interference* with fundamental rights. This conceptualization informs our approach to systemic risk assessments under the DSA, where understanding and quantifying the severity of interferences is essential for evaluating design choices, algorithmic effects, and mitigation measures. Interpreting these rights in socio-technical contexts demands interdisciplinary approaches. For example understanding freedom of expression entails studying content moderation; privacy relates to data infrastructures; and equality requires attention to bias studies. These systemic risk assessments shift fundamental-rights protection from a state-individual model to a multi-actor setting where VLOPs mediate rights at scale, regulators impose duties on private actors, and users experience interferences through platform design, ranking, and enforcement.

In line with Malgieri and Santos [494], assessing risks to fundamental rights requires a clear conceptual alignment between infringements, interferences, violations, and harms. We adopt a normative lens that locates risks to fundamental rights in the potential *infringement of positive legal rules*, that is, non-compliance with statutory provisions or jurisprudential standards that safeguard dignity, autonomy, and other Charter-protected values. Building on this, we distinguish between (i) *infringements* of rules implementing fundamental rights, (ii) *interferences*, which cover a continuum from negligible effects to serious encroachments, and (iii) *violations*, which represent clear breaches of fundamental rights as defined in law. Finally, we differentiate these from *harms* or adverse effects, which refer to measurable physical, psychological, economic, or social consequences. To bridge risk-based and rights-based approaches, we apply a three-tiered assessment of severity. First, the *objective* dimension evaluates the normative seriousness of an infringement, considering the rules violated, its scope, duration, and reversibility. Second, the *subjective and inter-subjective* dimension assesses how individuals, groups, and society perceive the severity of the interference. Third, the *real-life consequences* dimension examines tangible impacts on affected persons or communities, including changes in well-being, financial loss, scale of impact, and the reversibility of those effects. Together, these tiers enable a granular and legally grounded assessment of severity that is essential for evaluating systemic risks under the DSA.

Besides, our toolkit is explicitly designed to satisfy the DR’s core requirements for auditing systemic risks under Art. 34(1)(b) DSA by translating them into a concrete *audit procedure* (Art. 16 DR) with a clearly delimited *audit scope* aligned with Art. 35 DSA (e.g., content moderation systems, recommender systems, advertising delivery, and related data practices) [226].

⁸Such as *Digital Rights Ireland*, *Google Spain*, *Tele2 Sverige*, *Ministerio Fiscal*, and *La Quadrature du Net*, as well as by the ECtHR in *Zakharov*.

Methodologically, it combines *quantitative and qualitative* components (Art. 10 DR), linking rights-risk hypotheses to measurable indicators while retaining contextual interpretation of rights interferences. To support auditable conclusions, the toolkit operationalizes *materiality* by requiring explicit thresholds per risk category — and, where relevant, per affected users in vulnerable positions (Art. 10 DR)—so distributional impacts are made visible, not incidental. In line with the DR’s sequencing logic, identifying the groups impacted within situations that produce vulnerability precedes defining the audit scope per risk category (Art. 2 DR), e.g., specifying risks to LGBTQ+ politicians during electoral campaigns before selecting the subsystems and governance levers to examine [444]. Finally, the toolkit links sampling choices to these scope decisions by requiring that samples are demonstrably adequate for the affected populations and the specific risk tested (Art. 12(12)(f) DR).

13.2.2. VULNERABILITY AS A HEIGHTENED RISK TO FUNDAMENTAL RIGHTS

Vulnerability is a core concept in understanding negative effects on fundamental rights and a measure explicitly asked for in tests under the DR for sampling methodologies in Art. 12(f) DR.⁹

Building on Rebrean and Malgieri [613], we define vulnerability as heightened fundamental-rights risk arising from power-imbalanced dependencies and shaped by personal and structural resilience—not an inherent deficit, but a cumulative condition and governance lens indicating where support or empowerment is needed. This aligns with HCI work treating vulnerability as a design and assessment lens [115, 599].

First, the notion of vulnerability as the *higher risk to rights*, which expands the Commission’s notion of vulnerability as susceptibility to behavioral or decisional influence is broadly consistent with the DSA’s focus on systemic risks and harms, including protections for individuals in “vulnerable positions” (e.g., Rec. 3, 9). Second, the component of power-imbalanced dependencies can be seen through potentially harmful dependency between users and services (e.g. negative effects on freedom of expression), but also through e.g. negative effects on privacy through the lack of data protection [495]. Third, the DSA relies on both structural and personal resilience, since platform mechanisms and users’ active engagement jointly enable processes such as user reporting [668].

Consequently, and aligning with other scholarship (e.g., [351], the experience of vulnerability in digital environments is contextual and situational, rather than forming on the basis of individual traits alone, as e.g. outlined in anti-discrimination law [10, 11]. Additionally, the view of “users in a vulnerable position” connects with Value Sensitive Design’s conception of stakeholders as contextual roles rather than fixed entities [276], enabling analysis of dynamic user–system relations and showing how different roles can surface different risks. For Art. 34(1)(b) DSA, these perspectives support treating “vulnerable positions” as relational and situational: assessments should identify platform-mediated dependencies and contextual factors that heighten rights risks, and justify mitigation in terms of both harm reduction and user protection, support, and empowerment.

⁹ “[T]he representation and appropriate analysis of concerns related to particular groups as appropriate, such as minors or vulnerable groups and minorities, in relation to the audited obligation or commitment.”

13.2.3. RISK ASSESSMENT FRAMEWORKS AND TOOLKITS

Many EU governance frameworks (e.g., the GDPR, the AI Act, and the Cyber Resilience Act) require structured risk assessment processes. To enforce the right to the protection of personal data, the GDPR introduced in Art. 35 GDPR a formalised assessment mechanism that is frequently framed as a form of Fundamental Rights Impact Assessment (FRIA) [5, 699]. The AI Act [236] similarly adopts a risk-based approach to ensure the development and deployment of trustworthy AI in line with fundamental rights and European values [221, 279, 719]. Finally, the EU Cyber Resilience Act (CRA) introduces phased cybersecurity duties across the product lifecycle (including vulnerability handling, conformity assessment for certain products, and incident reporting) [236]. It anticipates overlap with the AI Act where products qualify as high-risk AI systems: CRA risk assessment should consider AI-specific vulnerabilities and, where relevant, fundamental-rights risks, while Art. 27 AI Act requires a deployment-centred Fundamental Rights Impact Assessment. Together, these instruments establish a complementary logic: cyber-resilient products as a baseline, and context-specific FRIA at deployment to evidence (and mitigate) foreseeable rights-relevant harms.

Large systems (such as VLOPs) are typically not singular systems but *systems-of-systems* (SoS): collections of multiple subsystems [679], each posing distinct risks to stakeholders' fundamental rights. With SoS complexity and non-deterministic behaviour, risk assessment approaches have shifted [679]. To manage complexity, it can be advisable to assess risks at the subsystem level first, noting that risks compound at higher levels when subsystem risks interact [274]. In parallel, design and engineering traditions have moved from preventive to proactive, iterative approaches to risk management, including Boehm's spiral model [85], Privacy by Design [120], Value-Sensitive Design [276], and Participatory Design [545], each prescribing forms of assessment throughout system development.

Toolkits bundle practical instruments (e.g., templates, worksheets, flowcharts, and questionnaires) to support the application of regulatory frameworks and to facilitate interdisciplinary cooperation [187]. Fairness toolkits typically operationalise fairness through ready-to-use metrics to quantify undesirable bias in data or models, often alongside algorithmic techniques for measurement and mitigation [187].

Deng et al. (2022) identify three requirements for fairness toolkit design: (i) interfaces should align with comparable tools and provide guidance and training materials; (ii) toolkits should include interactive, deliberation-oriented activities to stimulate critical reflection and interdisciplinary discussion; and (iii) toolkits should provide context-sensitive alerts and checklists to support task completion and to prompt critical conversations among stakeholders [187].

Several recent toolkits operationalise legal requirements through decision-tree or template-based structures. Hanif et al. [339] propose a decision-tree framework to classify AI systems into AI Act risk categories¹⁰. Decision trees represent decisions, paths, and outcomes: internal nodes encode decisions, branches encode resulting paths, and leaves encode final outcomes. Their interpretability can support classification by non-specialist practitioners [339], including cases where one system may implicate multiple categories [339, 374].

¹⁰These risk categories are: unacceptable, high, limited and minimal.

Bogucka et al. (2024) [87] propose a standardised template to facilitate risk assessment aligned with the EU AI Act, NIST's AI RMF, and ISO 42001. The toolkit follows a five-step process: (1) system context (intended use, purpose, capabilities/components, implementation context, and stakeholders); (2) potential risks (grouped across technical components, human interaction/experience, and systemic impacts on society, economy, and environment); (3) mitigation strategies; (4) anticipated benefits and positive impacts; and (5) reporting mechanisms and governance responsibilities [87].

Similarly, the Algorithmic Equity Toolkit consists of three components [442]. A flowchart (decision tree) supports scoping—identifying whether a technology constitutes an automated decision system. A questionnaire then structures interrogation of algorithmic harms and bias dimensions, and a system map/worksheet supports disentangling intended purposes from misuse and clarifying how technical terms and components combine into an automated system [442].

Finally, Rao et al. (2025) [609] present RiskRAG, developed through (1) literature research and co-design to derive design requirements, (2) implementation based on those requirements, and (3) evaluation through user studies. Key requirements include real-world use cases and examples, recommended mitigation strategies, clear structuring of risks by impact or significance, and explicit links to use cases to support a contextualised understanding of relevance and impact.

13.3. METHODOLOGY

To develop our toolkit, we used a mixed-methods design-science approach to examine practical challenges systematically and create artefacts to address them [399]. Following the first step used by Rao et al. (2025) [609] to develop their toolkit, we conducted a literature review to identify the most important design requirements. Based on the resulting requirements, we developed an initial proposal for our toolkit. We then iteratively refined this proposal using findings from a design-thinking workshop [102], in which two domain experts applied the “rose, thorn, and bud” method [488]. The experts have backgrounds in the social sciences, sociology, social law, and digital law, and between 3 and 12 years of professional experience. Finally, we developed the toolkit and evaluated it through a walkthrough (see Sections 13.3.1 and I.2; cf. [609]).

13.3.1. AUTHOR-LED STRUCTURED USE CASE WALKTHROUGH

To evaluate the toolkit in the absence of a user study, we conducted an author-led structured walkthrough using the fictional vignette “Undressed Without Consent” as a worked case 13.3.1. The walkthrough serves as an analytic evaluation of (i) coverage (whether the tree supports end-to-end assessment across context, subsystem, stakeholder, risk and reporting steps), (ii) traceability (whether the tree yields a reviewable record of claims, evidence, and rationale), and (iii) failure modes (where the tree under-specifies decisions or requires information that is typically unavailable). To reduce interpretive degrees of freedom, we defined an evaluation protocol prior to application: we fixed a “facts-as-given” sheet, recorded uncertainty explicitly, and treated nodes requiring unavailable evidence as “insufficient evidence,” rather than inferring (See Annex I.1). We then went through the toolkit step-by-step, logging each decision node, branch selection, and intermediate

output (subsystem mapping, rights mapping, vulnerability situations, risk statements, mitigation options).

We created the vignette “Undressed Without Consent” using a scenario-based legal synthesis that combines (i) key elements from two intermediary-liability/content-moderation cases and (ii) recent reporting on generative-AI “undressing” tools. From *Glawischnig-Piesczek v Facebook Ireland* (C-18/18) [17], we took the enforceability challenge of removal orders that can extend beyond identical uploads to “equivalent” variants—directly relevant to re-uploads and near-duplicates. Furthermore, we use a case of “variants” and how they can look like building on another leading case (*LG Frankfurt am Main*, 2-03 O 188/21 [2]), which treats practical evasion tactics (e.g., altered layouts, added/removed text, typos, even minor pixel changes) as potentially still covered by a post-notice removal/locking obligation. Finally, we build on an early-January 2026 reporting on Grok’s (of the VLOP ‘X’) image tool being used for non-consensual sexualized edits and the subsequent restriction of in-reply image editing to paid users, which critics argued to be a major harm, especially for women and girls [26, 27, 107, 265, 283, 532]. The resulting use case is fictionalized but retains these legally salient dynamics, reporting/complaint escalation by an Austrian parliamentarian, recurrence via near-duplicates of AI generated deepfake nudes, and mitigation measures that were argued in the case of X of offering the generative AI tool only to paying users, to enable structured evaluation of the decision tree toolkit against DSA-relevant compliance questions.

Use Case Description ‘Undressed Without Consent’

A social platform operating in the EU launches an in-product AI image editor integrated into replies, alongside a standalone “creator” app. Within days, users begin generating and reposting non-consensual sexualised edits of real people’s photographs at scale. The platform responds by restricting image editing to paid subscribers, arguing that this reduces abuse; critics contend that it shifts, rather than resolves, the risk. Regulators and civil-society organisations demand rapid containment and transparency, while the platform struggles with copies, reuploads, and near-duplicates that are not byte-identical. One affected person, a member of the Austrian Parliament, asks the platform to remove these and similar images.

13.4. FINDINGS

The literature review shows that a decision tree-based toolkit is particularly suitable for non-lawyers due to its high level of comprehensibility (use of non-specialist language) and interpretability, especially in situations where a system (or subsystem) falls under several risk categories. [339, 374, 442]. With regard to the toolkit, other authors have pointed out that the user interface should be similar to comparable tools, should include context-sensitive alerts targeting common pitfalls and that guidelines and training materials can support practitioner [187]. Furthermore, it seems to be advisable to include checklists and questionnaire based interrogations. (Cf. [187, 442])

As a result of the ‘rose, thorn and bud’ method, we received 52 comments, of which 11.76% were requests for examples to improve understandability, 17.65% were references

to limitations, 25.49% were comments that the toolkit is good in its current form, and 45.10% were suggestions for improvement. While the requested examples and most of the suggested improvements have been incorporated directly into the reworked version of the toolkit described in this section, the limitations are addressed in the ‘Limitations and Future Work’ section 13.5.

Building on the step-by-step process described by Bogucka et al. (2024) [87], we have divided the toolkit into five different steps: 1) context information, 2) subsystem identification, 3) stakeholder identification, 4) risk identification and assessment, and 5) reporting. These steps are described in more detail below.

As a result of the redesign following the design thinking workshop, we provide a summary after each subsection and have gathered some background information on who uses the toolkit, such as compliance officers, researchers or developers.

Besides, here, we explain the approach by walking through the worked case (see Section 13.3.1) step-by-step, using the toolkit outputs to illustrate how scoping decisions, subsystem and stakeholder mapping, and risk articulation translate into a documentation-ready risk record. To keep the main text focused, we report only the minimal inputs and outputs required, and refer readers to the Annex section I.2 for the full walkthrough protocol and the complete worked-case sheets used in the evaluation.

13.4.1. DSA APPLICABILITY AND CONTEXT INFORMATION

In this step, it is first determined whether the system under investigation falls within the scope of the toolkit as proposed by Krafft et al. [442]. In our case, this means checking whether the DSA applies or is likely to apply in the foreseeable future. The decision tree first asks whether the service is a VLOP or VLOSE within the meaning of Art. 33 DSA [224]. Once applicability has been determined, open-ended questions gather contextual information about the system’s purpose, provider, domain, and core technical capabilities [87]. In line with the workshop findings, the toolkit provides examples of how to define system purpose. This information is a starting point for assessing the social, technical, cultural, and other structural conditions that shape system use [613] and for understanding stakeholders in vulnerable positions. Because services within the toolkit’s scope are treated as systems of systems, the next step examines their constituent subsystems.

We establish that the worked case is within scope of the toolkit by fixing (i) DSA applicability assumptions (VLOP threshold) and (ii) the service classification as a very large online platform, before anchoring the assessment to the concrete feature context (in-reply AI image editor plus companion creator app).

Table 13.1: Minimal inputs for the worked case (full details in Annex B.1–B.2).

Annex pointer	Minimal input referenced in §4.1
Annex B.1 (DSA applicability)	Treated as a VLOP (Art. 33 threshold assumed) and classified as an online platform.
Annex B.2 (Context)	Feature under assessment: in-reply AI image editor plus companion creator app (EU-facing social platform).

13.4.2. SUBSYSTEM IDENTIFICATION

In order to address the complexity of SOS in risk assessment, a series assessments must be carried out for each subsystem, for which these systems must be identified first [274]. The DSA names several subsystems that should be given special attention, including recommendation systems, content moderation systems, systems for enforcing terms of use, systems for delivering advertising, and their data-related systems.¹¹ In addition, several other common subsystems are shown, grouped into six layers that define their overall purpose. These can be selected in a multiple-choice process, or other subsystems not mentioned can be added via an open-text field. Based on feedback from the workshop, further descriptions were added to these levels and subsystems. The interaction layer – also known as the front end – comprises systems such as web front ends or desktop applications, which are essential for the user experience, displaying information and user input. All other layers are located in the backend. The data layer comprises systems such as database systems or event logging systems and is responsible for data management, while the logical layer comprises recommendation, matching or indexing systems that work with the collected data and are essential for decision-making. The monetization layer includes systems such as advertising, pricing (e.g. for advertising auctions), payment and other systems that are essential for generating revenue. The infrastructure layer includes all systems that are required as a technical basis, such as monitoring and content delivery systems. Finally, the governance level encompasses all systems responsible for enforcing rules, guidelines and standards, including systems such as content moderation and fraud detection systems. Once the relevant subsystems have been selected, a summary of these systems is created and, in the next step, a selection is made as to which of these systems are external services (i.e. not entirely under the control of a company). This was an addition based on suggestions for improvement made during the workshop.

For the worked case, we intentionally keep the main text lightweight and only record the minimal subsystem scope used in the walkthrough (Table 13.3).

Table 13.3: Minimal inputs for the worked case (full subsystem sheet in Annex B.3).

Annex pointer	Minimal input referenced in §4.2
Annex B.3 (Subsystem identification)	The VLOP is treated as a system-of-systems; the walk-through scopes to the subsystems most implicated in generation, dissemination, persistence, and remedy of NCSI: interaction/UI surfaces, data/logging & near-duplicate handling, algorithmic curation, content moderation, policy/ToS enforcement, monetization/access-control (pay-wall), and creator-app tooling.

¹¹ See Art 34(2) DSA (a) the design of their recommender systems and any other relevant algorithmic system; (b) their content moderation systems; (c) the applicable terms and conditions and their enforcement; (d) systems for selecting and presenting advertisements (e) data related practices of the provider.

13.4.3. STAKEHOLDER IDENTIFICATION

According to ISO 15288, stakeholders are all people who have “a legitimate right, share, claim, influence or interest in a system” ([388] p. 10). For an appropriate risk assessment, not only direct stakeholders, i.e., all persons who interact directly with a system, should be taken into account, but also indirect stakeholders who never or only rarely interact with a system but are nevertheless affected by it [276, 387]. Based on feedback from the workshop, clarification of the difference between direct and indirect stakeholders is provided, with examples.

The concept of “users in a vulnerable position” connects with Value Sensitive Design’s treatment of stakeholders as contextual roles rather than fixed categories [276]. We define vulnerability as heightened fundamental-rights risk arising from power-imbalanced dependencies [495]. Article 13(1)(a)(f) DR supports this perspective by requiring sampling methods to represent and analyse concerns affecting particular groups, including minors, vulnerable groups, and minorities. Because fundamental-rights impacts are shaped by both personal and structural resilience, the assessment considers both dimensions for each stakeholder [495]. Where the worked case does not provide evidence about stakeholders’ skills, knowledge, or circumstances, the analysis records its assumptions explicitly.

For the worked case, the main text records only the minimal stakeholder roles needed to follow the decision-tree paths in subsequent steps (Table 13.5).

Table 13.5: Minimal inputs for the worked case (full stakeholder sheet in Annex B.3.1).

Annex pointer	Minimal input referenced in §4.3
Annex B.3.1 (Stakeholder roles + vulnerability lens)	Stakeholders are specified as <i>roles</i> (direct/indirect) per in-scope subsystems; for the walkthrough we retain only the roles most implicated in NCSI generation, dissemination, persistence, and remedy (e.g., creators/uploaders/viewers, reporters/complainants, moderation/policy teams, curation/product teams, and affected/depicted persons incl. non-users), and we flag vulnerable situations via personal/structural resilience where relevant (e.g., targets incl. minors/public officials).

13.4.4. RISK IDENTIFICATION AND ASSESSMENT

For the final step—risk identification and assessment—we operationalize Article 52(1) of the Charter [248]. For a specific stakeholder, we analyse whether a platform subsystem (i) interferes with a fundamental right protected by the Charter and, if so, whether that interference (ii) is provided for by law, (iii) pursues a legitimate aim, and (iv) complies with the principle of proportionality [247].

The first step of this analysis consists of determining whether a subsystem of a platform interference with a right (see: Table 13.7) and limits its exercise, for example through practices such as content moderation.¹²

Table 13.7: Relevant Articles of the EU Charter of Fundamental Rights

Article	Title	Short Description
Art. 1	Human Dignity	Human dignity is inviolable. It must be respected and protected.
Art. 7	Respect for Private and Family Life	Everyone has the right to respect for their private and family life, home and communications.
Art. 8	Protection of Personal Data	Everyone has the right to the protection of personal data concerning them. Such data must be processed fairly, for specified purposes, and based on consent or other legitimate basis laid down by law.
Art. 11	Freedom of Expression and Information	Everyone has the right to freedom of expression. This right includes freedom to hold opinions and to receive and impart information and ideas without interference by public authority. Media freedom and pluralism shall be respected.
Art. 21	Non-discrimination	Any discrimination based on grounds such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.
Art. 24	The Rights of the Child	Children shall have the right to such protection and care as is necessary for their well-being. They may express their views freely, and such views shall be considered in matters which concern them.
Art. 38	Consumer Protection	Union policies shall ensure a high level of consumer protection.

Where an interference is found, the second step is to determine whether it is “provided for by law” within the meaning of Article 52(1) of the Charter. This requires identifying the measure’s normative basis and regulatory framework, including obligations under EU law and the platform’s terms and conditions as implemented pursuant to Article 14 DSA. The applicable rules must be publicly accessible and formulated with sufficient clarity to make their application foreseeable [8].

Third, the analysis must examine whether the measure pursues an objective of general interest recognised by the EU or protects the rights and freedoms of others. In a risk assessment, each mitigation measure should be linked to that objective and supported by a clear explanation of how it contributes to the aim—for example, by mitigating the dissemination of illegal content, safeguarding users’ rights, or protecting minors. If the stated reason cannot justify restricting a Charter right, the measure does not satisfy Article 52(1) and cannot be treated as a legitimate limitation of that right.

The fourth and final step is to verify that the measure complies with the principle of

¹²I.e. as listed in Art. 17(1)(a)-(d): restrictions of the visibility (removal of content, disabling access to content, or demoting content), suspension, termination or other restriction of monetary payments, of the service, or of the recipient of the service’s account.

proportionality. Under EU law, this criterion requires a structured assessment of whether a measure exceeds what is appropriate and necessary to achieve its legitimate objective [9]. If more than one suitable option is available, the least restrictive alternative should be preferred, and the adverse effects must not be disproportionate to the aim pursued. The CJEU's case law shows that the level of scrutiny depends on the right concerned and the gravity of the interference. Privacy and data-protection cases, for example, may require strict necessity and effective safeguards [14]. Where several rights or freedoms are engaged—such as freedom of expression and economic interests—the proportionality assessment should explain how the measure reconciles those interests [12].

For the use case, we apply the Art. 52(1) test in the same question order as the annex decision-tree record and document each step outcome (including uncertainty) as a traceable assessment record. We begin by fixing the unit of analysis, selecting the relevant subsystem(s), the affected stakeholder role, and the Charter right; for the freedom-of-expression test, this scopes the assessment to S3 (moderation) and S5 (data/similarity handling), with D2 reposting users as the restricted party and Art. 11 as the right under assessment.¹³

13.4.5. SCORING, WEIGHING AND REPORTING

This subsection explains how we translate decision-tree classifications into a documented risk assessment output that supports DSA systemic risk assessments (Art. 34 DSA) and is compatible with the Delegated Regulation's audit requirements (Art. 37 DSA; DR). For each outcome, the toolkit records the implicated subsystem(s), affected stakeholder roles (including vulnerable situations), relevant Charter rights, and the legal/proportionality rationale. We then separate *risk signal severity* (magnitude of the rights-relevant risk) from *assurance/confidence* (strength of the evidence and controls), and produce a short assessment summary. Severity is characterized via probability and impact (including differential distribution), evidence is mapped to claims and uncertainty is explicitly flagged where inputs are insufficient, and findings are weighed using audit-risk and materiality considerations to support reasonable (not absolute) assurance. These artefacts support audit scoping by making uncertainty explicit and by indicating where additional testing and provider evidence would be required under the DR.

We translate the subsystem and stakeholder mapping into a small set of reviewable risk statements by treating risks as subsystem-mediated pathways to interferences with Charter rights and assigning an initial severity profile (objective seriousness, subjective/experienced severity, and real-life consequences). Where the vignette does not provide Member State legality determinations or internal platform metrics, we avoid definitive illegality claims and instead record uncertainty explicitly (including “insufficient evidence” nodes) to preserve auditability. Table 13.11 therefore reports only the minimal risk framing and outputs needed in the main text.

¹³We intentionally limit the testing to selected parts of the worked case because the decision-tree is designed to be applied *per subsystem* and *per stakeholder-right pairing*, rather than as a single, holistic judgment about the entire platform or incident. The worked case spans multiple sociotechnical stages (generation, distribution, persistence, notice-and-action, mitigation design, and remedies), and each stage is implemented by different subsystems that raise distinct interference mechanisms and evidentiary requirements. Limiting the test scope therefore improves traceability, avoids over-claiming beyond the vignette, and produces actionable outputs for future data collection and audit scoping.

The resulting output of this step is a documentation-ready risk assessment record that (i) preserves the decision-tree path and its legal rationale, (ii) records severity and distributional considerations, (iii) links each conclusion to evidence types consistent with the DR (Art. 13(3) DR; Art. 11 DR), and (iv) explicitly states the level of assurance, audit risks, and materiality considerations (Art. 2(8)–(13) DR; Art. 9 DR). This provides the empirical and procedural grounding required for proportional mitigation discussions and supports consistent, reviewable assessments across assessors and over time. Finally, reporting mechanisms and responsibilities for the governance of use are outline [87].

To keep the text readable and within space constraints, we surface only the identifiers and short labels of the worked-case outputs (Table 13.13). The complete risk-output table—including subsystem links, Charter-rights mapping, and the severity/assurance tags used to support auditability—is reported in Annex B.5 (Table 3).

13.5. LIMITATIONS AND FUTURE WORK

Our contribution is bounded in four ways. First, there is regulatory and interpretive uncertainty about how Art. 34(1)(b) DSA fundamental-rights impacts should be operationalized in practice. As a result, the toolkit necessarily embeds normative assumptions that may change as case law and audit practice develop. Second, the toolkit inherits the abstraction limits of a decision-tree representation: it simplifies continuous, context-dependent socio-technical dynamics and cannot replace legal and technical expertise in hard cases. Third, the walkthrough is evidence-constrained. When a step requires provider-internal artefacts or data, we record *insufficient evidence* rather than inferring, so conclusions may remain provisional without audit-accessible documentation. Fourth, our evaluation relies on an author-led structured walkthrough of a curated set of fictionalized social-platform use cases. This supports analytic assessment but limits empirical claims about real-world uptake and generalizability across VLOP architectures, jurisdictions, and risk domains.

Dividing a VLOP – an SOS – into subsystems and evaluating them separately is a common approach in computer science. While this ensures that risks can be identified in detail and that it is clear where a potential risk mitigation strategy could be applied, some potential risks arising from the interaction of these systems could be overlooked. Furthermore, risks may not originate solely from a subsystem itself, but also from people and processes – socio-technical systems – involved in that system. A more in-depth socio-technical assessment on systemic risks may be valuable, but is very complex and time-consuming and would go beyond the scope of our work. Identifying subsystems and the associated direct and indirect stakeholders of complex VLOPs is anything but easy. A good starting point for this would be a concept of operations (ConOps) document, if available, as it lists the subsystems and their objectives, including a description of the stakeholders involved and a description of the application context of the VLOP under investigation [388, 679].

13.6. DISCUSSION

To answer our core research question, we interpret the results of our author-led walkthrough as analytic evidence about the toolkit's capacity to structure and document Art. 34(1)(b) fundamental-rights risk assessments for VLOPs/VLOSEs, while recognising

the limits of curated cases and non-deployment evaluation.

The author-led structured walkthrough should be understood as an analytic evaluation of the toolkit's accountability value rather than a demonstration of field effectiveness (see Section 13.5). Applied to a curated set of fictionalized use cases, the walkthrough indicates that the decision-tree structure can yield a traceable assessment path—forcing structured assessment (subsystems, stakeholder roles, and situational vulnerability) and producing reporting outputs that align with Art. 34/37 expectations. Because the walkthrough was author-led and based on curated cases, our findings mainly show what the toolkit can cover, how traceable the outputs are, and where the approach breaks down. We do not claim evidence on real-world uptake, or how well the toolkit generalizes across different VLOP designs and risk domains; these questions require further testing, development, and assessment.

Our approach bridges legal interpretation and socio-technical evaluation, offering a methodological pathway for assessing structural harms, vulnerable positions [557, 775] as central concerns in FAccT research. Based on interdisciplinary work in computer science [136], including algorithmic auditing and accountability [137, 692, 730], human-computer interaction (HCI) [279], and legal compliance this toolkit enables users to identify [88], analyze, and assess the systemic risk of negative effects on fundamental rights.

The toolkit's vulnerability lens is intended to avoid reifying a static understanding of “vulnerability”. Instead, it adopts flexible, situational concepts for assessing systemic risks under the DSA. This framing makes distributional impacts and materiality judgements explicit. In practice, this matters for Art. 34(1)(b) DSA because the challenge is not merely to list Charter rights, but to establish who is placed in a vulnerable position by a given subsystem pathway, and with what severity and distribution.

To our knowledge, this decision-tree approach is the first attempt to operationalize a systemic-risk assessment methodology specifically for Art. 34(1)(b) DSA. It translates interdisciplinary assessment practices into concrete, traceable steps. Such structure is needed because platform systemic-risk assessments currently differ in methodology, limiting cross-platform comparison and the identification of good practices by providers, regulators, auditors, researchers, and civil society. The toolkit offers an initial method for decomposing complex legal assessments for interdisciplinary use, including by audit practitioners evaluating VLOPs and VLOSEs. The toolkit remains an initial design. Future work should test it across multiple cases with affected stakeholders and domain experts, then refine it for changing platform features, stakeholder relationships, and societal contexts within complex systems of systems.

13.7. CONCLUSION

We presented a decision-tree toolkit that operationalises the DSA's systemic-risk obligation for negative effects on fundamental rights (Art. 34(1)(b) DSA) into a traceable walkthrough: context information, subsystem and stakeholder identification, risk identification and assessment, scoring, weighing and reporting aligned with the DSA's audit logic. In the worked-case walkthrough, the toolkit produced a compact risk record and made evidentiary gaps explicit via “insufficient evidence” nodes, supporting reviewability rather than over-claiming. Future work should validate use in practitioner settings and

extend coverage across additional platform architectures and risk domains.

Table 13.9: Pointers to annex testing steps and the minimal input referenced in §4.3 for the worked case (Part I: Annex B.3.1–B.3.5).

Annex pointer	Minimal input referenced in §4.3
Annex B.3.1 (Stakeholder roles + vulnerability lens)	Stakeholders are specified as roles (direct/indirect) per in-scope subsystems; for the walkthrough we retain only roles implicated in NCSI generation, dissemination, persistence, and remedy (e.g., creators/uploaders/viewers, re-posters/complainants, moderation/policy teams, curation teams, affected/depicted persons incl. public officials), and we flag vulnerability where relevant (e.g., personal/structural resilience, minors).
Annex B.3.2 (Subsystem scoping + interference mechanism)	Subsystem(s) are selected for the specific platform action under test (e.g., S3 moderation; S5 similarity/persistence; S6 access-control). For each, we record the concrete interference mechanism (e.g., removal, disabling access, demotion, account restriction, monetisation restriction, “equivalent” suppression), and the stakeholder for whom the limitation is assessed.
Annex B.3.3 (Interference check per Charter right)	For each selected right, we answer whether the subsystem action limits the exercise of the right for the selected stakeholder (yes/no/unclear). For the worked case, takedown and “equivalent” near-duplicate suppression (S3/S5) are treated as an interference with Art. 11 for reposting users, while simultaneously serving protective aims for Arts. 1/7/8 for the depicted person.
Annex B.3.4 (Severity and probability prompts)	We apply the severity prompts (scope, duration, reversibility) and the probability prompt (likelihood of interference/harm) and record tiered outputs. Where the vignette does not provide required facts (e.g., reach, recurrence, error rates, appeal outcomes), we mark the step as underdetermined and attach an assurance tag.
Annex B.3.5 (Objective consequences checklist)	We document real-life consequences using the annex prompts (immaterial vs. material harm; well-being/participation impacts; where applicable financial loss or market impacts). If a consequence category is not supported by the vignette, we record <i>N/A</i> or <i>insufficient evidence (IE)</i> rather than inferring.

Table 13.10: Pointers to annex testing steps and the minimal input referenced in §4.3 for the worked case (Part II: Annex B.3.6–B.3.10).

Annex pointer	Minimal input referenced in §4.3
Annex B.3.6 (Art. 52(1) “provided for by law”)	We identify the normative basis invoked for the restriction (EU obligations plus platform ToS/policy clauses, including Art. 14 DSA implementation where applicable) and assess accessibility/foreseeability (clarity of scope, definitions such as “equivalent,” user-facing notice/SoR). If the invoked clauses or user notice are not provided, we mark <i>IE</i> and list the missing artefacts.
Annex B.3.7 (Art. 52(1) legitimate aim)	We state the objective of general interest and/or protection of the rights and freedoms of others that the measure pursues, and we connect this explicitly to the worked case (e.g., protecting the depicted person’s dignity, private life, and data protection in NCSI scenarios).
Annex B.3.8 (Art. 52(1) necessity + least restrictive means)	We document whether the measure is necessary and whether less restrictive alternatives could be comparably effective, noting where this requires provider-side evidence (comparative effectiveness, operational constraints, alternative mitigation set). Absent such evidence in the vignette, we record <i>IE</i> and specify the minimum required inputs.
Annex B.3.9 (Art. 52(1) proportionality balancing + safeguards)	We record the balancing rationale across competing rights (e.g., protecting Arts. 1/7/8 while limiting Art. 11) and the safeguards relied upon (statement of reasons quality, appeal/redress pathways, error controls, transparency). Where safeguards cannot be evidenced (missing SoR text, appeal stats, thresholds, false positive/negative rates), we mark <i>IE</i> .
Annex B.3.10 (Evidence map + evidence-dependent steps)	For every underdetermined step, we list the minimum evidence required (e.g., ToS clauses invoked; SoR templates; appeal outcomes; similarity thresholds; false positives/negatives; time-to-action; recurrence/suppression rates). We also compile a short list of steps explicitly marked evidence-dependent for audit scoping and future data collection planning.

Table 13.11: Step 4.4 minimal inputs/outputs for the worked case (full risk record in Annex B.4–B.5).

Annex pointer	Minimal content referenced in §4.4
Annex B.4 (Risk identification & assessment)	Risks are framed as subsystem-mediated pathways to <i>interferences with Charter rights</i> ; severity is assessed qualitatively across objective seriousness, experienced severity, and real-life consequences, with explicit “insufficient evidence” flags where the vignette does not determine legality/metrics.
Annex B.5 (Documentation output)	Worked-case outputs are condensed to the four risk statements (R1–R4) with severity/assurance tags; the complete evidence map and rationale remain in the Annex.

Table 13.13: Worked-case risk outputs referenced in the paper (full table with tags in Annex B.5 / Table 3).

ID	Risk output (label only; details in Annex)
R1	NCSI at scale (generation → distribution → persistence)
R2	Near-duplicate persistence after notice
R3	Paywall mitigation (displacement + collateral effects)
R4	Remedy/process bottlenecks (notice-and-action, escalation, statement of reasons)
M1	Art. 52(1) proportionality check for paywall mitigation

14

EVALUATING THE SYSTEMIC RISK OF NEGATIVE EFFECTS ON THE ELECTORAL PROCESS CONTENT ON ONLINE PLATFORMS UNDER THE DIGITAL SERVICES ACT

“Some curious linkages here, Gordon.”

Dale Cooper, Season II, Episode 18, Scene 5 continued, David Lynch and Mark Frost

Safeguarding democratic elections is hard. Social media plays a vital role in the discourse around elections and during electoral campaigns. The following article provides an analysis of the ‘systemic electoral risks’ created by Twitter and Facebook and the mitigation strategies employed by the platforms. It is based on the 2020 proposal by the European Commission for the new Digital Services Act (DSA) in the context of the 2021 German federal elections. This article focuses on Twitter and Facebook and their roles during the German federal elections that took place on 26 September 2021. We analysed three systemic electoral risk categories: (1) the dissemination of illegal content, (2) negative effects on electoral rights, and (3) the influence of disinformation and developed systematic categories for this purpose. In conclusion, we discuss how to respond to these challenges as well as avenues for future research.

This chapter is based on: Kübler, J., Sekwenz, M.-T., Rachinger, F., König, A., Gsenger, R., Pírková, E., Wagner,

14.1. INTRODUCTION

Safeguarding democratic elections is hard. Although frequently taken for granted, it is so central to democratic governance, and yet so difficult to ensure. Safeguarding democratic elections is not simply about ensuring that votes are counted correctly. The media environment around elections also plays a critical role. As the media environment has been changing rapidly in the past decades, the risks to free and fair elections are also evolving rapidly.

Social media plays a central role in these media environments in many parts of the world. This article investigates in detail the systemic risks posed by social media in the context of elections, and the ways in which these risks can be mitigated. It also assesses the extent to which social media platforms, such as Facebook and Twitter, sufficiently reduce these risks or whether they could be doing more.

In this context, the 2020 proposal of the European Commission for the new Digital Services Act (DSA) is an important piece of legislation that promises to significantly strengthen the European accountability regime for online platforms. Article 26 of the DSA forces very large online platforms (VLOPs) to identify significant systemic risks stemming from the operation of their platforms, and Article 27 proposes mitigation measures that these platforms should implement. Very large online platforms are defined by the DSA as those having more than 45 million recipients of the service, which is the equivalent of 10% of the European Union's population [19].

However, what would a concrete analysis of risks of an online platform based on the proposed DSA look like in practice? To address this question, this article proposes to conduct an external analysis of relevant electoral risks without access to internal platform data. Of course, such an analysis is very difficult. Consequently, the article was only able to conduct a much smaller version than the risk assessment than would be legally necessary under the proposed DSA. However, we believe our work can serve as an initial demonstration of what such analysis of risks for electoral processes could look like and contribute to the debate on how to implement them in practice.

The analysis was carried out in the context of the German federal elections that took place on 26 September 2021, taking into consideration the two large online platforms mentioned previously: Twitter and Facebook. The article focused exclusively on 'systemic electoral risks' rather than examining other areas of systemic risks also raised by large online platforms. Lastly, it should be noted that the whole of the EU DSA remains a legislative proposal, including Article 26 on 'risk assessment' and Article 27 on 'mitigation of risks.' As such, we hope that our experience in conducting a concrete analysis of risks in practice can contribute to further development of the DSA. Thus, we trust that this report can contribute to a better understanding of the degree to which the DSA is effective in safeguarding European elections and where more still needs to be done.

14.2. THEORETICAL AND METHODOLOGICAL APPROACH

Online platforms, such as social media sites, have become major avenues for the distribution of information and debates on politics, especially in the context of elections. Misinformation, disinformation, and propaganda tactics are not unique to our era [385]. The 2016 US presidential election, which installed Donald J. Trump as President, and

the UK's referendum decision to leave the European Union ('Brexit') the same year were watershed moments for the public perception of the online platforms' role in elections [386]. In fact, researchers subsequently found that during the US presidential election, 25% of tweets containing a link to news outlets spread either fake or extremely biased news [89]. Another article on the use of political bots during the UK referendum found, based on a sample of more than 1.5 million tweets, that less than 1% of sampled accounts generated almost a third of all messages, making the role of bots during Brexit small but strategic [368].

Whereas online platforms initially rejected holding any responsibility for the content published on their sites, they subsequently established a combination of human-driven and automated editorial processes to promote or remove certain content types. The so-called 'content moderation' is the systematic practice of a social media platform of screening content to ensure compliance with community guidelines, user agreements, laws, and regulations, and norms of appropriateness for a certain locality and its cultural context [621].

To respond to the need to examine and curate a large amount of data, online platforms have developed content moderation systems. For instance, the main strategies regarding content moderation that Facebook and Twitter have employed over the years are: fact-checking, deletion of content, and ban/suspend user accounts [624].

These content moderation systems rely heavily on the removal of harmful and, otherwise, undesirable content. However, there are growing concerns regarding the impact of these platforms' decisions on human rights and individuals' freedom of expression and information. Many community-led platforms offer alternatives to these challenges [425]. For instance, as an alternative to deleting undesirable content, some community-led platforms use systems that enable users to downvote/upvote content and/or other users. While each site uses a slightly different reputation system, they generally track the behavior of members by giving users "karma" points for their posts and other activities, as well as the ability to upvote (and, usually, also downvote) other's contributions. When a post is upvoted or downvoted by fellow members of a community, the poster receives or loses points [746]. This method of reducing the visibility of certain content is used by platforms such as Slashdot.

Apart from content moderation strategies, the design choices that online platforms make affect which information is available, how it is displayed, and how people communicate. A recent study showed that implementing changes in platform design to promote different forms of appropriate behavior within specific communities may be particularly effective in getting users to change their behavior [743].

14.2.1. THEORETICAL FRAMEWORK

This section discusses the theoretical framework built for this article. It exposes how the systemic risk assessment required by the DSA proposal to Very Large Online Platforms (VLOPs) was tailored to the specific risks identified in the context of elections (what we named in this article 'systemic electoral risks'), considering the negative impacts they might have on free and fair elections.

14.2.2. DSA, ONLINE PLATFORMS, AND SYSTEMIC RISKS

Social media platforms control the flow of information shared on their platforms through rules codified in their algorithms. These platforms choose to promote certain content above others to keep their websites appealing to users as part of their business model. They also screen (or moderate) content to guarantee its compliance with laws and regulations, community guidelines, and user agreements.

Within the context of the new EU DSA, a draft of which was published by the European Commission in December 2020, the platforms play an important role in safeguarding fundamental rights. The role of large platforms is particularly important in the context of Article 26 of the DSA, which argues that very large online platforms must take measures to prevent creating ‘systemic risks’ [19].

The term ‘systemic risk’ rose to prominence in discussions related to the 2008 economic crisis, when failing large financial firms with complex businesses caused ripple effects in the larger economy. Systemic risk thus describes risks that emerge from complex system failure, where the failure of a single component leads to systemic knock-on effects.

Similarly, in the DSA proposal, the European Commission recognizes that VLOPs cause significant societal risks due to the large number of recipients of the service and their role in facilitating public debate, economic transactions, and the dissemination of information, opinions, and ideas, and in influencing how recipients obtain and communicate information online [19]. As the number of users of social media platforms has soared, online activity has become increasingly central to offline cultural and political events, such as the UK’s Brexit vote and the 2016 US presidential election, as mentioned previously. In this context, the online networks’ potential to splinter the public into informational echo chambers, induce ingroup/outgroup hostilities, and make participants vulnerable to misinformation and propaganda dominated the headlines [618, 680].

14.2.3. SYSTEMIC ELECTORAL RISKS AND CATEGORIES OF ANALYSIS

Article 26 of the DSA 2020 proposal considers that VLOPs “shall identify, analyse and assess, [...] any significant systemic risks stemming from the functioning and use made of their services in the Union” [19]. However, risks in the context of elections are mentioned but not detailed in the current DSA proposal. Therefore, this article intends to tailor the areas of risk proposed by the DSA for VLOPs to the context of elections in the European Union. In this article, we call this type of risk ‘systemic electoral risks,’ which refers to the impacts of systemic risks—stemming from the functioning and use of VLOPs services—on democratic elections. These systemic risks may vary from disinformation or manipulative and abusive activities and may impact the ability to safeguard free and fair elections.

To discuss systemic electoral risks, the dimensions/categories proposed in the EU DSA were adapted for this article. Thus, in this article, systemic risks are defined as primarily falling into the following three categories:

- Dissemination of illegal content
- Negative effects on electoral rights
- Disinformation

This article developed a codebook consisting of three distinct parts corresponding to each category mentioned above. We identified different types of systemic risks for each category—so-called subcategories—which received a specific code, covering legal clauses, classifications of infringements to electoral rights during election campaigns, and various forms of disinformation. The subcategories of illegal content are based on the existing categories developed by Tiedeke et al. (2020) [703]. The electoral rights subcategories were developed based on work by the authors based on the state of the art. Lastly, the disinformation subcategories are based on Kapantai et al. (2021) [406].

A. DISSEMINATION OF ILLEGAL CONTENT

Content that is shared and published on social media platforms might fall under the restrictions of speech, such as libel, incitement of hatred, or defamation. Such illegal content might also fall under the category of disinformation. In this regard, its wide dissemination can influence elections and infringe on individuals' electoral rights. A prominent example is the 2016 elections in the United States, where disinformation and hateful content dominated the electoral process and reportedly influenced the election's outcome. Beyond manifestly illegal content, social networks remove content in contravention of their own Terms of Service (ToS), a legal document a person must agree to abide by when registering an account.

This article designed 63 codes for the 'illegal content category' based on the comprehensive taxonomy of German national and international law developed by Tiedeke et al. (2020) [703]. Created to evaluate the quality of content governance decisions in online forums in Germany and Austria, the taxonomy also includes relevant aspects of platform ToS that can lead to the deletion of content. Given the focus of the present article, we used the categories related to German and international law, and the relevant ToS categories.

B. NEGATIVE EFFECTS ON ELECTORAL RIGHTS

The Universal Declaration of Human Rights, accepted by the United Nations General Assembly in 1948, enshrines the rights and freedoms of all human beings. The right to free and fair elections is also rooted in the founding values of the European Union: respect for human dignity, freedom, democracy, equality, the rule of law, and respect for human rights. Hence, the European Commission has sought to enhance transparency, protect free and fair elections, and promote the democratic participation of all European citizens in various ways, for example, through its electoral package for the 2019 European Parliament election. Therefore, it seems sensible that category (b) proposed by Article 26 of the DSA, "any negative effects for the exercise of the fundamental rights" would include electoral rights and the right to free and fair elections.

The 'electoral rights category' captures the various dimensions by which disinformation can affect an election. This category was grouped into three overarching subcategories: procedural disinformation, disinformation on parties and candidates, and integrity of elections. The subcategories cover the entire lifecycle of an election, including voter registration, voter identification, election campaign, election day, counting, and the publication of the results. In total, 20 codes were designed for the 'electoral rights' category.

C. DISINFORMATION

Disinformation can be defined as the dissemination of false information with the aim of influencing public opinion, groups, or individuals serving political or economic interests. Contrary to misinformation, whose inaccuracies are unintended, disinformation is false information spread intentionally. This information is often disseminated covertly and is intended to obscure the truth. The related term ‘fake news’, however, is a political expression used to criticize a news story or media outlet. Online platforms implement different strategies to deal with disinformation.

The ‘disinformation category’ created for this article is based on Kapantai et al.’s (2021) comprehensive literature review of disinformation taxonomies [406]. It comprises 11 elements distinguishing the various forms disinformation takes in practice. A test run on 50 tweets and Facebook posts revealed that two subcategories were either irrelevant to our data or introduced noise, namely “biased” and “fake reviews”. This is because the way the two subcategories were defined was insufficiently precise in order to be able to accurately code them into the data systematically. Therefore, these subcategories were removed, leaving nine disinformation codes.

14.2.4. RESEARCH DESIGN, METHODOLOGY, AND CASE SELECTION

Taking into consideration the three DSA categories adapted to discuss systemic electoral risks and the possibility of online platforms adopting different approaches to assess and mitigate systemic risks, this article explored the following questions:

1. In the context of elections, what would a risk assessment in VLOPs look like in practice, considering the dissemination of illegal content, negative effects on electoral rights, and disinformation?
2. What measures are VLOPs taking to reduce systemic electoral risks through content moderation, design choices, and online advertising?
3. Are these measures and the approaches taken by VLOPs to assess and mitigate systemic electoral risks effective?
4. To what extent is the current DSA proposal, especially Articles 26 and 27, effective in protecting European elections?

Answering these questions was particularly challenging as the authors did not have the same amount of data as the platforms did. For example, we were unable to access content that the platforms had previously moderated and thus was no longer publicly available. This would have been helpful to understand platforms’ moderation behavior and the degree to which their actions in moderating content reduced the amounts of problematic content. We were only able to address these questions based on public information. Without inside privileged access to all relevant data, our methodology is necessarily limited in scope to the data sources that we are able to access as external researchers, and a full analysis by one of the VLOPs would need to be much more expansive. Notably, we were not able to access sufficient advertising data to be able to analyze the platforms’ systems for selecting and displaying advertisements, although we hope to be able to do so in future research projects. Nevertheless, we believe it is possible to make an initial

attempt at what a credible risk analysis could look like, while acknowledging that due to our lack of all relevant data, such an attempt is necessarily incomplete.

In reports and statements, social media companies are keen to stress the effectiveness of their measures in limiting the prevalence of illegal and misleading information on their platforms. In the absence of an independent validation of these reports, however, the public and policymakers are currently unable to assess the veracity of these claims. To explore the scope of illegal and misleading content, as well as content infringing on electoral rights, this article analyzed empirical data from two VLOPs operating in Germany.

This article relied on a quantitative research design. To explore the three dimensions or categories of potential threats emanating from social media platforms during the 2021 German federal elections, this article applied a quantitative analysis of organic user-generated content on selected social media platforms in the context of a major election to explore the scope of illegal content, disinformation, and content infringing on electoral rights on said platforms.

HOW TO CONDUCT AN ANALYSIS OF ELECTORAL RISKS

First, to assess the prevalence of potentially harmful content on social media platforms during election campaigns, we developed a codebook with subcategories (identified with codes) on dissemination of illegal content, infringements on electoral rights, and disinformation (see Annex). Coding is a common technique for condensing data into identifiable topics. A code is a distilled topic applied to a text segment illustrating that topic. By using codes, researchers can search for topics across data and thereby identify patterns. Based on these coded data, the article estimated the proportion of data that matched our categories and were present across the respective platforms.

Subsequently, we collected the data samples necessary for this article on Twitter and Facebook (see section ‘Data Collection’ below). This was followed by a coding test conducted on 50 random tweets to assess the appropriateness of the subcategories, allowing for fine-tuning definitions with the coders, and assessing the intercoder reliability for each subcategory. With subcategories and codes fine-tuned, a random representative sample of 1101 tweets and 1101 Facebook posts were coded. The coding process was done in parallel by two different groups of coders. Thus, each sample was coded twice. The data were then merged and the intercoder reliability for each subcategory discussed and, when necessary, the coding was adjusted. In fact, Facebook uses a similar but distinct method to estimate the prevalence of misinformation and other harmful content on its platform, relying on random sampling and manual labelling.

CASE SELECTION

As part of the research design, this study looks at two cases of VLOPs to assess the systemic risks online platforms pose to democratic elections. Therefore, we conducted data analysis on two VLOPs. Facebook and Twitter were chosen because they are the most relevant VLOPs globally, as well as in Germany. Thus, we believe that it is possible to better understand how an analysis of electoral risks could or should be done by studying these specific platforms.

DATA COLLECTION

This article analyzed the three systemic electoral risk categories created (dissemination of illegal content, negative effects on electoral rights, and disinformation) on two global VLOPs operating in Germany: Facebook and Twitter. The assessment was based on representative samples of public data. This approach should enable us to make more reliable statements, allowing for meaningful comparisons of online platforms rather than the anecdotal data that is mostly used at present. The data collection period covered the second half of May 2021, from day 15 to day 31.

Every social media platform is organized in a slightly different way. For instance, on Twitter, hashtags are used to specify the topic or intended audience of a tweet and allow a user to engage a much larger potential audience than only his or her immediate followers. Hence, data to study the public debate of elections on Twitter can be collected through the use of one or more relevant hashtags and the subsequent analysis of the resulting universe of messages. Facebook also allows the use of hashtags; however, it is not a primary feature of the platform, and they are not as routinely used as on Twitter.

To achieve a comparable sample of posts on both platforms, we collected data combining keywords and hashtags. Given that datasets collected using keywords risk introducing noise from the large number of messages using the keywords without actually referring to the intended topic(s), we chose keywords that referred uniquely to the election at hand, namely “Bundestagswahl” (federal election) and the abbreviations “BTW2021” and “BTW21”. Given that the data was collected relatively early into the campaign devoid of major mediated events, we included hashtags of all political parties—Christian Democratic Union of Germany (CDU), Christian Social Union in Bavaria (CSU), Free Democratic Party (FDP), Grüne, Linkspartei, Social Democratic Party of Germany (SPD), and Alternative for Germany (AfD)—with a realistic chance of passing the electoral threshold of 5% required for representation in the Bundestag. Furthermore, we included the names of each party’s lead candidate (“Spitzenkandidat”), and hashtags already in use to refer to the election (#btw2021 and #btw21), as well as more general terms, such as #bundestagswahl and #wahlkampf.

On Twitter, we collected tweets from the platform’s application programming interface (API) using the Python script *twarc2* and the Search endpoint. Our search returned 358,667 tweets in the period between 15 to 31 May 2021. Our Facebook dataset contained 6,712 posts and 38,685 comments for the same time period.

Based on our previous research we estimated a response distribution of between 2%–3%. We thus estimated that with a sample size of 1101 we would be able to attain a margin of error of 1% or lower with a 95% confidence interval. As a result, both datasets were subsequently transformed into randomized samples of 1101 entries each, and finally coded using the codebook previously created. The deductive codebook was based on the state of the art in the respective academic fields, and is included in the Annex. The codebook includes detailed categories in the area of dissemination of illegal content, electoral rights, and disinformation. In addition, the categories we developed are not perfectly distinct and at times overlap, with multiple categories fitting to a single post. For example, one Facebook post can both be disinformation and an electoral rights violation at the same time.

Three coders were involved in the coding process. All of the legal aspects of the coding

were law students who had studied law for at least three years. They coded the material separately using Microsoft Excel. For cases where there was a disagreement between at least two coders, an additional group of four senior researchers (with legal and social science backgrounds) came together with the coders to discuss the individual cases and mutually agree on an appropriate outcome. For Facebook, the coders agreed that a post was problematic 93.10% of the time. For Twitter, the coders agreed that a post was problematic 92.55% of the time. These rates of intercoder reliability are within a good range that suggests reliable coding. Finally, it is important to mention that during the data collection phase, several difficulties arose in accessing data from platforms. This situation makes it unnecessarily difficult, and sometimes impossible, for researchers to access reliable data to conduct research, undermining the capacity of third-party auditing of what happens on platforms and how effectively platforms enforce their policies.

14.3. APPLYING THE DSA'S RISK ASSESSMENT AND MITIGATION FRAMEWORK TO THE GERMAN FEDERAL ELECTIONS

The analysis of our representative samples of Facebook and Twitter data found a significant number of problematic posts and tweets. For the Facebook sample, 6.72% of all election-related posts were potentially illegal, disinformation, or infringements of electoral rights. As this data is based on the coding of a sample, it is possible that our sample overrepresents or underrepresents the underlying population data. However, with a confidence level of 95%, we can say that the underlying population data is within a margin of error of 1.46% of our sample. Of the problematic posts on Facebook, 4.05% were likely illegal under German law, 35.14% violated the platform's community standards or Terms of Service, 46.65% were violations of electoral rights, and 93.24% could be considered disinformation.

Similarly, for the Twitter sample, 5.63% were found to be problematic. As these results are also based on the coding of a sample, here too our sample might be overrepresenting or underrepresenting certain categories within the overall population of Twitter data we analyzed. However, we can say with a confidence level of 95% that there is a margin of error of 1.34% between the Twitter sample we coded and the overall population being studied. Of these problematic posts on Twitter, 14.52% broke platform rules, 51.61% infringed on electoral rights, and 100% were considered disinformation.

14.3.1. DISSEMINATION OF ILLEGAL CONTENT

Of the items flagged, 3 items (4.05%) in the Facebook sample and none in the Twitter sample were coded as likely illegal under German law. With regards to infringements of the service's Terms of Service or community standards, the article identified 35.14% on the Facebook sample, and 14.52% on the Twitter sample. Therefore, even after undergoing content moderation processes, there remained three potentially illegal posts on Facebook. We identified one post as 'malicious gossip (Üble Nachrede)' as described in §186 German Criminal Code, one post as 'disturbing public peace by threatening to commit offences' as described in §126 German Criminal Code, and one post as 'incitement of masses' (Volksverhetzung) based on §130 German Criminal Code, which punishes incitement to hatred against segments of the population and refers to calls for violent or arbitrary

measures against them. We did not find any illegal content in the Twitter sample.

14.3.2. NEGATIVE EFFECTS ON ELECTORAL RIGHTS

Of the problematic content found, 46.65% were violations of electoral rights on Facebook and 51.61% on Twitter. Within the electoral rights category, 'Candidates - Electoral campaign' was the most common. On Facebook, this subcategory accounted for 41.89% of all flagged posts. On Twitter, they accounted for all posts flagged as infringing electoral rights, that is, 51.61% of all problematic content. The prevalence of this category underlines that spreading disinformation by actors interested in harming or promoting certain candidates or parties, or increasing social and political divisions in society, is a common strategy to harm certain candidates.

On Facebook, our coders furthermore registered the presence of 2.70% of posts coded under the subcategory 'Integrity – Electoral results', defined as election losers and their supporters making undocumented claims on electoral fraud to justify electoral defeat, delegitimize democratic election, and encourage electoral protests. Also, 1.35% were coded in the subcategory 'Integrity – Counting and notification', which includes claims on lost ballot boxes, non-counted votes, or manipulation of vote counts and election protocols to justify electoral defeat, question electoral results, and delegitimize elections.

Moreover, 1.35% were coded as subcategory 'Procedural – Vote count', where actors interested in delegitimizing the elections spread disinformation on procedures of the vote count to disturb the electoral process, confuse voters, and to prevent certain voters from voting. Lastly, 1.35% were coded as subcategory 'Candidates – Election polls', where actors interested in (de-)legitimizing the elections or harming/promoting certain candidates or parties publish fictitious, false, or supportive election polls to (de-)mobilize voters and/or influence both voter turnout and voters' decisions.

Major risks to electoral rights identified by the interviewees include outdated electoral laws unfit for the online sphere and a lack of institutional oversight, as well as platforms playing favours with politicians, third-party interference, limited capacities of platforms to adequately respond to local specificities, and the very design of platform algorithms.

14.3.3. DISINFORMATION

Disinformation is the most common form of problematic content found by the coders in both Facebook and Twitter samples. Of all content flagged as problematic, 93.24% was considered disinformation on Facebook, and 100% on Twitter. Within the disinformation category, 'trolling', defined as the act of deliberately posting offensive or inflammatory content to an online community with the intent of provoking readers or disrupting conversation, was by far the most prevalent in both datasets, with 47.30% of problematic Facebook posts and 57.68% of tweets coded as such.

Other disinformation items found were 'rumours', referring to stories whose truthfulness is ambiguous or never confirmed (gossip, innuendo, unverified claims), with 31.08% of problematic posts coded as rumours on Facebook and 29.03% on Twitter. There were also 13.51% of Facebook posts and 14.52% of problematic tweets coded as 'conspiracy theories', which are stories without factual base as there is no established baseline for truth, often explaining important events as secret plots by government or powerful individuals. In addition, 4.05% of Facebook posts and 4.84% of tweets were coded as

‘fabricated’, meaning stories that completely lack any factual base and are 100% false, with the intention to deceive and cause harm.

Only on the Facebook sample, the article found 5.41% of problematic posts coded as ‘pseudo-science’, which promotes information that misrepresents real scientific studies with dubious or false claims. A lower amount of content was coded as ‘hoaxes’, which are relatively complex and large-scale fabrications presented as legitimate facts, intended to cause material loss or harm to the victim, with 1.35% of problematic Facebook posts. Further, 1.35% of Facebook posts were coded as ‘imposter’, where genuine sources are impersonated with false, made-up sources to support a false narrative. This can be very misleading, since the source or author is considered a great criterion for verifying credibility.

14.4. CONCLUSION

Safeguarding democratic elections is hard. We acknowledge that online platforms and their regulators have an enormously difficult task ahead of them in trying to safeguard elections. However, this acknowledgement should not detract from the fact that private platforms and public regulators’ current efforts to safeguard elections are simply not sufficient. As a result, democratic elections will continue to suffer from disinformation and continuous breaches of electoral rights.

Social media platforms are not a mirror of society, even if they often like to claim so. Their presence in society has effects that cannot be taken for granted, nor are they likely to go away any time soon. Regulators need to acknowledge the central role of these platforms in elections and systematically develop institutions that are adequately able to respond to the issues discussed. These institutions urgently need to be strengthened both in Germany and at the EU level. The large prevalence of problematic content in our analysis suggests that online platforms are not currently doing enough to respond to the challenge of problematic content around elections.

The EU DSA can undoubtedly contribute to improving the mitigation of systemic risks from the platforms. In particular, Article 26 and Article 27 of the DSA studied here create a valuable regulatory framework to push these platforms in the right direction. However, without expanded external audits of the platforms, they will continue to run rings around regulators and election observers. As one leading election observer acknowledged: “they’re playing us,” even as he spent his days “running after the tech companies.”

Importantly, the idea frequently stated by current and former Facebook staff that elections are ‘on balance’ better than they previously were before social media lacks empirical foundation. We do not know what democratic elections would look like without social media. Still, we can legitimately claim that elections would not be democratic elections if social media were not present or completely censored. The relevant question is not whether democratic elections are compatible with social media but rather how online platforms can be developed further to be more supportive of free and fair elections. This will likely require considerable resources and probably take some time, but it is definitely not impossible. If anything, it seems that these platforms are not sufficiently considering the vast body of knowledge that already exists, and even some of their internal research. If this is not taken seriously, safeguarding democratic elections is essentially impossible.

However, it does not have to be this way. We know that different performances by online platforms are possible by comparing how well the large platforms perform and are even more possible by considering many of the smaller online platforms that do a better job. The question is whether platforms and their regulators will be willing to take the systemic risks around elections seriously and take meaningful steps to mitigate them. These platforms should not simply be doing this a little here and there before each election campaign but instead systematically building more sustainable platforms.

IV

SYNTHESIS

15

EU DIGITAL REGULATION AS INDUSTRY SHAPING POLICY – THE DSA, BRUSSELS EFFECT, AND GLOBAL COMPETITIVENESS

"I love going into another world, and I love mysteries. So I don't really like to know very much ahead of time. I like the feeling of discovery."

David Lynch, *Catching the Big Fish: Meditation, Consciousness, and Creativity*

The European Union's recent wave of digital regulations, especially the Digital Services Act (DSA) [246], highlights a strategic use of law as an instrument of industrial policy and global digital influence. While China has explicitly cultivated national technology champions through state-led industrial policy, the United States has largely relied on market-driven innovation and occasional strategic interventions in areas of national interest such as AI and semiconductor supply chains. In contrast, the EU has pursued regulatory leadership to assert "digital sovereignty" and shape international standards [42, 223, 235]. This paper argues that the DSA functions not only as a legal framework for safer online spaces but also as a lever of industrial strategy and competitiveness.

Building on Schumpeter's notion of creative destruction and Aghion's work on competition and innovation [32, 650], we examine how the DSA's regulatory approach may simultaneously support complementary innovation (for example, in compliance and risk-assessment tools) while constraining more radical product innovation through heightened procedural and liability requirements. In this sense, the DSA becomes a case study in how industrial policy and innovation dynamics interact within a regulated digital economy.

The paper shows how the Brussels Effect – the EU's ability to export its rules globally – is amplified by the DSA, incentivizing global platforms to adopt European norms in content moderation and platform design [69, 90]. We discuss the strategic incentives driving harmonization of platform practices (liability risks, cost-efficiency, scale of moderation) and the resulting implications for innovation, market access, and long-term competitiveness.

15.1. INTRODUCTION

The European Union's recent wave of digital regulations, especially the Digital Services Act (DSA) [246], highlights a strategic use of law as an instrument of industrial policy and global digital influence [42, 223, 235]. While the United States and China have focused on cultivating technological champions and market dominance, the EU has pursued regulatory leadership to assert “digital sovereignty” and shape international standards [90, 472]. This paper argues that the DSA functions not only as a legal framework for safer online spaces but also as a lever of industrial strategy and competitiveness. While the DSA's long-term effects on markets and innovation are still unfolding, its design and early implementation already position it as a key mechanism through which the EU projects its industrial ambitions and global regulatory behaviour [90].

While the DSA exemplifies Europe's ambition to project regulatory power globally, it also raises a forward-looking question: how will such a sovereignty-driven framework influence innovation dynamism within the EU itself? Understanding whether the DSA enables or constrains innovation will be central to assessing its long-term policy impact.

Yet, beyond questions of sovereignty and competitiveness, regulation itself can influence innovation trajectories. Classic work in innovation economics shows that institutional frameworks shape the direction and intensity of inventive effort, raising the question of how the DSA may affect innovation dynamism within the EU. Regulation does not merely enforce compliance; it can also channel the direction of innovation. As Arrow, Nelson, and Winter demonstrated, institutional environments and rule-making structures influence, especially under uncertainty, where inventive effort is concentrated, which economically feasible routines persist, and which are abandoned [51, 558]. In this sense, Schumpeter's notion of creative destruction offers a useful lens: capitalist innovation is not a steady process of improvement, but an evolutionary one driven by the “perennial gale” of technological and organizational change that incessantly “revolutionizes the economic structure from within, destroying the old one, incessantly creating a new one” [650]. This raises a broader question: to what extent does the DSA foster innovation dynamism within Europe, or risk constraining it through heightened proceduralization?

A core process governed under the DSA is content moderation. Content moderation balances freedom of speech rights and provides legal reasons for moderation decisions, divided in legal reasons (e.g. national criminal law) or contractual reasons (Terms of Service). This distinction between legal and contractual grounds also resonates with recent U.S. regulatory debates. There, antitrust and consumer protection enforcement, particularly through the FTC, has increasingly looked to Terms of Service (ToS) as an entry point for intervention, emphasizing transparency and accountability in how platforms enforce their own rules [263, 464]. Such approaches reflect a shared regulatory aspiration: to discipline platform power not only through substantive law but also through the governance of contractual standards.

ToS are standard contracts that platforms conclude with their users when signing up for the platforms or services. The ToS automatically implement the platform's Community Standards in the contract between the user and the platform to define what is allowed or restricted on the platform. For example, a platform could have a clause in its Community Standards that could sanction the posting of unicorn content as part of its right to contractual freedom. The dominance of ToS-based governance also has implications for

innovation. From a complexity perspective [411], self-regulatory systems operating at scale can develop adaptive path dependencies: the rules and enforcement architectures they embed shape the evolution of future technologies (moderation tools and organizational routines) [96, 691]. As Kauffman observed, complex adaptive systems evolve at the “edge of chaos” [411], where internal order and external adaptation co-produce new forms of stability. In such systems, selection acts not only on outcomes but also on the architectures that constrain future adaptation, a dynamic equally relevant to institutional self-regulation or to fringe cases of violation/non-violation of platform rules and the potential virality of such content [194].

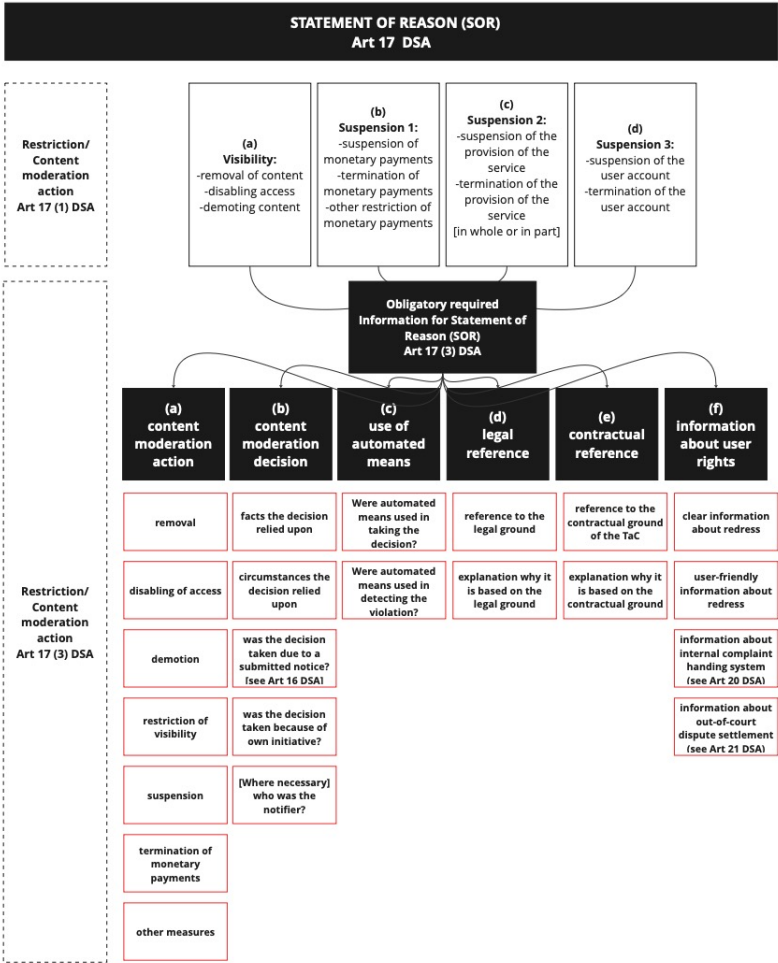
When faced with strict EU rules that carry hefty penalties [246], major tech firms often implement the required changes globally rather than maintain separate regimes, effectively exporting EU norms abroad [69]. The EU explicitly leverages its single market, 450 million consumers, as geopolitical and economic leverage to project its regulatory preferences worldwide [706].

Regulation is often seen as a constraint on industry, but in the European Union’s digital strategy, it also serves an enabling, market-shaping role, e.g. through the creation of safer spaces [480, 676]. Here, “market-shaping” refers to the EU’s use of regulation not merely to restrict harmful conduct but to structure incentives, coordinate expectations, and open new domains for competition and innovation. As the United States and China secure competitive advantage through the creation of technology, whether consumer platforms, advanced AI, or national champions, the EU is asserting its advantage in the creation of rules for the digital economy [66].

This paper examines the DSA’s dual role as a legal framework for online safety and as a de facto instrument of industry-shaping policy. We analyze empirical trends from the DSA’s Transparency Database to illustrate enforcement patterns and platform adaptations and discuss why platforms are incentivized to favour global Community Standards over fragmented national laws. This reliance on ToS-based enforcement also complicates empirical assessment: if platforms categorize most removals as contractual rather than legal violations, the resulting transparency data may underrepresent the scale of illegal content and give regulators a distorted picture of compliance under the DSA, posing the risk of marginalizing platforms’ problems with legally relevant violations.

15.2. EMPIRICAL TRENDS IN DSA ENFORCEMENT: INSIGHTS FROM THE SOR TRANSPARENCY DATABASE

A unique feature of the DSA is its emphasis on transparency in platform enforcement actions [659]. The EU provides a Transparency Database and dashboard for these State-ments of Reasons (SoRs), offering a first empirical glimpse into how platforms are implementing the DSA and adjusting their practices [148]. Figure 15.1 provides an overview of what Article 17 demands from online platforms as information in their SoR entries, for example, the norm that the platform based the moderation decision on according to Article 17(3)(d) for legal grounds and (e) for contractual references like the Terms of Service (ToS). The data was collected on 13 August 2025 and shows entries in the database from 25 September 2023 to the date of data collection.



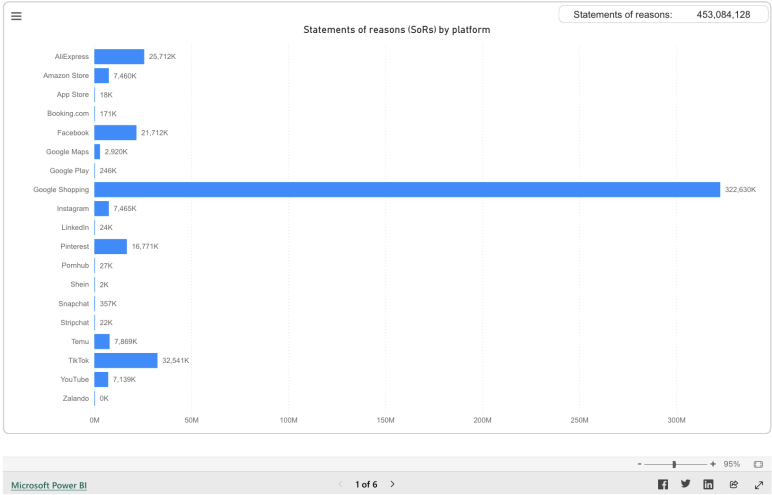


Figure 15.2: Volume of Statements of Reasons for Terms of Service violations issued by Very Large Online Platforms and Very Large Search Engines.

As illustrated above, the volume of content removals (with accompanying SoRs) varies widely across platforms, and so does their enforcement preference between Terms of Service (ToS) and legal removal grounds.

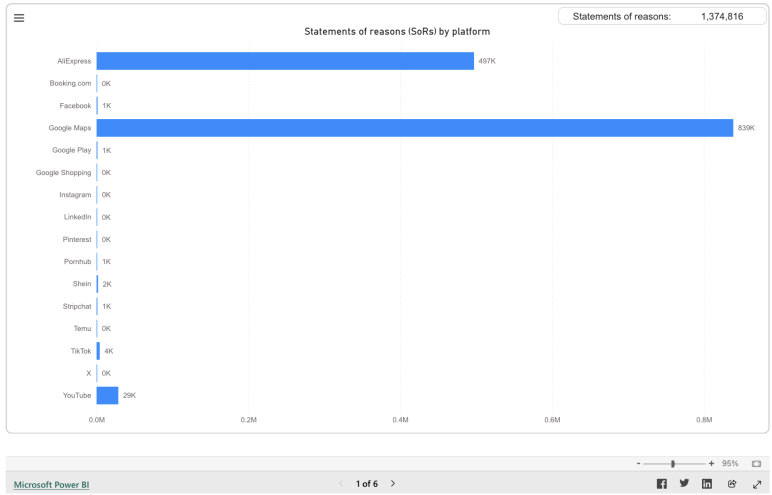


Figure 15.3: Volume of Statements of Reasons for legal violations issued by Very Large Online Platforms and Very Large Search Engines.

The data shows all Very Large Online Platforms and Very Large Search Engines (Article 33 DSA) have a clear predominance of ToS in their SoR database entries over actions grounded in specific legal violations, with 33,220,337,157 entries for SoRs based on ToS

violations and only 14,614,473 based on illegal content (see Figures 15.2 and 15.3). This means that for every SoR entry based on illegal content, there are about 2,274 SoRs based on ToS violations, or, expressed in percent, 99.96% of all entries are based on contractual grounds.

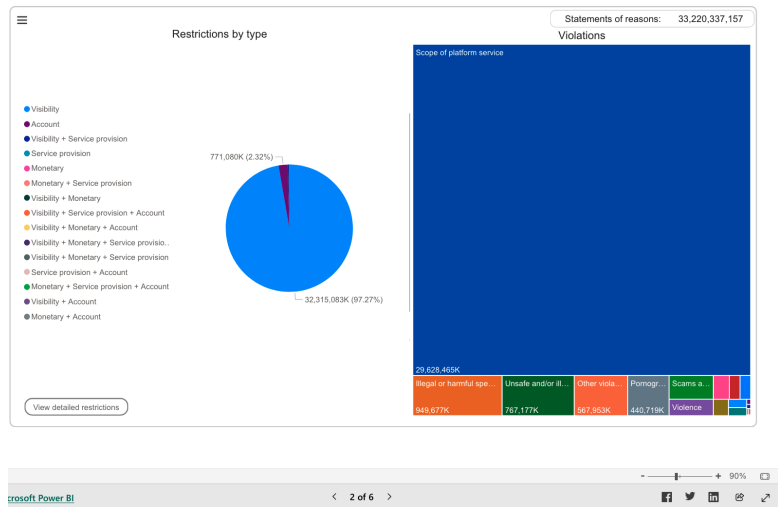


Figure 15.4: Breakdown of content restriction actions by type and violation category for Terms of Service violations issued by Very Large Online Platforms and Very Large Search Engines.

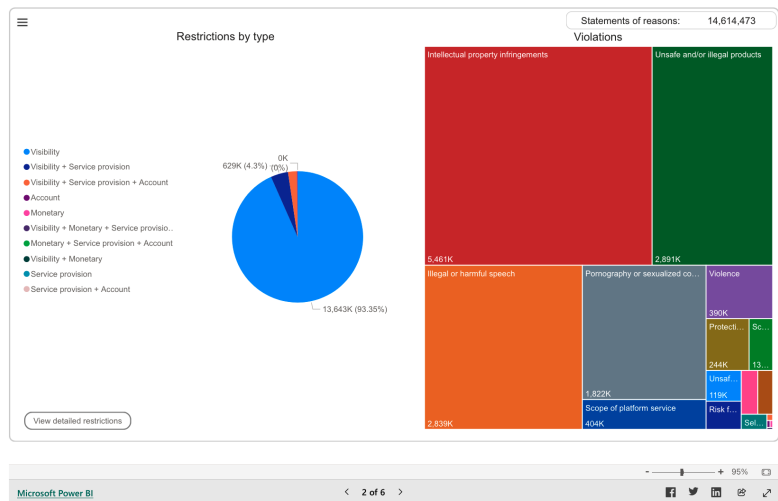


Figure 15.5: Breakdown of content restriction actions by type and violation category for legal violations issued by Very Large Online Platforms and Very Large Search Engines.

Also, data indicates that “Scope of platform service” is the dominant category for

SoR under ToS, with 29,628,465K entries. When comparing the “Illegal/harmful speech” category across enforcement bases, the difference is stark. Under ToS-based SoRs, this category accounts for 949,677,000 entries, representing 2.86% of all ToS-based SoRs. In contrast, under the illegal content classification, “Illegal or harmful speech” comprises only 2,839,000 entries, or 19.43% of all illegal-content SoRs. In absolute terms, this means there are roughly 335 times more entries for “Illegal/harmful speech” in the ToS category than in the illegal content category. This divergence highlights how platforms overwhelmingly handle such speech through ToS enforcement, where the definitional scope is broader and procedural requirements are lighter, rather than classifying it as a legal violation, which entails more formal obligations.

This distribution highlights a form of behavioural harmonization: platforms are predominantly enforcing private norms rather than the fragmented landscape of illegality under Member State laws. This has been demonstrated for content moderation on small and medium platforms, and for Facebook and Twitter under the predecessor of the DSA in Germany – the Network Enforcement Act (NetzDG) [15, 739]. The DSA’s transparency regime compels platforms to categorize each takedown, but the prevalence of the ToS implies that for most content removals, companies rely on their global Community Standards as the basis [667]. In effect, a piece of content might be illegal under some EU national law, but the platform often chooses to moderate it under an equivalent provision of its own rules, which can be applied uniformly [740].

This choice also reflects a strategic business calculus. By relying on Community Standards rather than specific legal provisions, platforms retain greater flexibility to interpret and adjust enforcement in line with their operational and reputational priorities. So too are the platform’s design choices led by differences in complexity for user reporting under the DSA [667]. In that sense, the application of private norms is both a compliance strategy and a business decision; one that may not always align with the European Commission’s vision of harmonized, law-based governance under the DSA.

15.3. STRATEGIC INCENTIVES: GLOBAL STANDARDS VS. NATIONAL FRAGMENTATION

Why do online platforms prefer to enforce one set of rules globally rather than deal with divergent national laws? The DSA’s implementation has cast a spotlight on this question, which goes to the heart of the Brussels Effect [90].

SCALABILITY AND COST-EFFICIENCY

The foremost reason is efficiency. It is vastly simpler and cheaper for a platform like Facebook or YouTube to have a single, uniform code of conduct/Community Standard that applies to all content worldwide. Moderation, especially the review of millions of pieces of user- or AI-generated content, is resource-intensive. Training moderation staff or building AI filters for each country’s unique legal definitions is costly. Platforms instead distil common elements into their ToS and enforce those uniformly – in set theory terms, the ToS become the superset of fragmented legal laws.

REDUCTION OF LEGAL UNCERTAINTY AND LIABILITY

Choosing to remove content under a platform's own rules rather than waiting to assess its illegality under law can also reduce liability exposure. Under the DSA (and previously the e-Commerce Directive) [242, 246], platforms are shielded from liability for user content if they lack “actual knowledge” of illegal material or act expeditiously when notified (see Article 8 DSA). However, if a platform deliberately analyzes content and determines it to be illegal under a specific national law, it might be argued that the platform has knowledge and must remove it everywhere that law applies – or even face potential responsibility for similar content [18]. By contrast, if the platform removes content as a breach of its ToS, it sidesteps an official determination of illegality.

STABILITY, INNOVATION INCENTIVES, AND PATH DEPENDENCE

Beyond cost and liability arguments, uniform global rules also reduce uncertainty and lower transaction costs, which can in turn spur innovation. As Aghion et al. note, predictable regulatory frameworks foster investment and experimentation by clarifying the boundaries of competition and reducing coordination frictions [32]. In this sense, the move toward harmonized rules may not only streamline compliance but also reshape the conditions for innovation in moderation technologies and digital governance through the DSA if uncertainties in enforcement and definitions can be overcome.

At the same time, complexity economists such as Arthur and Perez remind us that harmonization can cut both ways. Uniform frameworks can stabilize expectations and enable broad technological diffusion, but they can also produce path dependence, e.g. locking systems into prevailing technological or organizational trajectories through shaping technologies from a social point of view; as, for example, demanding certain standardization practices in the design of illegal content reporting mechanisms under Article 16 DSA [52, 591]. In the context of the DSA, such dynamics raise the question of whether global compliance architectures will foster adaptive innovation or entrench existing moderation paradigms.

CONTENT MODERATION AT SCALE

Content moderation at scale is a related point. Instagram alone has 1.3 billion images shared daily [404]. Moderating “at scale” is often cited as a near-impossible task [289]; adding a layer of jurisdiction-specific rules multiplies the challenge [264]. The harmful aspects of being exposed to horrible content for human moderators, the need for speed, the complexity of distribution and sheer volume, as well as the DSA's reporting obligations, further encourage investments in automated moderation and compliance tools [202].

RESOURCE CONCENTRATION AND EXPERTISE

Resource concentration and expertise also matter. Handling content under private rules allows platforms to concentrate expertise in-house rather than relying on 27 different national legal systems to guide moderation [110, 552]. Platforms can develop global guidelines and internal expertise that can be updated rapidly in response to new harms (e.g. “we now ban deepfake videos of non-existing news events”). This more streamlined approach uses novel regulations as legal basis for process innovation and moderation standardization.

MARKET ENTRY BARRIERS AND COMPETITIVE STRUCTURE

Becoming or being compliant furthermore is a market entry barrier for smaller firms. As Joskow notes, regulatory regimes often interact with existing cost structures in ways that favor incumbents with established compliance infrastructures and sunk investments [400]. Regulatory compliance imposes fixed and recurring costs, such as monitoring, reporting, and certification, that scale less efficiently for smaller entrants, effectively reinforcing monopolistic market structures rather than opening them to competition.

In essence, the DSA has shone a light on a pre-existing trend: platforms governing themselves transnationally. It forces them to document and justify that governance (via statements of reasons, data access, audits), but it does not fundamentally force a renunciation of global rules in favour of local ones – in fact, it arguably also privileges global rules by making the alternative (case-by-case national compliance) more demanding [138].

The broad definition of illegal content as anything illegal in any Member State (see Article 3(h) DSA) suggests that if a platform is notified under the DSA about such content, it will remove or disable it EU-wide (to avoid “accessibility” in that – and more – Member States) [246]. From a global competitiveness perspective, harmonized moderation simplifies the DSA-aligned process, which is part of the EU’s broader attempt to set uniform rules in Big Tech and foster a more competitive technological playing field for the European market [271, 648].

The DSA’s role as a quasi-digital-industrial policy instrument carries significant implications for how platforms are built and operated, who can enter or remain in the EU market, and the pace and direction of innovation.

15.4. INNOVATION CLIMATE

The debate over Europe’s heavy digital regulation often centers on its impact on innovation. Critics warn of a chilling effect: companies might shy away from developing new features or business models if they anticipate regulatory hurdles or fines [603, 708]. However, it is equally plausible that clear rules of the road enable innovation by providing legal certainty [600]. The DSA harmonizes what is expected of platforms across all Member States, replacing a patchwork of national laws (like Germany’s NetzDG or Austria’s Communication Platforms Law) with one framework [15, 616]. A company now knows that if it complies with the DSA, it can operate EU-wide. This legal clarity and uniformity can reduce compliance costs in the aggregate and allow a company to scale across Europe more easily than before. Furthermore, it aims to create value for users by creating a safer online space and a more transparent handling of content moderation processes within platforms.

Yet this harmonization also produces a dual effect: the DSA could generate both innovation in compliance technologies and innovation deterrence in new platform models that fall outside the standardized regulatory template (e.g. smaller platforms that do not have to conform to systemic risk assessments and more stringent reporting obligations).

These dynamics underscore a broader trade-off at the heart of the EU’s regulatory strategy. By prioritizing legal certainty, uniform standards, and accountability, the DSA lowers barriers for some types of innovation, particularly those aligned with compliance and oversight, but may simultaneously raise barriers for exploratory or decentralized models. The challenge for European policy is thus not simply to balance regulation and

innovation, but to ensure that the forms of innovation encouraged by regulation remain diverse enough to sustain long-term competitiveness.

From an innovation-economics perspective, such regulatory harmonization can shape not only the volume but also the direction of innovation. By standardizing procedures and risk-assessment requirements, the DSA may encourage complementary innovation; for instance, in compliance and auditing technologies, while discouraging more radical or product-level experimentation that conflicts with strict liability or procedural constraints [32, 51, 558]. As Galasso and Schankerman show, institutional design conditions where firms channel inventive effort, often steering it toward adaptive rather than disruptive innovation [278].

Another aspect is innovation in compliance technology itself. The DSA could spur development of better AI content filters (to meet the requirements without squelching too much speech) [59, 177, 416, 700], more sophisticated age verification and parental control systems (as platforms must consider protection of minors) [480, 585], and enhanced auditing tools (for the mandated annual risk assessments) [655]. In this sense, the DSA illustrates the dual nature of policy in innovation governance, supporting complementary innovation while constraining radical technological change. European firms might nonetheless take the lead in these compliance-oriented domains, potentially turning regulatory alignment itself into a competitive export advantage.

15.5. CONCLUSION

The Digital Services Act exemplifies the European Union's distinctive approach to digital governance using law and regulatory power as levers to steer the digital economy in line with European values and interests. Far from being merely a set of legal constraints, the DSA functions as a form of industrial policy by other means – one aimed at recalibrating the relationship between global tech platforms and the societies they serve.

Yet this ambition also raises a critical question: given the sheer scale of content decisions and the predominance of firm-specific enforcement under ToS, to what extent do these moderation outcomes actually reflect “European values and interests,” or simply the operational and commercial logics of global platforms adapting to regulatory pressure? The DSA thus operates within an uneasy balance between public oversight and private norm enforcement. Through the DSA, the EU asserts influence in a domain where it lacks native corporate giants, effectively turning regulatory leadership into a competitive asset.

In brief, the DSA demonstrates that regulation can indeed be a form of industrial strategy – setting rules is a way of setting direction. It has made the EU a central actor in digital policy debates and given it a seat at the table to which even the largest corporations must pay heed.

15.6. DISCLAIMER

Since this article was written several geopolitical changes have also influenced the international effects described in this chapter. One of them is the current reception of the US government towards the DSA. Exemplified by JD Vances reaction in December 2025 by telling off the EU and stating that “attacking American companies over garbage” was what the EU Commission was doing through issuing the first fine against the VLOP X [163, 431].

Besides a recent "Directive to Prevent the Unfair Exploitation of American Innovation" issued by President Donald J. Trump states that the US is "review whether any act, policy, or practice in the European Union or United Kingdom incentivizes U.S. companies to develop or use products and technology in ways that undermine free speech or foster censorship" and "Regulations that dictate how American companies interact with consumers in the European Union, like the Digital Markets Act and the Digital Services Act, will face scrutiny from the Administration" [570]. This geopolitical pressure also has a member state (in this case the Netherlands' Digital Service Coordinator – Authority for Consumer and Market) and NGO component as shown in the case of "EDRi member[s and] Bits of Freedom and Justice for Prosperity, among others, are called 'censorious NGOs'" by the United States House Committee on the Judiciary in July 2025 – framing the DSA as a tool of censorship of the EU [214, 367]. Besides, these acts of pressure were also issued against individuals for example against Thierry Breton as a former Commissioner and one of the founding fathers of the DSA, Imran Ahmed from the Center for Countering Digital Hate, Anna-Lena von Hoderberg and Josephine Ballon from HateAid, and Clare Melford of the Global Disinformation Index, who were banned from traveling to the US and issuing visa bans [56, 560, 587, 630].

By undermining the rule of law and the EU's sovereignty in regulating technology companies the dimensions of power, resistance to change, and potential to international controversy are showing the stark contrast between the pathways the US and the EU are stepping foot on.

These actions show the limitations of the Brussels effect as described in this chapter and show how hard the current US government tries to turn it into a *Mar-a-Lago Effect* that pulls in the different direction of the DSA, trying to hamper any regulatory action on US tech. While this chapter makes the point of opportunities of regulation, this disclaimer also shows the more recent developments that might lead to a different outcome for the DSA on an international stage. The strong US reaction however is only one side of the global effects regulation in the style of the DSA has which can be shown in the UK's Online Safety Act [569], the endeavors of Switzerland to regulate online platforms [319], or Canada's proposed Online Harms Bill [356]. These political endeavors paint a picture of the need and opportunity of regulation of online platforms for which the DSA can be seen one of the first models (next to Germany's NetzDG and Austria's KoPLG [15, 57]).

These developments reveal both the reach and the fragility of the Brussels Effect within the DSA's context. While the DSA has emerged as a powerful reference model for platform regulation, its international trajectory still unclear. The strong reaction from the US shows that regulatory influence can trigger equally forceful counter-mobilisation with the potential of geopolitical conflict. Yet the fact that other jurisdictions continue to pursue platform regulation in their own forms demonstrates that the concerns addressed by the DSA have become (shared) global governance questions. In this sense, the DSA remains a highly consequential reference point of platform governance: not because it will be copied uniformly, but because it has helped define the terms on which the regulation, instruments, and degree of governing online platforms is discussed on a main stage globally.

16

REGULATING THROUGH EVIDENCE: HOW THE DSA TURNS PLATFORM GOVERNANCE INTO AUDIT WORK

“Wer weiß’, sagt Renate und zuckt mit den Schultern. ‘Wer weiß, ob es jemals klappt.’ ‘Natürlich klappt es’, erwidert Daria reflexartig. ‘Sag das nicht immer. Ich verliere langsam den Verstand bei diesem unbegründeten Optimismus von allen Seiten.’”

“‘Who knows,’ says Renate with a shrug. ‘Who knows whether it will ever work out.’ ‘Of course it will work out,’ Daria replies reflexively. ‘Stop saying that all the time. I’m slowly losing my mind because of all this unfounded optimism from every side.’”

Paola López, *Die Summe unserer Teile*, p. 117

The EU Digital Services Act (DSA) grants the European Commission direct supervisory powers over designated very large online platforms and search engines (VLOPs/VLOSEs). Since late 2023, the Commission has published a steady stream of enforcement communications—most prominently requests for information (RFIs), but also preliminary findings and (in at least one instance) a monetary fine. These materials matter for safety and security practice because they translate legal duties into concrete evidentiary demands: what risks are prioritized, what documentation and measurements are requested, and how investigatory attention shifts over time. We compile a structured dataset of publicly available Commission enforcement communications concerning DSA supervision of VLOPs/VLOSEs, collected from official Commission channels. We provide a concise descriptive analysis focused on the safety & security problem space (e.g., illegal content and goods, protection of minors, recommender and engagement risks, election integrity, and security/privacy). Rather than evaluating mitigation effectiveness, we characterise enforcement signals: the

distribution of action types over time, recurring platform/service targets, and the substantive themes and evidence requests emphasised in public communications. We discuss implications for auditors, trusted flaggers, and compliance teams regarding risk scoping, documentation readiness, and evidence collection under regulator timelines.

16.1. INTRODUCTION

Platform safety and security work is increasingly shaped by regulation, not only by internal trust & safety policies. The EU Digital Services Act (DSA) creates due diligence obligations for online intermediaries and grants the European Commission direct supervisory and enforcement powers over designated very large online platforms (VLOPs) and very large online search engines (VLOSEs) (Art. 33) [224]. A core practical challenge is interpretive: what *exactly* regulators ask platforms to demonstrate (and how) when investigating safety- and security-relevant risks. Teams must translate legal duties into operational routines and tests, data traces, documentation and audit artefacts, and user-facing interfaces that remain legible under scrutiny. In practice, compliance is not a single deliverable but a distributed system of work—conducted by stakeholders like legal counsel, policy and safety staff, product and design, engineers, data scientists, and external auditors [114, 320] — whose outputs must align over time [665]. This Late-Breaking Work provides a compact, practitioner-oriented mapping of European Commission DSA enforcement communications as signals for day-to-day safety and security compliance work. We ask:

1. What action types dominate Commission enforcement communications, and how do they evolve over time?
2. Which platform/service categories recur as enforcement targets?
3. Which safety/security themes and evidentiary demands are most salient in public-facing communications?

16.2. BACKGROUND AND RELATED WORK

In safety- and security-sensitive domains, enforcement communications function as boundary objects between legal obligations and operational compliance work [681]. RFIs, in particular, translate abstract duties (e.g., systemic risk mitigation, recommender accountability, transparency) into concrete questions about systems, processes, and data. Preliminary findings reveal where the Commission believes legal non-compliance may exist and can foreshadow remedial pathways [160, 161]. A fine represents an escalation where the Commission determines a sanction is warranted as outlined in Article 74 DSA [163].

The DSA establishes a layered accountability architecture that combines platform-facing duties with public reporting infrastructures. For VLOPs/VLOSEs, key obligations include assessing and mitigating systemic risks (Arts. 34–35), undergoing independent audits (Art. 37) [692], and maintaining procedural mechanisms that enable reporting, explanation, and contestation of moderation decisions (Arts. 16–17–20) [629, 668]. Compliance is further enacted through infrastructures intended to enable external scrutiny, including transparency reporting templates [638], the Terms and Conditions database (Art. 14) [604], the Statement of Reasons database (Art. 17) [710], advertising repositories (Art. 39) [209, 760], and researcher access mechanisms (Art. 40) [475, 727]. This design makes accountability dependent on documentation and interoperability rather than on any single “compliance output.” Risk assessments, decision logs, transparency reports, and audit artefacts operationalise overlapping phenomena using different categories,

evidentiary thresholds, and time windows [665]. As a result, platforms can produce internally coherent documents that remain difficult to reconcile across the broader DSA reporting stack. For HCI audiences, this foregrounds a socio-technical governance problem: legal duties are realised through systems of categorisation, traceability, and interface design—each of which can introduce friction, ambiguity, or selective visibility. These enforcement actions matter because they influence:

- **Risk scoping:** which harms are treated as salient (e.g., minors’ protection, illegal goods, election integrity) [153, 162, 164, 165].
- **Evidence work:** what documentation, metrics, and audit trails are expected.
- **Operational readiness:** how teams prioritize investigations, remediation, and reporting.

16.3. METHOD

Dataset and Approach (Descriptive and Explorative, Not Evaluative) Our corpus consists of publicly available Commission communications related to DSA enforcement, including requests for information (RFIs), announcements of formal proceedings under Art. 66, procedural updates, preliminary findings, and commitments-related communications where present. We analyse these materials in relation to the DSA compliance stack (e.g., Art. 17 Statement of Reasons database; Arts. 39–40 transparency infrastructures) to identify which infrastructures are positioned as enforcement-relevant.

Unit of analysis and data extraction. Each record corresponds to a publicly available European Commission communication concerning DSA supervision/enforcement of a designated VLOP/VLOSE (e.g., RFI notice, opening of proceedings, preliminary findings, fine). For each record we extracted: publication date, platform/service, action type, headline, URL, and (where explicitly stated) cited DSA provisions and the main safety/security issue named in the text. **Descriptive topic labelling.** To support aggregation, we assigned each record one primary “target” label using the Commission’s own framing (headline and lead summary), e.g., protection of minors, illegal goods/marketplaces, advertising transparency, researcher access, recommender systems, election integrity, generative AI risks, security/privacy. This labelling AI-supported labelling is used for descriptive counts and timelines rather than interpretive claims about underlying causes. **Limitations** This work captures *published enforcement communications* and does not observe internal Commission case files (except the leaked decision [163]), platform submissions, or outcomes beyond what is publicly documented. Accordingly, we do not infer causal impact or mitigation effectiveness.

16.4. FINDINGS

RFIs dominate early enforcement and form thematic “waves.” The majority of actions in the corpus are RFIs. Over time, RFIs shift from broad compliance questions toward more specific safety/security themes (e.g., minors’ protection, illegal goods on marketplaces, recommender risks, and election-related integrity).

AI-supported topic labelling. To support descriptive aggregation of enforcement themes, we used an AI-assisted labelling workflow for the primary “target” field (e.g.,

protection of minors, illegal goods/marketplaces, advertising transparency, researcher access, recommender systems, election integrity, generative AI risks, security/privacy). For each Commission item, we provided the ChatGPT4 with the record metadata (date, platform/service, action type, headline, and lead summary text) and instructed it to assign *exactly one* primary target label from a fixed label set, prioritising the Commission's own framing in the headline/lead paragraphs. The model was also instructed to return a short justification quoting only the Commission text, and to output "UNCLEAR" if the public communication did not provide sufficient information. We reviewed the outputs and corrected labels where the assigned target did not match the Commission's framing or where multiple themes were present but one was clearly primary. This workflow is used to enable reproducible descriptive counts and timelines; it is not a qualitative coding study and does not support interpretive claims beyond the published communications (See Appendix J.0.1).

Table 16.1: European Commission DSA enforcement actions by category.

Category	Count
Requests for information (RFI)	76
Preliminary findings	8
Fines	1

Enforcement attention clusters around a small set of platform categories. Across actions, the Commission repeatedly targets large social platforms (e.g. Meta, X), video/content-sharing services (e.g. YouTube, TikTok), online marketplaces (e.g. Shein, Amazon), and adult-content services (e.g. XVideos, Pornhub). This clustering suggests recurring safety/security priorities and repeatable evidence expectations for those service types. **“Targets” emphasize evidentiary readiness.** Across themes, the Commission communications frequently ask for information that implies operational measurement and traceability (e.g., recommender logic and controls, content moderation capacity, transparency mechanisms, risk mitigation processes, and monitoring).

16.5. DISCUSSION

Nevertheless, the communication data of the Commission is only an external signal to understanding enforcement and compliance better. For example the communication does not always link to DSA provisions which can make analysis for compliance oriented teams difficult (e.g. several documents do only mention Articles 66 and 74(2) DSA for defining the RFI and not the detailed links to the potential infringements [147, 150, 151, 154, 155]). This suggests variability in how legally specific the public communications are: some enumerate multiple Articles (and occasionally paragraphs) [146, 157–159, 164], while others articulate concerns with limited or more explicit citation detail in the text segments captured.

Enforcement as a Future-of-Work Case Study Across Commission enforcement communications, enforcement functions less like “compliance checks” and more like an ongoing specification of what must be made evidentially and contextually provided

Table 16.2: How many actions targeted each platform/service.

Platform/service	RFIs	Preliminary findings	Fines	Total
TikTok	8	3	0	11
Facebook	8	1	0	9
Instagram	8	1	0	9
X	4	1	1	6
Snapchat	5	0	0	5
YouTube	5	0	0	5
App Store	4	0	0	4
Bing	4	0	0	4
Google Search	4	0	0	4
Amazon Store	3	0	0	3
Shein	3	0	0	3
AliExpress	2	1	0	3
Temu	2	1	0	3
Booking.com	2	0	0	2
Google Play	2	0	0	2
LinkedIn	2	0	0	2
Pornhub	2	0	0	2
Stripchat	2	0	0	2
XVideos	2	0	0	2
Google Maps	1	0	0	1
Google Shopping	1	0	0	1
Pinterest	1	0	0	1
Zalando	1	0	0	1

on a case-by-case basis: communications are front-loaded toward evidence extraction (RFIs/investigatory steps), repeatedly foreground governance domains like systemic risk categories and platform targets (VLOP/VLOSE), and increasingly articulate assessable domain objects (e.g. data access [149, 152, 163, 167]), while explicit Article-level anchoring is comparatively uneven in the public-facing layer (Compare e.g. [152] as an example with mentioning of the Articles 25, 39, and 40(12) and [156] with no legal anchoring). This configuration shifts compliance work to maintaining *evidence*: the standing capacity to assemble and reconcile interface behavior, measurement practices, documentation artefacts, and organizational responses on regulator timelines. Through an organizational lens, cross-functional coordination is needed (legal, policy/safety, product/design, engineering, data) and strengthens demand for cooperation across the DSA reporting stack (e.g., user-facing mechanisms, transparency infrastructures, audit artefacts).

X's DSA Fine as an Industry Case Study On 5 December 2025, the European Commission adopted a non-compliance decision and imposed fines on the provider of X (formerly Twitter) for three DSA infringements: Article 25(1) (deceptive interface design), Article 39 (advertising repository), and Article 40(12) (data access for qualified researchers) [163]. The Commission found that X's design and operation of its "verified" status feature deceived recipients into believing that verified accounts were meaningfully vetted, thereby impairing informed decision-making about account authenticity and reliability. In parallel, the Commission concluded that X's advertising repository did not meet DSA accessibility expectations and that empirical assessments indicated substantial incompleteness (e.g., studies finding that only roughly half of ads expected to be retrievable

Table 16.3: What Commission actions targeted.

Primary target	Count
Researcher access / data transparency	22
Protection of minors / age assurance	16
Illegal products / marketplace compliance	11
Ads transparency / ad repository / monetisation	10
Generative AI risks	9
Election integrity / disinformation / crisis response	5
Financial scams	5
Recommender systems transparency	3
User reporting / complaints / dark patterns	2
Content moderation resources	1
Addictive design features	1
Total (RFI + preliminary findings + fines)	85

appeared in repository results).

Across all three obligations, the Commission treated the infringements as ongoing at the date of adoption and traced their start to 28 August 2023 (the date the relevant VLOP obligations became applicable to X following designation) [146]. It concluded that the non-compliance lasted for at least 27 months. As enforcement outcomes, the Commission imposed administrative fines totalling EUR 120 million, broken down as EUR 45 million (Art. 25(1)), EUR 35 million (Art. 39), and EUR 40 million (Art. 40(12)) [163].

Finally, the leaked decision operationalized a compliance pathway through the DSA's enhanced supervision logic: X must implement measures to terminate the infringements and report on implementation, and it must submit an action plan (including an independent audit commitment) within 90 working days for remedies relating to Articles 39 and 40(12), with specific expectations such as removing unnecessary obstacles to qualified-researcher access and refraining from contractually prohibiting independent access to publicly accessible data (e.g., via scraping) through its terms.

Evidence base and reasoning. The Commission frames its findings as the product of an “overall body of evidence” rather than reliance on any single artefact, and it repeatedly rejects the provider's attempts to narrow what counts as probative material. For Article 25(1), it treats the deception claim as a straightforward inference from the interface's design and operation: conferring a “verified” signifier on accounts that are not meaningfully vetted is considered, on its face, liable to mislead recipients about authenticity and reliability. On that basis, the Commission argues that it need not conduct statistical consumer testing in every case and that publicly available/open-source material can suffice where the infringement is obvious (e.g. [163, 9, 46-47, 59-60, 71-72, 81]); it also weighs third-party studies by methodological soundness and factual relevance rather than academic seniority (e.g. relying on preprints [163, 30-34, 37, 39, 42, 87]), emphasising that such studies are only one element among multiple evidentiary strands.

For Article 39, the Commission triangulates its own trial interactions/tests of the repository's searchability and functionality and independent empirical studies on repository completeness, explicitly rebutting the argument that it relied mainly on a single “stress test” report. It discusses why certain third-party evidence is not methodologically

comparable while still using large-sample research (e.g., an ISC-PIF study [163, 59-61, 64, 78]) to support the conclusion that a substantial share of ads expected to be retrievable were missing from search results.

For Article 40, the decision relies on a wide evidentiary base that includes exchanges with more than 20 researchers and experts (with interviews with more than 15) ([163, 134]), extensive review of internal documents, and publicly available materials. Substantively, the Commission's reasoning turns on the interpretation that compliance is not only about offering access via provider-controlled channels (e.g., APIs) but also about refraining from contractual barriers that prevent qualified researchers from independently accessing publicly accessible data (including by automated means such as scraping/crawling, Art. 40 (12) DSA). It treats prior-authorization requirements and contractual prohibitions as unnecessary obstacles that create undue delay and chill legitimate research, reinforcing the finding through concrete application-handling shortcomings and generic rejection rationales.

In summary, the leaked document provides a clearer account of how the Commission assesses compliance internally than its earlier public communications. However, the Commission's criteria for selecting evidentiary sources and expert input remain opaque, creating continuing challenges for practitioners.

16.5.1. IMPLICATIONS FOR KEY STAKEHOLDERS

Auditors For auditors, the decisions and documents suggests that assurance and evidence gathering increasingly resembles constant learning, integration, and reconciliation rather than tick the box exercises. Besides, auditors must validate how platforms translate enforcement domains and risk objects into controls, mitigation, and evidence, and whether those translations remain consistent across artefacts and time.

This implies an auditing practice oriented toward collecting evidence and testing methods. It requires traceability from claims to logs and outputs under the Delegated Act [22]; consistency testing across artefacts, time, and systems; method stability across monitoring cycles; and contestability by regulators and affected parties, such as Trusted Flaggers and out-of-court dispute settlement bodies. This expands the demand for interdisciplinary auditors who can understand legal obligations, interrogate data, apply platform-specific methods, and evaluate socio-technical systems.

Regulators (Commission and DSCs) From a future-of-work angle, Commission enforcement communications function like an industry-shaping signal: they specify the forms of evidence the market must be able to produce. RFIs and investigations operate as upstream standard-setting in practice, incentivising platforms to build evidence production pipelines, and incentivising auditors to offer compliance services and specialized tooling (logging, monitoring, risk registers, repository auditors, audit-ready information and tools). On the other hand future communication would greatly benefit if the Commission would be more detailed in their communication, as well as in how it selects evidence grounding sources and experts input.

Researchers and civil society Researchers and civil society actors form an external actor around compliance scrutiny: they evaluate claims, produce counter-evidence, and pressure both regulators and platforms. Our findings suggest that their effectiveness depends less on "more disclosure" than on infrastructures that enable tailored research

to support ongoing investigations and opened proceedings. Where enforcement foregrounds domains such as minors protection, advertising transparency, or recommender-related risks, external actors need access mechanisms and artefacts that support independent testing, as shown in the opened proceedings linked to Article 40's data access demands [149, 152, 163, 167].

However, uneven legal specificity in public communications can increase asymmetric workloads: external actors must perform additional work, often without access to internal evidence or funding. This points to a tension: Testing compliance becomes concentrated among well-resourced organisations, potentially narrowing potential risk assessment. Besides, the current opened investigations are not yet explained, nor roughly evenly distributed across platforms, if researchers and civil society should provide essential work on the compliance an testing of such platforms distributing research and funding access across the palette of services and platforms to study, since domain expertise takes time and the possibility to conduct research over a longer time span.

16.6. CONCLUSION

European Commission enforcement actions under the DSA provide actionable signals about regulator priorities in platform safety and security. By structuring and summarizing these actions, we provide a compact resource for auditing and compliance work that must anticipate evidence demands and shifting enforcement focus.

17

CONCLUSION

In this final chapter, we further synthesize the findings developed across Parts 1–3 of the dissertation and bring them together to answer the overarching research question of how very large online platforms can be assessed to create DSA-compliant online spaces. We then reflect on the broader impact of this work, outlining its implications for compliance practices, platform governance, user rights, design research, auditing methodologies, and the societal functioning of digital environments. Finally, we identify key directions for future research that build on the conceptual, methodological, and empirical contributions of this dissertation and support the continued development of rigorous, interdisciplinary approaches to DSA implementation and evaluation.

17.1. EMPIRICAL FINDINGS

This thesis identifies multiple problem areas for DSA compliance across Parts 1–3 and synthesizes how the preceding chapters jointly answer the overarching research question: *How can online platforms be assessed to create DSA-compliant online spaces?* The summary below follows the dissertation's part structure and the condensed chapter-specific research questions in Table 1.2.

PART 1: CONCEPTUAL FOUNDATIONS OF DSA COMPLIANCE (RQ-A, RQ-B, RQ-C)

Part 1 conceptualises DSA compliance as a *proceduralised, multi-actor governance infrastructure* rather than a checklist. It shows how the DSA's institutional mechanisms and transparency obligations restructure platform governance (RQ-A). **A key finding of Chapter 2 is that the DSA creates a dense governance architecture in which transparency mechanisms, user rights, systemic-risk obligations, and oversight actors are structurally interconnected, making compliance a multi-layered rather than isolated legal exercise.** This Part also highlights how content moderation becomes a legally structured socio-technical process with new rights, duties, and interface-requirements (RQ-B). **A central finding of Chapter 3 is that content moderation under the DSA is best understood**

as a staged socio-legal process involving multiple actors, interfaces, and accountability points, rather than as a single platform decision. Additionally, Part 1 explains why the DSA institutionalizes structural complexity through interdependent artefacts, obligations, and socio-technical multiactor processes (RQ-C). **A core finding of Chapter 4 is that the complexity of DSA compliance is structural: it arises from the interdependence of legal rules, socio-technical systems, reporting artefacts, and auditing demands, which cannot be reduced to a simple checklist without loss of accountability.**

PART 2: METHODOLOGICAL FRAMEWORKS FOR ASSESSING DSA COMPLIANCE (RQ-D, RQ-E, RQ-F, RQ-J)

Part 2 develops methods for making compliance assessable. Chapter 5 introduces a cross-artefact *consistency* framework for oversight and auditing (RQ-D), showing that compliance gaps often appear as inconsistencies between reporting instruments. Chapter 6 develops mixed-method sampling and legal-content-analysis approaches for systemic-risk assessments and independent audits (RQ-E). Its central finding is that credible DSA auditing requires sampling, qualitative assessment, and explicit evidentiary standards because no single method can capture the varied and adaptive character of systemic risk. Chapter 7 proposes a domain-specific legal taxonomy and codebook for electoral and political content, including advertising content (RQ-F), translating abstract regulatory requirements into analysable categories. Chapter 8 evaluates where LLMs can support audit workflows and where they fail (RQ-J). It shows that LLMs can help scale annotation and analysis only when embedded in legally grounded, explainable, task-specific, human-supervised processes shaped by interdisciplinary expertise.

PART 3: EMPIRICAL EVALUATION OF DSA COMPLIANCE (RQ-G, RQ-K, RQ-L, RQ-M, RQ-I, RQ-H)

Part 3 empirically tests the conceptual and methodological contributions across systemic risks and user-facing compliance mechanisms. Chapter 9 shows that platform moderation decisions frequently diverge from legally coded ground truth, revealing persistent tensions between legal accuracy, contextual interpretation, and time-constrained moderation (RQ-G). Chapter 10 finds that reporting interfaces remain difficult to navigate and legally opaque; formal notice mechanisms therefore do not automatically translate into effective user rights under Art. 16 DSA (RQ-K). Chapter 11 shows that LLM assistance can improve aspects of reporting quality but can also distort legal categorisation and statements of reasons, making AI-supported user empowerment inherently ambivalent (RQ-L). Chapter 12 finds that Trusted Flaggers have become important intermediary compliance actors, although platform frictions, capacity constraints, coordination problems, funding needs, and ambiguity concerning evidence and accountability limit their practical effectiveness (RQ-M). Chapter 13 shows how fundamental-rights risks under Art. 34(1)(b) DSA can be operationalised through a structured assessment process involving context specification, subsystem mapping, stakeholder identification, and risk scoring (RQ-I). Chapter 14 demonstrates how a DSA-aligned empirical risk assessment can translate systemic risks into concrete categories, including illegal content, negative effects on electoral rights, and disinformation affecting electoral processes, thereby making DSA risk assessment empirically tractable (RQ-H).

PART 4: SYNTHESIS, GOVERNANCE IMPLICATIONS, AND ENFORCEMENT LESSONS (RQ-N, RQ-O)

Part 4 consolidates the dissertation's findings into a multidimensional account of DSA compliance and situates them in the broader governance context. Chapter 15 shows that the DSA does not merely regulate platforms retrospectively; it also reshapes industry incentives, governance strategies, and accountability infrastructures across jurisdictions, situating platform governance within an internationally contested context (RQ-N). Chapter 16 finds that Commission enforcement remains fragmented but nevertheless steers platform governance by signalling which risks, metrics, and forms of documentation must become part of routine compliance practice (RQ-O). The overall implication is that robust compliance assessment depends on interoperable artefacts, standardised interdisciplinary audit methods, and sustained empirical monitoring of systemic risks.

17.2. METHODS AND MEASUREMENTS FOR ONLINE PLATFORMS, REGULATORS, AND AUDITORS

Ensuring meaningful compliance with the Digital Services Act (DSA) requires methods that are both legally grounded and empirically workable across diverse sets of obligations and actors. This section synthesizes the methodological contributions developed throughout this dissertation and outlines how they can be applied by online platforms, regulators, and independent auditors to measure, monitor, and evaluate compliance in a systematic manner. Together, these methods form a multi-dimensional toolkit for assessing systemic risks, auditing platform behaviour, analyzing political and electoral content, and integrating AI-assisted evaluation responsibly.

First, the dissertation advances a **mixed-methods framework for systemic risk assessment and independent audits**. Because the DSA mandates the identification, analysis, and mitigation of systemic risks (Arts. 34–35) and requires regular independent audits (Art. 37), platforms and auditors must combine quantitative and qualitative approaches. The framework developed here integrates:

- *statistical sampling*, necessary to ensure representativeness and detect platform-wide patterns;
- *legal content analysis*, enabling evaluators to classify risks with reference to established literature, national, and EU legal standards;
- *qualitative evaluation*, including contextual review and annotation; and
- *consistency analysis*, comparing reporting artefacts across internal, external, and historical dimensions.

This multi-layered approach enables a richer and more reliable assessment of platform compliance than any single method could achieve on its own.

Second, the dissertation provides a **legal codebook for electoral and political content** especially relevant for advertising content, addressing a gap in systemic risk assessment research. Political and election-related risks are central to the DSA's systemic risk framework, yet platforms and auditors lack standardized categorization schemes aligned with

EU law, the regulation for Transparency and Targeted Advertising (TTPA) [236], and the Commission's Guidelines on electoral risks [229]. The codebook introduced here translates abstract legal obligations into operational annotation categories, enabling researchers, platforms, and auditors to conduct systematic evaluations of political content, electoral integrity risks, and related moderation decisions, and to create other legal-based codebooks. This tool also supports the construction of transparent and reproducible audit trails, strengthening the evidence creation of risk assessments.

Third, the dissertation evaluates the **role of Large Language Models (LLMs) as auditing and user-support instruments**. While LLMs hold the potential to scale up parts of the auditing workflow, such as coding explanations, summarizing artefacts, or assisting users in reporting illegal content—the findings emphasize the need for strict safeguards. LLMs demonstrate measurable improvements in supporting users in reporting content tasks, but they also introduce risks such as hallucination, and error propagation. The methods developed in this dissertation therefore outline conditions under which LLMs can be incorporated into DSA compliance processes: constrained prompting schemes, human-in-the-loop oversight, explanation quality controls, and systematic evaluation of processes. This provides different stakeholders with guidance on how AI assistance can augment, rather than undermine, compliance efforts.

Taken together, the methods and measurement strategies introduced in this dissertation provide the foundations for a **rigorous, reproducible, and scalable approach to DSA compliance evaluation**. By integrating legal reasoning, empirical techniques, and emerging AI capabilities, these methods help operationalize the DSA's overarching goal: the creation of safer, transparent, and rights-respecting online environments.

17.3. SOCIETAL IMPACT

The societal impact of this dissertation follows directly from the novelty and contribution claims in Chapter 1: it strengthens the DSA's capacity to deliver real-world public value by turning an ambitious legal regime into something that can be *implemented, measured, and contested* in practice. At a moment when the DSA is moving from rule-making to active enforcement, the thesis reduces the gap between law-on-the-books and governance-in-action by providing evaluation logics and empirical evidence that help users, civil society, regulators, and auditors scrutinise platform behaviour beyond self-reporting.

A first societal contribution is to **support enforceable user rights and due process online**. By treating design and interface choices as compliance-critical (rather than as “mere UX”), the dissertation shows where reporting pathways, explanation practices, and redress mechanisms fail to be accessible, comprehensible, or usable—and why this undermines the DSA's promise of meaningful participation in content moderation. This work is societally relevant because it translates procedural rights into practical conditions: whether ordinary users can submit actionable notices, understand the basis of decisions, and contest outcomes. These insights directly matter for the fairness, predictability, and contestability of moderation decisions that affect everyday speech, safety, and dignity online.

Building on the dissertation's novel emphasis on *legal design and proceduralisation*, the findings show where notice-and-action, explanation-of-reasons, and complaint path-

ways fail to deliver accessibility, clarity, and due process as envisioned by the DSA. The contribution is not only diagnostic: the work translates user-rights obligations into evaluative criteria and design-relevant evidence, including insights on when AI assistance can improve legal categorisation and when it introduces automation bias or error propagation. This supports more meaningful user agency in content moderation processes that have historically been opaque and difficult to contest.

The dissertation's empirical work on Trusted Flaggers and reporting infrastructures clarifies how new roles operate under real constraints—fragmented channels, feedback gaps, and resource pressures, and where accountability blind spots emerge in existing transparency artefacts. This matters societally because it affects whether civil society can reliably function as an enforcement multiplier and whether the DSA's multi-actor model delivers meaningful participation beyond formal status.

Second, the dissertation contributes to **public accountability and democratic oversight** by strengthening the evidentiary basis on which enforcement can occur. It engages with the interpretive and infrastructural work that makes the DSA governable—through involvement in classification and transparency infrastructures and by aligning methodological development with emerging enforcement questions. The societal effect is a reduction of information asymmetries: instead of platforms' transparency functioning primarily as disclosure, the thesis advances approaches that allow verification across artefacts, reporting cycles, and institutional interfaces. This supports regulators and auditors in asking “the evidence” questions, and supports researchers and civil society in making more robust, evidence-based claims about platform responsibility.

The dissertation's core methodological novelty—cross-artefact *consistency analysis* combined with mixed-method sampling and legal content analysis—enables stronger verification of platform reporting and more robust evidence production for audits. By demonstrating how compliance gaps emerge across outcomes *and* across mechanisms (reporting interfaces, transparency artefacts, risk assessments), the thesis supports a shift from trust-by-disclosure to accountability-by-verification. The evaluation of LLMs further contributes by specifying the safeguards required for responsible integration of AI into compliance and auditing workflows.

Extending the thesis's novelty in *domain-specific operationalisation*, the dissertation provides a legal codebook and empirical assessment approach for election-related and political content risks. This makes the DSA's systemic-risk obligations for democratic processes actionable: regulators, auditors, and civil society can use these instruments to identify recurring patterns, assess different forms of risks, and scrutinise whether platforms' claims align with observable evidence during electoral contexts. The impact is practical oversight capacity, not only descriptive analysis.

Third, the dissertation has impact through its **interdisciplinary translation of the DSA into operational governance practice**. By bridging legal doctrine with HCI/legal design, organisational studies, and empirical auditing methods, it contributes tools that can be used by different actors in the compliance ecosystem: design-relevant criteria for platforms, audit-ready methods for independent evaluators, and interpretable evidence for public authorities. This matters societally because DSA implementation depends on coordination across communities that do not naturally share concepts, metrics, or standards. The thesis provides common analytical “interfaces” that make it easier to

align technical practice with fundamental-rights obligations, and to evaluate whether platforms' measures actually mitigate systemic risks rather than merely re-describing them.

Where AI is used to support reporting or auditing, the dissertation's novelty lies in specifying *both* opportunities and challenges: LLMs can support users in reporting illegal content under legal uncertainty, but also introduce automation bias and error propagation. This contributes to more accountable adoption of AI in compliance ecosystems, aligning "scaling up" with due-process and audit requirements rather than undermining them.

Taken together, the dissertation's societal impact lies in making online governance more reliable and trustworthy: user rights become more actionable, oversight becomes more evidence-based, and enforcement becomes more feasible across the complex, multi-actor infrastructures the DSA creates. By contributing methods and empirical insights that fit the realities of Commission-led supervision and platform-scale socio-technical systems, the work supports safer online environments, more accountable platform power, and more resilient democratic public spheres.

17.4. RECOMMENDATIONS

The DSA can only deliver the aspiration of meaningful compliance if compliance is assessed and implemented as a socio-technical governance problem rather than as a purely legal one. This dissertation shows that compliance is produced through the interaction of legal rules, interface and system design, socio-technical processes, and oversight infrastructures (Chapters 2, 3, 4). An actionable interdisciplinary agenda therefore needs to generate evidence, methods, and standards that different actors in the DSA compliance ecosystem can actually use. The recommendations below specify priorities by stakeholder.

For **platforms**, the recommendation is to operationalise DSA procedures into *consistent compliance reporting* and maintain the evidence needed to demonstrate them. Reporting and redress should be implemented as coherent parts of the moderation process: users should be able to distinguish illegality from ToS reporting pathways, understand what information is required, and receive structured outcomes and reasons in a predictable format (Chapter 10). If AI is used to support user reporting, it should function as decision support with explicit uncertainty signals and it should be routinely evaluated for automation bias and error propagation (Chapter 11). Platforms should also maintain an auditable trace from the legal obligation to the design of reporting mechanisms, so that compliance can be evidenced to auditors and authorities (Chapters 4, 10, 5).

For **Digital Service Coordinators (DSCs) and the European Commission**, the priority is to stabilise shared evaluation logics that reduce uncertainties. This includes developing enforceable, evidence-based expectations and other forms of standards and templates to assess compliance (Chapter 10). For systemic risk oversight, regulators should require audit-evidence sections in systemic risk assessment and independent audit reports with declared sampling frames, documented measurement choices, and cross-artefact consistency checks to test whether reporting artefacts align with observed platform behaviour (Chapters 5, 6, 14). Where LLMs are used in auditing or monitoring, authorities should mandate minimum verification and documentation standards to ensure that AI-assisted

outputs remain contestable (Chapter 8). Regulators should also develop better support strategies for civil-society organisations, Trusted Flaggers, and researchers when relying on them to perform essential oversight and evidence-gathering tasks.

For **independent auditors**, the main recommendation is to prioritize methods that generate replicable, contestable evidence. The dissertation demonstrates that mixed-method approaches, sampling plus legal content analysis plus cross-artefact consistency testing, are necessary to evaluate systemic risk assessments and to test the plausibility of platform reporting (Chapters 5, 6). Where research targets electoral and political content, domain-specific operationalisations (codebooks and risk taxonomies) are essential to make results interpretable for regulators and other stakeholders (Chapter 7). When using LLMs to scale analysis, auditors should treat them as assistive components within controlled pipelines, benchmarking performance and documenting failure in a transparent way (Chapter 8).

For **Trusted Flaggers and civil society organisations**, the priority is to leverage their role to produce structured evidence of problems on platforms. Trusted Flaggers should document recurring challenges with platforms, e.g., unavailable categories, inconsistent feedback, obscure prioritization, and interface problems, and translate them into constant communication to DSCs and the Commission as governance-relevant evidence – beyond the DSA’s yearly Trusted Flagger reports (Chapter 10). Besides, Trusted Flaggers should come together in ongoing exchanges across their domains and member states to coordinate and strategize for impact.

For **end users**, the practical priority is that DSA rights are only meaningful if people can actually use them. Platforms and regulators should therefore evaluate reporting and appeal pathways from the user’s perspective: can users find the right channel, understand what to submit, and receive reasons and outcomes they can act on? Chapter 10 shows how fragmented entry points and unclear legal categories undermine this in practice; Chapter 11 further indicates that support tools can improve performance but also introduce new mistakes and reliance effects. The implication is straightforward: reduce procedural friction (fewer steps, clearer legal categories, clearer explanations) and treat user experience evidence as part of compliance evidence linked to the systemic risk of disseminating illegal content (Chapters 10, 11).

For **researchers**, the recommendation is to produce compliance evidence that is comparable, reproducible, and usable for oversight. Chapter 5 motivates cross-artefact consistency checks as a way to test whether platforms’ claims align across reporting instruments, while Chapter 6 and Chapter 7 provide the methodological basis for generating auditable measurements (sampling frames, operational definitions, and domain-specific coding schemes). Empirically, Chapters 14, 9, 13, illustrates how these components can be combined in a concrete assessment, and Chapter 8 clarifies when LLMs can responsibly scale analysis and when they compromise verifiability. Researchers should therefore prioritize sharing best practices and protocols, codebooks, sampling rationales, and documented pipelines so that findings can be replicated, challenged, and integrated into regulatory and audit evidence (Chapters 5, 6, 7, 14, 8).

Taken together, these role-specific recommendations point to a practical interdisciplinary agenda: develop shared evaluation protocols that different actors can apply to the same compliance problem and that produce evidence that can travel across contexts

and assessment domains. In this dissertation, the legal foundations and governance logic (Part 1) are translated into assessment methodologies (Part 2) and then tested in empirical studies (Part 3), showing how doctrinal interpretation, design evaluation, and quantitative/qualitative auditing methods can be combined into workflows that regulators, auditors, platforms, researchers, and civil society can actually use. This matters because DSA effectiveness will depend on whether compliance claims can be checked consistently across platforms and over time, and whether findings can be integrated into supervision and enforcement rather than remaining discipline-bound research outputs.

17.5. FUTURE WORK AND LIMITATIONS

17.6. LIMITATIONS

This dissertation is necessarily time-bounded by the DSA's early implementation phase. Several obligations central to the thesis—particularly systemic risk assessment and the assessment of impacts on fundamental rights—remain operationally under-specified and are likely to be clarified through guidance, standardization, audit practice, and enforcement decisions. As a result, parts of the dissertation translate legal duties into evaluative criteria that require interpretive judgment (e.g., what counts as “sufficient” evidence for a risk claim, what level of specificity is needed to make mitigations testable, and how vulnerable positions should be operationalized). These translation choices are transparent and motivated, but they may need revision as regulatory practice matures and as shared baselines for evidentiary adequacy emerge.

The empirical work is also constrained by the persistent asymmetry of access between platforms and external evaluators. Many compliance-relevant questions depend on provider-controlled artefacts—internal logging, taxonomies, measurement pipelines, and organizational documentation—that are not consistently available to researchers and that cannot be reconstructed from public-facing transparency materials alone. This limits the ability to validate certain claims end-to-end (for example, whether reported metrics are derived in consistent ways across time, teams, and product surfaces, or how particular notice-and-action inputs travel through internal classification and decision systems).

A further limitation follows from the dissertation's methodological design choice to privilege auditability and comparability. Framework-based consistency analysis, structured coding, and decision-tree representations make complex compliance claims inspectable and reproducible, but they also necessarily abstract from the continuous, adaptive, and interactive character of platform governance. Socio-technical effects often arise from interactions between recommender systems, moderation operations, policy shifts, user adaptation, and external events, and these interactions can be difficult to capture fully in discrete categories or staged assessments. This means that some nuanced contextual factors may be underrepresented, and some interaction effects across subsystems may only be partially captured.

Compliance is also not an entirely objective property waiting to be measured. Decisions about what counts as adequate transparency, acceptable risk, sufficient evidence, or effective mitigation reflect legal interpretation, institutional priorities, and political choices. The methods developed in this dissertation make those choices more explicit and contestable, but they do not eliminate them. Measurements should therefore be

read as situated assessments whose categories, thresholds, and evidentiary assumptions remain open to scrutiny by regulators, platforms, researchers, civil society, and affected users.

Relatedly, the generalizability of specific empirical findings is bounded by the chosen cases, platforms, and contexts. While the cases are selected to be analytically meaningful for DSA obligations and to stress-test evaluative methods, platform architectures, product surfaces, and Member State contexts differ substantially. User-facing mechanisms are especially sensitive to language, interface conventions, and localised legal and organisational settings. The dissertation therefore aims to generalize primarily at the level of evaluative method—what to check, how to demand evidence, and how to make claims auditable — rather than asserting that observed patterns will replicate uniformly across all platforms or jurisdictions.

The dissertation does not provide a comprehensive doctrinal account of freedom of expression or of all fundamental rights affected by content moderation. It focuses instead on making selected DSA duties empirically assessable. Future work should examine more directly how competing rights claims shape the classification of harmful and illegal content, the proportionality of moderation measures, and the design of fundamental-rights impact assessments. This would connect the dissertation's evaluative methods more closely to the substantive rights that give those assessments legal meaning.

The studies also observe compliance through selected public artefacts and interactions; they cannot isolate regulation as the sole cause of a platform practice. Business models, monetisation incentives, advertising markets, organisational capacity, reputational pressures, geopolitical conflict, and strategic responses to enforcement may all shape the same outcomes. These conditions are not background noise: they can influence which risks are measured, which mitigations are funded, and which evidence is disclosed. The dissertation identifies where these factors may affect interpretation, but a full causal account of their influence lies outside its empirical design.

Vulnerability should likewise be examined not only as a contextual attribute of users but also as a potential object of monetisation. Engagement-based business models may profit from attention generated by polarising, harmful, or otherwise risky content, thereby shaping both exposure to harm and the incentives to mitigate it. Assessing compliance therefore requires attention to how platform revenue models interact with risk identification, moderation priorities, and the distribution of protection across user groups.

Finally, where the dissertation engages with automation and LLM-supported components, it inherits well-known limitations of such systems, including sensitivity to prompts and framing, evaluation boundaries that may not reflect long-run use, and the risk that automation shifts the locus of contestability from substantive decisions to the design of the evidence pipeline itself. If automated tools are used to scale compliance work, then their own logging, governance, and validation become part of what must be audited. Across these limitations, the dissertation's contribution is best understood as providing a rigorous, transparent scaffolding for evidence-based, contestable compliance assessment under evolving conditions, rather than a final and stable set of compliance thresholds.

17.7. FUTURE WORK

While this dissertation lays conceptual, methodological, and empirical foundations for assessing Digital Services Act (DSA) compliance, it also highlights several areas where sustained research is necessary. The DSA introduces long-term governance challenges that will evolve as platforms adapt, regulatory expectations mature, and new technologies reshape online environments. Future work must therefore build on the insights developed here by expanding methodological capacity, deepening empirical evidence, and strengthening interdisciplinary collaboration.

First, there is a pressing need for research on the **responsible scaling of AI and large language models (LLMs)** in compliance ecosystems. As shown in this dissertation, LLMs can assist users in reporting illegal content under Article 16 DSA and support parts of auditing workflows. They can expand capacity, but they also introduce new risks, may provide error-prone advice, and can obscure how tasks are performed. Future research should develop standards for AI-assisted auditing, documentation practices, reliability thresholds, and meaningful human oversight.

Second, future studies should pursue **longitudinal monitoring of systemic risks** across multiple election cycles, regulatory reporting periods, and platform-policy changes, including policies for generative-AI content. The DSA's systemic-risk obligations (Arts. 34–35 and 37) are bound to the time and context of evaluation and require continuous assessment and adaptation. Longitudinal testing can reveal how platform mitigation techniques evolve, how risks shift in response to social and technical dynamics, and where mitigation strategies succeed or fail. Such monitoring is critical for determining whether DSA compliance measures produce sustained changes in platform safety and accountability.

Third, there is substantial value in developing **cross-jurisdictional comparisons**. As other regions introduce legislation inspired by or parallel to the DSA, such as the UK Online Safety Act [569], comparative analyses can illuminate how different regulatory architectures shape platform compliance. Research should explore how compliance practices transfer across jurisdictions, how multinational platforms navigate competing obligations, and where international alignment or divergence emerges in transparency reporting, systemic-risk mitigation, interface-design requirements, and audit obligations.

Fourth, future work must address persistent challenges of **data access and auditability**. Effective oversight depends on researchers, regulators, civil society, and auditors obtaining meaningful data, including content samples, sufficiently complete datasets, access for studying complex algorithmic systems such as recommender systems, risk-assessment rationales, and thorough documentation. Secure research infrastructures, privacy-preserving access mechanisms, standardised reporting schemas, and auditable logs will be central to enabling scrutiny. Collaboration with regulators and civil society will be essential to ensure that the DSA's data-access provisions translate into practical and safe research opportunities.

Fifth, future research should examine the **political economy of compliance**. Comparative studies could test how advertising and subscription models, monetisation incentives, organisational resources, market concentration, enforcement exposure, and geopolitical pressure affect investment in moderation, risk assessment, and evidence production. Combining compliance data with organisational, economic, and policy evidence would

help distinguish regulatory effects from changes driven by business strategy or international politics.

Ongoing dialogue with platforms and regulators is also needed to understand context-dependent challenges and the evidence used in concrete enforcement cases. Where confidentiality permits, greater methodological transparency about evidence selection and analysis would help researchers concentrate on material questions and align research agendas with genuine oversight needs.

Finally, future research should expand the interdisciplinary agenda initiated in this dissertation by developing **shared methods, evaluation standards, and open tools** for DSA compliance assessment. Harmonising sampling strategies, content-analysis codebooks, consistency metrics, and user-interface evaluation frameworks will support comparability across institutions and jurisdictions. Experimental work on user rights, empirical studies of governance processes, and technical research on algorithmic systems will remain vital for strengthening the evidence base on platform accountability.

In sum, the DSA marks the beginning of a new phase of platform governance. Evaluating its effects requires ongoing research that is empirically grounded, methodologically rigorous, and attentive to the interplay of legal, technical, economic, political, and societal factors.

17.8. FINAL REFLECTIONS

This dissertation set out to answer *How can online platforms be assessed to create DSA-compliant online spaces?* Its core insight is that DSA compliance is not a static state but a socio-technical governance practice that must be made visible, measurable, and aligned with legal rules, platform design, socio-technical processes, and oversight infrastructures. The path forward is therefore empirical and design-centred: developing shared assessment templates and standards, testing them across platforms and systemic-risk contexts, and translating findings into enforceable evidentiary requirements so that the DSA's transparency, due-process, and systemic-risk obligations produce verifiable improvements in online environments.

BIBLIOGRAPHY

- [1] [n. d.]. API and Schema - DSA Transparency Database. <https://transparency.dsa.ec.europa.eu/page/api-documentation>
- [2] [n. d.]. Bürgerservice Hessenrecht - 2-03 O 188/21 | LG Frankfurt 3. Zivilkammer | 08.04.2022 | Urteil. <https://www.rv.hessenrecht.hessen.de/bshe/documnt/LARE220002783>
- [3] [n. d.]. Consumption of Online News Rises in Popularity. <https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20220824-1>.
- [4] [n. d.]. General Data Protection Regulation (GDPR) – Legal Text. <https://gdpr-info.eu/>.
- [5] [n. d.]. Guidelines 3/2025 on the Interplay between the DSA and the GDPR | European Data Protection Board. https://www.edpb.europa.eu/our-work-tools/documents/public-consultations/2025/guidelines-32025-interplay-between-dsa-and-gdpr_en.
- [6] [n. d.]. Home - DSA Transparency Database. <https://transparency.dsa.ec.europa.eu/>
- [7] 1881. Loi du 29 juillet 1881 sur la liberté de la presse - Légifrance.
- [8] 1979. The Sunday Times v. the United Kingdom (No. 1). European Court of Human Rights. <https://hudoc.echr.coe.int/eng?i=001-57584> Application No. 6538/74, Judgment of 26 April 1979, Series A no. 30, § 49.
- [9] 1990. The Queen v Minister of Agriculture, Fisheries and Food and Secretary of State for Health, ex parte: Fedesa and others. Court of Justice of the European Union. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A61988CJ0331> Case C-331/88, Judgment of 13 November 1990, ECLI:EU:C:1990:391, ECR 1990 I-04023, para. 13.
- [10] 2000. Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin. <http://data.europa.eu/eli/dir/2000/43/oj>
- [11] 2000. Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation. <http://data.europa.eu/eli/dir/2000/78/oj>

- [12] 2003. Eugen Schmidberger, Internationale Transporte und Planzüge v Republik Österreich. Court of Justice of the European Union. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A62000CJ0112> Case C-112/00, Judgment of 12 June 2003, ECLI:EU:C:2003:333, ECR 2003 I-05659, paras 77, 81–82.
- [13] 2010. Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32010L0013>
- [14] 2014. Digital Rights Ireland Ltd v Minister for Communications, Marine and Natural Resources and Others and Kärntner Landesregierung and Others. Court of Justice of the European Union. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A62012CJ0293> Joined Cases C-293/12 and C-594/12, Judgment of 8 April 2014, ECLI:EU:C:2014:238, paras 52, 54–55.
- [15] 2017. Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken – Network Enforcement Act.
- [16] 2018. Directive (EU) 2018/1808 of the European Parliament and the Council of 14 November 2018 amending Directive 2010/13/EU on the Coordination of Certain Provisions Laid down by Law, Regulation or Administrative Action in Member States Concerning the Provision of Audiovisual Media Services (Audiovisual Media Services Directive) in View of Changing Market Realities.
- [17] 2019. Eva Glawischnig-Piesczek tegen Facebook Ireland Limited. <https://eur-lex.europa.eu/legal-content/NL/TXT/?uri=CELEX:62018CJ0018>
- [18] 2019. Glawischnig-Piesczek v. Facebook Ireland Limited, Case C-18/18. Court of Justice of the European Union.
- [19] 2020. Proposal for Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM(2020) 825 final, 2020/0361 (COD). <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN>
- [20] 2021. Revised version of the German Network Enforcement Act (NetzDGÄndG).
- [21] 2021. Senate Bill 7072 (2021) - The Florida Senate. <https://www.flsenate.gov/Session/Bill/2021/7072/>
- [22] 2023. Commission Delegated Regulation (EU) 2024/436 of 20 October 2023 supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council, by laying down rules on the performance of audits for very large online platforms and very large online search engines. http://data.europa.eu/eli/reg_del/2024/436/oj Legislative Body: CNECT, COM.

- [23] 2024. Commission on Guidelines for providers of Very Large Online Platforms and Very Large Online Search Engines on the mitigation of systemic risks for electoral processes pursuant to the Digital Services Act. https://ec.europa.eu/newsroom/repository/document/2024-13/C_2024_2121_1_EN_annexe_acte_autonome_cp_part1_v3_tpHHZgYyBGFMF8J5rE00R1GdOis_103911.pdf
- [24] 2024. Regulation (EU) 2024/900 of the European Parliament and of the Council of 13 March 2024 on the transparency and targeting of political advertising (Text with EEA relevance). <http://data.europa.eu/eli/reg/2024/900/oj/eng> Legislative Body: EP, CONSIL.
- [25] 2025. Texas House Bill 20. <https://capitol.texas.gov/tlodocs/872/billtext/html/HB00020F.HTM>
- [26] 2026. Grok limits AI image editing to paid users after nudes backlash. *Le Monde* (Jan. 2026). https://www.lemonde.fr/en/pixels/article/2026/01/09/grok-limits-ai-image-editing-to-paid-users-after-nudes-backlash_6749258_13.html
- [27] 2026. Grok's nudes test Europe's resolve to stand up to US tech. <https://www.politico.eu/article/grok-nude-test-europe-resolve-us-tech-elon-musk/>
- [28] ACLU, EFF, Irina Raicu, Sarah T. Roberts, CDT, Open Technology Institute, Nicolas Suzor, and Sarah Myers West. 2018. Santa Clara Principles on Transparency and Accountability in Content Moderation. <https://santaclaraprinciples.org/images/scp-og.png>
- [29] AG Bonn. 2016. Urteil vom 10.08.2016. <https://openjur.de/u/2154100.html>
- [30] AG Hannover. 2020. Urteil vom 03.02.2020. <https://openjur.de/u/2316520.html>
- [31] BfJ Press agency. 2019. Federal Office of Justice Issues Fine against Facebook. https://www.bundesjustizamt.de/DE/Presse/Archiv/2019/20190702_EN.html;jsessionid=2991FA58C1ACD667E932F1E75AC007FB.2_cid370?nn=3451904
- [32] Philippe Aghion et al. 2005. Competition and Innovation: An Inverted-U Relationship. *The Quarterly Journal of Economics* 120, 2 (2005), 701–728.
- [33] Muhammad Ali, Piotr Sapiezynski, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2021. Ad Delivery Algorithms: The Hidden Arbiters of Political Messaging. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 13–21. <https://doi.org/10.1145/3437963.3441801>
- [34] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. Understanding the

- Effect of Deplatforming on Social Networks. In *13th ACM Web Science Conference 2021*. ACM, Virtual Event United Kingdom, 187–195. <https://doi.org/10.1145/3447535.3462637>
- [35] AliExpress. 2024. *Digital Services Act Risk Assessment and Mitigation Measures Report*. Risk Assessment Report. AliExpress. <https://files.alicdn.com/tps/service/3679fe26faf49a587280bd28a1c70020.pdf>
- [36] Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan D. Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2024. Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science* 8, 1 (Dec. 2024), 17. <https://doi.org/10.1007/s42001-024-00345-9>
- [37] Lulwah Alkulaib, Abdulaziz Alhamadani, Taoran Ji, and Chang-Tien Lu. 2020. Collect Ethically: Reduce Bias in Twitter Datasets. In *Information Management and Big Data (Communications in Computer and Information Science)*, Juan Antonio Lossio-Ventura, Nelly Condori-Fernandez, and Jorge Carlos Valverde-Rebaza (Eds.). Springer International Publishing, Cham, 106–114. https://doi.org/10.1007/978-3-030-46140-9_11
- [38] Garrett Allen, Mike Beijen, David Maxwell, and Ujwal Gadiraju. 2023. In a Hurry: How Time Constraints and the Presentation of Web Search Results Affect User Behaviour and Experience. In *Web Engineering: 23rd International Conference, ICWE 2023, Alicante, Spain, June 6–9, 2023, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 221–235. https://doi.org/10.1007/978-3-031-34444-2_16
- [39] Rea Alonzo, Junayd Hussain, Saverio Stranges, and Kelly K Anderson. 2021. Interplay between social media use, sleep quality, and mental health in youth: A systematic review. *Sleep medicine reviews* 56 (2021), 101414.
- [40] Heldt (Amélie). 2019. Reading between the lines and the numbers: an analysis of the first NetzDG reports. <https://doi.org/10.14763/2019.2.1398>
- [41] Matthew Andreotta, Robertus Nugroho, Mark J. Hurlstone, Fabio Boschetti, Simon Farrell, Iain Walker, and Cecile Paris. 2019. Analyzing Social Media Data: A Mixed-Methods Framework Combining Computational and Qualitative Text Analysis. *Behavior Research Methods* 51, 4 (Aug. 2019), 1766–1781. <https://doi.org/10.3758/s13428-019-01202-8>
- [42] A. Annoni et al. 2025. Artificial Intelligence: A European Perspective. *AI Flagship Report* (2025). European Commission.
- [43] Anthropic. 2025. Claude Language Model. <https://claude.ai/>
- [44] Naomi Appelman and Paddy Leerssen. 2022. On "Trusted" Flaggers. *Yale Journal of Law & Technology* 24 (2022), 453–474. https://yjolt.org/sites/default/files/0_-_appelman_leerssen_-_on_trusted_flaggers.pdf

- [45] Niels Appermont. 2025. 'A conceptual framework on legal complexity'. *The Theory and Practice of Legislation* 13, 2 (May 2025), 236–263. <https://doi.org/10.1080/20508840.2025.2515804> Publisher: Routledge.
- [46] National Archives. 1791. The Constitution of the United States. , Amendment I pages. www.archives.gov/founding-docs/bill-of-rights
- [47] Carolina Are. 2024. The assemblages of flagging and de-platforming against marginalised content creators. *Convergence* 30, 2 (April 2024), 922–937. <https://doi.org/10.1177/13548565231218629> Publisher: SAGE Publications Ltd.
- [48] Carolina Are. 2025. Flagging as a silencing tool: Exploring the relationship between de-platforming of sex and online abuse on Instagram and TikTok. *New Media & Society* 27, 6 (June 2025), 3577–3595. <https://doi.org/10.1177/14614448241228544> Publisher: SAGE Publications.
- [49] Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 2024. The Silicon Ceiling: Auditing GPT's Race and Gender Biases in Hiring. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. <https://doi.org/10.1145/3689904.3694699>
- [50] Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameya Bhatawdekar, Guillaume Bouchard, and Isabelle Augenstein. 2023. Detecting Harmful Content on Online Platforms: What Platforms Need vs. Where Research Efforts Go. *ACM Comput. Surv.* 56, 3 (Oct. 2023), 72:1–72:17. <https://doi.org/10.1145/3603399>
- [51] Kenneth J. Arrow. 1962. Economic Welfare and the Allocation of Resources for Invention. *The Rate and Direction of Inventive Activity* (1962), 609–625.
- [52] W. Brian Arthur. 2009. *The Nature of Technology: What It Is and How It Evolves*. Free Press.
- [53] Kelly A. Aschbrenner, Emily R. Haines, Gina R. Kruse, Ayotola O. Olugbenga, Annette N. Thomas, Tanveer Khan, Stephanie Martinez, Karen M. Emmons, and Stephen J. Bartels. 2024. Applying cognitive walkthrough methodology to improve the usability of an equity-focused implementation strategy. *Implementation Science Communications* 5, 1 (Sept. 2024), 95. <https://doi.org/10.1186/s43058-024-00630-8>
- [54] Marco Aspromonte, Andrea Ferraris, Federico Galli, and Giuseppe Contissa. 2024. LLMs to the Rescue: Explaining DSA Statements of Reason with Platform's Terms of Services. In *Proceedings of the Natural Legal Language Processing Workshop 2024* (2024). Association for Computational Linguistics, Miami, FL, USA, 205–215. <https://doi.org/10.18653/v1/2024.nllp-1.17>
- [55] Evelyn Mary Aswad. 2018. The Future of Freedom of Expression Online. *Duke Law & Technology Review* 17 (2018), 26.

- [56] Evelyn Austin and Jelle Postma. 2026. Our reaction to US pressure on the Digital Services Act. <https://www.bitsoffreedom.nl/2026/02/04/our-reaction-to-us-pressure-on-the-digital-services-act/>
- [57] Austrian Federal Government. 2020. Bundesgesetz über Maßnahmen zum Schutz der Nutzer auf Kommunikationsplattformen (Kommunikationsplattformen-Gesetz – KoPl-G – Communication Platforms Act). <https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20011415>
- [58] Brooke Auxier, Monica Anderson, et al. 2021. Social media use in 2021. *Pew Research Center* 1, 1 (2021), 1–4.
- [59] Ivan Bakulin et al. 2025. FLAME: Flexible LLM-Assisted Moderation Engine. arXiv preprint arXiv:2502.09175.
- [60] Jack M. Balkin. 2014. Old-School/New-School Speech Regulation. *Harvard Law Review* 127, 8 (2014), 2296–2342.
- [61] Sharon Bar-Ziv. 2021. A Content Analysis Approach to Intellectual Property Research. In *Handbook of Intellectual Property Research: Lenses, Methods, and Perspectives*, Irene Calboli and Maria Lilla Montagnani (Eds.). Oxford University Press, Oxford. <https://doi.org/10.1093/oso/9780198826743.003.0031>
- [62] Dipto Barman, Kevin Koidl, and Owen Conlan. 2024. Discerning Individual Preferences for Identifying and Flagging Misinformation on Social Media. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24)*. Association for Computing Machinery, New York, NY, USA, 110–119. <https://doi.org/10.1145/3627043.3659545>
- [63] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2020. Fairness and machine learning. *Recommender systems handbook* 1 (2020), 453–459.
- [64] Nina Baur and Jörg Blasius. 2022. Methods of empirical social research (Methoden der empirischen Sozialforschung). In *Handbook of methods of empirical social research (Handbuch Methoden der empirischen Sozialforschung)*, Nina Baur and Jörg Blasius (Eds.). Springer Fachmedien, Wiesbaden, 1–32. https://doi.org/10.1007/978-3-658-37985-8_1
- [65] Bayern OLG. 2020. Beschluss vom 14.02.2020. <https://openjur.de/u/2274641.html>
- [66] Guillaume Beaumier and Lars Gjesvik. 2025. Digital Governance in a Rubber Band: Structural Constraints in Governing a Global Digital Economy. *Global Studies Quarterly* 5, 1 (2025).
- [67] Catalina-Gabriela Belgiu and Iuliana Constantin. 2017. Political Marketing Campaign. *Logos Universality Mentality Education Novelty: Political Sciences & European Studies* 4, 2 (Dec. 2017), 25–32. <https://doi.org/10.18662/lumenpses.3> Number: 2.

- [68] Luca Belli. 2022. Structural power as a critical element of social media platforms' private sovereignty. *Constitutionalising Social Media. (Hart 2022)* (2022).
- [69] Annegret Bendiek and Isabella Stuerzer. 2023. The Brussels Effect, European Regulatory Power and Political Capital. *Digital Society* 2, 5 (2023).
- [70] Astrid Bertrand, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2022. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*. Association for Computing Machinery, New York, NY, USA, 78–91. <https://doi.org/10.1145/3514094.3534164>
- [71] BGH. 2013. Beschluss vom 12.11.2013. <https://openjur.de/u/662251.html>
- [72] BGH. 2017. Beschluss vom 01.02.2017. <https://openjur.de/u/2116511.html>
- [73] BGH. 2018. Beschluss vom 19.06.2018. <https://openjur.de/u/2121414.html>
- [74] BGH. 2018. Beschluss vom 22.03.2018. <https://openjur.de/u/2110505.html>
- [75] BGH. 2021. Beschluss vom 11.08.2021. <https://openjur.de/u/2383741.html>
- [76] Savita Bhat and Vasudeva Varma. 2023. Large Language Models As Annotators: A Preliminary Evaluation For Annotating Low-Resource Language Content. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, Daniel Deutsch, Rotem Dror, Steffen Eger, Yang Gao, Christoph Leiter, Juri Opitz, and Andreas Rücklé (Eds.). Association for Computational Linguistics, Bali, Indonesia, 100–107. <https://doi.org/10.18653/v1/2023.eval4nlp-1.8>
- [77] Xavier Javines Bilon and Jose Antonio R. Clemente. 2020. Evaluation of Sampling Methods for Content Analysis of Facebook Data. *Philippine Journal of Science* 69, 1 (2020), 67–76.
- [78] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Sam Hottois, Latanya Sweeney, and Hanna Wallach (Eds.). PMLR, New York, NY, USA, 149–159. <https://proceedings.mlr.press/v81/binns18a.html> ISSN: 2640-3498.
- [79] Reuben Binns. 2020. Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance. *Big Data & Society* 7, 1 (2020).
- [80] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In *Social Informatics (Lecture Notes in Computer Science)*, Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri (Eds.). Springer International Publishing, Cham, 405–415. https://doi.org/10.1007/978-3-319-67256-4_32

- [81] Kean Birch, DT Cochrane, and Callum Ward. 2021. Data as asset? The measurement, governance, and valuation of digital personal data by Big Tech. *Big Data & Society* 8, 1 (2021), 20539517211017308. <https://doi.org/10.1177/20539517211017308> arXiv:<https://doi.org/10.1177/20539517211017308>
- [82] Shreyan Biswas, Alexander Erlei, and Ujwal Gadiraju. 2025. Mind the Gap! Choice Independence in Using Multilingual LLMs for Persuasive Co-Writing Tasks in Different Languages. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–20. <https://doi.org/10.1145/3706598.3713201>
- [83] Shreyan Biswas, Alexander Erlei, and Ujwal Gadiraju. 2026. Belief Updating and Delegation in Multi-Task Human–AI Interaction: Evidence from Controlled Simulations. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–23.
- [84] European Data Protection Board. 2023. Binding Decision 3/2022 on the Dispute Submitted by the Irish SA on Meta Platforms Ireland Limited and Its Facebook Service (Art. 65 GDPR). https://edpb.europa.eu/system/files/2023-01/edpb_bindingdecision_202203_ie_sa_meta_facebookservice_redacted_en.pdf
- [85] Barry Boehm and Wilfred J Hansen. 2000. *Spiral development: Experience, principles, and refinements*. Technical Report.
- [86] Daniel James Bogiatzis-Gibbons. 2024. Beyond Individual Accountability: (Re-)Asserting Democratic Control of AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 74–84. <https://doi.org/10.1145/3630106.3658541>
- [87] Edyta Bogucka, Marios Constantinides, Sanja Šćepanović, and Daniele Quercia. 2024. Co-Designing an AI Impact Assessment Report Template with AI Practitioners and AI Compliance Experts. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 168–180.
- [88] Edyta Paulina Bogucka, Marios Constantinides, Sanja Šćepanović, and Daniele Quercia. 2025. Impact Assessment Card: Communicating Risks and Benefits of AI Uses. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–42.
- [89] Alexandre Bovet and Hernán A. Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications* 10, 1 (2019), 7. <https://doi.org/10.1038/s41467-018-07761-2>
- [90] Anu Bradford. 2020. *The Brussels Effect: How the European Union Rules the World*. Oxford University Press. Google-Books-ID: mZXHDwAAQBAJ.
- [91] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

- [92] Virginia Braun and Victoria Clarke. 2021. Conceptual and Design Thinking for Thematic Analysis. *Qualitative Psychology* 9 (May 2021), 3—26. <https://doi.org/10.1037/qap0000196>
- [93] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology* 18, 3 (July 2021), 328–352. <https://doi.org/10.1080/14780887.2020.1769238>
- [94] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology* 18, 3 (2021), 328–352.
- [95] Virginia Braun and Victoria Clarke. 2021. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative research in sport, exercise and health* 13, 2 (2021), 201–216.
- [96] Timothy F. Bresnahan and Manuel Trajtenberg. 1995. General Purpose Technologies “Engines of Growth”? *Journal of Econometrics* 65, 1 (1995), 83–108.
- [97] Joseph C. Bright, Joseph B. Kadane, and Daniel S. Nagin. 1988. Statistical Sampling in Tax Audits. *Law & Social Inquiry* 13, 2 (1988), 305–333. arXiv:828448
- [98] Harry Brignull. 2019. Dark Patterns. <https://www.darkpatterns.org/>
- [99] Or Brook. 2021. Politics of Coding: On Systematic Content Analysis of Legal Text. <https://doi.org/10.2139/ssrn.3887536>
- [100] Sally Broughton Micova and Andrea Calef. 2023. Elements for Effective Systemic Risk Assessment Under the DSA. <https://doi.org/10.2139/ssrn.4512640>
- [101] Nina Brown. 2021. Regulatory Goldilocks: Finding the Just and Right Fit for Content Moderation on Social Platforms. *Texas A&M Law Review* 8, 3 (May 2021), 451–494. <https://doi.org/10.37419/LR.V8.I3.1>
- [102] Tim Brown et al. 2008. Design thinking. *Harvard business review* 86, 6 (2008), 84.
- [103] Jim Buatti, Lead Counsel, and Wikimedia Foundation and Aeryn Palmer. 2017. Wikimedia Foundation releases first transparency report of 2017. <https://diff.wikimedia.org/2017/02/13/sixth-transparency-report/>
- [104] B. Buchanan, A. Lohn, Micah Musser, and K. Šedová. 2021. Truth, Lies, and Automation: How Language Models Could Change Disinformation. (2021). <https://doi.org/10.51593/2021ca003>
- [105] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW1 (2021), 1–21.

- [106] Bundesrepublik Deutschland. 2017. Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz – NetzDG). Bundesgesetzblatt Teil I Nr. 61, 01.09.2017, p. 3352–3355. <https://www.gesetze-im-internet.de/netzdg/> Amended by Article 1 of the law of 03.06.2021 (BGBl. I S. 1444).
- [107] Matt Burgess. [n. d.]. Grok Is Pushing AI ‘Undressing’ Mainstream. *Wired* ([n. d.]). <https://www.wired.com/story/grok-is-pushing-ai-undressing-mainstream/> Section: tags.
- [108] Dylan Butts. 2025. Meta says it fixed ‘error’ after Instagram users report a flood of graphic and violent content. *CNBC* (Feb. 2025). <https://www.cnbc.com/2025/02/27/meta-apologizes-after-instagram-users-see-graphic-and-violent-content.html>
- [109] Julian Bäuml, Helen Bader, Marc-André Kaufhold, and Christian Reuter. 2025. Towards Youth-Sensitive Hateful Content Reporting: An Inclusive Focus Group Study in Germany. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI ’25)*. Association for Computing Machinery, New York, NY, USA, 1–22. <https://doi.org/10.1145/3706598.3713542>
- [110] Jie Cai et al. 2024. Content Moderation Justice and Fairness on Social Media: Comparisons Across Different Contexts and Platforms. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM.
- [111] Jie Cai, Aashka Patel, Azadeh Naderi, and Donghee Yvette Wohn. 2024. Content Moderation Justice and Fairness on Social Media: Comparisons Across Different Contexts and Platforms. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA ’24)*. Association for Computing Machinery, New York, NY, USA, 1–9. <https://doi.org/10.1145/3613905.3650882>
- [112] California State Assembly. 2022. Assembly Bill No. 587. https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB587
- [113] Alessandra Calvi and Dimitris Kotzinos. 2023. Enhancing AI fairness through impact assessment in the European Union: a legal and computer science perspective. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’23)*. Association for Computing Machinery, New York, NY, USA, 1229–1245. <https://doi.org/10.1145/3593013.3594076>
- [114] Malcolm Campbell-Verduyn and Marc Lenglet. 2023. Imaginary failure: RegTech in finance. *New Political Economy* 28, 3 (May 2023), 468–482. <https://doi.org/10.1080/13563467.2022.2140795> _eprint: <https://doi.org/10.1080/13563467.2022.2140795>.
- [115] Rachele Carli and Amro Najjar. 2023. A Vulnerability-oriented Impact Assessment for the Development of Human-Centred and Fundamental Rights-Empowering Social Robots.. In *Proceedings of the 11th International Conference on Human-Agent*

- Interaction (HAI '23)*. Association for Computing Machinery, New York, NY, USA, 404–406. <https://doi.org/10.1145/3623809.3623941>
- [116] Astrid Carolus, Martin J. Koch, Samantha Straka, Marc Erich Latoschik, and Carolin Wienrich. 2023. MAILS - Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Computers in Human Behavior: Artificial Humans* 1, 2 (2023), 100014. <https://doi.org/10.1016/j.chbah.2023.100014>
- [117] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. ACM, Rio de Janeiro, Brazil, 2254–2272. <https://doi.org/10.1145/3630106.3659037>
- [118] Céline Castets-Renard. 2020. Algorithmic Content Moderation on Social Media in EU Law: Illusion of Perfect Enforcement. *SSRN Electronic Journal* (2020). <https://papers.ssrn.com/abstract=3535107>
- [119] Caroline Cauffman and Catalina Goanta. 2021. A New Order: The Digital Services Act and Consumer Protection. *European Journal of Risk Regulation* 12, 4 (Dec. 2021), 758–774. <https://doi.org/10.1017/err.2021.8>
- [120] Ann Cavoukian et al. 2009. Privacy by design: The 7 foundational principles. *Information and privacy commissioner of Ontario, Canada* 5, 2009 (2009), 12.
- [121] Edoardo Celeste, Nicola Palladino, Dennis Redeker, and Kinfe Yilma. 2023. *The Content Governance Dilemma: Digital Constitutionalism, Social Media and the Search for a Global Standard*. Springer International Publishing, Cham. <https://doi.org/10.1007/978-3-031-32924-1>
- [122] Cengiz and Others v Turkey. [n. d.]. ECtHR 48226/10, 14027/11, 2015.
- [123] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Aaron Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. “You can’t stay here”: The efficacy of Reddit’s 2015 ban examined through hate speech. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, Vol. 1. 1–22.
- [124] Edward Y. Chang. 2024. SocraSynth: Multi-LLM Reasoning with Conditional Statistics. <https://doi.org/10.48550/arXiv.2402.06634> arXiv:2402.06634 [cs].
- [125] Edward Y. Chang. 2024. SocraSynth: Multi-LLM Reasoning with Conditional Statistics. arXiv:cs.AI/2402.06634 <https://arxiv.org/abs/2402.06634>

- [126] Kyla Chasalow and Karen Levy. 2021. Representativeness in statistics, politics, and machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 77–89.
- [127] Xiang Cheng, Raveesh Mayya, and João Sedoc. 2025. From Human Annotation to LLMs: SILICON Annotation Workflow for Management Research. <https://doi.org/10.48550/arXiv.2412.14461> arXiv:2412.14461 [cs].
- [128] Simon Chesterman. 2024. Lawful but Awful: Evolving Legislative Responses to Address Online Misinformation, Disinformation, and Mal-Information in the Age of Generative AI. *The American Journal of Comparative Law* 72, 4 (2024). <https://doi.org/10.1093/ajcl/avaf020>
- [129] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding. <https://doi.org/10.48550/arXiv.2306.14924> arXiv:2306.14924 [cs].
- [130] Amit K. Chopra and Munindar P. Singh. 2018. Sociotechnical Systems and Ethics in the Large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New Orleans LA USA, 48–53. <https://doi.org/10.1145/3278721.3278740>
- [131] Yi Ting Chua, Simon Parkin, Matthew Edwards, Daniela Oliveira, Stefan Schiffner, Gareth Tyson, and Alice Hutchings. 2019. Identifying unintended harms of cybersecurity countermeasures. In *2019 APWG Symposium on Electronic Crime Research (eCrime)*. IEEE, 1–15.
- [132] Lorenzo Cima, Alessio Miaschi, Amaury Trujillo, Marco Avvenuti, Felice Dell'Orletta, and Stefano Cresci. 2025. Contextualized counterspeech: Strategies for adaptation, personalization, and evaluation. In *Proceedings of the ACM on Web Conference 2025*. 5022–5033.
- [133] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology* 12, 3 (2017), 297–298.
- [134] Conor Clune and Emma McDaid. 2024. Content moderation on social media: constructing accountability in the digital space. *Accounting, Auditing & Accountability Journal* 37, 1 (Jan. 2024), 257–279. <https://doi.org/10.1108/AAAJ-11-2022-6119>
- [135] Jennifer Cobbe. 2021. Algorithmic Censorship by Social Platforms: Power and Resistance. *Philosophy & Technology* 34 (2021), 739.
- [136] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2021. Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 598–609. <https://doi.org/10.1145/3442188.3445921>

- [137] Jennifer Cobbe, Michael Veale, and Jatinder Singh. 2023. Understanding accountability in algorithmic supply chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1186–1197. <https://doi.org/10.1145/3593013.3594073>
- [138] Julie E. Cohen. 2019. Networks, Standards, and Transnational Governance Institutions. In *Between Truth and Power: The Legal Constructions of Informational Capitalism*. Oxford University Press.
- [139] Julie E. Cohen. 2019. The Regulatory State in the Information Age. In *Between Truth and Power: The Legal Constructions of Informational Capitalism*, Julie E. Cohen (Ed.). Oxford University Press, Oxford, UK, 0. <https://doi.org/10.1093/oso/9780190246693.003.0007>
- [140] Kevin Collier. 2023. U.S. celebrities were tricked into recording videos later doctored into anti-Zelenskyy propaganda. <https://www.nbcnews.com/tech/misinformation/celebrities-tricked-cameo-videos-russia-ukraine-propaganda-elijah-wood-rcna128331>
- [141] Commission. 2024. Supervision of the designated very large online platforms and search engines under DSA. <https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses>
- [142] European Commission. [n. d.]. Supervision of the designated very large online platforms and search engines under DSA | Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses>
- [143] European Commission. 2018. Code of Practice on Disinformation. <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation> Library Catalog: ec.europa.eu.
- [144] European Commission. 2020. COMMISSION STAFF WORKING DOCUMENT IMPACT ASSESSMENT REPORT Accompanying the document Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on contestable and fair markets in the digital sector (Digital Markets Act). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52020SC0363>
- [145] European Commission. 2022. Overview of the designated Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs). <https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses>
- [146] European Commission. 2023. Commission opens formal proceedings against X under the DSA. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6709

- [147] European Commission. 2023. The Commission sends request for information to X under DSA. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_4953
- [148] European Commission. 2023. Statements of Reasons - DSA Transparency Database. <https://transparency.dsa.ec.europa.eu/statement>
- [149] European Commission. 2024. Commission opens formal proceedings against TikTok under DSA. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_6487
- [150] European Commission. 2024. Commission opens proceedings against TikTok under the DSA. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2227
- [151] European Commission. 2024. Commission requests information from online marketplaces Temu and Shein on compliance with the Digital Services Act | Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/news/commission-requests-information-online-marketplaces-temu-and-shein-compliance-digital-services-act>
- [152] European Commission. 2024. Commission sends preliminary findings to X for breach of DSA. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_3761
- [153] European Commission. 2024. Commission sends request for information on illegal content and protection of minors to Pornhub, XVideos and Stripchat under the Digital Services Act | Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/news/commission-sends-request-information-illegal-content-and-protection-minors-pornhub-xvideos-and>
- [154] European Commission. 2024. Commission sends request for information to Meta under the Digital Services Act | Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/news/commission-sends-request-information-meta-under-digital-services-act-2>
- [155] European Commission. 2024. Commission sends request for information to TikTok regarding the launch of TikTok Lite in France and Spain | Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/news/commission-sends-request-information-tiktok-regarding-launch-tiktok-lite-france-and-spain>
- [156] European Commission. 2024. Commission sends requests for information to 17 Very Large Online Platforms and Search Engines under the Digital Services Act | Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-17-very-large-online-platforms-and-search-engines-under>

- [157] European Commission. 2024. DSA: Commission opens formal proceedings against AliExpress. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1485
- [158] European Commission. 2024. DSA: Commission opens formal proceedings against Meta. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2664
- [159] European Commission. 2024. DSA: Commission opens formal proceedings against TikTok. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_926
- [160] European Commission. 2025. Commission accepts commitments offered by AliExpress under the DSA and takes further action. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_1551
- [161] European Commission. 2025. Commission accepts TikTok's commitments on advertising transparency under the Digital Services Act | Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/news/commission-accepts-tiktoks-commitments-advertising-transparency-under-digital-services-act>
- [162] European Commission. 2025. Commission endorses the integration of the voluntary Code of Practice on Disinformation into the Digital Services Act | Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/news/commission-endorses-integration-voluntary-code-practice-disinformation-digital-services-act>
- [163] European Commission. 2025. Commission fines X €120 million under the Digital Services Act. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2934
- [164] European Commission. 2025. Commission opens investigations to safeguard minors from pornographic content under the DSA. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_1339
- [165] European Commission. 2025. Commission requests information from Shein on illegal products and its recommender system | Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/news/commission-request-s-information-shein-illegal-products-and-its-recommender-system>
- [166] European Commission. 2025. Dashboard for Statement of Reason Reports under the Digital Services Act. <https://transparency.dsa.ec.europa.eu/dashboard?lang=en>
- [167] European Commission. 2025. DSA - Commission preliminarily finds TikTok and Meta in breach of their transparency obligations. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2503

- [168] European Commission. 2025. Trusted flaggers under the Digital Services Act (DSA) | Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/policies/trusted-flaggers-under-dsa>
- [169] Zachary Cooper, William Lehr, and Volker Stocker. 2024. The New Age: Legal & Economic Challenges to Copyright and Creative Economies in the Era of Generative AI. <https://doi.org/10.2139/ssrn.5022340>
- [170] Sara Giro Correia. 2019. Responding to victimisation in a digital world: a case study of fraud and computer misuse reported in Wales. *Crime Science* 8, 1 (2019), 1–12.
- [171] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1571–1583. <https://doi.org/10.1145/3531146.3533213>
- [172] Jenna Course-Choi and Linda Hammond. 2021. Social media use and adolescent well-being: A narrative review of longitudinal studies. *Cyberpsychology, Behavior, and Social Networking* 24, 4 (2021), 223–236.
- [173] Court of Justice of the European Union. 2019. Fashion ID GmbH & CoKG v Verbraucherzentrale NRW eV. <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:62017CJ0040>
- [174] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (March 2016), 410–428. <https://doi.org/10.1177/1461444814543163> Publisher: SAGE Publications.
- [175] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428.
- [176] Bart Custers and Gianclaudio Malgieri. 2022. Priceless data: Why the EU fundamental right to data protection is at odds with trade in personal data. *Computer Law & Security Review* 45 (2022), 105683.
- [177] Brandon Dang, Martin J. Riedl, and Matthew Lease. 2020. But Who Protects the Moderators? The Case of Crowdsourced Image Moderation. arXiv preprint arXiv:1804.10999.
- [178] H. Justin Davidson. 1959. Accuracy in Statistical Sampling. *The Accounting Review* 34, 3 (1959), 356–365. arXiv:241272
- [179] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13, 3 (1989), 319. <https://doi.org/10.2307/249008>

- [180] Giovanni De Gregorio. 2020. Democratising online content moderation: A constitutional framework. *Computer Law & Security Review* 36 (April 2020), 105374. <https://doi.org/10.1016/j.clsr.2019.105374>
- [181] Paul B. de Laat. 2012. Coercion or Empowerment? Moderation of Content in Wikipedia as “Essentially Contested” Bureaucratic Rules. *Ethics and Information Technology* 14 (2012), 123.
- [182] Mathias-Felipe de Lima-Santos and Arwa Kooli. 2021. *Instagrammable Data: Using Visuals to Showcase More Than Numbers on AJ Labs Instagram Page*. Office for Official Publications of the European Communities, 20, rue de Reims, L-2417 Luxembourg, LUXEMBOURG. <https://doi.org/10.31124/advance.14876853.v1>
- [183] Alexandre de Streel and Michèle Ledger. 2021. 3. Regulating the moderation of illegal online content. (2021).
- [184] Deloitte. 2024. *Independent Practitioner’s Assurance Report on Zalando SE’s Compliance with the Digital Services Act*. Assurance Report. Deloitte. https://corporate.zalando.com/sites/default/files/media-download/Zalando-SE_DSA_Independent-Practitioners-Assurance-Report.pdf
- [185] Deloitte. 2024. *Microsoft Bing DSA Audit Report*. Assurance Report. Deloitte. <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft%20Bing%20DSA%20Audit%20Report%20-%20August%202024.pdf>
- [186] Jean-Yves Delort, Bavani Arunasalam, and Cecile Paris. 2011. Automatic Moderation of Online Discussion Sites. *International Journal of Electronic Commerce* 15, 3 (April 2011), 9–30. <https://doi.org/10.2753/JEC1086-4415150302>
- [187] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try to) Use Fairness Toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 473–484.
- [188] Yang Deng, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam, and Tat-Seng Chua. 2025. Proactive Conversational AI: A Comprehensive Survey of Advancements and Opportunities. *ACM Transactions on Information Systems* 43, 3 (2025), 1–45. <https://doi.org/10.1145/3715097>
- [189] Daria Dergacheva, Vasilisa Kuznetsova, Rebecca Scharlach, and Christian Katzenbach. 2023. One Day in Content Moderation: Analyzing 24 h of Social Media Platforms’ Content Decisions through the DSA Transparency Database. (Dec. 2023). <https://doi.org/10.26092/elib/2707>
- [190] Mercy Sharon Devadas, R. Sabitha, K. Shankar, A.P. Edlabadkar, H J Aljawawdeh, and K. Pramitha. 2024. An Experimental Evaluation of Flagging Sensitive Contents Over Social Networking Environment Using Elevated Classification Logic. In *2024 International Conference on Innovative Computing, Intelligent Communication and*

- Smart Electrical Systems (ICSES)*. 1–7. <https://doi.org/10.1109/ICSES63760.2024.10910843>
- [191] Fabiana Di Porto and Alexander Egberts. 2023. The collective welfare dimension of dark patterns regulation. *European Law Journal* 29, 1-2 (2023), 114–141. <https://doi.org/10.1111/eulj.12478> _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/eulj.12478>.
- [192] Thiago Dias Oliva. 2020. Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression. *Human Rights Law Review* 20, 4 (Dec. 2020), 607–640. <https://doi.org/10.1093/hrlr/ngaa032>
- [193] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture* 25, 2 (April 2021), 700–732. <https://doi.org/10.1007/s12119-020-09790-w>
- [194] Carlos Diaz Ruiz. 2025. Disinformation on Digital Media Platforms: A Market-Shaping Approach. *New Media & Society* 27 (2025), 2188–2207.
- [195] Patrick Dieuaide and Christian Azaïs. 2020. Platforms of Work, Labour, and Employment Relationship: The Grey Zones of a Digital Governance. *Frontiers in Sociology* 5 (Feb. 2020), 2. <https://doi.org/10.3389/fsoc.2020.00002>
- [196] Lisa M. DiSalvo, Gabriela Viviana Saenz, W. Eric Wong, and Dongcheng Li. 2022. Social Media Safety Practices and Flagging Sensitive Posts. In *2022 IEEE 22nd International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*. 8–15. <https://doi.org/10.1109/QRS-C57518.2022.00012> ISSN: 2693-9371.
- [197] Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard choices in artificial intelligence. *Artificial Intelligence* 300 (Nov. 2021), 103555. <https://doi.org/10.1016/j.artint.2021.103555>
- [198] Leyla Dogruel, Dominique Facciorusso, and Birgit Stark. 2020. ‘I’m still the master of the machine.’ Internet users’ awareness of algorithmic decision-making and their perception of its effect on their autonomy. *Information, Communication & Society* (Dec. 2020), 1311–1332. <https://doi.org/10.1080/1369118X.2020.1863999>
- [199] Domo, Inc. 2023. Domo Resource – Data Never Sleeps 11.0. <https://www.domo.com/learn/infographic/data-never-sleeps-11>
- [200] Evelyn Douek. 2021. Governing Online Speech: From “Posts-as-Trumps” to Proportionality and Probability. *Columbia Law Review Forum* 121 (2021), 1–27. <https://columbialawreview.org/content/governing-online-speech-from-posts-as-trumps-to-proportionality-and-probability/>
- [201] Evelyn Douek. 2021. The Rise of Content Cartels. *Columbia Law Review* 121, 6 (2021), 1369–1442. <https://columbialawreview.org/content/the-rise-of-content-cartels/>

- [202] Evelyn Douek. 2022. *Content Moderation as Administration*. Technical Report. SSRN Scholarly Paper ID 4005326. <https://papers.ssrn.com/abstract=4005326>
- [203] Evelyn Douek. 2022. Content Moderation as Systems Thinking. *Harvard Law Review* 36, 2 (Dec. 2022), 528–606. <https://harvardlawreview.org/print/vol-136/content-moderation-as-systems-thinking/>
- [204] Judit Dr. Bayer, Dr Bernd Holznagel Prof., Lubianiec Katarzyna, Pinteá Adela, B. Schmitt Dr. Josephine, Judit Szakács, and Erik Uszkiewicz. 2021. *Disinformation and propaganda: impact on the functioning of the rule of law and democratic processes in the EU and its Member States: 2021 update*. European Parliament Policy Department for External Relations. <https://doi.org/10.2861/752863> Accepted: 2021-05-13T06:53:33Z.
- [205] Chiara Drolsbach and Nicolas Pröllochs. 2023. Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database. <http://arxiv.org/abs/2312.04431> arXiv:2312.04431 [cs].
- [206] Chiara Patricia Drolsbach and Nicolas Pröllochs. 2024. Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24)*. Association for Computing Machinery, New York, NY, USA, 939–942. <https://doi.org/10.1145/3589335.3651482>
- [207] Edgar Dubourg, Valentin Thouzeau, and Nicolas Baumard. 2024. A step-by-step method for cultural annotation by LLMs. *Frontiers in Artificial Intelligence* 7 (May 2024), 1–10. <https://doi.org/10.3389/frai.2024.1365508> Publisher: Frontiers.
- [208] Shiran Dudy, Thulasi Tholeti, Resmi Ramachandranpillai, Muhammad Ali, Toby Jia-Jun Li, and Ricardo Baeza-Yates. 2025. Unequal Opportunities: Examining the Bias in Geographical Recommendations by Large Language Models. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 1499–1516. <https://doi.org/10.1145/3708359.3712111>
- [209] Bram Duivenvoorde and Catalina Goanta. 2023. The regulation of digital advertising under the DSA: A critical assessment. *Computer Law & Security Review* 51 (Nov. 2023), 105870. <https://doi.org/10.1016/j.clsr.2023.105870>
- [210] Mary T Dzindolet, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. 2002. The perceived utility of human and automated aids in a visual detection task. *Human factors* 44, 1 (2002), 79–94.
- [211] William Echikson and Olivia Knodt. 2018. Germany's NetzDG: A Key Test for Combatting Online Hate. <https://papers.ssrn.com/abstract=3300636>

- [212] Niklas Eder. 2024. Making Systemic Risk Assessments Work: How the DSA Creates a Virtuous Loop to Address the Societal Harms of Content Moderation. *German Law Journal* 25, 7 (Oct. 2024), 1197–1218. <https://doi.org/10.1017/glj.2024.24>
- [213] Editorial. 2025. The Guardian view on content moderation: Meta did far too little. Now it wants to do even less. *The Guardian* (Jan. 2025). <https://www.theguardian.com/commentisfree/2025/jan/08/the-guardian-view-on-content-moderation-meta-did-far-too-little-now-it-wants-to-do-even-less>
- [214] EDRI. [n. d.]. US pressure on the Digital Services Act in the Netherlands. <https://edri.org/our-work/us-pressure-on-the-digital-services-act-in-the-netherlands/>
- [215] Michel Eeten and Milton Mueller. 2013. Where Is the Governance in Internet Governance. *New Media & Society* 15 (2013), 720.
- [216] Upol Ehsan, Q Vera Liao, Samir Passi, Mark O Riedl, and Hal Daumé III. 2024. Seamful xai: Operationalizing seamful design in explainable ai. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–29.
- [217] Mohamed Elaraby, Diane Litman, Xiang Lorraine Li, and Ahmed Magooda. 2024. Persuasiveness of Generated Free-Text Rationales in Subjective Decisions: A Case Study on Pairwise Argument Ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, Miami, Florida, USA, 14311–14329. <https://doi.org/10.18653/v1/2024.findings-emnlp.836>
- [218] Micheline Elias and Anastasia Bezerianos. 2012. Annotating BI visualization dashboards: needs & challenges. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1641–1650. <https://doi.org/10.1145/2207676.2208288>
- [219] Alexander Erlei, Federico Maria Cau, Radoslav Georgiev, Sagar Chethan Kumar, Kilian Bizer, and Ujwal Gadiraju. 2026. When Life Gives You AI, Will You Turn It Into A Market for Lemons? Understanding How Information Asymmetries About AI System Capabilities Affect Market Outcomes and Adoption. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–24.
- [220] Alexander Erlei, Abhinav Sharma, and Ujwal Gadiraju. 2024. Understanding Choice Independence and Error Types in Human-AI Collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. <https://doi.org/10.1145/3613904.3641946>
- [221] AI EU. 2021. European Commission white paper on artificial intelligence—a European approach to excellence and trust. URL https://ec.europa.eu/info/sites/info/files/commissionwhitepaper-artificial-intelligence-feb2020_en.pdf (2021).

- [222] European Commission. 2008. *Survey Sampling Reference Guidelines*. European Communities, Luxembourg. <https://ec.europa.eu/eurostat/documents/64157/4374310/Survey+sampling+reference+guidelines.pdf>
- [223] European Commission. 2022. Digital Rights and Principles. https://ec.europa.eu/commission/presscorner/detail/en/ip_22_7683.
- [224] European Commission. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance). <http://data.europa.eu/eli/reg/2022/2065/oj/eng> Legislative Body: EP, CONSIL.
- [225] European Commission. 2023. Commission Delegated Regulation (EU) .../... of 20 October 2023 supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council, by laying down rules on the performance of audits for very large online platforms and very large online search engines. *Official Journal of the European Union*, C(2023)6807final Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32023R2852> [Accessed: 20 May 2025].
- [226] European Commission. 2023. Delegated Regulation Supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council by Laying Down Rules on the Performance of Audits for Very Large Online Platforms and Very Large Online Search Engines. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32023R2713> C/2023/6807.
- [227] European Commission. 2024. Annex II of the Delegated Regulation for DSA Transparency Reporting (Implementing Regulation laying down templates concerning the transparency reporting obligations of providers of online platforms). <https://digital-strategy.ec.europa.eu/en/library/implementing-regulation-laying-down-templates-concerning-transparency-reporting-obligations>
- [228] European Commission. 2024. Commission Delegated Regulation (EU) .../... of XXX Supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council by Laying Down the Technical Conditions and Procedures under which Providers of Very Large Online Platforms and of Very Large Online Search Engines are to Share Data Pursuant to Article 40 of Regulation (EU) 2022/2065. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14044-Digital-Services-Act-rules-on-data-access-for-researchers_en Proposal under Article 40 DSA.
- [229] European Commission. 2024. Commission Guidelines for Providers of Very Large Online Platforms and Very Large Online Search Engines on the Mitigation of Systemic Risks for Electoral Processes Pursuant to the Digital Services Act. C(2024)2121final, Brussels, 26March2024 Annex to the Guidelines: C(2024)2121 final.

- [230] European Commission. 2024. Commission Implementing Regulation (EU) .../... of 4 November 2024 laying down templates concerning the transparency reporting obligations of providers of intermediary services and of providers of online platforms under Regulation (EU) 2022/2065 of the European Parliament and of the Council. [C\(2024\)7005final, Brussels, 4November2024](#) Official templates and rules for transparency reports under the Digital Services Act (DSA).
- [231] European Commission. 2024. Commission Opens Formal Proceedings against Temu under the DSA. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_5622
- [232] European Commission. 2024. Commission opens formal proceedings against TikTok. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_926
- [233] European Commission. 2024. Commission opens formal proceedings under DSA. https://ec.europa.eu/commission/presscorner/detail/en/IP_24_2373
- [234] European Commission. 2024. Digital Services Act / Transparency Database / dsa-tdb · GitLab. <https://code.europa.eu/dsa/transparency-database/dsa-tdb>
- [235] European Commission. 2024. Europe's Digital Decade. <https://digital-strategy.ec.europa.eu/en/policies/europes-digital-decade>.
- [236] European Commission. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). <http://data.europa.eu/eli/reg/2024/1689/oj/eng> Legislative Body: CONSIL, EP.
- [237] European Commission. 2025. Digital Services Act: Commission concludes that TikTok breached obligations on advertising transparency. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_1223 Accessed: 20 May 2025.
- [238] European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance). <https://eur-lex.europa.eu/eli/reg/2016/679/oj> OJ L 119, 4.5.2016, p. 1–88.
- [239] European Parliament and Council of the European Union. 2021. Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on Addressing the Dissemination of Terrorist Content Online (Text with EEA Relevance).

- <https://eur-lex.europa.eu/eli/reg/2021/784/oj> OJ L 172, 17.5.2021, p. 79–109.
- [240] European Parliament and Council of the European Union. 2022. Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector and Amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (Text with EEA Relevance).
- [241] European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations EC No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA Relevance).
- [242] European Parliament and Council. 2000. Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce). Official Journal of the European Communities. , 16 pages. <https://eur-lex.europa.eu/eli/dir/2000/31/oj> OJ L 178, 17.07.2000, p. 1–16.
- [243] European Parliament and Council. 2016. General Data Protection Regulation. <http://data.europa.eu/eli/reg/2016/679/oj/eng> Legislative Body: EP, CONSIL.
- [244] European Parliament and Council. 2018. Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) (codified version) (Text with EEA relevance)Text with EEA relevance. <http://data.europa.eu/eli/dir/2010/13/2018-12-18/eng> Legislative Body: OP_DATPRO.
- [245] European Parliament and Council. 2022. Decision (EU) 2022/2481 of the European Parliament and of the Council of 14 December 2022 establishing the Digital Decade Policy Programme 2030. Official Journal of the European Union. , 4–26 pages. <https://eur-lex.europa.eu/eli/dec/2022/2481/oj> OJ L 323, 19.12.2022, p. 4–26.
- [246] European Parliament and Council. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). Official Journal of the European Union. , 102 pages. <https://eur-lex.europa.eu/eli/reg/2022/2065/oj> OJ L 277, 27.10.2022, p. 1–102.
- [247] European Union. 2007. Explanations relating to the Charter of Fundamental Rights. Official Journal of the European Union. <https://eur-lex.europa.eu/legal-c>

- ontent/EN/TXT/?uri=oj%3AJOC_2007_303_R_0017_01 (2007/C 303/02), OJ C 303, 14 December 2007, pp. 17–35 (see in particular the Explanation on Article 52).
- [248] European Union. 2012. Charter of Fundamental Rights of the European Union. Official Journal of the European Union. , 391–407 pages. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT> OJ C 326, 26.10.2012, p. 391–407.
- [249] EY. 2024. *Independent Audit of the App Store*. Technical Report. EY. https://www.apple.com/legal/dsa/20242708_app-store_reasonable-assurance-report_non-confidential.pdf
- [250] EY. 2024. *Independent Audit on Facebook*. Assurance Report. EY. https://scontent-ams4-1.xx.fbcdn.net/v/t39.8562-6/468556126_544803875088152_2594679734771781492_n.pdf?_nc_cat=103&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=hY8niIVCit8Q7kNvgHVmk2J&_nc_zt=14&_nc_ht=scontent-ams4-1.xx&_nc_gid=AhBwGz_K-urR47LNVPIVnqF&oh=00_AYCsE5VnLS579fLSR4nziZ55H-UkWxiaCj66u9kbhJaiTw&oe=678C2AD3
- [251] EY. 2024. *Independent Audit on Google Maps*. Assurance Report. EY. https://storage.googleapis.com/transparencyreport/report-downloads/dsa-audit-google-maps_2023-8-28_2024-5-31_en_v1.pdf
- [252] EY. 2024. *Independent Audit on Google Play*. Technical Report. EY. https://storage.googleapis.com/transparencyreport/report-downloads/dsa-audit-google-play_2023-8-28_2024-5-31_en_v1.pdf
- [253] EY. 2024. *Independent Audit on Google Search*. Technical Report. EY. https://storage.googleapis.com/transparencyreport/report-downloads/dsa-audit-google-search_2023-8-28_2024-5-31_en_v1.pdf
- [254] EY. 2024. *Independent Audit on Google Shopping*. Technical Report. EY. https://storage.googleapis.com/transparencyreport/report-downloads/dsa-audit-google-shopping_2023-8-28_2024-5-31_en_v1.pdf
- [255] EY. 2024. *Independent Audit on Instagram*. Technical Report. EY. https://scontent-ams4-1.xx.fbcdn.net/v/t39.8562-6/468562318_2548652918858554_7920553818496711968_n.pdf?_nc_cat=107&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=VOFEXjsBkScQ7kNvgHHzNwJ&_nc_zt=14&_nc_ht=scontent-ams4-1.xx&_nc_gid=An2ZVz3-6UBR4U9pDvXKsbg&oh=00_AYBoZVeg8CIoe74USWXFSccaZCmNaeSvh-BGMYtVrvK9bw&oe=678C057B
- [256] EY. 2024. *Independent Audit on Pinterest*. Technical Report. EY. <https://help.pinterest.com/sites/pinhelp/files/dsa/Pinterest-2024-DSA-EY-Independent-Audit-Report-with-Redactions.pdf>
- [257] EY. 2024. *Independent Audit on Snapchat*. Technical Report. EY. https://assets.ctfassets.net/kw9k15xxztrs/5CrFmNWQztRkDcvhZ8TwyC/6014cd3df6544a2a68a56dfad1712b56/DSA_Independent_Audit_Report_-_Snapchat_-_August_2024.pdf

- [258] EY. 2024. *Independent Audit on YouTube*. Technical Report. EY. https://storage.googleapis.com/transparencyreport/report-downloads/dsa-audit-youtube_2023-8-28_2024-5-31_en_v1.pdf
- [259] Florian Eyert and Paola Lopez. 2023. Rethinking Transparency as a Communicative Constellation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 444–454. <https://doi.org/10.1145/3593013.3594010>
- [260] Florian Eyert and Paola Lopez. 2023. Rethinking transparency as a communicative constellation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Chicago, IL, USA, 444–454.
- [261] Matteo Fabbri and Ludovico Boratto. 2025. Auditing Recommender Systems for User Empowerment in Very Large Online Platforms under the Digital Services Act. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25)*. Association for Computing Machinery, New York, NY, USA, 51–61. <https://doi.org/10.1145/3705328.3748074>
- [262] Gregory Falco, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, Alwyn Goodloe, Jerry Gupta, Christopher Hart, Marina Jirotko, Henric Johnson, Cara LaPointe, Ashley J. Llorens, Alan K. Mackworth, Carsten Maple, Sigurður Emil Pálsson, Frank Pasquale, Alan Winfield, and Zee Kin Yeong. 2021. Governing AI Safety through Independent Audits. *Nature Machine Intelligence* 3, 7 (jul 2021), 566–571. <https://doi.org/10.1038/s42256-021-00370-7>
- [263] Federal Trade Commission. 2024. AI (and Other) Companies: Quietly Changing Your Terms of Service Could Be Unfair or Deceptive. <https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/02/ai-other-companies-quietly-changing-your-terms-service-could-be-unfair-or-deceptive>.
- [264] Lenka Fiala and Martin Husovec. 2022. Using Experimental Evidence to Improve Delegated Enforcement. *International Review of Law and Economics* 71 (2022), 106079.
- [265] Hayden Field. 2026. Grok is undressing children — can the law stop it? <https://www.theverge.com/ai-artificial-intelligence/855832/grok-undressing-children-csam-law-x-elon-musk>
- [266] Mikkel Flyverbom. 2016. Digital Age Transparency: Mediation and the Management of Visibilities. *International Journal of Communication* 10, 0 (2016), 13. <https://ijoc.org/index.php/ijoc/article/view/4490>
- [267] Mikkel Flyverbom. 2016. Disclosing and Concealing: Internet Governance, Information Control and the Management of Visibility. *Internet Policy Review* 5 (2016).
- [268] Mikkel Flyverbom. 2016. Transparency: Mediation and the management of visibilities. 10 (Jan. 2016), 110–122.

- [269] Raymond Fok and Daniel S Weld. 2024. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine* 45, 3 (2024), 317–332.
- [270] Pat Foran CTV News Toronto Consumer Alert Videojournalist Follow |Contact. 2024. Ontario man loses \$12K to deepfake scam involving Prime Minister Justin Trudeau. <https://toronto.ctvnews.ca/ontario-man-loses-12k-to-deepfake-scam-involving-prime-minister-justin-trudeau-1.6823827> Section: Toronto.
- [271] Paul Fraioli. 2025. Big Tech and the Brussels Effect. *Survival* 67, 2 (2025), 173–190.
- [272] Mirko Franco, Ombretta Gaggi, and Claudio E. Palazzi. 2025. Integrating Content Moderation Systems with Large Language Models. *ACM Trans. Web* 19, 2 (May 2025), 18:1–18:21. <https://doi.org/10.1145/3700789>
- [273] Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. 2025. Argumentative Large Language Models for Explainable and Contestable Claim Verification. arXiv:cs.CL/2405.02079 <https://arxiv.org/abs/2405.02079>
- [274] James W Freeman, Thomas C Darr, and Richard B Neely. 1997. Risk Assessment for Large Heterogeneous Systems. In *Proceedings 13th Annual Computer Security Applications Conference*. IEEE, 44–52.
- [275] French Republic. 2004. Loi n° 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique.
- [276] Batya Friedman and David G Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. Mit Press.
- [277] Ujwal Gadiraju and Agathe Balayn. 2025. The Enterprising and Elusive Prospects of Human-AI Collaboration. In *Enterprise AI*. Springer, 211–243.
- [278] Alberto Galasso and Mark Schankerman. 2015. Patents and Cumulative Innovation: Causal Evidence from the Courts. *The Quarterly Journal of Economics* 130, 1 (2015), 317–369.
- [279] Miguel Garcia-Godinez. 2025. A default framework for fundamental rights impact assessment. *AI and Ethics* (2025), 1–15.
- [280] R Kelly Garrett and Shannon Poulsen. 2019. Flagging Facebook Falsehoods: Self-Identified Humor Warnings Outperform Fact Checker and Peer Warnings. *Journal of Computer-Mediated Communication* 24, 5 (Sept. 2019), 240–258. <https://doi.org/10.1093/jcmc/zmz012>
- [281] Christophe Geiger and Bernd Justin Jütte. 2021. Platform Liability under Article 17 of the Copyright in the Digital Single Market Directive, Automated Filtering and Fundamental Rights: An Impossible Match.

- [282] R. Stuart Geiger and Aaron Halfaker. 2017. Operationalizing Conflict and Cooperation between Automated Software Agents in Wikipedia: A Replication and Expansion of 'Even Good Bots Fight'. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (Dec. 2017), 1–33. <https://doi.org/10.1145/3134684>
- [283] Amelia Gentleman, Helena Horton, and Dan Milmo. 2026. Grok AI still being used to digitally undress women and children despite suspension pledge. *The Guardian* (Jan. 2026). <https://www.theguardian.com/technology/2026/jan/05/elon-musk-grok-ai-digitally-undress-images-of-women-children>
- [284] German Federal Ministry of Justice. 2017. German Network Enforcement Act (NetzDG). https://www.bmj.de/DE/Themen/FokusThemen/NetzDG/NetzDG_EN_node.html
- [285] Ysabel Gerrard. 2018. Beyond the Hashtag: Circumventing Content Moderation on Social Media. *New Media & Society* 20, 12 (Dec. 2018), 4492–4511. <https://doi.org/10.1177/1461444818776611>
- [286] Ysabel Gerrard. 2018. Beyond the Hashtag: Circumventing Content Moderation on Social Media. *New Media & Society* 20 (2018), 4492.
- [287] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.
- [288] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. <https://doi.org/10.12987/9780300235029> Journal Abbreviation: Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media Pages: 288 Publication Title: Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.
- [289] Tarleton Gillespie. 2020. Content Moderation, AI, and the Question of Scale. *Big Data & Society* 7, 2 (2020).
- [290] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (July 2020), 2053951720943234. <https://doi.org/10.1177/2053951720943234> Publisher: SAGE Publications Ltd.
- [291] Tarleton Gillespie. 2022. Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society* 8, 3 (July 2022), 20563051221117552. <https://doi.org/10.1177/20563051221117552> Publisher: SAGE Publications Ltd.
- [292] Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candès, and Dan Jurafsky. 2025. Can Unconfident LLM Annotations Be Used for Confident Conclusions? <https://doi.org/10.48550/arXiv.2408.15204> arXiv:2408.15204 [cs].
- [293] Catalina Goanta, Thales Bertaglia, and Adriana Iamnitchi. 2022. The Case for a Legal Compliance API for the Enforcement of the EU's Digital Services Act on Social Media Platforms. In *2022 ACM Conference on Fairness, Accountability, and*

- Transparency*. ACM, Seoul Republic of Korea, 1341–1349. <https://doi.org/10.1145/3531146.3533190>
- [294] J W Goethe-Universität. [n. d.]. Friedrich Carl von Savigny, the Legal Method, and the Modernity of Law. *Friedrich Carl von Savigny* ([n. d.]).
- [295] Juan Felipe Gomez, Caio Machado, Lucas Monteiro Paes, and Flavio Calmon. 2024. Algorithmic Arbitrariness in Content Moderation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2234–2253. <https://doi.org/10.1145/3630106.3659036>
- [296] Juan Felipe Gomez, Caio Machado, Lucas Monteiro Paes, and Flavio Calmon. 2024. Algorithmic Arbitrariness in Content Moderation. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Rio de Janeiro, Brazil, 2234–2253.
- [297] Emma Goodman. 2013. *Online Comment Moderation: Emerging Best Practices*. Technical Report. World Association of Newspapers and News Publishers (WAN-IFRA), Darmstadt, Germany. https://netino.fr/wp-content/uploads/2013/10/WAN-IFRA_Online_Commenting.pdf
- [298] Google. [n. d.]. About the YouTube Priority Flagging program. <https://support.google.com/youtube/answer/7554338?hl=en>
- [299] Google. 2024. *Report of Systemic Risk Assessments: Google Maps*. Systemic Risk Assessment Report. Google. https://storage.googleapis.com/transparencyreport/report-downloads/dsa-risk-assessment_2023-8-28_2023-8-28_en_v1.pdf
- [300] Google. 2024. *Report of Systemic Risk Assessments Google Play*. Technical Report. Google. https://storage.googleapis.com/transparencyreport/report-downloads/dsa-risk-assessment_2023-8-28_2023-8-28_en_v1.pdf
- [301] Google. 2024. *Report of Systemic Risk Assessments Google Search*. Technical Report. Google. https://storage.googleapis.com/transparencyreport/report-downloads/dsa-risk-assessment_2023-8-28_2023-8-28_en_v1.pdf
- [302] Google. 2024. *Report of Systemic Risk Assessments Google Shopping*. Technical Report. Google. https://storage.googleapis.com/transparencyreport/report-downloads/dsa-risk-assessment_2023-8-28_2023-8-28_en_v1.pdf
- [303] Google. 2024. *Report of Systemic Risk Assessments Youtube*. Technical Report. Google. https://storage.googleapis.com/transparencyreport/report-downloads/dsa-risk-assessment_2023-8-28_2023-8-28_en_v1.pdf
- [304] Robert Gorwa. 2021. Elections, Institutions, and the Regulatory Politics of Platform Governance: The Case of the German NetzDG. *Telecommunications Policy* 45 (2021), 102145.

- [305] Robert Gorwa and Timothy G Ash. 2019. Democratic Transparency in the Platform Society. <https://doi.org/10.31235/osf.io/ehcy2>
- [306] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 205395171989794. <https://doi.org/10.1177/2053951719897945>
- [307] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Platform Governance: A New Research Agenda. *Internet Policy Review* 9, 2 (2020).
- [308] Sreeprasad Govindankutty and Avneesh Kumar. 2024. Design and Implementation of Automated Content Moderation Systems in Social Media. *Integrated Journal for Research in Arts and Humanities* 4, 6 (Nov. 2024), 380–402. <https://doi.org/10.55544/ijrah.4.6.27>
- [309] Anton Grafström and Lina Schelin. 2014. How to Select Representative Samples. *Scandinavian Journal of Statistics* 41, 2 (2014), 277–290. arXiv:24586652
- [310] Colin M. Gray, Ritika Gairola, Nayah Boucaud, Maliha Hashmi, Shruthi Sai Chivukula, Ambika R Menon, and Ja-Nae Duane. 2024. Legal Trouble?: UX Practitioners' Engagement with Law and Regulation. In *Companion Publication of the 2024 ACM Designing Interactive Systems Conference (DIS '24 Companion)*. Association for Computing Machinery, New York, NY, USA, 106–110. <https://doi.org/10.1145/3656156.3663698>
- [311] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 534:1–534:14. <https://doi.org/10.1145/3173574.3174108> event-place: Montreal QC, Canada.
- [312] Colin M. Gray, Thomas Mildner, and Ritika Gairola. 2025. Getting Trapped in Amazon's "Iliad Flow": A Foundation for the Temporal Analysis of Dark Patterns. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 225, 10 pages. <https://doi.org/10.1145/3706598.3713828>
- [313] Colin M Gray, Cristiana Santos, Nataliia Bielova, Michael Toth, and Damian Clifford. 2021. Dark patterns and the legal requirements of consent banners: An interaction criticism perspective. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. Association for Computing Machinery, Yokohama, Japan, 1–18.
- [314] Colin M. Gray, Cristiana Teixeira Santos, Nataliia Bielova, and Thomas Mildner. 2024. An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing

- Machinery, New York, NY, USA, 1–22. <https://doi.org/10.1145/3613904.3642436>
- [315] Giovanni De Gregorio and Pietro Dunn. 2022. The European risk-based approaches: Connecting constitutional dots in the digital age. *Common Market Law Review* 59, 2 (April 2022). <https://kluwerlawonline.com/api/Product/CitationPDFURL?file=Journals/COLA/COLA2022032.pdf>
- [316] Rachel Griffin. 2025. Governing platforms through corporate risk management: the politics of systemic risk in the Digital Services Act. *European Law Open* (2025), 1–31.
- [317] James Grimmelman. 2017. The Virtues of Moderation. <https://doi.org/10.31228/osf.io/qwxf5>
- [318] Samuel Groesch, Alena Birrer, Natascha Just, and Florian Saurwein. 2026. Big data, small answers: How the DSA Transparency Database falls short of its regulatory objectives. *Telecommunications Policy* 50, 1 (Jan. 2026), 103088. <https://doi.org/10.1016/j.telpol.2025.103088>
- [319] Muriel Binder Groom, John. 2026. Swiss Platform Regulation: Is DSA Compliance Enough? <https://connectontech.bakermckenzie.com/swiss-platform-regulation-is-dsa-compliance-enough/>
- [320] Rohan Grover. 2024. Encoding Privacy: Sociotechnical Dynamics of Data Protection Compliance Work. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. <https://doi.org/10.1145/3613904.3642872>
- [321] Lara Groves, Jacob Metcalf, Alayna Kennedy, Briana Vecchione, and Andrew Strait. 2024. Auditing Work: Exploring the New York City algorithmic bias audit regime. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 1107–1120. <https://doi.org/10.1145/3630106.3658959>
- [322] Florence G'sell. 2023. The Digital Services Act (DSA): A General Assessment. <https://doi.org/10.2139/ssrn.4403433>
- [323] Rita Gsenger. 2025. Platform governance under the Digital Services Act: a perspective on disinformation. *Information, Communication & Society* (Nov. 2025), 1–19. <https://doi.org/10.1080/1369118X.2025.2590561>
- [324] Rita Gsenger and Marie-Therese Sekwenz (Eds.). 2025. *Digital Decade: How the EU Shapes Digitalisation Research* (1. auflage ed.). Number 3 in Normsetzung und Entscheidungsverfahren – Schriftenreihe des Weizenbaum-Instituts für normative Wissenschaften. Nomos Verlagsgesellschaft mbH & Co. KG, Baden-Baden. <https://doi.org/10.5771/9783748943990>
- [325] Feng Gu, Zongxia Li, Carlos Rafael Colon, Benjamin Evans, Ishani Mondal, and Jordan Lee Boyd-Graber. 2025. Large Language Models Are Effective Human Annotation Assistants, But Not Good Independent Annotators. <https://doi.org/10.48550/arXiv.2503.06778> arXiv:2503.06778 [cs].

- [326] Greg Guest and Kathleen M. MacQueen. 2007. *Handbook for Team-Based Qualitative Research*. Rowman Altamira, Lanham, MD, USA.
- [327] Akshit Gupta, Debadeep Basu, Ramya Ghantasala, Sihang Qiu, and Ujwal Gadiraju. 2022. To Trust or Not To Trust: How a Conversational Interface Affects Trust in a Decision Support System. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*. Association for Computing Machinery, New York, NY, USA, 3531–3540. <https://doi.org/10.1145/3485447.3512248>
- [328] Abhishek Gupta, Iga Kozłowska, and Nga Than. 2022. The Golden Circle: Creating Socio-technical Alignment in Content Moderation. <https://doi.org/10.48550/arXiv.2202.13500> arXiv:2202.13500 [cs].
- [329] Hana Habib and Lorrie Faith Cranor. 2022. Evaluating the Usability of Privacy Choice Mechanisms. In *Proceedings of the Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*. USENIX Association, Boston, MA, USA, 273–289. <https://www.usenix.org/conference/soups2022/presentation/habib>
- [330] Philipp Hacker, Atoosa Kasirzadeh, and Lilian Edwards. 2025. AI, Digital Platforms, and the New Systemic Risk. <https://doi.org/10.48550/arXiv.2509.17878> arXiv:2509.17878 [cs].
- [331] Philipp Hacker, Ralf Krestel, Stefan Grundmann, and Felix Naumann. 2020. Explainable AI under Contract and Tort Law: Legal Incentives and Technical Challenges. *Artificial Intelligence and Law* 28, 4 (2020), 415–439. <https://doi.org/10.1007/s10506-020-09260-6> Published online 19 January 2020; print December 2020.
- [332] Pieter Haek. 2025. Pressure mounts on EU to use legal weapons against Musk's interference. <https://www.politico.eu/article/elon-musk-interference-alice-weidel-x-livestream-afd-germany-eu-digital-services-act/>
- [333] Blayne Haggart and Clara Iglesias Keller. 2021. Democratic Legitimacy in Global Platform Governance. *Telecommunications Policy* 45 (2021), 102152.
- [334] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 466:1–466:35. <https://doi.org/10.1145/3479610>
- [335] Mark A. Hall and Ronald F. Wright. 2008. Systematic Content Analysis of Judicial Opinions. *California Law Review* 96, 1 (2008), 63–122. arXiv:20439171
- [336] Wendy Hall, Ramine Tinati, and Will Jennings. 2018. From Brexit to Trump: Social Media's Role in Democracy. *Computer* 51, 1 (Jan. 2018), 18–27. <https://doi.org/10.1109/MC.2018.1151005> Conference Name: Computer.

- [337] Andrew Halterman and Katherine A. Keith. 2025. Codebook LLMs: Evaluating LLMs as Measurement Tools for Political Science Concepts. <https://doi.org/10.48550/arXiv.2407.10747> arXiv:2407.10747 [cs].
- [338] Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd. 2025. DeepSeek. <https://www.deepseek.com/>
- [339] Hilmy Hanif, Jorge Constantino, Marie-Therese Sekwenz, Michel van Eeten, Jolien Ubacht, Ben Wagner, and Yury Zhauniarovich. 2024. Navigating the EU AI Act Maze Using a Decision-Tree Approach. *ACM Journal on Responsible Computing* 1, 3 (2024), 1–16.
- [340] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*. Vol. 52. Elsevier, Amsterdam, 139–183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- [341] Ayako A Hasegawa, Daisuke Inoue, and Mitsuaki Akiyama. 2024. How {WEIRD} is Usable Privacy and Security Research?. In *33rd USENIX Security Symposium (USENIX Security 24)*. 3241–3258.
- [342] Sam Havron, Diana Freed, Rahul Chatterjee, Damon McCoy, Nicola Dell, and Thomas Ristenpart. 2019. Clinical computer security for victims of intimate partner violence. In *28th USENIX Security Symposium (USENIX Security 19)*. 105–122.
- [343] Gaole He, Nilay Aishwarya, and Ujwal Gadiraju. 2025. Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 907–924. <https://doi.org/10.1145/3708359.3712133>
- [344] Gaole He, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. 2024. Opening the Analogical Portal to Explainability: Can Analogies Help Laypeople in AI-assisted Decision Making? *Journal of Artificial Intelligence Research* 81 (2024), 117–162.
- [345] Gaole He, Abri Bharos, and Ujwal Gadiraju. 2024. To err is AI! Debugging as an intervention to facilitate appropriate reliance on AI systems. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*. 98–105.
- [346] Gaole He, Gianluca Demartini, and Ujwal Gadiraju. 2025. Plan-Then-Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents As A Daily Assistant. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Number 414 in CHI '25. Association for Computing Machinery, New York, NY, USA, 1–22. <https://doi.org/10.1145/3706598.3713218>
- [347] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*

- (CHI '23). Association for Computing Machinery, New York, NY, USA, 1–18. <https://doi.org/10.1145/3544548.3581025>
- [348] Marjorie Heins. 2013. The Brave New World of Social Media Censorship. *Harvard Law Review Forum* 127 (2013), 325.
- [349] Natali Helberger and Pamela Samuelson. 2024. The Digital Services Act as a Global Transparency Regime. *Verfassungsblog* (March 2024). <https://verfassungsblog.de/the-digital-services-act-as-a-global-transparency-regime/>
- [350] N. Helberger, M. Sax, J. Strycharz, and H.-W. Micklitz. 2022. Choice Architectures in the Digital Economy: Towards a New Understanding of Digital Vulnerability. *Journal of Consumer Policy* 45, 2 (June 2022), 175–200. <https://doi.org/10.1007/s10603-021-09500-5>
- [351] Natali Helberger, Marijn Sax, Joanna Strycharz, and H-W Micklitz. 2022. Choice architectures in the digital economy: Towards a new understanding of digital vulnerability. *Journal of Consumer Policy* 45, 2 (2022), 175–200.
- [352] Amélie Heldt. 2019. Let's Meet Halfway: Sharing New Responsibilities in a Digital Age. *Journal of Information Policy* 9 (2019), 336.
- [353] Amélie Heldt. 2019. Reading between the lines and the numbers: an analysis of the first NetzDG reports. *Internet Policy Review* 8, 2 (June 2019), 1–19. <https://policyreview.info/articles/analysis/reading-between-lines-and-numbers-analysis-first-netzdg-reports>
- [354] Amélie Heldt and Stephan Dreyer. 2021. Competent Third Parties and Content Moderation on Platforms: Potentials of Independent Decision-Making Bodies From A Governance Structure Perspective. *Penn State University Press* 11 (2021), 266.
- [355] John Allen Hendricks and Dan Schill. 2017. The Social Media Election of 2016. In *The 2016 US Presidential Campaign: Political Communication and Practice*, Robert E. Denton Jr (Ed.). Springer International Publishing, Cham, 121–150. https://doi.org/10.1007/978-3-319-52599-0_5
- [356] Canadian Heritage. 2024. Proposed Bill to address Online Harms. <https://www.canada.ca/en/canadian-heritage/services/online-harms.html> Last Modified: 2025-05-14.
- [357] Johanna Herman. 2024. Dark patterns: EU's regulatory efforts. *SECURITY AND PRIVACY* 7, 6 (2024), e441. <https://doi.org/10.1002/spy2.441> _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/spy2.441>.
- [358] Alex Hern. 2020. How Hate Speech Campaigners Found Facebook's Weak Spot. *The Guardian* (2020). <https://www.theguardian.com/technology/2020/jun/29/how-hate-speech-campaigners-found-facebooks-weak-spot>

- [359] Mireille Hildebrandt. 2018. Algorithmic regulation and the rule of law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2128 (Aug. 2018), 20170355. <https://doi.org/10.1098/rsta.2017.0355> Publisher: Royal Society.
- [360] Wolfgang Hoffmann-Riem. 2016. Legal Technology/Computational Law. (2016).
- [361] Franz Hofmann and Benjamin Raue. 2023. *Digital Services Act Gesetz über digitale Dienste*. Vol. 1. Nomos Verlagsgesellschaft mbH & Co. KG. <https://www.nomos-shop.de/de/p/digital-services-act-978-3-8487-7882-9>
- [362] Alexander Hohlfeld. 2024. DSA: Risk Assessment & Audit Database. https://docs.google.com/spreadsheets/d/12hJWpCFmHJMQLz1qkd60gGsMW82YcSgWJHXD7BHVps/htmlview?usp=embed_facebook
- [363] Mirya R. Holman, Monica C. Schneider, and Kristin Pondel. 2015. Gender Targeting in Political Advertisements. *Political Research Quarterly* 68, 4 (2015), 816–829. <https://www.jstor.org/stable/24637818> Publisher: [University of Utah, Sage Publications, Inc.].
- [364] Daniel Holznagel. 2024. Follow Me to Unregulated Waters!: Are Major Online Platforms Violating the DSA’s Rules on Notice and Action? *Verfassungsblog* (May 2024). <https://verfassungsblog.de/follow-me-to-unregulated-waters/>
- [365] Viktoria Horn, <https://orcid.org/0000-0002-2572-6575>, View Profile, Johannes Hirth, <https://orcid.org/0000-0001-9034-0321>, View Profile, Julian Holfeld, <https://orcid.org/0009-0003-8406-641X>, View Profile, Jens Hendrik Behmenburg, <https://orcid.org/0009-0004-4616-2010>, View Profile, Claude Draude, <https://orcid.org/0000-0001-8467-195X>, View Profile, Gerd Stumme, <https://orcid.org/0000-0002-0570-7908>, and View Profile. 2024. Disclosing Diverse Perspectives of News Articles for Navigating between Online Journalism Content. In *Proceedings of the 13th Nordic Conference on Human-Computer Interaction*. Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3679318.3685414>
- [366] Sudhamshu Hosamane, Alyvia Walters, Yao Lyu, and Shagun Jhaver. 2025. "I thought it was my mistake, but it's really the design": A Critical Examination of the Accessibility of User-Enacted Moderation Tools on Facebook and X. <https://doi.org/10.48550/arXiv.2509.10789> arXiv:2509.10789 [cs].
- [367] The White House. 2025. Fact Sheet: President Donald J. Trump Issues Directive to Prevent the Unfair Exploitation of American Innovation. <https://www.whitehouse.gov/fact-sheets/2025/02/fact-sheet-president-donald-j-trump-issues-directive-to-prevent-the-unfair-exploitation-of-american-innovation/>
- [368] Philip N. Howard and Bence Kollanyi. 2016. Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum. *SSRN Electronic Journal* (2016). <https://doi.org/10.2139/ssrn.2798311>

- [369] Louis M. Hsu and Ronald Field. 2003. Interrater Agreement Measures: Comments on Kappan, Cohen's Kappa, Scott's π , and Aickin's α . *Understanding Statistics* 2, 3 (Aug. 2003), 205–219. https://doi.org/10.1207/S15328031US0203_03
- [370] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55.
- [371] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55. <https://doi.org/10.1145/3703155>
- [372] Tao Huang. 2025. Content moderation by LLM: from accuracy to legitimacy. *Artificial Intelligence Review* 58, 10 (July 2025), 320. <https://doi.org/10.1007/s10462-025-11328-1>
- [373] Victor Huang, Yifan Hu, and Yaohua Li. 2022. A Systematic Literature Review of New Trends in Self-expression Caused by Emojis and Memes:. In *Proceedings of the Conference*. Atlantis Press, Sanya, China, 75–79. <https://doi.org/10.2991/as-sehr.k.220105.016>
- [374] Isabelle Hupont, Songül Tolan, Hatice Gunes, and Emilia Gómez. 2022. The landscape of facial processing applications in the context of the European AI Act and the development of trustworthy systems. *Scientific Reports* 12, 1 (2022), 10688.
- [375] Martin Husovec. 2018. The Promises of Algorithmic Copyright Enforcement: Take-down or Staydown? Which Is Superior? And Why? *JIPITEC* 9, 1 (Jan. 2018), 2–20. <https://www.jipitec.eu/issues/jipitec-9-1-2018/4683>
- [376] Martin Husovec. 2020. The Liability System of the Digital Services Act and the Use of Automation. *Common Market Law Review* 57 (2020), 1921–1954.
- [377] Martin Husovec. 2022. Will the DSA work? *Verfassungsblog* (Nov. 2022). <https://doi.org/10.17176/20221109-095546-0>
- [378] Martin Husovec. 2023. How to Facilitate Data Access under the Digital Services Act. <https://papers.ssrn.com/abstract=4452940>
- [379] Martin Husovec. 2024. The Digital Services Act's red line: what the Commission can and cannot do about disinformation. *Journal of Media Law* 16, 1 (Jan. 2024), 47–56. <https://doi.org/10.1080/17577632.2024.2362483> _eprint: <https://doi.org/10.1080/17577632.2024.2362483>.
- [380] Martin Husovec and Martin Husovec. 2024. *Principles of the Digital Services Act*. Oxford University Press, Oxford, New York.

- [381] Zaheer Hussain and Mark D Griffiths. 2021. The associations between problematic social networking site use and sleep quality, attention-deficit hyperactivity disorder, depression, anxiety and stress. *International Journal of Mental Health and Addiction* 19, 3 (2021), 686–700.
- [382] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 48:1–48:27. <https://doi.org/10.1145/3392854>
- [383] Basileal Imana, Aleksandra Korolova, and John Heidemann. 2023. Having Your Privacy Cake and Eating it Too: Platform-supported Auditing of Social Media Algorithms for Public Interest. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (apr 2023), 1–33. <https://doi.org/10.1145/3579610>
- [384] Petros Iosifidis and Nicholas Nicoli. 2020. *Digital Democracy, Social Media and Disinformation*. Routledge, London. <https://doi.org/10.4324/9780429318481>
- [385] Cherilyn Ireton, Julie Posetti, and UNESCO. 2018. *Journalism, 'fake news' et disinformation: Handbook for journalism education and training*. UNESCO. <http://unesdoc.unesco.org/images/0026/002655/265552E.pdf>
- [386] Mike Isaac. 2016. Facebook, in Cross Hairs After Election, Is Said to Question Its Influence. *The New York Times* (2016). <https://www.nytimes.com/2016/11/13/technology/facebook-in-cross-hairs-after-election-is-said-to-question-its-influence.html>
- [387] ISO. 2022. ISO/IEC/IEEE 24748-7000 Systems and software engineering — Life cycle management — Part 7000: Standard model process for addressing ethical concerns during system design. (2022).
- [388] IEC ISO. 2015. ISO/IEC/IEEE 15288:2015 Systems and software engineering — System life cycle processes. (2015).
- [389] Elena Izyumenko, Martin Senftleben, Nynke Schutte, Edith G. Smit, Guda van Noort, and Lieselotte van Velzen. 2024. Online behavioural advertising, consumer empowerment and fair competition: Are the DSA transparency obligations the right answer? <https://doi.org/10.2139/ssrn.4729118>
- [390] Jack Jamieson, Toru Hara, and Mitsuaki Akiyama. 2025. Flagging Emotional Manipulation: Impacts of Manipulative Content Warnings on Sharing Intentions and Perceptions of Health-Related Social Media Posts. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–9. <https://doi.org/10.1145/3706599.3719948>

- [391] Jun Li Jeung and Janet Yi-Ching Huang. 2023. Correct Me If I Am Wrong: Exploring How AI Outputs Affect User Perception and Trust. In *Computer Supported Cooperative Work and Social Computing* (2023-10-14). ACM, New York, NY, USA, 323–327. <https://doi.org/10.1145/3584931.3606997>
- [392] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, Vol. 3. CSCW, 1–35.
- [393] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X Zhang. 2023. Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–33.
- [394] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–33. <https://doi.org/10.1145/3610080>
- [395] Shagun Jhaver and Amy X. Zhang. 2025. Do users want platform moderation or individual control? Examining the role of third-person effects and free speech support in shaping moderation preferences. *New Media & Society* 27, 5 (May 2025), 2930–2950. <https://doi.org/10.1177/14614448231217993> Publisher: SAGE Publications.
- [396] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. Moderation Challenges in Voice-based Online Communities on Discord. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 55:1–55:23. <https://doi.org/10.1145/3359157>
- [397] Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. 2023. A Trade-off-centered Framework of Content Moderation. *ACM Transactions on Computer-Human Interaction* 30, 1 (Feb. 2023), 1–34. <https://doi.org/10.1145/3534929>
- [398] Huo Jingnan. 2024. AI-generated images have become a new form of propaganda this election season. *NPR* (Oct. 2024). <https://www.npr.org/2024/10/18/nx-s1-5153741/ai-images-hurricanes-disasters-propaganda>
- [399] Paul Johannesson and Erik Perjons. 2021. Knowledge types and forms. In *An introduction to design science*. Springer, 23–40.
- [400] Paul L. Joskow. 2007. Regulation of Natural Monopoly. In *Handbook of Law and Economics*. Vol. 2. Elsevier.
- [401] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the Looking Glass: Study of Transparency in Reddit’s Moderation Practices. *Proceedings of the ACM on Human-Computer Interaction* 4, GROUP (Jan. 2020), 1–35. <https://doi.org/10.1145/3375197>

- [402] Natascha Just and Florian Saurwein. 2024. Enhancing Social-Media Regulation through Transparency? Examining the New Transparency Regime in the EU. *TechREG Chronicle* 2024, 2 (Feb. 2024), online. <https://doi.org/10.5167/uzh-257668>
- [403] Katharina Kaesling. 2023. Liability and Platform Governance under the Digital Services Act. *Common Market Law Review* 60, 2 (2023), 551–578. <https://doi.org/10.54648/COLA2023065> Discusses expansion of the e-Commerce Directive scope under the DSA.
- [404] Kalum. 2025. Instagram Statistics in 2025 Every Marketer Should Know. Metricool Blog.
- [405] Eleni Kapantai, Androniki Christopoulou, Christos Berberidis, and Vassilios Peristeras. 2021. A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society* 23, 5 (May 2021), 1301–1326. <https://doi.org/10.1177/1461444820959296>
- [406] Emmanouil Kapantai, Christos Christos, Fatima Mekkaoui, and Bin He. 2021. A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society* 23, 9 (2021), 2601–2626.
- [407] Joel Kaplan. 2025. More Speech and Fewer Mistakes. <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>
- [408] Maria Karakiza. 2015. The impact of social media in the public sector. *Procedia-social and behavioral Sciences* 175 (2015), 384–392.
- [409] Matthew Katsaros, Jisu Kim, and Tom Tyler. 2024. Online Content Moderation: Does Justice Need a Human Face? *International Journal of Human-Computer Interaction* 40, 1 (Jan. 2024), 66–77. <https://doi.org/10.1080/10447318.2023.2210879> Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/10447318.2023.2210879>.
- [410] Matthew Katsaros, Kathy Yang, and Lauren Fratamico. 2021. Reconsidering Tweets: Intervening During Tweet Creation Decreases Offensive Content. <http://arxiv.org/abs/2112.00773> arXiv:2112.00773 [cs].
- [411] Stuart A. Kauffman. 1993. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press.
- [412] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. 2022. Sensible AI: Re-imagining interpretability and explainability using sensemaking theory. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 702–714.
- [413] Rishabh Kaushal, Jacob Van De Kerkhof, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. 2024. Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database. In *Proceedings of the*

- 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). Association for Computing Machinery, New York, NY, USA, 1121–1132. <https://doi.org/10.1145/3630106.3658960>
- [414] Rishabh Kaushal, Jacob Van De Kerkhof, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. 2024. Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Rio de Janeiro, Brazil, 1121–1132.
- [415] Daphne Keller. [n. d.]. Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users. <https://lawreviewblog.uchicago.edu/2022/06/28/keller-control-over-speech/>
- [416] Daphne Keller. 2020. Facebook Filters, Fundamental Rights, and the CJEU's Glawischnig-Piesczek Ruling. *GRUR International* 69, 7 (2020), 616–626.
- [417] Daphne Keller. 2022. The DSA's Industrial Model for Content Moderation. <https://verfassungsblog.de/dsa-industrial-model/>
- [418] Daphne Keller. 2023. The Rise of the Compliant Speech Platform. <https://www.lawfaremedia.org/article/the-rise-of-the-compliant-speech-platform>
- [419] Daphne Keller. 2023. Rushing to Launch the EU's Platform Database Experiment. <https://cyberlaw.stanford.edu/blog/2023/07/rushing-launch-eus-platform-database-experiment/>
- [420] Daphne Keller and Paddy Leerssen. 2021. Some Humility About Transparency. <https://cyberlaw.stanford.edu/blog/2021/03/some-humility-about-transparency>
- [421] Thomas F. Keller. 1968. Discussion of The Auditing Standard of Consistency. *Journal of Accounting Research* 6 (1968), 18–22. <https://doi.org/10.2307/2490064> Publisher: [Accounting Research Center, Booth School of Business, University of Chicago, Wiley].
- [422] Jakko Kemper and Daan Kolkman. 2019. Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society* 22, 14 (Dec. 2019), 2081–2096. <https://doi.org/10.1080/1369118X.2018.1477967>
- [423] Matthias Kettemann, Felicitas Rachinger, and Marie-Therese Sekwenz. 2022. Deplatforming. In *Regulierung von Kommunikationsplattformen: Aktuelle Fragen der Umsetzung*, Christoph Grabenwarter, Michael Holoubek, and Marcella Rami (Eds.). MANZ Verlag, Vienna, 78–89. <https://doi.org/10.5771/9783214164935-78>
- [424] Matthias C. Kettemann and M.T. Sekwenz. 2022. Pandemics and Platforms: Private Governance of (Dis)Information in Crisis Situations. In *Pandemocracy in Europe*.

- Hart Publishing, 263–282. <https://www.bloomsbury.com/uk/pandemocracy-in-europe-9781509946365/>
- [425] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating Behavior in Online Communities. In *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press, 125–178.
- [426] Kyungdong Kim, Min Han Kim, Youngmin Cho, et al. 2025. The mobilization effect of social media use: an instrumental variable approach. *Journal of Policy Studies* 40, 1 (2025), 65–87.
- [427] Kyung-Sun Kim, Sei-Ching Joanna Sin, and Tien-I Tsai. 2014. Individual Differences in Social Media Use for Information Seeking. *The Journal of Academic Librarianship* 40, 2 (March 2014), 171–178. <https://doi.org/10.1016/j.acalib.2014.03.001>
- [428] John Kingston. 2017. Using Artificial Intelligence to Support Compliance with the General Data Protection Regulation. *Artificial Intelligence and Law* 25, 4 (2017), 429–443. <https://doi.org/10.1007/s10506-017-9206-9> Published online 1 September 2017.
- [429] Michał Kłincewicz, Mark Alfano, and Amir Ebrahimi Fard. 2025. Slopaganda: The interaction between propaganda and generative AI. <https://doi.org/10.48550/arXiv.2503.01560> arXiv:2503.01560 [cs].
- [430] Kate Klonick. 2018. The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review* 131, 6 (2018), 1598–1670.
- [431] Ferdinand Knapp. 2025. JD Vance: EU should not be ‘attacking American companies over garbage’. <https://www.politico.eu/article/jd-vance-eu-should-not-be-attacking-american-companies-over-garbage-x-elon-musk/>
- [432] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019-05-02). ACM, Glasgow, Scotland, UK, 1–14. <https://doi.org/10.1145/3290605.3300641>
- [433] Moritz Koerber. 2019. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018): Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics*, S. Bagnara, R. Tartaglia, T. Albolino, T. Alexander, and Y. Fujita (Eds.). Springer, Cham, 13–30.
- [434] Emmanouela Kokolaki, Evangelia Daskalaki, Katerina Psaroudaki, Meltini Christodoulaki, and Paraskevi Fragopoulou. 2020. Investigating the dynamics of illegal online activity: The power of reporting, dark web, and related legislation. *Computer Law & Security Review* 38 (Sept. 2020), 105440. <https://doi.org/10.1016/j.clsr.2020.105440>

- [435] Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. LLM-Mod: Can Large Language Models Assist Content Moderation?. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–8. <https://doi.org/10.1145/3613905.3650828>
- [436] Lisa Kusters and Oskar Josef Gstrein. 2023. TikTok and Transparency Obligations in the EU Digital Services Act (DSA) – A Scoping Review. <https://doi.org/10.2139/ssrn.4652543>
- [437] Ivica Kostric, Krisztian Balog, and Ujwal Gadiraju. 2025. Should We Tailor the Talk? Understanding the Impact of Conversational Styles on Preference Elicitation in Conversational Recommender Systems. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25)*. Association for Computing Machinery, New York, NY, USA, 164–173. <https://doi.org/10.1145/3699682.3728353>
- [438] Yubo Kou and Xinning Gui. 2021. Flag and Flagability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3411764.3445279>
- [439] Yubo Kou and Xinning Gui. 2021. Flag and flagability in automated moderation: The case of reporting toxic behavior in an online game community. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. Association for Computing Machinery, Yokohama, Japan, 1–12.
- [440] Yubo Kou, Renkai Ma, Zinan Zhang, Yingfan Zhou, and Xinning Gui. 2024. Community Begins Where Moderation Ends: Peer Support and Its Implications for Community-Based Rehabilitation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. <https://doi.org/10.1145/3613904.3642675>
- [441] Vlasios Koutsos, Xiangnan Tian, Dimitrios Papadopoulos, and Dimitris Chatzopoulos. 2024. Cross Ledger Transaction Consistency for Financial Auditing. In *6th Conference on Advances in Financial Technologies (AFT 2024) (Leibniz International Proceedings in Informatics (LIPIcs))*, Rainer Böhme and Lucianna Kiffer (Eds.), Vol. 316. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 4:1–4:25. <https://doi.org/10.4230/LIPIcs.AFT.2024.4> ISSN: 1868-8969.
- [442] PM Krafft, Meg Young, Michael Katell, Jennifer E Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, et al. 2021. An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 772–781.
- [443] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2017. Accountable Algorithms. *University of*

- Pennsylvania Law Review* 165, 3 (2017), 633–705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/
- [444] Johanne Kübler, Marie-Theres Sekwenz, Felicitas Rachinger, Anna König, Rita Gsenger, Eliška Pírková, Ben Wagner, Matthias C. Kettemann, Michael Krennerich, and Carolina Ferro. 2023. The 2021 German Federal Election on Social Media: Analysing Electoral Risks Created by Twitter and Facebook. In *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS 55)*. University of Hawai'i at Mānoa, Maui, Hawaii, USA, 4036–4045.
- [445] Deepak Kumar, Yousef AbuHashem, and Zakir Durumeric. 2024. Watch Your Language: Investigating Content Moderation with Large Language Models. arXiv:cs/2309.14517
- [446] Deepak Kumar, Yousef AbuHashem, and Zakir Durumeric. 2024. Watch Your Language: Investigating Content Moderation with Large Language Models. <https://doi.org/10.48550/arXiv.2309.14517> arXiv:2309.14517 [cs].
- [447] Alisa Küper and Nicole Krämer. 2025. Psychological traits and appropriate reliance: Factors shaping trust in AI. *International Journal of Human–Computer Interaction* 41, 7 (2025), 4115–4131.
- [448] Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2024. Large Language Models in Law: A Survey. *AI Open* 5 (2024), 181–196. <https://doi.org/10.1016/j.aiopen.2024.09.002>
- [449] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [450] Daniele Lain, Kari Kostinen, and Srdjan Čapkun. 2022. Phishing in organizations: Findings from a large-scale and long-term study. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, IEEE, San Francisco, CA, USA, 842–859.
- [451] Khoa Lam, Benjamin Lange, Borhane Blili-Hamelin, Jovana Davidovic, Shea Brown, and Ali Hasan. 2024. A Framework for Assurance Audits of Algorithmic Systems. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 1078–1092. <https://doi.org/10.1145/3630106.3658957>
- [452] Landgerichts Berlin. 2020. Beschluss vom 11.03.2020. <https://openjur.de/u/2253932.html>
- [453] Kyle Langvardt. 2017. Regulating Online Content Moderation. *Georgetown Law Journal* 106 (2017), 1353.
- [454] Candice Lanius, Ryan Weber, and William I. MacKenzie. 2021. Use of bot and content flags to limit the spread of misinformation among social networks: a behavior and attitude survey. *Social Network Analysis and Mining* 11, 1 (March 2021), 32. <https://doi.org/10.1007/s13278-021-00739-x>

- [455] Benjamin Laufer, Sameer Jain, A. Feder Cooper, Jon Kleinberg, and Hoda Heidari. 2022. Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 401–426. <https://doi.org/10.1145/3531146.3533107>
- [456] Johann Laux, Sandra Wachter, and Brent Mittelstadt. 2021. Taming the few: Platform regulation, independent audits, and the risks of capture created by the DMA and DSA. *Computer law & Security review* 43 (2021), 105613.
- [457] Stephanie Law. 2024. Effective enforcement of the Online Safety Act and Digital Services Act: unpacking the compliance and enforcement regimes of the UK and EU's online safety legislation. *Journal of Media Law* 16, 2 (July 2024), 263–300. <https://doi.org/10.1080/17577632.2025.2459441> Publisher: Routledge _eprint: <https://doi.org/10.1080/17577632.2025.2459441>.
- [458] Chelsea Lee and Robert L. Ostergard. 2017. Measuring Discrimination Against LGBTQ People: A Cross-National Analysis. *Human Rights Quarterly* 39, 1 (2017), 37–72. arXiv:44488971
- [459] Hee-Eun Lee, Tatiana Ermakova, Vasilis Ververis, and Benjamin Fabian. 2020. Detecting Child Sexual Abuse Material: A Comprehensive Survey. *Forensic Science International: Digital Investigation* 34 (Sept. 2020), 301022. <https://doi.org/10.1016/j.fsidi.2020.301022>
- [460] Paddy Leerssen. 2023. An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation. *Computer Law & Security Review* 48 (April 2023), 105790. <https://doi.org/10.1016/j.clsr.2023.105790>
- [461] Paddy Leerssen. 2024. Outside the Black Box: From Algorithmic Transparency to Platform Observability in the Digital Services Act. *Weizenbaum Journal of the Digital Society* 4, 2 (May 2024). <https://doi.org/10.34669/wi.wjds/4.2.3> Number: 2.
- [462] Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Mädche, Gerhard Schwabe, and Ali Sunyaev. 2024. HILL: A Hallucination Identifier for Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024-05-11). ACM, Honolulu, HI, USA, 1–13. <https://doi.org/10.1145/3613904.3642428>
- [463] Mark Leiser and Cristiana Santos. 2023. Dark Patterns, Enforcement, and the emerging Digital Design Acquis: Manipulation beneath the Interface. *Dark Patterns, Enforcement, and the emerging Digital Design Acquis: Manipulation beneath the Interface* (May 2023), 1–31. Preprint, not yet assigned to a journal volume or number. Publisher: SSRN.

- [464] Mark A. Lemley. 2025. Protecting Consumers in a Post-Consent World. *SSRN Working Paper* (2025).
- [465] Francesco Paolo Levantino and Federica Paolucci. 2025. Advancing the Protection of Fundamental Rights through AI Regulation: How the EU and the Council of Europe Are Shaping the Future. In *European Yearbook on Human Rights 2024*. Brill Nijhoff, 3–37. https://doi.org/10.1163/9789004708389_002 Section: European Yearbook on Human Rights 2024.
- [466] LG Bamberg. 2023. Urteil vom 15.12.2023. <https://openjur.de/u/2505424.html>
- [467] LG Brandenburg. 2023. Landesrechtsportal Brandenburg | Entscheidungsdatenbank der Gerichte in Brandenburg. <https://gerichtsentscheidungen.brandenburg.de/gerichtsentscheidung/22167>
- [468] LG Frankfurt am Main. 2018. vom 13.09.2018. <https://openjur.de/u/2181591.html>
- [469] LG Meiningen. 2022. Beschluss vom 05.08.2022. <https://openjur.de/u/2448105.html>
- [470] LG Stuttgart. 2005. Urteil vom 15.06.2005. <https://openjur.de/u/354308.html>
- [471] LG Würzburg. 2017. Urteil vom 07.03.2017. <https://openjur.de/u/2281788.html>
- [472] Feng Li. 2021. Sustainable Competitive Advantages via Temporary Advantages: Insights from the Competition between American and Chinese Digital Platforms in China. *British Journal of Management* (2021). Online first.
- [473] Yifang Li and Kelly Caine. 2022. Obfuscation Remedies Harms Arising from Content Flagging of Photos. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–25. <https://doi.org/10.1145/3491102.3517520>
- [474] Yueqi Li and Sanjay Goel. 2024. Making It Possible for the Auditing of AI: A Systematic Review of AI Audits and AI Auditability. *Information Systems Frontiers* 26, 3 (jul 2024), 1–31. <https://doi.org/10.1007/s10796-024-10508-8>
- [475] Anna Liesenfeld. 2025. The Legal Significance of Independent Research based on Article 40 DSA for the Management of Systemic Risks in the Digital Services Act. *European Journal of Risk Regulation* 16, 1 (March 2025), 184–196. <https://doi.org/10.1017/err.2024.61>
- [476] Daniel Link and Bernd Hellingrath. 2016. A Human-is-the-Loop Approach for Semi-Automated Content Moderation. (2016), 13.

- [477] LinkedIn. 2024. *LinkedIn 2023 Risk Assessment Report*. Technical Report. LinkedIn. https://content.linkedin.com/content/dam/help/tns/en/LinkedIn_2023_Risk_Assessment_Report.pdf
- [478] Gabriel Lins De Holanda Coelho, Paul H. P. Hanel, and Lukas J. Wolf. 2020. The Very Efficient Assessment of Need for Cognition: Developing a Six-Item Version. *Assessment* 27, 8 (2020), 1870–1885. <https://doi.org/10.1177/1073191118793208>
- [479] Mingli Liu, Kimberly E Kamper-DeMarco, Jie Zhang, Jia Xiao, Daifeng Dong, and Peng Xue. 2022. Time spent on social media and risk of depression in adolescents: a dose–response meta-analysis. *International journal of environmental research and public health* 19, 9 (2022), 5164.
- [480] Sonia Livingstone, Giovanna Mascheroni, et al. 2017. Maximizing Opportunities and Minimizing Risks for Children Online: The Role of Digital Skills in Emerging Strategies of Parental Mediation. *Journal of Communication* 67, 1 (2017), 82–105.
- [481] Emma J Llansó. 2020. No amount of “AI” in content moderation will solve filtering’s prior-restraint problem. *Big Data & Society* 7, 1 (Jan. 2020), 2053951720920686. <https://doi.org/10.1177/2053951720920686> Publisher: SAGE Publications Ltd.
- [482] Paola Lopez. 2023. Power and Resistance in the Twitter Bias Discourse. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’23)*. Association for Computing Machinery, New York, NY, USA, 603. <https://doi.org/10.1145/3593013.3594027>
- [483] Philipp Lorenz-Spreen, Lisa Oswald, Stephan Lewandowsky, and Ralph Hertwig. 2023. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour* 7, 1 (Jan. 2023), 74–101. <https://doi.org/10.1038/s41562-022-01460-1>
- [484] Juniper L Lovato, Antoine Allard, Randall Harp, Jeremiah Onaolapo, and Laurent Hébert-Dufresne. 2022. Limits of individual consent and models of distributed consent in online social networks. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Seoul, Republic of Korea, 2251–2262.
- [485] Juniper L. Lovato, Antoine Allard, Randall Harp, Jeremiah Onaolapo, and Laurent Hébert-Dufresne. 2022. Limits of Individual Consent and Models of Distributed Consent in Online Social Networks. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’22)*. Association for Computing Machinery, New York, NY, USA, 2251–2262. <https://doi.org/10.1145/3531146.3534640>
- [486] Zhuoran Lu, Dakuo Wang, and Ming Yin. 2024. Does more advice help? the effects of second opinions in AI-assisted decision making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–31.

- [487] Yao Lyu, Jie Cai, Anisa Callis, Kelley Cotter, and John M. Carroll. 2024. "I Got Flagged for Supposed Bullying, Even Though It Was in Response to Someone Harassing Me About My Disability.": A Study of Blind TikTokers' Content Moderation Experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3613904.3642148>
- [488] Maggie Ma. 2021. Rose, Thorn, Bud.
- [489] Renkai Ma and Yubo Kou. 2022. "I'm not sure what difference is between their content and mine, other than the person itself": A Study of Fairness Perception of Content Moderation on YouTube. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (Nov. 2022), 425:1–425:28. <https://doi.org/10.1145/3555150>
- [490] Stuart Macdonald and Katy Vaughan. 2024. Moderating borderline content while respecting fundamental values. *Policy & Internet* 16, 2 (2024), 347–361. <https://doi.org/10.1002/poi3.376> _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/poi3.376>.
- [491] Sapna Maheshwari. 2023. TikTok Pushes Back Against Claims It Fuels Antisemitism. *The New York Times*, Business section. <https://www.nytimes.com/2023/11/02/business/tiktok-antisemitism-claims-israel-palestinians.html> Accessed July 5, 2024.
- [492] Gianclaudio Malgieri. 2025. Scalable Fairness: The legal tool against power. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 2127–2137.
- [493] Gianclaudio Malgieri and Bart Custers. 2018. Pricing privacy – the right to know the value of your personal data. *Computer Law & Security Review* 34, 2 (April 2018), 289–303. <https://doi.org/10.1016/j.clsr.2017.08.006>
- [494] Gianclaudio Malgieri and Cristiana Santos. 2025. Assessing the (severity of) impacts on fundamental rights. *Computer Law & Security Review* 56 (April 2025), 106113. <https://doi.org/10.1016/j.clsr.2025.106113>
- [495] Gianclaudio Malgieri and Daniel Susser. 2025. Digital Dependency: Needs, Vulnerability, and Power in the Platform Economy. <https://doi.org/10.2139/ssrn.5196081>
- [496] Haroon Malik and Zifeng Tian. 2017. A Framework for Collecting YouTube Meta-Data. *Procedia Computer Science* 113 (Dec. 2017), 194–201. <https://doi.org/10.1016/j.procs.2017.08.347>
- [497] Chris Marsden. 2022. Platform law in Europe: combatting digital harms through co-regulation. *Journal of Law & Economic Regulation* 15, 1 (2022), 34–66. <https://research.monash.edu/en/publications/platform-law-in-europe-combatting-digital-harms-through-co-regula/> Publisher: Seoul National University.

- [498] Chris Marsden, Trisha Meyer, and Ian Brown. 2020. Platform values and democratic elections: How can the law regulate digital disinformation? *Computer Law & Security Review* 36 (2020), 105373. <https://doi.org/10.1016/j.clsr.2019.105373>
- [499] Christopher T. Marsden, Ian Brown, and Michael Veale. 2020. Responding to Disinformation: Ten Recommendations for Regulatory Action and Forbearance. <https://doi.org/10.31228/osf.io/aerw9>
- [500] Franziska Martini, Paul Samula, Tobias R Keller, and Ulrike Klinger. 2021. Bot, or Not? Comparing Three Methods for Detecting Social Bots in Five Political Discourses. *Big Data & Society* 8, 2 (July 2021), 205395172110335. <https://doi.org/10.1177/20539517211033566>
- [501] Arunesh Mathur, Gunes Acar, Michael J Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–32.
- [502] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What makes a dark pattern... dark? Design attributes, normative considerations, and measurement methods. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. Association for Computing Machinery, Yokohama, Japan, 1–18.
- [503] Tara Matthews, Elie Bursztein, Patrick Gage Kelley, Lea Kissner, Andreas Kramm, Andrew Oplinger, Andreas Schou, Manya Sleeper, Stephan Somogyi, Dalila Szostak, et al. 2025. Supporting the Digital Safety of At-Risk Users: Lessons Learned from 9+ Years of Research and Training. *ACM Transactions on Computer-Human Interaction* 32, 3 (2025), 1–39.
- [504] Samuel Mayworm, Shannon Li, Hibby Thach, Daniel Delmonaco, Christian Paneda, Andrea Wegner, and Oliver L. Haimson. 2024. The Online Identity Help Center: Designing and Developing a Content Moderation Policy Resource for Marginalized Social Media Users. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 1–30. <https://doi.org/10.1145/3637406>
- [505] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The Cost of Reading Privacy Policies. *I/S: A Journal of Law and Policy for the Information Society* 4, 3 (2008), 543–568.
- [506] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.
- [507] Bethan McKernan. 2023. Israel and Hamas at war after surprise attacks from Gaza Strip. <https://www.theguardian.com/world/2023/oct/07/hamas-launches-surprise-attack-on-israel-as-palestinian-gunmen-reported-in-south> World news.

- [508] Who Targets Me. 2020. A Goldilocks Zone for political ads. <https://whotargets.me/en/a-goldilocks-zone-for-political-ads/>
- [509] Who Targets Me. 2021. How to correctly identify political ads (while acknowledging you can't). <https://whotargets.me/en/how-to-correctly-identify-political-ads/>
- [510] Morgan Meaker. [n. d.]. Slovakia's Election Deepfakes Show AI Is a Danger to Democracy. *Wired* ([n. d.]). <https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/> Section: tags.
- [511] Siddharth Mehrotra, Ujwal Gadiraju, Eva Bittner, Folkert van Delden, Catholijn M. Jonker, and Myrthe L. Tielman. 2025. "Even explanations will not help in trusting [this] fundamentally biased system": A Predictive Policing Case-Study. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. 51–62.
- [512] Siddharth Mehrotra, Ujwal Gadiraju, Eva Bittner, Folkert van Delden, Catholijn M. Jonker, and Myrthe L. Tielman. 2025. "Even explanations will not help in trusting [this] fundamentally biased system": A Predictive Policing Case-Study. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25)*. Association for Computing Machinery, New York, NY, USA, 51–62. <https://doi.org/10.1145/3699682.3728343>
- [513] Manisha Mehta and Fausto Giunchiglia. 2025. Understanding Gen Alpha's Digital Language: Evaluation of LLM Safety Systems for Content Moderation. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 2863–2873. <https://doi.org/10.1145/3715275.3732184>
- [514] Colten Meisner. [n. d.]. The weaponization of platform governance: Mass reporting and algorithmic punishments in the creator economy. *Policy & Internet* n/a, n/a ([n. d.]). <https://doi.org/10.1002/poi3.359>
- [515] Mentimeter. 2025. Menti Online Voting. <https://www.menti.com/>
- [516] Meta. 2020. Measuring Our Progress Combating Hate Speech. <https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech/> Accessed 2 May 2023.
- [517] Meta. 2023. Internet Regulations | Meta. <https://about.meta.com/regulations/>
- [518] Meta. 2024. *Instagram Transparency Report April-September*. Technical Report. Meta. https://scontent-ams4-1.xx.fbcdn.net/v/t39.8562-6/466943155_1291701138400105_7867447844898917200_n.pdf?_nc_cat=103&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=a77fZ3TJmhEQ7kNvgGXTNZg&_nc_zt=14&_nc_ht=scontent-ams4-1.xx&_nc_gid=ArDwCYpyo0t0ChniCYwXUqz&oh=00_AYBr_PrhwGnLiFMy7gJl_1Qcou0J4C_ZPdR6F3nBZInLnA&oe=6781B09D

- [519] Meta. 2024. *Systemic Risk Assessment and Mitigation Report for Facebook*. Systemic Risk Assessment Report. Meta Platforms, Inc. https://scontent-ams4-1.xx.fbcdn.net/v/t39.8562-6/468433223_2965672840272736_5366479269132269710_n.pdf?_nc_cat=109&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=4I-o7iu9Df8Q7kNvgHwdJ-s&_nc_zt=14&_nc_ht=scontent-ams4-1.xx&_nc_gid=A5U0EkvemFseq7DBtR4eTnZ&oh=00_AYA50bYAlIgpIHM9AYJmHP5rz0X9keQ622_o iXwQXH9uXw&oe=678C28A5
- [520] Meta. 2024. *Transparency Report for Instagram*. Technical Report. Meta. https://scontent-ams4-1.xx.fbcdn.net/v/t39.8562-6/446849908_1440652056655341_3316525698930778970_n.pdf?_nc_cat=103&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=kiNdYQXaOUQq7kNvgFH8Sc0&_nc_ht=scontent-ams4-1.xx&oh=00_AYCWrlinlydNTgUj1BsdtLsWpTeQsVx8AlCfV7BaMhx1TQ&oe=66627760
- [521] Meta AI. 2025. Large Language Model Meta AI. <https://www.llama.com/>
- [522] Meta Platforms Ireland Limited. 2023. Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Instagram: First Report. https://scontent-ams2-1.xx.fbcdn.net/v/t39.8562-6/447971060_1481740992549061_1404827436118992020_n.pdf?_nc_cat=100&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=BeeXozqT008Q7kNvwGG086K&_nc_oc=AdmBsKP4zr6sQu_YT7SqYT9jCcbjBvLl9WRdG0gQC1ZDHNfiZ0b9QmxV3p77_lvKZ1k&_nc_zt=14&_nc_ht=scontent-ams2-1.xx&_nc_gid=N85RCBBKvEYojc2_pYcTYw&oh=00_AfLx57-917GcM3jki7_kP_lmDi0Hv-8n6SL3cizrI3mAvg&oe=683221EC Accessed 20 May 2025.
- [523] Meta Platforms Ireland Limited. 2024. Instagram Systemic Risk Assessment under the Digital Services Act. <https://about.instagram.com/dsa/systemic-risk-assessment-2024> Accessed May 2025.
- [524] Meta Platforms Ireland Limited. 2024. Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Instagram: Second Report. https://scontent-ams2-1.xx.fbcdn.net/v/t39.8562-6/448110749_316590818169033_3098716353127960367_n.pdf?_nc_cat=111&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=0xOPmvNylbYQ7kNvwHHNk8&_nc_oc=AdnDwlVXksVltm6a8SSSQfLCPgpcvFpv0VZDgvmM5-I8c4kCoECWdl5e-W8lyT49pQE&_nc_zt=14&_nc_ht=scontent-ams2-1.xx&_nc_gid=Q9cdNG7sQ00Uj-gNXrrEoQ&oh=00_AfKuv60EqelLhzTlhSPF8UYFm-LU11h0480ubN7I4RVf6w&oe=68322ED7 Accessed 20 May 2025.
- [525] Meta Platforms Ireland Limited. 2024. Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Instagram: Third Report. https://scontent-ams4-1.xx.fbcdn.net/v/t39.8562-6/466943155_1291701138400105_7867447844898917200_n.pdf?_nc_cat=103&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=QCLU7-Zi7nwQ7kNvwGBp4n0&_nc_oc=AdmTcf41M0OtGngJT4kfeTrhqnxVTtA0QwB0awSSXzvvE_vZUYRf-kESXbPE_bKMT0c&_nc_zt=14&_nc_ht=scontent-ams4-1.xx&_nc_gid=AfWBhcnPSxK_2Z5VSIynFg&oh=00_AfKbtd_GUNsOVuwLtOtrilwqvW5XhvpHplyJulyJpWKzRg&oe=6832215D Accessed 20 May 2025.

- [526] Meta Platforms Ireland Limited. 2025. Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Instagram: Fourth Report. https://scontent-ams4-1.xx.fbcdn.net/v/t39.8562-6/492672947_1857060258361688_2578453831459945466_n.pdf?_nc_cat=102&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=L11hY0Ey6IoQ7kNvWEjfPOW&_nc_oc=Adk7gPIkfQ45GnY8UNvZu07d9zY5PH003dyfbAk0wEAx8qrGikPpH2AJ1tGBHOU81-E&_nc_zt=14&_nc_ht=scontent-ams4-1.xx&_nc_gid=h-hu5ehYDfSv2w1RXk1iCQ&oh=00_AfL67o0untYG1FdZNsBZfhiq9fwpMGQRJ79FaXhQQA-Cvg&oe=68320F3D Accessed 20 May 2025.
- [527] Susan Michie, Maartje M Van Stralen, and Robert West. 2011. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implementation science* 6 (2011), 1–12.
- [528] Sally Broughton Micova. 2021. What is the Harm in Size: Very Large Online Platforms in the Digital Services Act.
- [529] Katharine Miller. 2023. AI overreliance is a problem. Are explanations a solution. *HAI-Stanford University Human-Centered Artificial Intelligence* (2023).
- [530] Tim Miller. 2023. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. 333–342.
- [531] Dan Milmo. 2023. Social Media Platforms to Face £18m Fines for Publishing Animal Cruelty Content. *The Guardian*, Technology section. <https://www.theguardian.com/technology/2023/sep/07/social-media-platforms-to-face-18m-fines-for-publishing-animal-cruelty-content> Accessed July 5, 2024.
- [532] Dan Milmo, Amelia Gentleman, and Aisha Down. 2026. Grok being used to create sexually violent videos featuring women, research finds. *The Guardian* (Jan. 2026). <https://www.theguardian.com/technology/2026/jan/09/grok-ai-create-sexually-violent-videos-featuring-women-research-finds>
- [533] Ashlee Milton, Leah Ajmani, Michael Ann DeVito, and Stevie Chancellor. 2023. “I see me here”: mental health content, community, and algorithmic curation on TikTok. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–17.
- [534] Ministry of Electronics and Information Technology, Government of India. 2023. Intermediary Guidelines and Digital Media Ethics Code India. <https://www.meity.gov.in/writereaddata/files/Information%20Technology%20%28Intermediary%20Guidelines%20and%20Digital%20Media%20Ethics%20Code%29%20Rules%2C%202021%20%28updated%2006.04.2023%29-.pdf>
- [535] Miro. 2025. Miro Collaboration Platform. <https://miro.com/>
- [536] Tanner Mirrlees. 2021. GAFAM and Hate Content Moderation: Deplatforming and Deleting the Alt-right. In *Media and Law: Between Free Speech and Censorship*, Mathieu Deflem and Derek M. D. Silva (Eds.). Sociology of Crime, Law and Deviance,

- Vol. 26. Emerald Publishing Limited, 81–97. <https://doi.org/10.1108/S1521-613620210000026006>
- [537] Saeedeh Mohammadi and Taha Yasseri. 2025. AI Feedback Enhances Community-Based Content Moderation through Engagement with Counterarguments. *arXiv preprint arXiv:2507.08110* (2025).
- [538] Jorge Morais Carvalho, Martim Ferreira dos Santos Farinha, and Francisco Argal e Lima. 2021. Introduction to the Digital Services Act, Content Moderation and Consumer Protection. (2021). <http://hdl.handle.net/10362/165152>
- [539] Rachel Elizabeth Moran, Joseph Schafer, Mert Bayar, and Kate Starbird. 2025. The End of Trust and Safety?: Examining the Future of Content Moderation and Upheavals in Professional Online Safety Efforts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3706598.3713662>
- [540] Patricia L. Moravec, Antino Kim, and Alan R. Dennis. 2020. Appealing to Sense and Sensibility: System 1 and System 2 Interventions for Fake News on Social Media. *Info. Sys. Research* 31, 3 (Aug. 2020), 987–1006. <https://doi.org/10.1287/isre.2020.0927>
- [541] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose. *Proceedings of the International AAAI Conference on Web and Social Media* 7, 1 (2013), 400–408. <https://doi.org/10.1609/icwsm.v7i1.14401>
- [542] Kathleen L Mosier and Linda J Skitka. 2018. Human decision makers and automated decision aids: Made for each other? In *Automation and human performance*. CRC Press, Boca Raton, FL, 201–220.
- [543] Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing LLM and human annotations of conversational safety. <https://doi.org/10.48550/arXiv.2406.06369> arXiv:2406.06369 [cs].
- [544] Hamdy Mubarak, Samir Abdaljalil, Azza Nassar, and Firoj Alam. 2023. Detecting and Reasoning of Deleted Tweets before They Are Posted. <https://doi.org/10.48550/arXiv.2305.04927> arXiv:cs/2305.04927
- [545] Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Commun. ACM* 36, 6 (1993), 24–28.
- [546] Ralf Müller-Terpitz and Markus Köhler (Eds.). 2024. *Digital Services Act (DSA): Gesetz über digitale Dienste. Kommentar*. C.H. Beck.
- [547] John Murray and Mikkel Flyverbom. 2021. Datafied corporate political activity: Updating corporate advocacy for a digital era. *Organization* 28, 4 (July 2021), 621–640. <https://doi.org/10.1177/1350508420928516>

- [548] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2024. Auditing large language models: a three-layered approach. *AI and Ethics* 4, 4 (Nov. 2024), 1085–1115. <https://doi.org/10.1007/s43681-023-00289-2>
- [549] Coimisiún na Meán. 2024. *Article 22 Trusted Flaggers Application Form and Guidance*. Technical Report. https://www.cnam.ie/app/uploads/2024/02/2024_0216_Article22_GuidanceForm_Branded_vF_KW.pdf
- [550] Teresa K Naab, Anja Kalch, and Tino GK Meitz. 2018. Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society* 20, 2 (Feb. 2018), 777–795. <https://doi.org/10.1177/1461444816670923> Publisher: SAGE Publications.
- [551] Teresa K Naab, Anja Kalch, and Tino GK Meitz. 2018. Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society* 20, 2 (2018), 777–795.
- [552] Mia Nahrgang et al. 2025. Written for Lawyers or Users? Mapping the Complexity of Community Guidelines. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 19. 1295–1307.
- [553] Philip M. Napoli. 2020. Social Media and the Public Interest: Media Regulation in the Disinformation Age. *Columbia Journal of Law & the Arts* 43, 3 (2020), 529–538.
- [554] Herman Narula. 2022. *Virtual society: The metaverse and the new frontiers of human experience*. Crown Currency.
- [555] Arbi Haza Nasution and Aytuğ Onan. 2024. ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. *IEEE Access* 12 (2024), 71876–71900. <https://doi.org/10.1109/ACCESS.2024.3402809> Conference Name: IEEE Access.
- [556] L. Naudts, N. Helberger, M. Veale, and M. Sax. 2025. A Right to Constructive Optimization: A Public Interest Approach to Recommender Systems in the Digital Services Act. *Journal of Consumer Policy* 48, 3 (Sept. 2025), 269–296. <https://doi.org/10.1007/s10603-025-09586-1>
- [557] Nataliya Nedzhvetskaya and JS Tan. 2024. No Simple Fix: How AI Harms Reflect Power and Jurisdiction in the Workplace. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 422–432. <https://doi.org/10.1145/3630106.3658915>
- [558] Richard R. Nelson and Sidney G. Winter. 1982. *An Evolutionary Theory of Economic Change*. Harvard University Press.
- [559] Lynnette Hui Xian Ng, Iain J. Cruickshank, and Kathleen M. Carley. 2022. Cross-platform information spread during the January 6th capitol riots. *Social Network Analysis and Mining* 12, 1 (Sept. 2022), 133. <https://doi.org/10.1007/s13278-022-00937-1>

- [560] Duc Minh Nguyet (Moon) Nguyen and Philip Luck. 2026. The New Containment Doctrine: How the United States Is Using Trade to Stop Digital Regulation. (March 2026). <https://www.csis.org/analysis/new-containment-doctrine-how-united-states-using-trade-stop-digital-regulation>
- [561] Gabriel Nicholas and Aliya Bhatia. 2023. Toward Better Automated Content Moderation in Low-Resource Languages. *Journal of Online Trust and Safety* 2, 1 (Sept. 2023), 1–11. <https://doi.org/10.54501/jots.v2i1.150>
- [562] Janni Nielsen, Torkil Clemmensen, and Carsten Yssing. 2002. Getting access to what goes on in people's heads? reflections on the think-aloud technique. In *Proceedings of the second Nordic conference on Human-computer interaction (NordiCHI '02)*. Association for Computing Machinery, New York, NY, USA, 101–110. <https://doi.org/10.1145/572020.572033>
- [563] Chris Norval, Richard Cloete, and Jatinder Singh. 2023. Navigating the Audit Landscape: A Framework for Developing Transparent and Auditable XR. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1418–1431. <https://doi.org/10.1145/3593013.3594090>
- [564] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrira Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 340–350.
- [565] Dawn Carla Nunziato. 2023. *The Digital Services Act and the Brussels Effect on Platform Content Moderation*. Legal Studies Research Paper 2023-28. George Washington University Law School.
- [566] Jonathan A. Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (2020), 128–147. <https://doi.org/10.1080/1369118X.2018.1486870>
- [567] Ursula Oberst, Elisa Wegmann, Benjamin Stodt, Matthias Brand, and Andrés Chamarro. 2017. Negative consequences from heavy social networking in adolescents: The mediating role of fear of missing out. *Journal of adolescence* 55 (2017), 51–60.
- [568] OECD. 2020. *Current approaches to terrorist and violent extremist content among the global top 50 online content-sharing services*. Technical Report. OECD, Paris. <https://doi.org/10.1787/68058b95-en>
- [569] House of Commons. 2023. Online Safety Act. <https://www.legislation.gov.uk/ukpga/2023/50> Publisher: Statute Law Database.
- [570] Washington District Office. 2025. The Foreign Censorship Threat: How the European Union's Digital Services Act Compels Global Censorship and Infringes on

- American Free Speech. <http://judiciary.house.gov/media/press-releases/foreign-censorship-threat-how-european-unions-digital-services-act-compels>
- [571] Fife Ogunde. 2024. Generative AI in American and Canadian Courts: A “Training” Approach to Regulation. *Law, Innovation and Technology* 16, 2 (2024), 715–740. <https://doi.org/10.1080/17579961.2024.2392930> Online first.
- [572] OLG Braunschweig. 2014. Urteil vom 02.04.2014. <https://openjur.de/u/691127.html>
- [573] OLG Braunschweig. 2021. Urteil vom 05.02.2021. <https://openjur.de/u/2318910.html>
- [574] OLG Braunschweig. 2022. Urteil vom 05.10.2022. <https://openjur.de/u/2452863.html>
- [575] OLG Hamm. 2017. Beschluss vom 07.09.2017. <https://openjur.de/u/2151334.html>
- [576] OLG Karlsruhe. 2017. Beschluss vom 11.05.2017. <https://openjur.de/u/2239957.html>
- [577] OLG München. 2020. Endurteil vom 08.12.2020. <https://openjur.de/u/2309762.html>
- [578] OLG Stuttgart. 2014. Beschluss vom 04.02.2014. <https://openjur.de/u/747206.html>
- [579] Will Oremus. 2024. Analysis | TikTok fails a test of its ban on political ads. *Washington Post* (Oct. 2024). <https://www.washingtonpost.com/politics/2024/10/17/tiktok-political-ads-ban-report-global-witness/>
- [580] Marie Ozanne, Aparajita Bhandari, Natalya N Bazarova, and Dominic DiFranzo. 2022. Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society* 9, 2 (July 2022), 20539517221115666. <https://doi.org/10.1177/20539517221115666> Publisher: SAGE Publications Ltd.
- [581] Michalis Pachilakis, Panagiotis Papadopoulos, Nikolaos Laoutaris, Evangelos P. Markatos, and Nicolas Kourtellis. 2022. YourAdvalue: Measuring Advertising Price Dynamics without Bankrupting User Privacy. In *Abstract Proceedings of the 2022 ACM SIGMETRICS/IFIP PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS/PERFORMANCE '22)*. Association for Computing Machinery, New York, NY, USA, 41–42. <https://doi.org/10.1145/3489048.3522629>
- [582] Konstantina Palla, José Luis Redondo García, Claudia Hauff, Francesco Fabbri, Henrik Lindström, Daniel R. Taber, Andreas Damianou, and Mounia Lalmas. 2025. Policy-as-Prompt: Rethinking Content Moderation in the Age of Large Language Models. arXiv:cs.CY/2502.18695 <https://arxiv.org/abs/2502.18695>

- [583] Christina A. Pan, Sahil Yakhmi, Tara P. Iyer, Evan Strasnick, Amy X. Zhang, and Michael S. Bernstein. 2022. Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (March 2022), 1–31. <https://doi.org/10.1145/3512929> arXiv:2202.06393 [cs].
- [584] Cecilia Panigutti, Delia Fano Yela, Lorenzo Porcaro, Astrid Bertrand, and Josep Soler Garrido. [n. d.]. How to investigate algorithmic-driven risks in online platforms and search engines? A narrative review through the lens of the EU Digital Services Act. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (2025-06-23). ACM, 828–839. <https://doi.org/10.1145/3715275.3732052>
- [585] Kostantinos Papadamou et al. 2020. Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children. In *Proceedings of the International AAAI Conference on Web and Social Media*.
- [586] Charis Papaevangelou and Fabio Votta. 2025. Trading nuance for scale? Platform observability and content governance under the DSA. *Internet Policy Review* 14, 3 (2025), 1–31.
- [587] European Parliament. 2026. EP leaders reject the visa ban imposed on former Commissioner Breton | News | European Parliament. <https://www.europarl.europa.eu/news/en/press-room/20260122IPR32567/ep-leaders-reject-the-visa-ban-imposed-on-former-commissioner-breton>
- [588] Douglas A Parry, Brittany I Davidson, Craig JR Sewall, Jacob T Fisher, Hannah Mieczkowski, and Daniel S Quintana. 2021. A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour* 5, 11 (2021), 1535–1547.
- [589] Christopher Parsons. 2019. The (In)effectiveness of Voluntarily Produced Transparency Reports. *Business & Society* 58, 1 (2019), 103–131. <https://doi.org/10.1177/0007650317717957>
- [590] Christopher Parsons. 2019. The (In)Effectiveness of Voluntarily Produced Transparency Reports. *Business & Society* 58 (2019), 103.
- [591] Carlota Perez. 2002. *Technological Revolutions and Financial Capital: The Dynamics of Bubbles and Golden Ages*. Edward Elgar.
- [592] Sharon Lisseth Perez, Xiaoying Song, and Lingzi Hong. 2025. Analyzing the language of rejection: a study of user flagging responses to hate speech on Reddit. *Information Research an international electronic journal* 30, iConf (March 2025), 815–823. <https://doi.org/10.47989/ir30iConf47191>
- [593] Venkata Phanindra Peta. 2024. User-Friendly Design and Generative AI: A Powerful Combination for Improving Financial Literacy and Investment Decision-Making. <https://papers.ssrn.com/abstract=4958392>

- [594] Juergen Pfeffer, Angelina Mooseder, Jana Lasser, Luca Hammer, Oliver Stritzel, and David Garcia. 2023. This Sample Seems to Be Good Enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API. *arXiv:cs/2204.02290*
- [595] C. A. Piña-García, Carlos Gershenson, and J. Mario Siqueiros-García. 2016. Towards a Standard Sampling Methodology on Online Social Networks: Collecting Global Trends on Twitter. *Applied Network Science* 1, 1 (Dec. 2016), 1–19. <https://doi.org/10.1007/s41109-016-0004-1>
- [596] Pinterest. 2024. *Digital Services Act Risk Assessment and Mitigation Report 2024*. Systemic Risk Assessment Report. Pinterest. <https://help.pinterest.com/sites/pinhelp/files/dsa/2024-Pinterest-DSA-Risk-Assessment-and-Mitigation-Report.pdf>
- [597] Jean-Christophe Plantin, Carl Lagoze, Paul N Edwards, and Christian Sandvig. 2018. Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society* 20, 1 (2018), 293–310. <https://doi.org/10.1177/1461444816661553>
- [598] Aneta Podkalicka and Martin Fredriksson. 2023. Mediatized marketplaces: Platforms, places, and strategies for trading material goods in digital economies. *Convergence* 29, 5 (2023), 1352–1368. <https://doi.org/10.1177/13548565231192103> arXiv:<https://doi.org/10.1177/13548565231192103>
- [599] Kristina Popova, Rachael Garrett, Claudia Núñez-Pacheco, Airi Lampinen, and Kristina Höök. 2022. Vulnerability as an ethical stance in soma design processes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3491102.3501994>
- [600] Aurelien Portuese, Orla Gough, and Joseph Tanega. 2017. The Principle of Legal Certainty as a Principle of Economic Efficiency. *European Journal of Law and Economics* 44, 1 (2017), 131–156.
- [601] Associated Press. 2020. 'Anti-corona' extremists try to storm German parliament. *The Observer* (Aug. 2020). <https://www.theguardian.com/world/2020/aug/29/berlin-braces-for-anti-coronavirus-protest-against-covid-19-restrictions>
- [602] Associated Press. 2024. Company that sent fake Biden robocalls in New Hampshire agrees to \$1m fine. *The Guardian* (Aug. 2024). <https://www.theguardian.com/technology/article/2024/aug/22/fake-biden-robocalls-fine-lingo-telecom>
- [603] Lorcán Price. 2025. Written Statement of Lorcán Price, Legal Counsel, Alliance Defending Freedom International. Submission to the U.S. Congress.
- [604] João Pedro Quintais, Naomi Appelman, and Ronan Fahy. 2022. Using Terms and Conditions to Apply Fundamental Rights to Content Moderation. <https://papers.ssrn.com/abstract=4286147>

- [605] Patrick Rafail. 2021. Nonprobability Sampling and Twitter. *Social Science Computer Review* 39, 6 (2021), 1236–1252. <https://doi.org/10.1177/0894439319840455>
- [606] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AI/ES '19)*. Association for Computing Machinery, New York, NY, USA, 429–435. <https://doi.org/10.1145/3306618.3314244>
- [607] Inioluwa Deborah Raji and Joy Buolamwini. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020), 145–151.
- [608] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 33–44. <https://doi.org/10.1145/3351095.3372873>
- [609] Pooja SB Rao, Sanja Šćepanović, Ke Zhou, Edyta Paulina Bogucka, and Daniele Quercia. 2025. RiskRAG: A Data-Driven Solution for Improved AI Model Risk Reporting. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–26.
- [610] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting Human-AI Collaboration in Auditing LLMs with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Montréal QC Canada, 913–926. <https://doi.org/10.1145/3600211.3604712>
- [611] Miriam Rateike, Isabel Valera, and Patrick Forré. 2024. Designing Long-term Group Fair Policies in Dynamical Systems. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 20–50. <https://doi.org/10.1145/3630106.3658538>
- [612] Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2541–2573. <https://doi.org/10.18653/v1/2023.emnlp-main.155>
- [613] Maria-Lucia Rebrean and Gianclaudio Malgieri. 2025. Vulnerability in the EU AI Act: building an interpretation. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 1985–1997.

- [614] REPHRAIN: Protecting citizens online. 2025. Privacy Clinics. <https://www.rephrain.ac.uk/privacy-clinics/>
- [615] Republik Österreich. 1974. Bundesgesetz vom 23. Jänner 1974 über die mit gerichtlicher Strafe bedrohten Handlungen (Strafgesetzbuch – StGB). RIS – Rechtsinformationssystem des Bundes. <https://www.ris.bka.gv.at/Gelten.deFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10002296> BGBl. Nr. 60/1974, idgF. Konsolidierte Fassung auf RIS (Gesetzesnummer 10002296).
- [616] Republik Österreich. 2021. Bundesgesetz über Maßnahmen zum Schutz der Nutzer auf Kommunikationsplattformen (Kommunikationsplattformen-Gesetz – KoPl-G). Bundesgesetzblatt Teil I Nr. 151/2020. <https://www.ris.bka.gv.at/Geltende.Fassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20011458> In force since 1 January 2021.
- [617] Reuters. 2019. Germany fines Facebook for under-reporting complaints. <https://www.reuters.com/article/us-facebook-germany-fine-idUSKCN1TX1IC>
- [618] Christopher Rhodes. 2021. Fake news and disinformation. *House of Commons Library* CBP 8749 (2021).
- [619] Horst W. J. Rittel and Melvin M. Webber. 1973. Dilemmas in a general theory of planning. *Policy Sciences* 4, 2 (June 1973), 155–169. <https://doi.org/10.1007/BF01405730>
- [620] Jan Ole Rixen, Luca-Maxim Meinhardt, Michael Glöckler, Marius-Lukas Ziegenbein, Anna Schlothauer, Mark Colley, Enrico Rukzio, and Jan Gugenheimer. 2023. The loop and reasons to break it: Investigating infinite scrolling behaviour in social media applications and reasons to stop. *Proceedings of the ACM on Human-Computer Interaction* 7, MHCI (2023), 1–22.
- [621] Sarah T. Roberts. 2017. Content Moderation. (2017), 1–4. https://doi.org/10.1007/978-3-319-32001-4_44-1
- [622] Sarah T. Roberts. 2018. Digital Detritus: 'Error' and the Logic of Opacity in Social Media Content Moderation. *First Monday* (2018).
- [623] Ana Rodríguez de las Heras Ballel. 2021. Contract Governance in the Platform Economy. *European Review of Private Law* 29, 1 (2021), 75–102. On the transformation of online platforms into contractual frameworks.
- [624] Richard Rogers. 2020. Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication* 35, 3 (2020), 213–229. <https://doi.org/10.1177/0267323120922066>
- [625] Eleonora Rosati. 2024. The DSA's Trusted Flaggers: Revolution, Evolution, or mere Gattopardismo? *Verfassungsblog* (Feb. 2024). <https://verfassungsblog.de/the-dsas-trusted-flaggers/>

- [626] Arianna Rossi, Rossana Ducato, Helena Haapio, and Stefania Passera. 2019. When Design Met Law: Design Patterns for Information Transparency. *Droit de la consommation* 1 (2019), 12–29. Issue No. 1, Université Catholique de Louvain, Centre de Droit de la Consommation.
- [627] Arianna Rossi and Simon Parkin. 2026. “What I’m Interested in is Something that Violates the Law”: Regulatory Practitioner Views on Automated Detection of Deceptive Design Patterns. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–18. <https://doi.org/10.1145/3772318.3790832>
- [628] Jonathan Rougier. 2019. Confidence in Risk Assessments. *Journal of the Royal Statistical Society Series A: Statistics in Society* 182, 3 (June 2019), 1081–1095. <https://doi.org/10.1111/rssa.12445>
- [629] Hannah Ruschemeier and João Pedro Quintais. 2025. Fundamental Rights in Out-of-Court Dispute Settlement under the Digital Services Act. <https://papers.ssrn.com/abstract=5900422>
- [630] Johnny Ryan. 2025. Former EU commissioner and activists barred from US in attack on European tech regulators. *The Guardian* (Dec. 2025). <https://www.theguardian.com/technology/2025/dec/24/us-state-department-visa-ban-former-eu-commissioner-europe>
- [631] Jeffrey Sablosky. 2021. “Dangerous organizations: Facebook’s content moderation decisions and ethnic visibility in Myanmar”. *Media, Culture & Society* (Jan. 2021), 0163443720987751. <https://doi.org/10.1177/0163443720987751> Publisher: SAGE Publications Ltd.
- [632] Arianna Sala, Lorenzo Porcaro, and Emilia Gómez. 2024. Social media use and adolescents’ mental health and well-being: an umbrella review. *Computers in Human Behavior Reports* 14 (2024), 100404.
- [633] Maryam Salehijam. 2018. The Value of Systematic Content Analysis in Legal Research. *Tilburg Law Review* 23, 1–2 (sep 2018), 34–42. <https://doi.org/10.5334/tilr.5>
- [634] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP ’23)*. Association for Computing Machinery, New York, NY, USA, 215–227. <https://doi.org/10.1145/3565472.3592959>
- [635] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2024. Dealing with uncertainty: Understanding the impact of prognostic versus diagnostic tasks on trust and reliance in human-AI decision making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.

- [636] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2024. Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–17. <https://doi.org/10.1145/3613904.3641905>
- [637] Joni Salminen, View Profile, Soon-Gyo Jung, View Profile, Bernard Jansen, and View Profile. 2022. Intentionally Biasing User Representation?: Investigating the Pros and Cons of Removing Toxic Quotes from Social Media Personas. In *Nordic Human-Computer Interaction Conference*. Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3546155.3546647>
- [638] P. Samuelson and N. Helberger. 2024. *The Digital Services Act as a Global Transparency Regime*. BerlinVerfassungsbooks. <https://doi.org/10.17176/20240429-081042-0>
- [639] Andy Sanchez, Rita Gsenger, Susannah Montgomery, Marie-Therese Sekwenz, Jason Pridmore, and Ben Wagner. 2025. Playing with Politics: Preliminary Results from Interactive Interventions on AI and Democracy in Five Countries with 2024 Elections. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–16. <https://doi.org/10.1145/3706599.3719861>
- [640] Cristiana Santos, Nataliia Bielova, Johanna Gunawan, Brennan Schaffner, Simone van der Hof, Arianna Rossi, and Marie-Therese Sekwenz. 2025. Feedback to the European Data Protection Board's Guidelines 3/2025 on the interplay between the DSA and the GDPR (Version 1.1). https://www.edpb.europa.eu/sites/default/files/webform/public_consultation_reply/feedback-gdpr-dsa-31-10-2025.pdf
- [641] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), 1668–1678.
- [642] Sheikh Muhammad Sarwar, Dimitrina Zlatkova, Momchil Hardalov, Yoan Dinkov, Isabelle Augenstein, and Preslav Nakov. 2022. A Neighborhood Framework for Resource-Lean Content Flagging. *Transactions of the Association for Computational Linguistics* 10 (May 2022), 484–502. https://doi.org/10.1162/tacl_a_00472
- [643] Adam Satariano and Jeanna Smialek. 2025. E.U. Expands Investigation of Musk's X Over Illicit Content. <https://www.nytimes.com/2025/01/17/business/eu-investigation-x-musk.html>
- [644] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *arXiv preprint arXiv:2204.06916* (2022).
- [645] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. 2022. A meta-analysis of the utility of explainable artificial intelligence in

- human-AI decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 617–626.
- [646] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.
- [647] Vera Schmitt, Luis-Felipe Villa-Arenas, Nils Feldhus, Joachim Meyer, Robert P. Spang, and Sebastian Möller. 2024. The Role of Explainability in Collaborative Human-AI Disinformation Detection. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 2157–2174. <https://doi.org/10.1145/3630106.3659031>
- [648] Thibault Schrepel. 2023. A Systematic Content Analysis of Innovation in European Competition Law. SSRN Working Paper.
- [649] Peter H. Schuck. 1992. Legal Complexity: Some Causes, Consequences, and Cures. *Duke Law Journal* 42, 1 (Oct. 1992), 1. <https://doi.org/10.2307/1372753>
- [650] Joseph A. Schumpeter. 1942. *Capitalism, Socialism and Democracy*. Harper & Brothers.
- [651] Sebastian Felix Schwemer. 2019. Trusted notifiers and the privatization of online enforcement. *Computer Law & Security Review* 35, 6 (Nov. 2019), 105339. <https://doi.org/10.1016/j.clsr.2019.105339>
- [652] Andrew Sears. 1997. Heuristic Walkthroughs: Finding the Problems Without the Noise. *International Journal of Human-Computer Interaction* 9, 3 (Sept. 1997), 213–234. https://doi.org/10.1207/s15327590ijhc0903_2
- [653] Marie Sekwenz. 2025. EU Digital Regulation as Industry Shaping Policy: The DSA, Brussels Effect, and Global Competitiveness.
- [654] Marie-Therese Sekwenz. 2026. Wicked by Design: Regulating and Auditing Platforms under the Digital Services Act. In *Mensch-Maschine-Kooperation im Cyberspace / Human-Machine Cooperation in Cyberspace: Proceedings of the 29th International Legal Informatics Symposium IRIS 2026*, Erich Schweighofer, Stefan Eder, Federico Costantini, and Felix Schmautzer (Eds.). Editions Weblaw, Bern, Switzerland, 441–450. https://iris-conferences.eu/wp-content/uploads/2026/02/TB_IRIS26.pdf
- [655] Marie-Therese Sekwenz et al. 2025. Can't LLMs Do That? Supporting Third-Party Audits under the DSA: Exploring Large Language Models for Systemic Risk Evaluation of the Digital Services Act in an Interdisciplinary Setting. In *Adjunct Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*. ACM.
- [656] Marie-Therese Sekwenz, Kyle Beadle, and Simon Parkin. 2026. “There is literally zero funding”: Understanding the Emerging Role of Trusted Flaggers under the EU

- Digital Services Act. *arXiv preprint arXiv:2603.29874* (2026). <https://doi.org/10.48550/arXiv.2603.29874>
- [657] Marie-Therese Sekwenz, Shreyan Biswas, Rita Hermann-Gsenger, and Ujwal Gadiraju. 2026. AI at the Front Lines of Platform Governance: Using LLMs to Support Illegal Content Reporting under the Digital Services Act. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26)*. Association for Computing Machinery, New York, NY, USA, 2848–2884. <https://doi.org/10.1145/3805689.3812301>
- [658] Marie-Therese Sekwenz and Rita Gsenger. 2025. The Digital Services Act: Online Risks, Transparency and Data Access. Nomos Verlagsgesellschaft mbH & Co. KG, 115–140. <https://doi.org/10.5771/9783748943990-115>
- [659] Marie-Therese Sekwenz and Rita Gsenger. 2025. The Digital Services Act: Online Risks, Transparency and Data Access. In *Digital Decade: How the EU Shapes Digitalisation Research*. Nomos.
- [660] Marie Therese Sekwenz and Rita Gsenger. 2025. Mapping Compliance: A Taxonomy for Political Content Analysis under the EU’s Digital Electoral Framework. In *IRIS 2025*. Vienna.
- [661] Marie-Therese Sekwenz, Rita Gsenger, Scott Dahlgren, and Ben Wagner. 2025. Doing Audits Right? The Role of Sampling and Legal Content Analysis in Systemic Risk Assessments and Independent Audits in the Digital Services Act. <https://doi.org/10.48550/arXiv.2505.03601> arXiv:2505.03601 [cs].
- [662] Marie-Therese Sekwenz, Rita Gsenger, Volker Stocker, Esther Görnemann, Dinara Talypova, Simon Parkin, Lea Greminger, and Georgios Smaragdakis. 2025. Can’t LLMs do that? Supporting Third-Party Audits under the DSA: Exploring Large Language Models for Systemic Risk Evaluation of the Digital Services Act in an Interdisciplinary Setting. In *Adjunct Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work (CHIWORK '25 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3707640.3731929>
- [663] Marie Therese Sekwenz, Rita Gsenger, and Ben Wagner. 2025. Forthcoming: Doing Audits Right? The Role of Sampling and Legal Content Analysis in Systemic Risk Assessments and Independent Audits in the Digital Services Act. (2025).
- [664] Marie-Therese Sekwenz, Daria Simons, and Alina Wundsam. 2025. Prompt template for a fictitious LLM agent in a content-flagging experiment. <https://doi.org/10.48550/arXiv.2507.21842> arXiv:2507.21842 [cs].
- [665] Marie-Therese Sekwenz, Ben Wagner, and Hans De Bruijn. 2025. From Reports to Reality: Testing Consistency in Instagram’s Digital Services Act Compliance Data. <https://doi.org/10.48550/arXiv.2507.01787> arXiv:2507.01787 [cs].

- [666] Marie-Therese Sekwenz, Ben Wagner, and Hans de Bruijn. 2026. From Reports to Reality: Testing Consistency in Instagram's Digital Services Act Compliance Data. *ACM Journal on Responsible Computing* (2026). <https://doi.org/10.1145/3821526>
- [667] Marie-Therese Sekwenz, Ben Wagner, and Simon Parkin. 2025. "It Is Unfair, and It Would Be Unwise to Expect the User to Know the Law!" – Evaluating Reporting Mechanisms under the Digital Services Act. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. ACM.
- [668] Marie-Therese Sekwenz, Ben Wagner, and Simon Parkin. 2025. "It is unfair, and it would be unwise to expect the user to know the law!" – Evaluating reporting mechanisms under the Digital Services Act. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 532–546. <https://doi.org/10.1145/3715275.3732036>
- [669] Farhana Shahid. 2024. Human-AI Collaboration to Facilitate Culturally-Aware Content Moderation. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW Companion '24)*. Association for Computing Machinery, New York, NY, USA, 9–12. <https://doi.org/10.1145/3678884.3682057>
- [670] Adam Sheingate, James Scharf, and Conner Delahanty. 2022. Digital Advertising in U.S. Federal Elections, 2004-2020. *Journal of Quantitative Description: Digital Media* (2022). <https://doi.org/10.51685/jqd.2022.026>
- [671] Yunhee Shim and Shagun Jhaver. 2024. Incorporating Procedural Fairness in Flag Submissions on Social Media Platforms. <https://doi.org/10.48550/arXiv.2409.08498> arXiv:2409.08498 [cs].
- [672] Donghoon Shin, Lucy Lu Wang, and Gary Hsieh. 2024. From Paper to Card: Transforming Design Implications with Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–15. <https://doi.org/10.1145/3613904.3642266>
- [673] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. 2010. How Useful Are Your Comments? Analyzing and Predicting Youtube Comments and Comment Ratings. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. Association for Computing Machinery, New York, NY, USA, 891–900. <https://doi.org/10.1145/1772690.1772781>
- [674] Herbert A Simon. 1956. Rational choice and the structure of the environment. *Psychological review* 63, 2 (1956), 129.
- [675] Munindar P. Singh. 2014. Norms as a basis for governing sociotechnical systems. *ACM Trans. Intell. Syst. Technol.* 5, 1 (Jan. 2014), 21:1–21:23. <https://doi.org/10.1145/2542182.2542203>

- [676] Nathalie A. Smuha (Ed.). 2025. *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*. Cambridge University Press.
- [677] Snapchat. 2024. *Snap DSA Report Risk Assessment Results and Mitigations*. Technical Report. Snapchat. https://downloads.ctfassets.net/kw9k15zzxtrs/55C3sV2gzvejT1ucl2xEvz/7c12ac42b6c8c19144c21b16524dbf0e/DSA_Risk_and_Mitigation_Assessment_Report_-_Snapchat_-_August_2023.pdf
- [678] Peter Snyder, Periwinkle Doerfler, Chris Kanich, and Damon McCoy. 2017. Fifteen minutes of unwanted fame: Detecting and characterizing doxing. In *Proceedings of the 2017 internet measurement conference*. 432–444.
- [679] Ian Sommerville. 2011. Software engineering (ed.). *America: Pearson Education Inc* (2011).
- [680] Dominic Spohr. 2017. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review* 34, 3 (2017), 150–160.
- [681] Susan Leigh Star. 2010. This is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology, & Human Values* 35, 5 (2010), 601–617. <https://doi.org/10.1177/0162243910377624> arXiv:<https://doi.org/10.1177/0162243910377624>
- [682] Miriah Steiger, Timir J. Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3411764.3445092>
- [683] H. V. Subramanian, C. Canfield, D. B. Shank, and M. Kinnison. 2023. Combining Uncertainty Information with AI Recommendations Supports Calibration with Domain Knowledge. *Journal of Risk Research* 26, 10 (2023), 1137–1152. <https://doi.org/10.1080/13669877.2023.2259406>
- [684] Daniel Susser. 2022. Decision Time: Normative Dimensions of Algorithmic Speed. In *2022 ACM Conference on Fairness Accountability and Transparency*. ACM, Seoul Republic of Korea, 1410–1420. <https://doi.org/10.1145/3531146.3533198>
- [685] John Sutton. 2001. *Law/Society: Origins, Interactions, and Change*. Vol. 474. Pine Forge Press.
- [686] Nicolas Suzor. 2019. *Lawless: The Secret Rules that Govern our Digital Lives*. Cambridge University Press (2019).
- [687] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansoor Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large Language Models for Data Annotation and Synthesis: A Survey. <https://doi.org/10.48550/arXiv.2402.13446> arXiv:2402.13446 [cs].

- [688] Mui Joo Tang and Eang Teng Chan. 2021. The Impact of Online Social Networking (Social Media) on Interpersonal Communication and Relationships. In *Intelligent Computing*, Kohei Arai (Ed.). Springer International Publishing, Cham, 624–640. https://doi.org/10.1007/978-3-030-80129-8_44
- [689] Adam Tapal, Ela Oren, Reuven Dar, and Baruch Eitam. 2017. The Sense of Agency Scale: A Measure of Consciously Perceived Control over One's Mind, Body, and the Immediate Environment. *Frontiers in Psychology* 8 (2017), 1552. <https://doi.org/10.3389/fpsyg.2017.01552>
- [690] Nathan TeBlunthuis, Benjamin Mako Hill, and Aaron Halfaker. 2021. Effects of Algorithmic Flagging on Fairness: Quasi-experimental Evidence from Wikipedia. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 56:1–56:27. <https://doi.org/10.1145/3449130>
- [691] David J. Teece, Gary Pisano, and Amy Shuen. 1997. Dynamic Capabilities and Strategic Management. *Strategic Management Journal* 18, 7 (1997), 509–533.
- [692] Petros Terzis, Michael Veale, and Noëlle Gaumann. 2024. Law and the Emerging Political Economy of Algorithmic Audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1255–1267. <https://doi.org/10.1145/3630106.3658970>
- [693] Petros Terzis, Michael Veale, and Noëlle Gaumann. 2024. Law and the Emerging Political Economy of Algorithmic Audits. <https://doi.org/10.48550/arXiv.2406.11855> arXiv:2406.11855 [cs].
- [694] Petros Terzis, Michael Veale, and Noëlle Gaumann. 2024. Law and the Emerging Political Economy of Algorithmic Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Rio de Janeiro, Brazil, 1255–1267.
- [695] Hibby Thach, Samuel Mayworm, Michaelanne Thomas, and Oliver L. Haimson. 2024. Trans-centered moderation: Trans technology creators and centering transness in platform and community governance: Trans-centered moderation. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 326–336. <https://doi.org/10.1145/3630106.3658909>
- [696] Stephen L Thaler. 1995. “Virtual input” phenomena within the death of a simple pattern associator. *Neural Networks* 8, 1 (1995), 55–65.
- [697] Jacob Thebault-Spieker, Sukrit Venkatagiri, Naomi Mine, and Kurt Luther. 2023. Diverse Perspectives Can Mitigate Political Bias in Crowdsourced Content Moderation. In *2023 ACM Conference on Fairness Accountability and Transparency*. 1280–1291. <https://doi.org/10.1145/3593013.3594080> arXiv:2305.14500 [cs].
- [698] Stefan Theil. 2025. Carefully Tailored: Doctrinal Methods and Empirical Contributions. *Oxford Journal of Legal Studies* 45, 4 (Dec. 2025), 1047–1075. <https://doi.org/10.1093/ojls/gqaf029>

- [699] Anna Thomaïdou and Konstantinos Limnïotis. 2025. Navigating Through Human Rights in AI: Exploring the Interplay Between GDPR and Fundamental Rights Impact Assessment. *Journal of Cybersecurity and Privacy* 5, 1 (2025), 7.
- [700] Kurt Thomas et al. 2024. Supporting Human Raters with the Detection of Harmful Content Using Large Language Models. arXiv preprint arXiv:2406.12800.
- [701] Stuart A. Thompson and Kate Conger. 2025. Meet the Next Fact-Checker, Debunker and Moderator: You. <https://www.nytimes.com/2025/01/07/technology/meta-facebook-content-moderation.html>
- [702] Anna-Sophia Tiedeke et al. 2020. *What Can Be Said Online in Germany and Austria? A Legal and Terms of Service Taxonomy*. Technical Report. Social Science Research Network. <https://papers.ssrn.com/abstract=3735932> SSRN Scholarly Paper ID 3735932.
- [703] Jan Philipp Tiedeke, Matthias C. Kettemann, Wolfgang Benedek, Markus Frischhut, and Frank Sauer. 2020. Taxonomy of Content Moderation. *International Review of Law, Computers & Technology* 34, 2 (2020), 203–225.
- [704] TikTok. 2023. *DSA Transparency report*. Technical Report 1. TikTok. [https://sf16-va.tiktokcdn.com/obj/eden-va2/zayvwly_fjulyhwzuyh\[/ljhWZthlaukjl kulzlp/DSA_H2_2024/TikTok-DSA-Transparency-Report-Jan-to-Jun-2024.pdf](https://sf16-va.tiktokcdn.com/obj/eden-va2/zayvwly_fjulyhwzuyh[/ljhWZthlaukjl kulzlp/DSA_H2_2024/TikTok-DSA-Transparency-Report-Jan-to-Jun-2024.pdf)
- [705] TikTok. 2024. *TikTok DSA Transparency Report Jan to Jun 2024*. Technical Report. TikTok. [https://sf16-va.tiktokcdn.com/obj/eden-va2/zayvwly_fjulyhwzuyh\[/ljhWZthlaukjl kulzlp/DSA_H2_2024/TikTok-DSA-Transparency-Report-Jan-to-Jun-2024.pdf](https://sf16-va.tiktokcdn.com/obj/eden-va2/zayvwly_fjulyhwzuyh[/ljhWZthlaukjl kulzlp/DSA_H2_2024/TikTok-DSA-Transparency-Report-Jan-to-Jun-2024.pdf)
- [706] Paul Timmers. 2022. Digital Industrial Policy for Europe. CERRE Report.
- [707] Helma Torkamaan, Steffen Steinert, Maria Soledad Pera, Olya Kudina, Samuel Kernan Freire, Himanshu Verma, Sage Kelly, Marie-Therese Sekwenz, Jie Yang, Karolien van Nunen, Martijn Warnier, Frances Brazier, and Oscar Oviedo-Trespalacios. [n. d.]. Challenges and future directions for integration of large language models into socio-technical systems. *Behaviour & Information Technology* 0, 0 ([n. d.]), 1–20. <https://doi.org/10.1080/0144929X.2024.2431068>
- [708] Ioanna Tourkochoriti. 2024. The Digital Services Act and the EU as the Global Regulator of the Internet. *Chicago Journal of International Law* (2024).
- [709] Chau Tran, Kejsi Take, Kaylea Champion, Benjamin Mako Hill, and Rachel Greenstadt. 2024. Challenges in Restructuring Community-based Moderation. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2 (Nov. 2024), 415:1–415:24. <https://doi.org/10.1145/3686954>
- [710] Amaury Trujillo, Tiziano Fagni, and Stefano Cresci. 2024. The DSA Transparency Database: Auditing Self-reported Moderation Actions by Social Media. <https://doi.org/10.48550/arXiv.2312.10269> arXiv:2312.10269 [cs].

- [711] Amaury Trujillo, Tiziano Fagni, and Stefano Cresci. 2025. The DSA Transparency Database: Auditing Self-reported Moderation Actions by Social Media. *Proc. ACM Hum.-Comput. Interact.* 9, 2 (May 2025), CSCW187:1–CSCW187:28. <https://doi.org/10.1145/3711085>
- [712] Roberto Truzoli, Lorena Magistrati, Caterina Viganò, Stella Conte, Lisa A Osborne, and Phil Reed. 2023. Social media users potentially experience different withdrawal symptoms to non-social media users. *International Journal of Mental Health and Addiction* 21, 1 (2023), 411–417.
- [713] Aina Turillazzi, Federico Casolari, Mariarosaria Taddeo, and Luciano Floridi. 2022. The Digital Services Act: An Analysis of Its Ethical, Legal, and Social Implications. <https://doi.org/10.2139/ssrn.4007389>
- [714] Jean M Twenge, Thomas E Joiner, Megan L Rogers, and Gabrielle N Martin. 2018. Increases in depressive symptoms, suicide-related outcomes, and suicide rates among US adolescents after 2010 and links to increased new media screen time. *Clinical psychological science* 6, 1 (2018), 3–17.
- [715] Twitter. 2023. The Twitter Rules: Safety, Privacy, Authenticity, and More. <https://help.twitter.com/en/rules-and-policies/twitter-rules>
- [716] Twitter, Inc. 2021. Permanent Suspension of @realDonaldTrump. https://blog.twitter.com/en_us/topics/company/2020/suspension
- [717] Petter Törnberg. 2023. How to use LLMs for Text Analysis. <https://doi.org/10.48550/arXiv.2307.13106> arXiv:2307.13106 [cs].
- [718] Petter Törnberg. 2024. Best Practices for Text Annotation with Large Language Models. <https://doi.org/10.48550/arXiv.2402.05129> arXiv:2402.05129 [cs].
- [719] E Union. 2021. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *COM/2021/206final* (2021), 1–107.
- [720] United States Congress. 1996. Communications Decency Act. <https://www.justice.gov/archives/ag/departments-justice-s-review-section-230-communications-decency-act-1996>
- [721] Aleksandra Urman and Mykola Makhortykh. 2023. How transparent are transparency reports? Comparative analysis of transparency reporting across on-line platforms. *Telecommunications Policy* 47, 3 (April 2023), 102477. <https://doi.org/10.1016/j.telpol.2022.102477>
- [722] Frank Peterson v Google LLC, Others, and Elsevier Inc v Cyando AG. [n. d.]. ECJ Joined Cases C-682/18 and C-683/18, 2021.

- [723] Jacob van de Kerkhof. 2024. Constitutional Aspects of Trusted Flaggers in The Netherlands. <https://doi.org/10.2139/ssrn.4943851>
- [724] Jacob Van De Kerkhof. 2025. Article 22 Digital Services Act: Building trust with trusted flaggers. *Internet Policy Review* 14, 1 (March 2025). <https://doi.org/10.14763/2025.1.1828>
- [725] Max van Drunen, Eva Groen-Reijman, Tom Dobber, Arman Noroozian, Paddy Leerssen, Natali Helberger, C. H. de Vreese, Fabio Votta, and N. Helberger. 2022. Transparency and (no) more in the Political Advertising Regulation. *Internet Policy Review* (2022), 1–3. <https://policyreview.info/articles/news/transparency-and-no-more-political-advertising-regulation/1616>
- [726] Max Zeno van Drunen, Natalie Helberger, and Ronan Ó Fathaigh. 2022. The beginning of EU political advertising law: unifying democratic visions through the internal market. *International Journal of Law and Information Technology* (Sept. 2022), eaac017. <https://doi.org/10.1093/ijlit/eaac017>
- [727] M. Z. van Drunen and A. Noroozian. 2024. How to design data access for researchers: A legal and software development perspective. *Computer Law & Security Review* 52 (April 2024), 105946. <https://doi.org/10.1016/j.clsr.2024.105946>
- [728] Mr. Sony Varghese and Ms. Mansi Agrawal. 2021. Impact of Social Media on Consumer Buying Behavior. *Saudi Journal of Business and Management Studies* 6, 3 (March 2021), 51–55. <https://doi.org/10.36348/sjbms.2021.v06i03.001>
- [729] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.
- [730] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3173574.3174014>
- [731] Andreas Veglis. 2014. Moderation Techniques for Social Media Content. In *Social Computing and Social Media*, Gabriele Meiselwitz (Ed.). Springer International Publishing.
- [732] Neil Vigdor. 2024. Trump Promotes A.I. Images to Falsely Suggest Taylor Swift Endorsed Him. *The New York Times* (Aug. 2024). <https://www.nytimes.com/2024/08/19/us/politics/trump-taylor-swift-ai-images.html>
- [733] João Vinagre, Lorenzo Porcaro, Silvia Merisio, Erasmo Purificato, and Emilia Gomez. 2025. Data Access for Recommender Systems Research: leveraging the EU's Digital Services Act. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25)*. Association for Computing Machinery, New York, NY, USA, 1425–1426. <https://doi.org/10.1145/3705328.3748004>

- [734] Pierpaolo Vivo, Daniel M. Katz, and J. B. Ruhl. 2024. A complexity science approach to law and governance. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 382, 2270 (April 2024), 20230166. <https://doi.org/10.1098/rsta.2023.0166>
- [735] Fabio Votta, Simon Kruschinski, Mads Hove, Natali Helberger, Tom Dobber, and Claes Vreese. 2024. Who Does(n't) Target You? Mapping the Worldwide Usage of Online Political Microtargeting. *Journal of Quantitative Description Digital Media* 4 (May 2024), 1–73. <https://doi.org/10.51685/jqd.2024.010>
- [736] Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2024. Disinformation Capabilities of Large Language Models. <https://doi.org/10.48550/arXiv.2311.08838> arXiv:2311.08838.
- [737] Ben Wagner. 2011. Freedom of Expression on the Internet: Implications for Foreign Policy. *Global Information Society Watch* (2011), 18.
- [738] Ben Wagner. 2016. Global Expression Governance: The Norms of a Global Regime. In *Global Free Expression - Governing the Boundaries of Internet Content*, Ben Wagner (Ed.). Springer International Publishing, Cham, 137–156. https://doi.org/10.1007/978-3-319-33513-1_7
- [739] Ben Wagner et al. 2020. Regulating Transparency? Facebook, Twitter and the German Network Enforcement Act. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM.
- [740] Ben Wagner et al. 2024. Mapping Interpretations of the Law in Online Content Moderation in Germany. *Computer Law & Security Review* 55 (2024), 106054.
- [741] Ben Wagner, Matthias C. Kettemann, Anna Sophia Tiedeke, Felicitas Rachinger, and Marie-Therese Sekwenz. 2023. Mapping Interpretations of the Law in Online Content Moderation in Germany. <https://doi.org/10.2139/ssrn.4673858>
- [742] Ben Wagner, Matthias C. Kettemann, Anna Sophia Tiedeke, Felicitas Rachinger, and Marie-Therese Sekwenz. 2024. Mapping interpretations of the law in online content moderation in Germany. *Computer Law & Security Review* 55 (2024), 106054. <https://doi.org/10.1016/j.clsr.2024.106054>
- [743] Ben Wagner and Anna Kubina. 2021. Platform design for behavior change: Lessons from community-led social media. *Policy & Internet* (2021).
- [744] Ben Wagner and Lubos Kuklis. 2021. Disinformation, Data Verification and Social Media: Verifying Data through Auditing Intermediaries. (2021).
- [745] Ben Wagner, Johanne Kübler, Rita Gsenger, and Marie Therese Sekwenz. 2025. Forthcoming: Creating Accountability in Electoral Contexts through the Facebook Ad Library. (2025).

- [746] Ben Wagner, Johanne Kübler, Eliška Pírková, Rita Gsenger, and Carolina Ferro. 2021. Reimagining content moderation and safeguarding fundamental rights: A study on community-led platforms. (2021). https://enabling-digital.eu/wp-content/uploads/2021/07/Alternative-content_web.pdf
- [747] Ben Wagner, Krisztina Rozgonyi, Marie-Therese Sekwenz, Jennifer Cobbe, and Jatinder Singh. 2020. Regulating transparency? Facebook, Twitter and the German Network Enforcement Act. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 261–271. <https://doi.org/10.1145/3351095.3372856>
- [748] Ben Wagner, Marie-Therese Sekwenz, Johanne Kübler, and Carolina Ferro. [n. d.]. Study on transparency reporting of online intermediary services. ([n. d.]). https://enabling-digital.eu/wp-content/uploads/2024/03/Final_Draft_Transparency_Report-5.pdf
- [749] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–21. <https://doi.org/10.1145/3613904.3641960>
- [750] Lynne Webb and Y. Wang. 2013. Techniques for Sampling Online Text-Based Data Sets. In *Handbook of Research on Technoself: Identity in a Technological Society*, Rocci Luppigini (Ed.). IGI Global, Hershey, PA, USA, 95–114. <https://doi.org/10.4018/978-1-4666-4699-5.ch005>
- [751] Johnny Tian-Zheng Wei, Frederike Zufall, and Robin Jia. 2025. Operationalizing Content Moderation "Accuracy" in the Digital Services Act. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*. AAAI Press, 1527–1538.
- [752] Miranda Wei, Sunny Consolvo, Patrick Gage Kelley, Tadayoshi Kohno, Tara Matthews, Sarah Meiklejohn, Franziska Roesner, Renee Shelby, Kurt Thomas, and Rebecca Umbach. 2024. Understanding Help-Seeking and Help-Giving on Social Media for Image-Based Sexual Abuse. *arXiv preprint arXiv:2406.12161* 0 (2024), 18. [arXiv:2406.12161](https://arxiv.org/abs/2406.12161) <https://arxiv.org/abs/2406.12161> Preprint, accessed July 5, 2024.
- [753] Miranda Wei, Sunny Consolvo, Patrick Gage Kelley, Tadayoshi Kohno, Franziska Roesner, and Kurt Thomas. 2023. "There's so much responsibility on users right now." Expert Advice for Staying Safer From Hate and Harassment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Hamburg, Germany, 1–17.
- [754] Miranda Wei, Sunny Consolvo, Patrick Gage Kelley, Tadayoshi Kohno, Franziska Roesner, and Kurt Thomas. 2023. "There's so much responsibility on users right now." Expert Advice for Staying Safer From Hate and Harassment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. <https://doi.org/10.1145/3544548.3581229>

- [755] Miranda Wei, Jaron Mink, Yael Eiger, Tadayoshi Kohno, Elissa M Redmiles, and Franziska Roesner. 2024. {SoK}(or {SoLK?}): On the Quantitative Study of Sociodemographic Factors and Computer Security Behaviors. In *33rd USENIX Security Symposium (USENIX Security 24)*. 7011–7030.
- [756] Hannes Werthner, Carlo Ghezzi, Jeff Kramer, Julian Nida-Rümelin, Bashar Nuseibeh, Erich Prem, and Allison Stanger (Eds.). 2024. *Introduction to Digital Humanism: A Textbook*. Springer Nature Switzerland, Cham. <https://doi.org/10.1007/978-3-031-45304-5>
- [757] Bryce Westlake, Martin Bouchard, and Richard Frank. 2012. Comparing Methods for Detecting Child Exploitation Content Online. In *Proceedings of the 2012 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, Odense, Denmark, 156–163. <https://doi.org/10.1109/EISIC.2012.25>
- [758] Barbara Wimmer. 2021. Twitter ignoriert Hass-im-Netz-Gesetz. <https://futurezone.at/netzpolitik/hass-im-netz-gesetz-twitter-meldepflicht-klage/401814856>
- [759] Chloe Wittenberg, Ziv Epstein, Adam J. Berinsky, and David G. Rand. 2024. Labeling AI-Generated Content: Promises, Perils, and Future Directions. *An MIT Exploration of Generative AI* (March 2024). <https://doi.org/10.21428/e4baedd9.0319e3a6> Publisher: MIT.
- [760] Pieter Wolters and Frederik Zuiderveen Borgesius. 2025. The EU Digital Services Act: what does it mean for online advertising and adtech? *International Journal of Law and Information Technology* 33 (Jan. 2025), eaaf004. <https://doi.org/10.1093/ijlit/eaaf004>
- [761] Liz Woolery, Ryan Budish, and Kevin Bankston. [n. d.]. The Transparency Reporting Toolkit – Guide and Template. Open Technology Institute and Harvard University Berkman Klein Center for Internet & Society 2016.
- [762] Lucas Wright. 2022. Automated Platform Governance Through Visibility and Scale: On the Transformational Power of AutoModerator. *Social Media + Society* 8, 1 (Jan. 2022), 20563051221077020. <https://doi.org/10.1177/20563051221077020> Publisher: SAGE Publications Ltd.
- [763] Bao-Jun Xia. 2025. Navigating user engagement and cultural transitions in entertainment technology and social media based on activity management. *Entertainment Computing* 52 (Jan. 2025), 100791. <https://doi.org/10.1016/j.entcom.2024.100791>
- [764] Majid Yar. 2018. A Failure to Regulate? The Demands and Dilemmas of Tackling Illegal Content and Behaviour on Social Media. *The International Journal of Cybersecurity Intelligence and Cybercrime* 1, 1 (Aug. 2018), 5–20. <https://doi.org/10.52306/01010318RVZE9940>

- [765] Karen Yeung, Martin Lodge, Karen Yeung, and Martin Lodge (Eds.). 2019. *Algorithmic Regulation*. Oxford University Press, Oxford, New York.
- [766] Jillian C. York. 2021. *Silicon Values: The Future of Free Speech Under Surveillance Capitalism*. Verso Books.
- [767] Ernst & Young and Meta Platforms Ireland Limited. 2024. *Instagram Independent Audit Report: Digital Services Act Systemic Risk Assessment 2024*. Technical Report. Meta Platforms Ireland Limited. https://scontent-ams4-1.xx.fbcdn.net/v/t39.8562-6/468562318_2548652918858554_7920553818496711968_n.pdf?_nc_cat=107&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=P0Iw04xC3o0Q7kNvwEUsrah&_nc_oc=Admw0Dy67uTnFP5g44vT18U3NcwwGB2xJJUsNbCDdKuw-CmGjnmkV4qD8cSoiheLWzc2ImKMZ23Mlg4tKL0AtVnd&_nc_zt=14&_nc_ht=scontent-ams4-1.xx&_nc_gid=9mVgiTn8cAlAU2R6jcL-TA&oh=00_AfLTTeXRQIA3qoUUd7VLgl5eqxNaA8MD8k5WiKtA0xfuaA&oe=6844977B Conducted by Ernst & Young.
- [768] YouTube. 2023. Legal Policies - YouTube Help. <https://support.google.com/youtube/topic/6154211?hl=en>
- [769] Mireia Yurrita, Himanshu Verma, Agathe Balayn, Kars Alfrink, Ujwal Gadiraju, and Alessandro Bozzon. 2025. Identifying Algorithmic Decision Subjects' Needs for Meaningful Contestability. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW234 (Oct. 2025), 29 pages. <https://doi.org/10.1145/3757415>
- [770] Mireia Yurrita, Himanshu Verma, Agathe Balayn, Ujwal Gadiraju, Sylvia C Pont, and Alessandro Bozzon. 2025. Towards Effective Human Intervention in Algorithmic Decision-Making: Understanding the Effect of Decision-Makers' Configuration on Decision-Subjects' Fairness Perceptions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21.
- [771] Reza Zafarani and Huan Liu. 2015. Evaluation without Ground Truth in Social Media Research. *Commun. ACM* 58, 6 (May 2015), 54–60. <https://doi.org/10.1145/2666680>
- [772] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Who Let The Trolls Out?: Towards Understanding State-Sponsored Trolls. In *Proceedings of the 10th ACM Conference on Web Science*. ACM, Boston Massachusetts USA, 353–362. <https://doi.org/10.1145/3292522.3326016>
- [773] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Who let the trolls out? towards understanding state-sponsored trolls. In *Proceedings of the 10th acm conference on web science*. 353–362.
- [774] Savvas Zannettou, Joel Finkelstein, Barry Bradlyn, and Jeremy Blackburn. 2020. A Quantitative Approach to Understanding Online Antisemitism. *Proceedings of the*

- International AAAI Conference on Web and Social Media* 14 (May 2020), 786–797. <https://doi.org/10.1609/icwsm.v14i1.7343>
- [775] Aurora Zhang and Anette Hosoi. 2024. Structural Interventions and the Dynamics of Inequality. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1014–1030. <https://doi.org/10.1145/3630106.3658952>
- [776] Alice Qian Zhang, Kaitlin Montague, and Shagun Jhaver. 2023. Cleaning Up the Streets: Understanding Motivations, Mental Models, and Concerns of Users Flagging Social Media Posts. *arXiv preprint arXiv:2309.06688* 0 (2023), 1–18. arXiv:2309.06688 <https://arxiv.org/abs/2309.06688> Preprint, accessed July 5, 2024.
- [777] Alice Qian Zhang, Kaitlin Montague, and Shagun Jhaver. 2024. Cleaning Up the Streets: Understanding Motivations, Mental Models, and Concerns of Users Flagging Social Media Content. <https://doi.org/10.48550/arXiv.2309.06688> arXiv:2309.06688 [cs].
- [778] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv:cs.CL/2309.01219 <https://arxiv.org/abs/2309.01219>
- [779] Benjamin Zi Hao Zhao, Muhammad Ikram, Hassan Jameel Asghar, Mohamed Ali Kaafar, Abdelberi Chaabane, and Kanchana Thilakarathna. 2019. A decade of mal-activity reporting: A retrospective analysis of internet malicious activity blacklists. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*. 193–205.
- [780] Chengshuai Zhao, Zhen Tan, Chau-Wai Wong, Xinyan Zhao, Tianlong Chen, and Huan Liu. 2025. SCALE: Towards Collaborative Content Analysis in Social Science with Large Language Model Agents and Human Intervention. arXiv:cs.AI/2502.10937 <https://arxiv.org/abs/2502.10937>
- [781] Xiabing Zheng and Matthew KO Lee. 2016. Excessive use of mobile social networking sites: Negative consequences on individuals. *Computers in Human Behavior* 65 (2016), 65–76.
- [782] Rebecca Zipursky. 2019. Nuts About NETZ: The Network Enforcement Act and Freedom of Expression. *Fordham International Law Journal* 42, 4 (2019), 1325. <https://ir.lawnet.fordham.edu/ilj/vol42/iss4/7>
- [783] Efroni Zohar. 2021. The Digital Services Act: risk-based regulation of online platforms. <https://policyreview.info/articles/news/digital-services-act-risk-based-regulation-online-platforms/1606>

A

APPENDIX FOR CHAPTER 5

A.1. OVERVIEW OF KEY LEGAL REFERENCES AND PROVISIONS

Table A.1: Key Legal Provisions Referenced in this Article

Provision	Jurisdiction	Summary and Relevance
NetzDG (2017)	Germany	Requires semi-annual transparency reports on content moderation; criticized for non-standardized categories and incomplete reporting. Fines for non-compliance.
KoPl-G (2020)	Austria	Mandates transparency reporting for communication platforms; modeled after NetzDG, with similar shortcomings.
DSA Article 14	EU	Requires platforms to provide clear terms and conditions; anticipates a centralised TaC Database.
DSA Article 15	EU	Sets general transparency requirements for intermediary services, including machine-readable, public reports.
DSA Article 16	EU	Requires notice-and-action mechanisms for reporting illegal content.
DSA Article 17	EU	Establishes the Statement of Reasons (SOR) mechanism: platforms must justify each moderation action, referencing legal or contractual grounds.
DSA Article 24	EU	Extends transparency requirements to online platforms.
DSA Article 34	EU	Introduces periodic systemic risk assessments, focusing on fundamental rights and public interest risks.
DSA Article 37	EU	Mandates independent audits of VLOP/VLOSE risk assessments and reporting mechanisms.
DSA Article 42	EU	Sets biannual reporting requirements for VLOPs and VLOSEs.
DSA Delegated Regulation for transparency reporting	EU	Lays down harmonized templates, taxonomies, and metadata for transparency reports, ensuring comparability and standardization across platforms (Arts. 15(3), 24(2), 42 DSA).
DSA Delegated Regulation for conducting audits	EU	Specifies requirements for conducting audits and systemic risk assessments under the DSA.
Charter of Fundamental Rights (CFR)	EU	Referenced in DSA Article 34(1)(b) as the basis for risk assessment focus on fundamental rights.

B

APPENDIX FOR CHAPTER 6

B.1. OVERVIEW OF KEY LEGAL REFERENCES AND PROVISIONS

Table B.1: Key Legal Provisions Referenced in this Article

Provision	Jurisdiction	Summary and Relevance
NetzDG (2017)	Germany	Requires semi-annual transparency reports on content moderation; criticized for non-standardized categories and incomplete reporting. Fines for non-compliance.
KoPl-G (2020)	Austria	Mandates transparency reporting for communication platforms; modeled after NetzDG, with similar shortcomings.
DSA Article 14	EU	Requires platforms to provide clear terms and conditions; anticipates a centralised TaC Database.
DSA Article 15	EU	Sets general transparency requirements for intermediary services, including machine-readable, public reports.
DSA Article 16	EU	Requires notice-and-action mechanisms for reporting illegal content.
DSA Article 17	EU	Establishes the Statement of Reasons (SOR) mechanism: platforms must justify each moderation action, referencing legal or contractual grounds.
DSA Article 24	EU	Extends transparency requirements to online platforms.
DSA Article 34	EU	Introduces periodic systemic risk assessments, focusing on fundamental rights and public interest risks.
DSA Article 37	EU	Mandates independent audits of VLOP/VLOSE risk assessments and reporting mechanisms.
DSA Article 42	EU	Sets biannual reporting requirements for VLOPs and VLOSEs.
DSA Delegated Regulation for transparency reporting	EU	Lays down harmonized templates, taxonomies, and metadata for transparency reports, ensuring comparability and standardization across platforms (Arts. 15(3), 24(2), 42 DSA).
DSA Delegated Regulation for conducting audits	EU	Specifies requirements for conducting audits and systemic risk assessments under the DSA.
Charter of Fundamental Rights (CFR)	EU	Referenced in DSA Article 34(1)(b) as the basis for risk assessment focus on fundamental rights.

C

APPENDIX FOR CHAPTER 7

C.0.1. CODE BOOK

C.0.2. DISINFORMATION CATEGORIES

Table C.1

Code	Source	Qualification	Definition	Example
D-1	Kapantai et al. 2021	Fabricated	Stories that completely lack of any factual base, 100% false. The intention is to deceive and cause harm (Wardle and Derekshan, 2017). One of the most severe types (Zannettou et al., 2018) as fabrication adopts the style of news articles so the recipients believe it is legitimate (Tandoc et al., 2017). Could be text but also in visual format (Ireton and Posetti, 2018). (Source: Kapantai et al. 2021)	Steven Heller. 2014. Bat Boy, Hillary Clinton's alien baby, and a tabloid's glorious legacy. https://www.theatlantic.com/entertainment/archive/2014/10/the-ingenious-sensationalism-of-the-weekly-world-new/381525/ .
D-2	Kapantai et al. 2021	Imposter	Genuine sources that are impersonated with false, made-up sources to support basically a false narrative. It is actually very misleading since source or author is considered great criteria of verifying credibility (House Of Commons, 2018; Zannettou et al., 2018; Wardle and Derekshan, 2017). (use of journalists name/logo /branding of mimic URLs)	Very recent and often used: so-called medical experts/doctors that warn about vaccines/ say Covid-19 is not dangerous etc. For example: Wolfgang Wodarg: https://www.rubikon.news/artikel/die-poppel-epidemiologie . I'd say if the word expert is used a lot and the titles of people are thrown in your face all the time, be careful.

Table C.1

Code	Source	Qualification	Definition	Example
D-3	Kapantai et al. 2021	Conspiracy theories	Stories without factual base as there is no established baseline for truth. They usually explain important events as secret plots by government or powerful individuals (Zannettou et al., 2018). Conspiracies are, by definition, difficult to verify as true or false, and they are typically originated by people who believe them to be true (Allcott and Gentzkow, 2017). Evidences that refute the conspiracy are regarded as further proof of the conspiracy (EAVI, 2018). Some conspiracy theories may have damaging ripple-effects.	QAnon: https://arxiv.org/abs/2105.13020 . Often uses codes such as specific signs and abbreviations. Claims of whistleblowing, assertions of truth, and instructions to “do the research” may point towards conspiracy narratives. These narratives are often connected to far-right symbolism (for instance, references to prominent Jewish people).

Table C.1

Code	Source	Qualification	Definition	Example
D-4	Kapantai et al. 2021	Hoaxes	Relatively complex and large-scale fabrications which may include deceptions that go beyond the scope of fun or scam and cause material loss or harm to the victim (Rubin et al., 2015). They contain facts that are either false or inaccurate and are presented as legitimate facts. This category is also known in the research community either as half-truth or factoid stories (Zannettou et al., 2018) able to convince readers of the validity of a paranoia-fueled story (Rashkin et al., 2017).	Twitter Death Hoaxes,

Table C.1

Code	Source	Qualification	Definition	Example
D-5	Kapantai et al. 2021	Biased or one-sided	Stories that are extremely biased toward a person, party, situation, or event, driving division and polarization. The context of this type of news information is extremely imbalanced (i.e. left or right wing), inflammatory, emotional and often riddled with untruths. They contain either a mixture of true and false or mostly false, thus misleading information designed to confirm a particular ideological view (Zannettou et al., 2018; Potthast et al., 2018).	Propaganda-style news, for example Breitbart: https://www.breitbart.com/politics/2018/07/13/delingpole-trump-just-dropped-the-mother-of-all-brexit-bombs-on-the-resa-may/ . The example uses emotionally charged and polarising language to portray political negotiations as betrayal and to discredit political opponents.

Table C.1

Code	Source	Qualification	Definition	Example
D-6	Kapantai et al. 2021	Rumors	Refers to stories whose truthfulness is ambiguous or never confirmed (gossip, innuendo, unverified claims). This kind of false information is widely propagated on online social networks (Peterson and Gist, 1951).	People on Twitter opposed to Trump were spreading a false rumor that White House press secretary Kayleigh McEnany deleted tweets from her official Twitter account. They claim she deleted tweets beginning on Oct. 31. But in fact, the reason there are no tweets is that she simply wasn't tweeting from that account during those days. The Wayback Machine archive of the @PressSec account confirms she was silent between Nov. 1 and Nov. 5. On Friday, Nov. 6, McEnany confirmed that the rumor was false.

Table C.1

Code	Source	Qualification	Definition	Example
D-7	Kapantai et al. 2021	Clickbait	Sources that provide generally credible or dubious factual content but deliberately use exaggerated, misleading, and unverified headlines and thumbnails (Rehm, 2018; Szpakowski, 2018) to lure readers open the intended Web page (Ghanem et al., 2019). The goal is to increase their traffic for profit, popularity, or sensationalization (Pujahari and Sisodia, 2019; Zannettou et al., 2018). Once the reader is there, the content rarely satisfies their interest (EAVI, 2018).	Clicking the link takes you not to a tell-all Kimye exposé, but to a Vote.gov link encouraging you to register to vote.

Table C.1

Code	Source	Qualification	Definition	Example
D-8	Kapantai et al. 2021	Misleading connection	Misleading use of information to frame an issue or individual. When headlines, visuals, or captions do not support the content. Separate parts of source information may be factual but are presented using wrong connection (context/content).	Donald Trump's first campaign ad purported to show migrants crossing the border from Mexico, when the footage was actually migrants crossing from Morocco to Melilla in North Africa. This content isn't fake, but the context is wrong. In the weeks leading up the election, a video emerged that appeared to show ballot-box stuffing. As Alastair Reid from First Draft News, of which I am a member, explained, the date stamp in the top left hand corner shows it was captured on September 18, the date of elections in Russia. A quick Google reverse image search also confirms the origin of the footage. Again, the content isn't fake, the context is wrong.

Table C.1

Code	Source	Qualification	Definition	Example
D-9	Kapantai et al. 2021	Trolling	The act of deliberately posting offensive or inflammatory content to an online community with the intent of provoking readers or disrupting conversation. Today, the term “troll” is most often used to refer to any person harassing or insulting others online. However, it has also been used to describe human-controlled accounts performing bot-like activities (Wardle et al., 2018).	Ryanair: what Trump buys in board a Rayanair flight vs what Biden buys on board a Ryanair flight
D-10	Kapantai et al. 2021	Pseudoscience	Information that misrepresents real scientific studies with dubious or false claims. Often contradicts experts (EAVI, 2018). Promotes metaphysics, naturalistic fallacies, and other (Guacho et al., 2018). The actors hijack scientific legitimacy for profit or fame (Forstrop, 2005).	The earth is flat. The attitude indicator in planes have been mechanical for years. There would be a constant drift if the earth was a sphere.

C.0.3. ELECTORAL RIGHTS CATEGORIES

PROCEDURAL DISINFORMATION FOR VOTERS

Table C.2

Code	Source	Qualification	Definition	Example
E-1	Kapantai et al. 2021	Fabricated	Stories that completely lack of any factual base, 100% false. The intention is to deceive and cause harm (Wardle and Derekshan, 2017). One of the most severe types (Zannettou et al., 2018) as fabrication adopts the style of news articles so the recipients believe it is legitimate (Tandoc et al., 2017). Could be text but also in visual format (Ireton and Posetti, 2018). (Source: Kapantai et al. 2021)	Steven Heller. 2014. Bat Boy, Hillary Clinton's alien baby, and a tabloid's glorious legacy. https://www.theatlantic.com/entertainment/archive/2014/10/the-ingenious-sensationalism-of-the-weekly-world-new/381525/ .

Table C.2

Code	Source	Qualification	Definition	Example
E-2	Krennerich 2021	Procedural - Voter registration	Actors interested in delegitimizing the elections or manipulating voter registration employ disinformation regarding the right to vote of certain groups to disturb the electoral process and prevent (certain) voters from voting.	“Translation by the authors: if I now order postal voting documents for a compatriot, check the post office, fill in the ballot paper and hand it in, how do you recognize the fraud? They must be able to explain if it's so safe and easy, right?” Targeted groups may include people with disabilities. People, who have been given a supervisor appointed by the court, for example because they are mentally or mentally impaired or in a coma, were not allowed to vote in Germany until 2019. To cast a vote, the person has to express a clear will, thus a person unconscious in a coma for the entire voting period will probably not be able to vote - even if they theoretically have the right to do so.

Table C.2

Code	Source	Qualification	Definition	Example
E-3	Krennerich 2021	Procedural - Voter identification	Actors interested in delegitimizing the elections or manipulating voter participation employ false information on required documents for voter registration to disturb the electoral process, confuse voters and prevent certain voters from voting.	Claims of apparent lax enforcement of voter identification at the polling station.
E-4	Krennerich 2021	Procedural - Election campaign	Actors interested in delegitimizing the elections or manipulating voter participation employ false information on organizational restrictions on election campaign events under pandemic conditions to disturb the electoral process, confuse voters and prevent (certain) voters from voting.	Claims that rallies are illegal or that only a certain number of people are allowed to attend. Possible disinformation regarding the violation of the principle of equal opportunities and the state's duty of neutrality in elections.
E-5	Krennerich 2021	Procedural - Election campaign and party funding	Disinformation on financing of candidates and political parties (Source: Krennerich 2021)	Claims that a candidate is financed by George Soros or Bill Gates, for example.

Table C.2

Code	Source	Qualification	Definition	Example
E-6	Krennerich 2021	Procedural - Polling station	Actors interested in delegitimizing the elections or manipulating voter participation might spread disinformation regarding opening hours, location of polling stations and polling site changes etc. to disturb the electoral process, confuse voters and to prevent (certain) voters from voting. (Source: Krennerich 2021)	Claims that polling stations close early, changed to a new place to allow social distancing etc.
E-7	Krennerich 2021	Procedural - Voting	Actors interested in delegitimizing the elections or manipulating voter participation spread disinformation pertaining to procedures of the postal vote to disturb the electoral process, confuse voters and to prevent (certain) voters from voting.	False instructions on how to apply for mail-in ballots, how to fill out the ballot, how and when to post it etc.
E-8	Krennerich 2021	Procedural - Vote count	Actors interested in delegitimizing the elections spread disinformation on procedures of the vote count to disturb the electoral process, confuse voters and to prevent (certain) voters from voting. (Source: Krennerich 2021)	Translation by the authors "What is the point of elections if the greens count the ballot papers?" However, it is not permissible for only representatives of one party or one ideology to evaluate ballot papers.

Table C.2

Code	Source	Qualification	Definition	Example
E-9	Krennerich 2021	Procedural - Election observation	False or misleading claims about the availability, permissibility, independence, or conduct of election observation.	Disinformation on the (im-)possibility of election observation.
E-10	Krennerich 2021	Procedural - Electoral system	False or misleading claims about the rules by which votes are translated into mandates, including representation and the size or composition of elected bodies.	Disinformation with regards to the translation of votes into mandates, combined with criticism of the electoral system and the increasing size of the Bundestag

DISINFORMATION ON CANDIDATES AND PARTIES

Table C.3

Code	Source	Qualification	Definition	Example
C-1	Krennerich 2021	Candidates - Right to stand for elections	Actors interested in restricting electoral competition to ensure electoral victory (of the government) might employ false claims that eligibility conditions are not given in order to exclude certain candidates.	Fabricated criminal allegations.
C-2	Krennerich 2021	Candidates - Electoral registration of candidates and parties	Actors interested in restricting electoral competition to ensure electoral victory (of the government) might make erroneous claims that registration requirements are not fulfilled to exclude certain candidates and parties.	Disinformation on changed conditions for the collection of signatures due to the pandemic

Table C.3

Code	Source	Qualification	Definition	Example
C-3	Krennerich 2021	Candidates - Electoral campaign	Actors interested in harming/promoting certain candidates or parties or increasing social and political divisions in society spread malinformation on the private lives of candidates, or disinformation on political intentions, connections and activities of candidates and parties, or false allegations of violating campaign rules in order to defame candidates and parties, manipulate public opinion or influence voting behavior.	Translation by the authors: "Martin Sichert: Billionaire George Soros is known to want to make the world "better" in his favor by imposing an Ecodictatorship on the people under his control. every year at the world economic forum in Davos..." [...] "Bearbock (German politician) undergoes 5-year training program at the world economic forum. There, influential people are brought in to implement a worldwide political and social "change" in the sense of Soros and co." Translation by the authors: "Attila Hildmann: Bearbock verurteilte angriffe der hamas auf israel: 'Wir stehen an der Seite der Israelis' [...] Bearbock stands on the side of Israel because Bearbock is Jewish"

Table C.3

Code	Source	Qualification	Definition	Example
E-4	Krennerich 2021	Candidates - Election polls	Actors interested in (de-)legitimizing the elections or harming/promoting certain candidates or parties publish fictitious, false or supportive election polls to (de-) mobilize voters and/or influence both voter turnout and voters' decisions.	Circulation of non-representative or fabricated election polls, or the intentional misinterpretation of representative polls

INTEGRITY OF ELECTIONS

Table C.4

Code	Source	Qualification	Definition	Example
I-1	Krennerich 2021	Integrity - Voter registration	Actors interested in delegitimizing the elections or actors unwilling to accept electoral results might make undocumented claims that for example “dead voters” are voting to create distrust in the electoral process and to questioning electoral results.	Translation by the authors: “If the federal election takes place by postal vote and many dead people can vote for him, as was the case with Biden, Scholz will certainly become the next chancellor.” This disinformation is inspired by similar claims made in the USA during the 2020 election, even though the probability of dead people voting is comparatively low in Germany due to the link between the voter registry and the local resident registry, which does not exist in the USA.

Table C.4

Code	Source	Qualification	Definition	Example
I-2	Krennerich 2021	Integrity - Voting	Actors interested in delegitimizing the election or influencing voting behavior might spread rumors of non-secrecy of voting, sanctions on (non-)voters, electoral violence, as well as last-minute disinformation on election day (e.g. drop-out of candidates), or rumors that voting system is not safe and can be manipulated to create distrust and an atmosphere of fear, influencing voting behaviour. After the election, electoral contenders, particularly defeated candidates, and their supporters can make undocumented claims on systematic vote buying, ballot-box stuffing etc. to create distrust, justifying electoral defeat, and questioning electoral results, as well as delegitimizing elections in general and encouraging electoral protests.	Claims that mail-in votes are inherently fraudulent.

Table C.4

Code	Source	Qualification	Definition	Example
I-3	Krennerich 2021	Integrity - Counting and notification	Elections losers and their supporters make undocumented claims on lost ballot boxes, and non-counted votes, or the manipulation of vote counts and election protocols etc. to justify electoral defeat, question electoral results and delegitimizing elections, encouraging electoral protests.	Under the hashtag #Wahlbetrug, it was alleged that election workers had devalued AfD votes, i.e. made them invalid. The former AfD parliamentary group leader of the state parliament of Saxony-Anhalt and now ex-AfD politician André Poggenburg published a corresponding tweet. To underline his claim, he shared the photo of another tweet in which a user pretends to be an election worker in Saxony-Anhalt and attaches a photo on which election workers can be seen counting votes.

Table C.4

Code	Source	Qualification	Definition	Example
				The photo was not taken during the vote count for the state election in Saxony-Anhalt, but comes from Washoe County, Nevada, taken during the US election in November 2020, as the dts news agency, from which the photo was taken, confirmed to #Faktenfuchs. This is evident in the community's logo on the white column. The logo is painted over in red in the tweet, as are the faces of the election workers. Translation by the authors: "Almost 18,000 invalid ballot papers!!!! I've been a poll worker several times. A maximum of 5 invalid votes per district. That stinks to high heaven! #Voting fraud"

Table C.4

Code	Source	Qualification	Definition	Example
I-4	Krennerich 2021	Integrity - Publishing of electoral results	Election losers and their supporters make undocumented claims that late publication or changes of results as reported on election night indicate electoral fraud to justify electoral defeat, delegitimize democratic elections and encourage electoral protests.	On election night, Trump appeared to have a strong “lead” in Pennsylvania, and he expressed confidence he would win the state. He even wrongly “claimed” Pennsylvania’s Electoral College votes in a tweet before the race had been decided. But in the days ahead, as more mail ballots were processed and counted, Joe Biden pulled ahead and ultimately won the state by 81,000 votes, or about 1% of the total. The slow counting of mail ballots, and the way it eroded Trump’s early advantage, was the direct and expected result of Pennsylvania’s election rules — not fraud. The counting process took longer than in other states because elections officials were prohibited from starting their work before Election Day.

Table C.4

Code	Source	Qualification	Definition	Example
I-5	Krennerich 2021	Integrity - Electoral results	Election losers and their supporters make undocumented claims on electoral fraud to justify electoral defeat, delegitimize democratic election and encourage electoral protests.	Translation by the authors: "letter election has risen from 14 to 30 percent and the #CDU has simultaneously risen to almost 37 percent. Coincidences do happen... Anyone who believes that still believes in Father Christmas. The establishment will stop at nothing, not even electoral fraud. #electionfraud #Itwsa21"
I-6	Krennerich 2021	Integrity - Electoral observation	Government and its national/international supporters; government-friendly media promote false claims by pro-government election observers to legitimize non-democratic elections. Partisan election observers and media; foreign governments make false claims by partisan election observers (e.g. exaggerating the importance of minor irregularities) to delegitimizing democratic elections.	Unsubstantiated claims of a lack of independent oversight of the voting process and the vote count.

C.1. REGULATION OF POLITICAL CONTENT

C.1.1. REGULATION OF ONLINE CONTENT

Table C.5: Regulation of online content

Code	Source	Qualification	Definition	Example
Ro-1	(Art 34(1)(d) DSA), Art. 4 (a)	Gender-based violence	“violence against women” means gender-based violence, that is directed against a woman or a girl because she is a woman or a girl or that affects women or girls disproportionately, including all acts of such violence that result in, or are likely to result in, physical, sexual, psychological or economic harm or suffering, including threats of such acts, coercion or arbitrary deprivation of liberty, whether occurring in public or in private life (Source: Art. 4 (a))	E.g. online content targeting a female candidate before an election to accuse her of not being qualified for the job based on a falsely claimed non-existent of an academic degree [444]
Ro-2	Art 3 (10) TTPA, Art 8 (1) (b) TTPA, Art 9 (1) (e) TTPA, Art 10 TTPA, Art 11 (1) (b), Art 12 (1) (a-b) TTPA	Sponsorship and identification of sponsors	Ensures all political ads display the sponsor's name and details needed to identify the person who paid for the ad.	A political ad without identifiable information about the sponsor.
Ro-3	Art 3 (11) TTPA, Art 8 (1) (f) TTPA, Art 10 TTPA, Art 11 (1) (d) TTPA, Art 12 (1) (c) TTPA, Art 18 TTPA, Art 19 TTPA, Art 35 (3) DSA	Targeting Criteria Disclosure	Platforms must disclose targeting criteria for political ads, allowing users to understand why they were targeted.	An ad targeting specific demographics without disclosing the criteria, e.g. women in a certain zip code area.

Table C.5: Regulation of online content

Code	Source	Qualification	Definition	Example
Ro-4	Art 13 TTPA, Art 39 DSA	Ad Repository Requirements	Platforms must maintain a publicly accessible repository of political ads, including sponsor and targeting details. Platforms must provide transparency around algorithms used in ad targeting, allowing users to understand how content is distributed.	A platform that doesn't archive political ads for public access.
Ro-5	Art 26 DSA, Art 39 DSA, Art 11 TTPA, Art 28 TTPA	Political Ad Labelling	Requires clear, prominent labeling of political ads in real-time to distinguish them as political content.	Political ads lacking labels, making it unclear they're paid content.
Ro-6	Guideline Sec. 3.2.1(f)	Influencer Political Content	Influencers must declare political content and platforms must label it, ensuring clear disclosure of sponsorships.	Influencer posts endorsing candidates without sponsorship disclosure.
Ro-7	DSA Art. 39, Guideline Sec. 3.2.3	Labelling of Synthetic Media	Platforms must label synthetic media (e.g., deep-fakes, AI-generated misinformation) in political ads to prevent misinformation during elections.	AI-generated videos of candidates without clear labeling as synthetic.
Ro-8	Rec. 14	Silence Period Enforcement	Platforms must respect national silence periods (ad bans before elections) to avoid undue influence on voters.	A political ad shown during the silence period prior to voting day.

Table C.5: Regulation of online content

Code	Source	Qualification	Definition	Example
Ro-9	Guideline Sec. 3.3	Monitoring and Fact-checking	Platforms must label content identified by fact-checkers as disinformation, helping users make informed decisions.	Content spreading misinformation without a fact-check label.

D

APPENDIX FOR CHAPTER 8

D.1. INFORMATION ABOUT WORKSHOP PARTICIPANTS

Table D.1: Participant Expertise and Years of Experience.

#	Area of Expertise	Experience
1	HCI	<2 years
2	Law, Information Systems, Journalism	2–5 years
3	Security, Privacy, Measurements	>8 years
4	Consumer IoT behaviours, Security Awareness, Security Economics	>8 years
5	Computational Social Science, Hate Speech, Shitposting	5–8 years
6	Transformers, Machine Learning	5–8 years
7	Data Access, Research Engineering	2–5 years
8	HCI, ML	2–5 years
9	HCI, UX, Qualitative Coding	>8 years
10	Digital Infrastructures, Platforms, and Services; Competition, Regulation, Innovation	>8 years
11	Platform Governance, Disinformation, DSA	2–5 years
12	Platform Regulation, Blogging	>8 years

D.2. THEMATIC ANALYSIS PROCESS AND CODING TRACE

D.2.1. THEMATIC ANALYSIS PROCESS

The thematic analysis followed Braun & Clarke’s reflexive framework [92, 93], using an inductive approach to derive themes from the Miro board outputs.

The analysis comprised the following steps:

1. **Familiarization with Data:** The lead author reviewed the Miro board content,

focusing on the color-coded outputs from the three interdisciplinary teams: Legal (white), Social Science (green), and Tech (yellow).

2. **Generating Initial Codes:** Codes were created inductively, capturing key concepts and recurring patterns within each team's contributions.
3. **Collating Codes into Themes:** Codes were clustered into higher-order themes that spanned across teams, reflecting shared or divergent insights.
4. **Reviewing Themes:** The initial themes were iteratively refined through discussions with a co-author specializing in platform governance and data regulation. These discussions challenged assumptions and ensured alignment with interdisciplinary perspectives.
5. **Defining and Naming Themes:** Final themes were labeled to reflect key concerns and actionable insights relevant to LLM-supported auditing under the DSA.
6. **Producing the Report:** Themes were integrated into the *Results* section and contextualized with relevant literature. Traceability from raw data to final themes is demonstrated in Table D.2.

D.2.2. DETAILED CODING TRACE

Table D.2: Sample of Coding Traces from Miro Board to Final Themes

Raw Miro Data (Post-it / Note)	Initial Code	Final Theme	Team
Use the LLM to update the codebook itself	Codebook Adaptation	Organizational Aspects	Tech
Annotate rare codes with a human to verify	Human-in-the-loop Verification	Organizational Aspects	Legal
Demand explanations in outputs	Explanation Requirement	Methodology & Quality Control	Social Sci.
Prompts should be tested repeatedly with varied phrasings	Prompt Testing Strategies	Methodology & Quality Control	Tech
Users shouldn't be responsible for explaining everything	User Accountability	User Aspects	Social Sci.
Being a skilled auditor is something separate	Professionalization of Auditing	User Aspects	Social Sci.
Are platforms even enforcing their own rules?	Enforcement Gaps	Platform-Related Aspects	Legal
Simplified legal guidance co-developed with experts	Legal Simplification	Regulations	Legal
Open-source models preferred to ensure reproducibility	Model Reproducibility	Technological Aspects	Tech
Set temperature to zero to reduce randomness	Parameter Tuning	Technological Aspects	Tech
Legal definitions from law work well but too complex terms can fail	Legal Framing of Prompts	Legal Input	Legal
Prompt scoping helps reduce hallucinations	Scope Limitation	Limitation of Scope	Tech

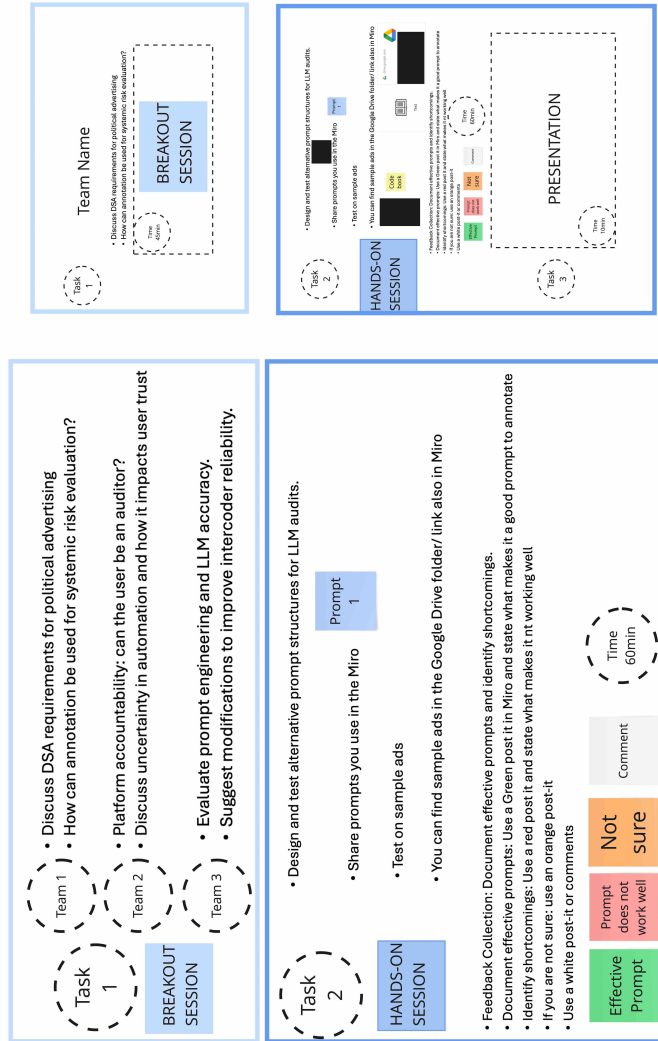


Figure D.1: Zoomed-in visual documentation of the Breakout and Hands-On Sessions via Miro board, showing the structure of tasks and layout from which thematic analysis was conducted. The board aggregates key ideas and discussion points on leveraging LLMs for systemic risk evaluation under the DSA.

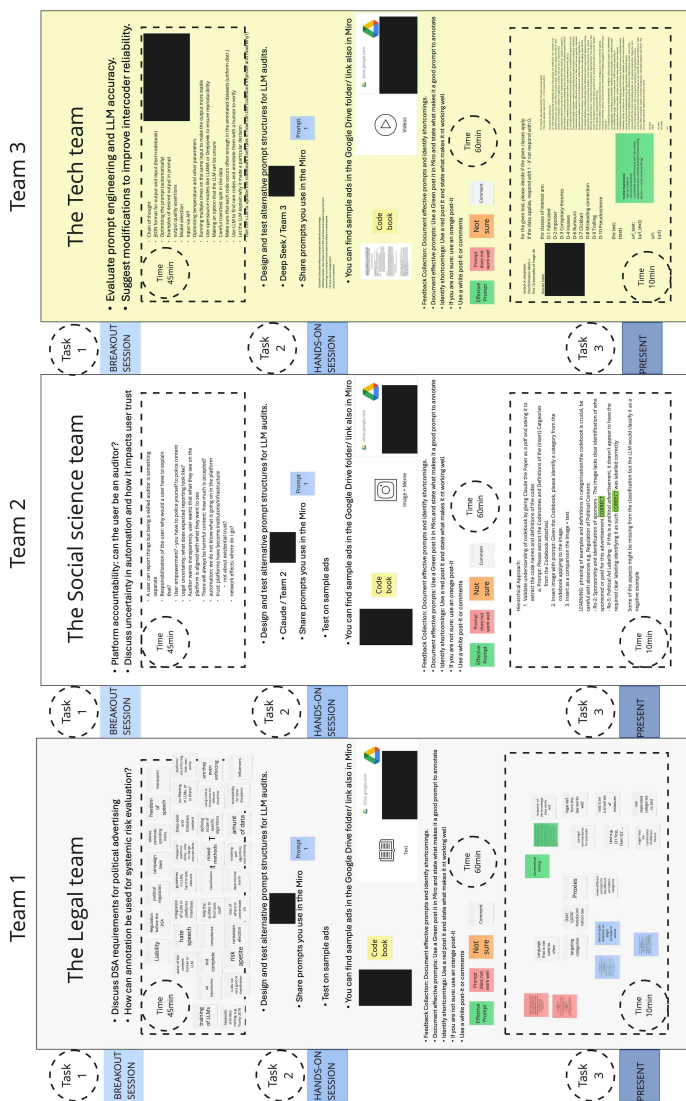


Figure D.2: Zoomed-out overview of all Breakout and Hands-On Sessions Miro boards, visualizing inputs from the interdisciplinary expert workshop on LLM-based systemic risk evaluation under the DSA.

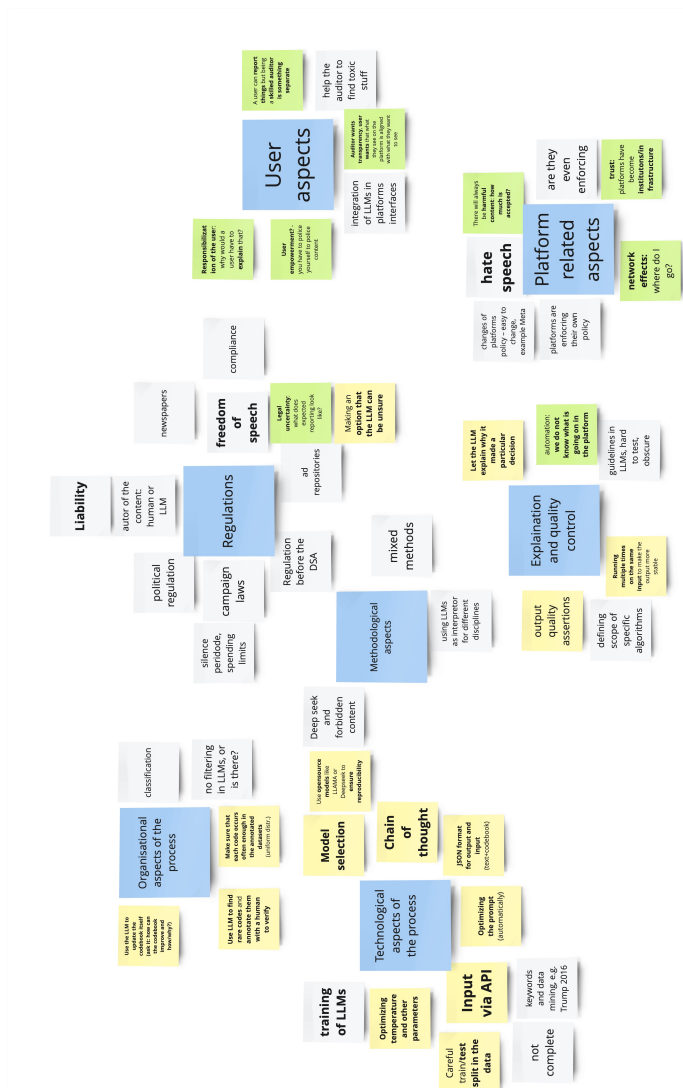


Figure D.3: Clustered and color-coded visualization of the Breakout Session discussions, as captured on the Miro board. These inputs were synthesized through thematic analysis, with key insights detailed in the text.

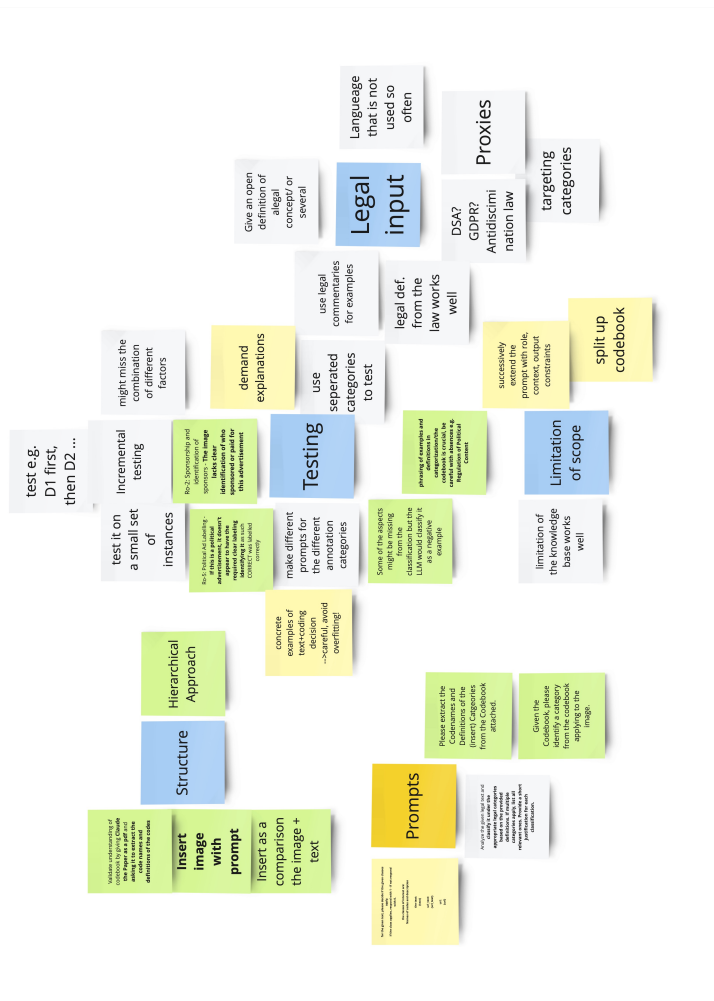


Figure D.4: Clustered and color-coded visualization of the Hands-On Session discussions, as captured on the Miro board. These inputs were synthesized through thematic analysis, with key insights detailed in the text.

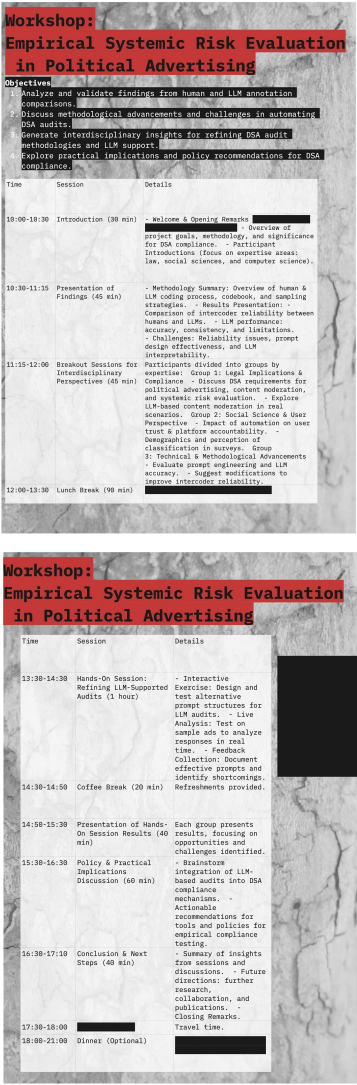


Figure D.5: Program for Expert Workshop.

E

APPENDIX FOR CHAPTER 9

E.1. CODEBOOK LEGAL AND NON-LEGAL CODES

Table E.1: Code 1-1

Field	Content
Legal Source	Security Council resolution 1566 (2004), Rec. 3 “(...) criminal acts, including against civilians, committed with the intent to cause death or serious bodily injury, or taking of hostages, with the purpose to provoke a state of terror in the general public or in a group of persons or particular persons, intimidate a population or compel a Government or an international organization to do or to abstain from doing any act”.
Qualification	Terrorism promotion
Notes on interpretation and application	E: The promotion or a call for a new form of terrorism has arisen, which is directed against IT services, in particular against state institutions, and is referred to as “cyber terrorism”. F: The Internet and social networks such as Facebook, LinkedIn, Video, Twitter and YouTube are increasingly being used by terrorist organizations to recruit, finance, train and incite Internet users to spread terrorism and commit terrorist acts.

Table E.2: Code 1-2

Field	Content
Legal Source	Article 2 (1) International Convention for the Suppression of the Financing of Terrorism (1999) “Any person commits an offence within the meaning of this Convention if that person by any means, directly or indirectly, unlawfully and wilfully, provides or collects funds with the intention that they should be used or in the knowledge that they are to be used, in full or in part, in order to carry out: (a) An act which constitutes an offence within the scope of and as defined in one of the treaties listed in the annex; or (b) Any other act intended to cause death or serious bodily injury to a civilian, or to any other person not taking an active part in the hostilities in a situation of armed conflict, when the purpose of such act, by its nature or context, is to intimidate a population, or to compel a government or an international organization to do or to abstain from doing any act.”
Qualification	Terrorism financing
Notes on interpretation and application	Promotion of terrorist financing through e.g. calls to action.

Continued on next page

Field	Content
	“Financing of terrorism: Terrorism activities take place in many different forms ranging from isolated acts to planned activities of organised groups. The forms of financing of terrorism therefore vary accordingly. It comprises not only the financing of the terrorist acts as such, but also any support to the criminal network. Terrorist organisations require significant funding, both for the actual undertaking of terrorist acts, but also to other issues: to maintain the functioning of the organisation, to provide for its basic technical necessities, as well as to cover costs related to spreading related ideologies (...)“

Table E.3: Code 1-3

Field	Content
Legal Source	Art. 24 International Covenant on Civil and Political Rights (ICCPR) “1. Every child shall have, without any discrimination as to race, colour, sex, language, religion, national or social origin, property or birth, the right to such measures of protection as are required by his status as a minor, on the part of his family, society and the State. 2. Every child shall be registered immediately after birth and shall have a name. 3. Every child has the right to acquire a nationality.” Declaration of the Rights of the Child, 1989
Qualification	Sexual exploitation of children / Child abuse material or anything objectionable involving minors, grooming or predation
Notes on interpretation and application	“Online sexual exploitation most commonly includes grooming, live streaming, consuming child sexual abuse material, and coercing and blackmailing children for sexual purposes.”
Code	1-4
Legal Source	United Nations Convention against Transnational Organized Crime and the Protocols, Protocol to Prevent, Suppress and Punish Trafficking in Persons, especially Women and Children, UNGA 55/25
Qualification	Promotion of and/or enabling of human trafficking, Offering or advertising human trafficking
Notes on interpretation and application	“Three constituent elements of human trafficking: The Act (What is done): Recruitment, transportation, transfer, harbouring or receipt of persons. The Means (How it is done): Threat or use of force, coercion, abduction, fraud, deception, abuse of power or vulnerability, or giving payments or benefits to a person in control of the victim.

Continued on next page

Field	Content
	The Purpose (Why it is done): For the purpose of exploitation, which includes exploiting the prostitution of others, sexual exploitation, forced labour, slavery or similar practices and the removal of organs."

Table E.4: Code 1-5

Field	Content
Legal Source	Article III Convention on the Prevention and Punishment of the Crime of Genocide, Article 25 (3) and Article 6 Rome Statute of the International Criminal Court
Qualification	Incitement to or promotion of genocide (jus cogens)
Notes on interpretation and application	"Statements that dehumanize as a 'genocidal hate speech' and contribute to a culture of hate and thus create the conditions for committing genocide. An example of genocidal hate speech would be an RTLM radio host stating that 'Tutsis are cockroaches.'" Dehumanizing rhetoric: "A campaign of hate and dehumanization of the Rohingya had been under way for months, and escalated after 8 June 2012, led by the Rakhine Nationalities Development Party (RNDP), various Rakhine organizations, radical Buddhist monk organizations, and several officials and influential figures. It was spread through anti-Rohingya or anti-Muslim publications, public statements, rallies and the boycott of Muslim shops. The Rohingya were labelled 'illegal immigrants' and 'terrorists' and portrayed as an existential threat that might 'swallow other races' with their 'incontrollable birth rates'. In November 2012, the RNDP, in Toe Thet Yay, an official publication, cited Hitler, arguing that 'inhuman acts' were sometimes necessary to 'maintain a race'."

Table E.5: Code 1-6

Field	Content
Legal Source	Art. 19 ICCPR, Art. 17 ECHR: Prohibition of abuse of rights "Nothing in this Convention may be interpreted as implying for any State, group or person any right to engage in any activity or perform any act aimed at the destruction of any of the rights and freedoms set forth herein or at their limitation to a greater extent than is provided for in the Convention."
Qualification	Qualified hate speech amounting to calls for violence

Continued on next page

Field	Content
Notes on interpretation and application	“Hate speech entails the advocacy, promotion or incitement to denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization or threat to such persons on the basis of a non-exhaustive list of personal characteristics or status that includes race, colour, language, religion or belief, nationality or national or ethnic origin, as well as descent, age, disability, sex, gender, gender identity and sexual orientation. A particular feature of hate speech is that it aims at inciting or can reasonably be expected to have the effect of inciting others to commit acts of violence, intimidation, hostility or discrimination against those it targets. Incitement to hatred can result from insulting, the holding up to ridicule or the slandering of specific groups of population where such forms of expression are exercised in an irresponsible manner, which might entail being unnecessarily offensive, advocating discrimination or using vexatious or humiliating language. Intent to incite might also be established where there is an unambiguous call by a person using hate speech to others to commit the impugned acts. It might also be inferred from the strength of the language used and other relevant circumstances in which hate speech is used. Member states must sanction or even prevent all forms of expression which spread, incite, promote or justify hatred, provided that any formalities, conditions, restrictions or penalties imposed are proportionate to the legitimate aim pursued.” “Incitement to hatred can result from insulting, the holding up to ridicule or the slandering of specific groups of population where such forms of expression are exercised in an irresponsible manner, which might entail being unnecessarily offensive, advocating discrimination or using vexatious or humiliating language. Intent to incite might also be established where there is an unambiguous call by a person using hate speech to others to commit the impugned acts. It might also be inferred from the strength of the language used and other relevant circumstances in which hate speech is used. Member states must sanction or even prevent all forms of expression which spread, incite, promote or justify hatred, provided that any formalities, conditions, restrictions or penalties imposed are proportionate to the legitimate aim pursued.”

Table E.6: Code 1-7

Field	Content
Legal Sources	Art. 17 ICCPR, CCPR General Comment No. 16, Consideration 4 GDPR, Article 8 ECHR, Article 11 IACHR
Qualification	Qualified breaches of personal information, defamatory personal content, Impersonation (fake accounts/profiles/pages), sexual objectification, unauthorized dissemination of intimate images (“revenge porn”)

Continued on next page

Field	Content
Notes on interpretation and application	“Article 17 provides for the right of every person to be protected against arbitrary or unlawful interference with his privacy, family, home or correspondence as well as against unlawful attacks on his honour and reputation. In the view of the Committee this right is required to be guaranteed against all such interferences and attacks whether they emanate from State authorities or from natural or legal persons. The obligations imposed by this article require the State to adopt legislative and other measures to give effect to the prohibition against such interferences and attacks as well as to the protection of this right.”

Table E.7: Code 1-8

Field	Content
Legal Sources	Art. 19 ICCPR, CCPR General Comment 34 (CCPR/C/GC/34), Art. 10 ECHR: Freedom of expression, Art. 13 IACHR: Freedom of Thought and Expression, Art 9 ACHPR (Banjul Charter)
Qualification	Qualified violations/misuses of freedom of expression
Notes on interpretation and application	Disinformation, sexually explicit content, nudity or pornography, content critical of religion (blasphemy/apostasy). Defamatory personal content, bullying, harassment (coordinated/organized damage, cyberstalking, deadnaming, doxing). Violent, graphic content, promotion or publication of crimes, content to organize violence or to support violent organizations, sexual violence and exploitation, aid to self-harm or suicide, disclosure of confidential or secret information, lèse majesté or critical comments on people in contemporary history.

Table E.8: Code 1-9

Field	Content
Legal Sources	Art. 20 ICCPR, General Comment No. 11: Prohibition of propaganda for war and inciting national, racial or religious hatred (Art. 20) 29/07/1983, CCPR General Comment No. 11. (General Comments).
Qualification	Prohibition of propaganda for war and inciting national, radical or religious hatred: Hate speech, violent extremist content.

Continued on next page

Field	Content
Notes on interpretation and application	Freedom of information is applicable not only to “information” or “ideas” that are favourably received or regarded as inoffensive or as a matter of indifference, but also to those that offend, shock or disturb the State or any sector of the population.

Table E.9: Code 2-G-1-1

Field	Content
Legal Source	§ 86 German Criminal Code Dissemination of propaganda material of unconstitutional organisations: “(1) Whoever disseminates in Germany or produces, stocks, imports or exports or makes publicly available through data storage media for dissemination in Germany or abroad the propaganda material. 1. of a political party which has been declared unconstitutional by the Federal Constitutional Court or a political party or organisation which has been held by final decision to be a surrogate organisation of such a party, 2. of an organisation which has been banned by final decision because it is directed against the constitutional order or against the concept of international understanding or which has been held by final decision to be a surrogate organisation of such a banned organisation, 3. of a government, organisation or institution outside the territorial scope of this statute which is actively pursuing the objectives of one of the political parties or organisations referred to in nos. 1 and 2 or 4. propaganda material the content of which is intended to further the activities of a former National Socialist organisation incurs a penalty of imprisonment for a term not exceeding three years or a fine. (2) Propaganda material within the meaning of subsection (1) is only material (section 11 (3)) whose content is directed against the free democratic basic order or the concept of international understanding. (3) Subsection (1) does not apply if the propaganda material or the act serves civic information, to prevent unconstitutional activities, to promote the arts or science, research or teaching, reporting about current or historical events, or similar purposes. (4) If the degree of guilt is minor, the court may dispense with imposing a penalty under this provision.”
Qualification	Dissemination of propaganda material of unconstitutional organisations
Notes on interpretation and application	Dissemination of propaganda. Propaganda means are writings (§ 11 III StGB) that contain content-promoting and support-oriented aspects. Organization must be prohibited (§§ 84, 84 StGB). Example:

Continued on next page

Field	Content
	(1) The setting of hyperlinks to forbidden Nazi websites does not constitute aiding and abetting sedition, if it is carried out in a recognisable manner, distancing oneself from the content of these sites within the framework of extensive documentation. (2) The offer on a homepage to have blocked websites read out to you over the telephone is not punishable if it is recognisably a satire.

Table E.10: Code 2-G-1-2

Field	Content
Legal Source	§ 86a German Criminal Code “(1) Whoever 1. disseminates the symbols of one of the political parties or organisations designated in section 86 (1) nos. 1, 2 and 4 in Germany or uses them publicly, in a meeting or in material (section 11 (3)) disseminated by themselves or 2. produces, stocks, imports or exports objects which depict or contain such symbols for dissemination or use in Germany or abroad in a manner referred to in no. 1, incurs a penalty of imprisonment for a term not exceeding three years or a fine. (2) Symbols within the meaning of subsection (1) are, in particular, flags, insignia, uniforms and their parts, slogans and forms of greeting. Symbols which are so similar as to be mistaken for those referred to in sentence 1 are deemed to be equivalent to them (...)”
Qualification	Use of symbols of unconstitutional organisations
Notes on interpretation and application	“By inserting swastikas into the ‘Aryan Music Group’ internet platform set up by him (...), the defendant made public use (Paragraph 86a (1) (1) of the German Criminal Code) of signs identifying unconstitutional organisations”.

Table E.11: Code 2-G-1-3

Field	Content
Legal Source	§ 89a German Criminal Code Preparation of serious violent offence endangering state

Continued on next page

Field	Content
	“(1) Whoever prepares a serious violent offence endangering the state incurs a penalty of imprisonment for a term of between six months and 10 years. A ‘serious violent offence endangering the state’ means an offence against life in the cases under section 211 or 212 or against personal liberty in the cases under section 239a or 239b which, under the circumstances, is intended and suited to undermine the continued existence or security of a state or of an international organisation, or to abolish, rob of legal effect or subvert the constitutional principles of the Federal Republic of Germany. (2) Subsection (1) only applies to offenders who prepare a serious violent offence endangering the state by 1. instructing another person or receiving instruction in the production or the use of firearms, explosives, explosive or incendiary devices, nuclear fission material or other radioactive substances, substances which contain or can produce poison, other substances which are detrimental to health, special devices which are necessary in the commission of the offence or other skills which may be of use in the commission of an offence under subsection (1), 2. producing, obtaining for themselves or another, storing or supplying to another weapons, substances or devices referred to in no. 1 or 3. obtaining or storing objects or substances which are essential for the production of weapons, substances or devices referred to in no. 1. (2a) Subsection (1) also applies to offenders who prepare a serious offence endangering the state by undertaking to leave Germany for the purpose of committing a serious offence endangering the state or the acts referred to in subsection (2) no. 1 in order to enter a state where people are given training and instructions within the meaning of subsection (2) no. 1. (3) Subsection (1) also applies if the preparations are made abroad. If the preparations are made outside the territory of the Member States of the European Union, the aforesaid only applies if the preparations are made by a German national or a foreign national whose livelihood is based in Germany, or the serious violent offence endangering the state so prepared is to be committed in Germany or against a German national. (...)”
Qualification	Preparation of serious violent offence endangering state

Continued on next page

Field	Content
Notes on interpretation and application	“A serious act of violence that endangers the state pursuant to § 89a (1) sentence 2 StGB exists if the realisation of an offence enumerated there is likely to impair the security of a state. The state protection clause covers both the Federal Republic of Germany and all foreign states. All states recognised under international law are protected, irrespective of whether they are regarded here as legal or criminal offences. The support or approval of German foreign policy is also irrelevant. The concept of the security of a state includes its internal and external security. Internal security is the state of relative safety of its existence and constitution against violent actions of internal forces, whereby the ability of a state to defend itself internally against disturbances is central. It will normally be compromised if the prepared act, as the perpetrator imagines it to be, is likely under the circumstances to affect the internal structure of a State. The necessary suitability must be determined objectively on the basis of the (quasi fictitious) circumstances of the prepared act. From a subjective point of view (‘determined’) it is a prerequisite that the perpetrator has included the possible consequences of the prepared act in his will. For this purpose, it is sufficient that he has recognised the factual circumstances which show the suitability for impairing the protected property and has included them in his will. On the other hand, it is not necessary to act in a targeted manner to impair internal security in the sense of an intention. With regard to the corresponding aptitude and determination, the concrete circumstances of the individual case must be taken into account; in this context, the question of state endangerment may depend on details such as the prominence of the victims, the publicity or symbolic significance of the location and the circumstances of the act (Senate, judgement of 27 October 2015 - 3 StR 218/15, BGHSt 61, 36. 10)”.

Table E.12: Code 2-G-1-4

Field	Content
Legal Source	§ 91 German Criminal Code Instructions for committing serious violent offence endangering state “(1) Whoever - extols or gives another person access to material (section 11 (3)) whose content is of such a nature as to serve as instructions for committing a serious violent offence endangering the state (section 89a (1)) if the circumstances of its dissemination are conducive to promoting or encouraging others’ preparedness to commit a serious violent offence endangering the state, - obtains material of the kind designated in no. 1 for the purpose of committing a serious violent offence endangering the state incurs a penalty of imprisonment for a term not exceeding three years or a fine. (2) Subsection (1) no. 1 does not apply if

Continued on next page

Field	Content
	1. the act serves the purpose of civic information, protection against anti-constitutional activities, the arts and science, research or teaching, reporting about current or historical events, or similar purposes or 2. the act solely serves the performance of lawful professional or official duties. (3) If the degree of guilt is minor, the court may dispense with imposing a penalty under this provision.”
Qualification	Instructions for committing serious violent offence endangering state
Notes on interpretation and application	Statements that lead to the commission of serious acts of violence that endanger the state. Instruction manual: Due to their content, they are suitable as instructions for an act within the meaning of 89a I of the penal code and to encourage the willingness of third parties to commit the offense. Spreading the instruction manual, by advertising or making it accessible. “(a) In the period between June 2014 and 21 November 2015, the defendant strongly lobbied his then life partner S. M. to leave for Syria, join the IS and carry out a suicide bombing with an explosive belt. It was important to him to supply the IS with a suicide bomber in order to strengthen it militarily by killing a large number of opponents (‘infidels’) (paragraph 1 part 1 of the executed warrant). (b) During the same period, the accused showed M. M., son of S. M., born on 5 August 2007, videos showing children fighting for the I.S., shooting at other people and being shot themselves. His aim was to win M. M. as a child soldier for the organisation and to make him receptive to its ideas and activities (paragraph 1 part 2 of the executed warrant). (c) On 15 July 2016 at about 10 p.m., the accused used a chat room to influence an unknown underage male person named H., presumably his younger brother living in Syria, to commit a suicide bombing with explosives for the IS. He stated that if his chat partner did not carry out an attack because he was ‘scared’ or thought he was too young, they would be ‘disgraced’ (paragraph 2 of the executed warrant). bb) With regard to the non-European terrorist organisation ‘Islamic State’ (IS), the urgent suspicion of a crime results from the evaluations of the Federal Public Prosecutor General in the introductory note of 12 December 2016, which - according to court experience - are based in particular on the expert opinions of the Islamic Studies expert Dr St. The defendant has so far denied the allegations, in particular his membership of the IS. During his police interrogation, he testified that one of his eight brothers had first been in Syria with the ‘Free Syrian Army’ (FSA) and had ‘then changed’, possibly to the I.S.; this brother had been ‘killed by the government’ in 2014. He - the defendant - never claimed to be a member of the IS.
<i>Continued on next page</i>	

Field	Content
	With regard to the membership in the IS, the urgent suspicion is based above all on the evaluation of the accused's digital communication, namely on the secured and transmitted chat contents. Thus, he described himself as a member of the 'Dawla' (chat of 21 January 2014) and declared that he had seen Abu Bakr al-Baghdadi (chat of 10 February 2014) and that he had 'pledged his loyalty to the ISIS' (chat of 6 July 2014). The extensive chat traffic is characterised by the Islamist attitude of the accused. In a large number of such messages, he exchanged information with various communication partners about matters related to the IS as well as suicide bombings and other acts of murder. For example, he told a caller with a Syrian telephone number who he advised to commit suicide bombing: 'I could let you join the Islamic State in Iraq and Syria. I could arrange it for you ... I'll have them arrange a martyr operation for you' (chat of 1 March 2014). Information on the circumstances of the accused's family members also finds its support in chat traffic. In particular, communication with them also provides 'inside information' on planned operations of the IS; for example, he pointed out to chat partners, presumably his aunt and uncle, in good time - out of concern for their safety - the planned attacks in Tartu and Raqqa (chats of 27 August 2013 and 14 May 2014). Moreover, the accused communicated, for example, with a person in Turkey speaking for the IS named 'S. N.' about the 'Pledge of Allegiance to the Caliph', 'Coalition Clauses' and the desired political future of Syria (chats from August 2013 to July 2014)."

Table E.13: Code 2-G-1-5

Field	Content
Legal Source	§100a German Criminal Code "(1) Whoever, despite knowing better, allows forged or falsified objects, reports concerning them or untrue assertions of a factual nature to come to the attention of another or to become known to the public, which, if they were genuine or true, would be of significance for the external security of the Federal Republic of Germany or its relations with a foreign power in order to deceive a foreign power into believing them to be genuine objects or facts, and thereby creates the risk of serious detriment to the external security of the Federal Republic of Germany or its relations with a foreign power, incurs a penalty of imprisonment for a term of between six months and five years. (2) Whoever produces such objects through forgery or falsification or procures them in order to allow them to come to the attention of another or to become known to the public in the manner indicated in subsection (1) in order to deceive a foreign power, and thereby creates the risk of serious detriment to the external security of the Federal Republic of Germany or its relations with a foreign power, incurs the same penalty."
Qualification	Treasonous forgery
Notes on interpretation and application	"3. counterfeiting on the grounds of treason (§ 100 a para. 2 StGB).

Continued on next page

Field	Content
	This offence is committed by anyone who intentionally allows facts, objects or messages about them which are false, falsified or untrue, but which, if they were genuine or true, would be state secrets, to reach an unauthorised person as genuine or true or makes them public and thereby endangers the welfare of the Federal Republic of Germany or one of its countries. The defendant has fulfilled this condition by deliberately making the claim, contrary to the truth, that there are political collateral agreements to the EDC Treaty which the Federal Government is keeping secret from the German people and the Bundestag, and that the Gehlen organisation is carrying out increased activity in France in connection with attempts by German generals to achieve hegemony in the EDC countries, 'disclosures', through which he is seriously endangering the political position of the Federal Republic, its reputation both internally and externally.
	<i>Continued on next page</i>

Field	Content
	It is not a case of betrayal of a genuine state secret within the meaning of Section 100 of the German Criminal Code, but rather the public disclosure of an 'invented state secret' in accordance with Section 100a (2) of the German Criminal Code. It is true that the concept of state secrets is defined for both § 100 and § 100a, according to the structure of the law, in § 99 StGB as 'facts, objects or knowledge etc., the secrecy of which is necessary for the welfare of the Federal Republic of Germany or one of its states before a foreign government'. The question of doubt as to whether a state or endeavour of the government that is contrary to the Basic Law can be a genuine state secret within the meaning of § 100 StGB, which has not yet been decided by the Federal Court of Justice, does not arise in the case of the counterfeiting of § 100a StGB in violation of the constitution. The genesis of this provision leaves no doubt that it is also - and precisely - intended to cover the passing on of news etc. about allegedly unlawful conditions and endeavours, i.e., 'illegal' state secrets. The Member of the Bundestag Dr. Arndt, who, as far as can be seen, was the only one who explicitly commented on this issue when § 100 a was discussed in the Bundestag, stated 'The invented state secrets can of course also include illegal facts. This will be the most dangerous and shameful betrayal in particular'. (Negotiations of the German Bundestag 1. parliamentary term, 158th session, Stenographic Reports S 6325). This was not contradicted during the deliberations in the Bundestag. This also corresponds to the meaning and purpose of § 100 a. If one were to interpret the concept of state secrets in § 100a StGB as meaning that the invented fact which is 'betrayed' must not be 'illegal', must not be contrary to the constitution, the meaning of this provision would be distorted and its effect paralysed, especially for the most important cases which the legislature had in mind. Facts which have just been invented and disseminated which are contrary to the Basic Law may considerably impair the welfare of the Federal Republic of Germany because they refer, and regularly do so in relation to foreign countries, to alleged unconstitutional events or circumstances. The more malicious the forgery, the greater the threat to the welfare of the Federal Republic of Germany would be posed by the nature of the invented 'secret', the lower would be its protection against this most dangerous form of foreign policy well poisoning. If the defendant in the Gehlen case had confined himself to the false assertion that the Gehlen organisation was carrying out increased activity in France, even the narrowest interpretation would undoubtedly have resulted in the facts of § 100 a (2). However, since he also stated that the Gehlen organisation had increased its activities in France in preparation for aggressive, illegal efforts by German generals, such an incorrect interpretation would remove the criminal liability of the offence under Paragraph 100a (2) and, under certain circumstances, the offence would remain unpunished at all. This is not the intention of the law and was not the will of the legislature. The 'facts' invented and disclosed by the defendant, which, with regard to both the secret agreements on the EDC contract and the Gehlen organisation, concerned allegedly unlawful acts and plans, were therefore state secrets within the meaning of § 100 a (2) StGB. They were presented to the public as true and secret and were considered state secrets according to § 100a."

Table E.14: Code 2-G-1-6

Field	Content
Legal Source	“§ 111 German Criminal Code (1) Whoever publicly, in a meeting or by disseminating material (section 11 (3)) incites the commission of an unlawful act incurs the same penalty as an abettor (section 26) (...)”
Qualification	Public incitement to commit offences
Notes on interpretation and application	The incitement for unlawful acts carried out in the manner described is punished. An incitement is a certain statement, beyond mere approval, that others should do or refrain from doing something. The general incitement is addressed to an indefinite number of people. “The conviction of the accused for public provocation to commit criminal offences pursuant to § 111 (1) and (2) StGB in case II 5 a) of the grounds for the judgement does not stand. a) According to the findings made in this regard, the defendant, after a dispute with the aggrieved party, praised K. After a dispute with the aggrieved party K., with whom he had previously entered into a relationship, the defendant offered to pay 200 euros for the killing of the aggrieved parties on his ‘Facebook’ page, which can be viewed by all users of this Internet platform. He also informed them of her whereabouts and attached several photographs of the victim. The entry did not result in any criminal offences against the victim. b) The conviction for public provocation to commit a crime does not stand up to legal review. The Regional Court correctly assumed the existence of the objective prerequisites of this criminal provision but did not deal with the defendant’s inner criminal side. It therefore remains unclear whether the defendant at least accepted that his summons would be taken seriously by others. The fact that the defendant has confessed to the commission of the offence does not replace the finding of a corresponding intention.”

Table E.15: Code 2-G-1-7

Field	Content
Legal Source	§ 126 German Criminal Code “(1) Whoever, in a manner which is suitable for causing a disturbance of the public peace, threatens to commit 1. breach of the peace as designated in section 125a sentence 2 nos. 1 to 4, 2. murder under specific aggravating circumstances (section 211), murder (section 212) or genocide (section 6 of the Code of Crimes against International Law) or a crime against humanity (section 7 of the Code of Crimes against International Law) or a war crime (section 8, 9, 10, 11 or 12 of the Code of Crimes against International Law),

Continued on next page

Field	Content
	3. grievous bodily harm (section 226), 4. an offence against personal liberty under section 232 (3) sentence 2, section 232a (3), (4) or (5), section 232b (3) or (4), section 233a (3) or (4), each to the extent that it represents a serious criminal offence, section 234, 234a, 239a or 239b, 5. robbery or extortion with use of force or threat of force (sections 249 to 251 or section 255), 6. a serious criminal offence constituting a public danger under sections 306 to 306c or section 307 (1) to (3), section 308 (1) to (3), section 309 (1) to (4), section 313, section 314 or section 315 (3), section 315b (3), section 316a (1) or (3), section 316c (1) or (3) or section 318 (3) or (4) or 7. a less serious criminal offence constituting a public danger under section 309 (6), section 311 (1), section 316b (1), section 317 (1) or section 318 (1) (...) incurs a penalty of imprisonment for a term not exceeding three years or a fine.”
Qualification	Disturbing public peace by threatening to commit offences. Expressions that are likely to disturb public peace.
Notes on interpretation and application	Public peace is an objective state of general legal security and the subjective consciousness of the population to live in peace and quiet. Disturbance of this state must be affected by the threat of criminal offences (catalogue). Peace is disturbed when a general disturbance of the population within Germany occurs, but at least below a not inconsiderable number of persons. Threat is the announcement that the commission of one of the offences mentioned in § 126 StGB is imminent, whereby the threatening party claims to have influence on this. The offence must already be substantiated in its essential features. “The threat of bomb attacks in a generally accessible medium is likely to frighten the population or in any case to create or intensify a state of insecurity and thereby disturb the public peace in the sense of the provision of § 126 (1) StGB. Even if the political conviction of an accused person cannot, in principle, justify the assumption of an unfavourable social prognosis on its own, this principle cannot apply without further ado in cases of the new phenomenon of radical Islamist perpetrators, even if there are doubts about the continued existence of the corresponding extremist ‘political’ conviction.” “An editor who reports on the website, via Facebook or Twitter from an objective perspective - deliberately against his better judgment - that the ‘largest terrorist attack to date in Western Europe’ has taken place in Mannheim, whereby it is not apparent to the reader of the freely accessible part of the article that it is an invented story, is about to disrupt the public peace by pretending that a murder, manslaughter or grievous bodily harm is imminent, in accordance with section 126 (2) in conjunction with section 126 (3) of the German Civil Code. para. 1 StGB.”

Table E.16: Code 2-G-1-8

Field	Content
Legal Source	§ 129 German Criminal Code “(1) Whoever forms an organisation or participates as a member in an organisation the objectives or activities of which are directed at the commission of offences which incur a penalty of a maximum term of imprisonment of at least two years incurs a penalty of imprisonment for a term not exceeding five years or a fine. Whoever supports such an organisation or recruits members or supporters for such an organisation incurs a penalty of imprisonment for a term not exceeding three years or a fine. (2) An organisation is a structured association of more than two persons, established to exist for a longer period of time, regardless of whether it has formally defined roles for its members, continuous membership or a developed structure and whose purpose is the pursuit of an overriding common interest. (3) Subsection (1) does not apply 1. if the organisation is a political party which the Federal Constitutional Court has not declared to be unconstitutional, 2. if the commission of offences is only one objective or an activity merely of subordinate importance or 3. to the extent that the objectives or activities of the organisation relate to criminal offences under sections 84 to 87. (...)”
Qualification	Forming criminal organisations
Notes on interpretation and application	This covers statements which contribute to the formation of a criminal organisation or encourage the formation of a criminal organisation or testify to the membership of a criminal organisation. An association is a long-term, voluntary, organizational union of at least three persons who pursue common goals when subordinated to the will of the entirety and feel themselves to be a unified association. The purpose of the association must be to commit offences/crimes. “1. membership of an association presupposes a certain degree of consensual integration of the offender into the organisation. It is still necessary for the offender to engage in an activity related to the organisation in order to further the criminal aims of the association, by which he promotes them from within and not only from without.”

Table E.17: Code 2-G-1-9

Field	Content
Legal Source	§ 129a German Criminal Code “(1) Whoever forms an organisation (section 129 (2)) whose objectives or activities are directed at the commission of
<i>Continued on next page</i>	

Field	Content
	1. murder under specific aggravating circumstances (section 211) or murder (section 212) or genocide (section 6 of the Code of Crimes against International Law) or a crime against humanity (section 7 of the Code of Crimes against International Law) or a war crime (section 8, 9, 10, 11 or 12 of the Code of Crimes against International Law) or 2. offences against personal liberty under section 239a or section 239b, (...) (2) Whoever forms an organisation whose objectives or activities are directed at 1. causing serious physical or mental harm to another person, especially of the kind designated in section 226, 2. committing offences under section 303b, section 305, section 305a or offences constituting a public danger under sections 306 to 306c or section 307 (1) to (3), section 308 (1) to (4), section 309 (1) to (5), section 313, 314 or section 315 (1), (3) or (4), section 316b (1) or (3) or section 316c (1) to (3) or section 317 (1), 3. committing offences against the environment under section 330a (1) to (3), 4. committing offences under section 19 (1) to (3), section 20 (1) or (2), section 20a (1) to (3), section 19 (2) no. 2 or (3) no. 2, section 20 (1) or (2), or section 20a (1) to (3), in each case also in conjunction with section 21, or under section 22a (1) to (3) of the War Weapons Control Act or 5. committing offences under section 51 (1) to (3) of the Weapons Act, or whoever participates as a member of such an organisation if one of the offences referred to in nos. 1 to 5 is intended to seriously intimidate the population, to unlawfully coerce an authority or an international organisation by force or threat of force, or to destroy or significantly impair the fundamental political, constitutional, economic or social structures of a state or of an international organisation and which, given the nature or consequences of such offences, may seriously damage a state or an international organisation incurs the same penalty. (...)"
Qualification	Forming terrorist organisations
Notes on interpretation and application	"The subject of the arrest warrant is the accusation that the defendant, who was 18 years old at the time of the offence, had in nine cases in June and July 2016 supported the Association 'Islamic State' (hereinafter: IS), whose purposes and activities were directed towards committing murder (section 211 of the German Criminal Code) or manslaughter (section 212 of the German Criminal Code), by sending him from Germany via the messenger service 'Telegram' to be informed by IS member M.s Messenger Service 'Telegram', either by translating texts relating to the IS into German or by checking the accuracy of existing translations into German, knowing that his communication partner belonged to the IS. Knowing the aims and actions of this organisation, he had intended to promote its propaganda through his activities (punishable under Section 129b (1) sentences 1 and 2 of the German Criminal Code in conjunction with Section 129a (5) sentence 1, Section 53 of the German Criminal Code and Sections 1, 105 of the German Judicial Reform Act)."

Table E.18: Code 2-G-1-10

Field	Content
Legal Source	§ 129b German Criminal Code “(1) Sections 129 and 129a apply to foreign organisations. If the offence relates to an organisation outside the Member States of the European Union, this only applies if the offence was committed by way of an activity carried out within the territorial scope of this statute or if the offender or the victim is a German national or is in Germany. (...)”
Qualification	Foreign criminal and terrorist organisations
Notes on interpretation and application	Compare: 2-G-1-8 and 2-G-1-9: with reference to criminal or terrorist organizations in other countries.

Table E.19: Code 2-G-1-11

Field	Content
Legal Source	§ 130 German Criminal Code “(1) Whoever, in a manner which is suitable for causing a disturbance of the public peace, 1. incites hatred against a national, racial, religious group or a group defined by their ethnic origin, against sections of the population or individuals on account of their belonging to one of the aforementioned groups or sections of the population, or calls for violent or arbitrary measures against them or 2. violates the human dignity of others by insulting, maliciously maligning or defaming one of the aforementioned groups, sections of the population or individuals on account of their belonging to one of the aforementioned groups or sections of the population incurs a penalty of imprisonment for a term of between three months and five years. (2) Whoever 1. disseminates material (section 11 (3)) or makes it available to the public, or offers, supplies or makes available to a person under 18 years of age material (section 11 (3)) which a) incites hatred against one of the groups referred to in subsection (1) no. 1, sections of the population or individuals on account of their belonging to one of the groups referred to in subsection (1) no. 1, or sections of the population, b) calls for violent or arbitrary measures against one of the persons or bodies of persons referred to in letter (a) or c) attacks the human dignity of one of the persons or bodies of persons referred to in letter (a) by insulting, maliciously maligning or defaming them,

Continued on next page

Field	Content
	2. makes content referred to in no. 1 (a) to (c) available to a person under 18 years of age or to the public through broadcasting or tele media services or 3. produces, purchases, supplies, stocks, offers, advertises or undertakes to import or export material (section 11 (3)) of such content referred to in no. 1 (a) to (c) in order to use it or parts obtained from it within the meaning of no. 1 or 2 or to facilitate such use by another incurs a penalty of imprisonment for a term not exceeding three years or a fine. (3) Whoever publicly or in a meeting approves of, denies or downplays an act committed under the rule of National Socialism of the kind indicated in section 6 (1) of the Code of Crimes against International Law in a manner which is suitable for causing a disturbance of the public peace incurs a penalty of imprisonment for a term not exceeding five years or a fine. (4) Whoever publicly or in a meeting disturbs the public peace in a manner which violates the dignity of the victims by approving of, glorifying or justifying National Socialist tyranny and arbitrary rule (...)”
Qualification	Incitement of masses (Volksverhetzung)
Notes on interpretation and application	A statement that is inflammatory. The incitement must be directed against parts of the population with references to national, racial, religious or ethnic groups. Incitement to hatred is a form of incitement to a hostile act which is aimed at the feelings of the addressees and goes beyond mere expressions of rejection and contempt. An appeal to violence - or any other arbitrary measure - presupposes an appeal character of the statement. Defamation is the making or spreading of knowingly false factual claims which lower the reputation of the population. Insulting is a particularly disparaging manifestation of disregard according to content or form. Malicious disparagement is the representation of others than despicable, inferior or unworthy for reprehensible reasons. “The Higher Regional Court of Hamm has decided that a professional soldier can be convicted of incitement of masses, if he insults foreigners and refugees in publicly accessible comments on Facebook as ‘Gesochse’, ‘Monkeys’, ‘Vermin’ and criminal ‘Pack’.”

Table E.20: Code 2-G-1-12

Field	Content
Legal Source	§ 131 German Criminal Code “(1) Whoever 1. takes material (section 11 (3)) which describes cruel or otherwise inhuman acts of violence against humans or humanoid beings in a manner which glorifies or downplays such acts of violence or which represents the cruel or inhuman aspects of the event in a manner which violates human dignity and

Continued on next page

Field	Content
	a) disseminates or makes it available to the public, b) offers, supplies or makes it available to a person under 18 years of age or 2. makes content referred to in no. 1 available through broadcasting or tele media services to a) a person under 18 years of age, b) the public or 3. produces, obtains, supplies, stocks, offers, advertises or undertakes to import or export material (section 11 (3)) of such content referred to in no. 1 in order to use it or parts obtained from it within the meaning of no. 1 (a) or (b), or no. 2, or to facilitate such use by another (...) (2) Subsection (1) does not apply if the act serves the reporting about current or historical events. (3) Subsection (1) sentence 1 no. 1 (b) and no. 2 (a) does not apply if it is the person having the duty of care and custody of another person who acts; this does not apply if the person having the duty of care and custody grossly breaches the duty of care and upbringing by offering, supplying or making available such material." Qualification Depictions of violence
Notes on interpretation and application	"1) to disseminate or publicly use a mark within the meaning of § 20 (1) sentence 1 no. 5 of the German Associations Act by sharing it in the publicly viewable chronicle of a Facebook account. 2) A dissemination or use covered by § 20 Paragraph 1 Sentence 1 No. 5 VereinsG does not exist if it is clearly evident from the content of the statement associated with the use of the trademark that the use of the trademark does not run counter to the protective purpose of § 20 Paragraph 1 Sentence 1 No. 5 VereinsG. 3) In the event of ambiguity in the content of the statement associated with the use of the prohibited sign, other possible interpretations with convincing reasons must not be excluded in advance in comparison with the interpretation leading to a conviction pursuant to § 20 (1) sentence 1 no. 5 Vereinsgesetz, unlike in the case of a criminal conviction for a mere statement. 4) to disseminate or make available to the public within the meaning of § 131 (1) sentence 1 no. 1 letter a StGB by sharing a video recording in the publicly viewable chronicle of a Facebook account. 5) The mere fact that a video recording shows cruel or inhuman acts of violence does not allow the conclusion to be drawn that the cruel or inhuman nature of the event is thereby depicted in a manner that violates human dignity within the meaning of section 131(1)(1)(3) of the StGB. 6) In the third variant of § 131, paragraph 1, no. 1 of the German Criminal Code, the presentation of violence as an offence presupposes that it is intended to create or reinforce an attitude in the viewer that denies the fundamental claim to value and respect that every human being has."

Table E.21: Code 2-G-1-13

Field	Content
Legal Source	§ 140 German Criminal Code “Whoever 1. rewards or 2. approves of publicly, in a meeting or by disseminating material (section 11 (3)) in a manner which is suitable for causing a disturbance of the public peace one of the unlawful acts referred to in section 138 (1) nos. 2 to 4 and no. 5 last alternative and in section 126 (1) or an unlawful act under section 176 (3), sections 176a and 176b, under section 177 (4) to (8) or section 178 after it has been committed or culpably attempted incurs a penalty of imprisonment for a term not exceeding three years or a fine.”
Qualification	Rewarding and approval of offences
Notes on interpretation and application	Reward means that the perpetrator rewards one of the above-mentioned illegal activities after it has been committed, i.e., directly or indirectly, after the fact, to a participant (or an association which is responsible for the act or has declared itself responsible for the act) by giving a previously unpromised advantage. The advantage does not have to be of a material nature, it can also lie in a demonstrative award. Approval means that the perpetrator approves the act publicly or in a meeting or by distributing writings, i.e., by means of a statement recognizably related to the concrete act. “(1) In order to assess whether a statement is to be understood as condoning offences in accordance with the restrictive meaning of the constituent element of the offence, its objective meaning must first be determined, taking into account the circumstances of the individual case, from the point of view of an average public which is impartial and reasonably well-informed. (2) In the light of the freedom of opinion protected by Article 5(1) sentence 1 of the Basic Law, no meaning may be attached to it which it objectively does not have, and in the event of ambiguity, the interpretation leading to a conviction may be assumed only if other possibilities of interpretation without penalty can be excluded on convincing grounds (para. 16). (3) The requirements of Article 140 of the Criminal Code are therefore not met if a statement (here: on arson attacks on asylum seekers' accommodation) made by a defendant is not exclusively descriptive in the sense of an explanatory model for the way in which citizens react in the event of an alleged lack of political participation (para. 22).”

Table E.22: Code 2-G-1-14

Field	Content
Legal Source	§ 166 German Criminal Code “(1) Whoever publicly or by disseminating material (section 11 (3)) reviles the religion or ideology of others in a manner which is suitable for causing a disturbance of the public peace incurs a penalty of imprisonment for a term not exceeding three years or a fine. (2) Whoever publicly or by disseminating material (section 11 (3)) reviles a church or other religious or ideological community in Germany or its institutions or customs in a manner which is suitable for causing a disturbance of the public peace incurs the same penalty.”
Qualification	Revilement of religious faiths and religious and ideological communities
Notes on interpretation and application	“When assessing the legality of the terms of use and Community standards applied by the operator of a social media platform (here: Facebook), a balance must be struck between the user's freedom of opinion on the one hand and the fundamental rights of Articles 2, 12 and 14 of the Basic Law in dispute for the operator of a social media platform on the other. The general freedom of action and freedom to exercise an occupation and the right to set up and run a business must be brought into balance with Article 5 of the Basic Law by way of practical concordance. In this respect, the fundamental rights do not have a direct effect, but only a radiating effect. (2) It would be incompatible with the need to balance the conflicting fundamental rights positions in accordance with the principle of practical concordance if the operator of a social media platform were to be allowed, on the basis of ‘virtual domiciliary rights’ on the platform provided by him, to delete a user's contribution which the operator considers to be in breach of its guidelines, even if the contribution does not exceed the limits of permissible expression of opinion.”

Table E.23: Code 2-G-1-15

Field	Content
Legal Source	§ 184b German Criminal Code “(1) Whoever 1. disseminates child pornographic material or makes it available to the general public, whereby pornographic material (section 11 (3)) is deemed to be child pornography if it relates to a) sexual acts performed by, on or in the presence of a person under 14 years of age (child), b) the reproduction of a child in a state of full or partial undress in an unnaturally sexual pose or c) the sexually provocative reproduction of a child's bare genitalia or bare buttocks,

Continued on next page

Field	Content
	2. undertakes to procure, for another, child pornographic material which reproduces actual or realistic acts, 3. produces child pornographic material which reproduces actual acts or 4. produces, obtains, supplies, stocks, offers, advertises or undertakes to import or export such child pornographic material in order to use it or parts thereof within the meaning of no. 1 or 2 or of section 184d (1) sentence 1, or to facilitate such use by another, unless the offence is subject to a penalty under no. 3, (...)” § 184d German Criminal Code “(1) Whoever makes pornographic content available to another person or to the public through broadcasting or tele media services incurs a penalty under sections 184 to 184c. In the cases under section 184 (1), sentence 1 does not apply to dissemination via tele media if technical or other measures are in place to ensure that the pornographic content is not accessible to persons under 18 years of age. Section 184b (5) and (6) applies accordingly. (2) Whoever undertakes to access child pornography via tele media also incurs a penalty under section 184b (3). Whoever undertakes to access youth pornography via tele media also incurs a penalty under section 184c (3); section 184c (4) applies accordingly. Section 184b (5) and (6) sentence 1 applies accordingly.”
Qualification	Dissemination, procurement and possession of child pornography, making pornographic content available through broadcasting or tele media services; accessing child or youth pornographic content via tele media
Notes on interpretation and application	An image is child pornographic if it involves sexual acts by and on persons under the age of 14 (child), shows a partially or completely unclothed child in an unnaturally sexually explicit position, or shows the unclothed genitals or buttocks of a child in a sexually suggestive manner. “Making child pornographic material publicly accessible on the Internet in accordance with § 184b, paragraph 1, no. 2 of the German Criminal Code (StGB) is to be affirmed if the addressee is given the opportunity to access the material, which is given by making the files available. However, the fact that the addressee has actually been granted (read) access does not require the facts to be fulfilled (continuation BGH, 27 June 2001, 1 StR 66/01, BGHSt 47, 55).”

Table E.24: Code 2-G-1-16

Field	Content
Legal Source	§ 185 German Criminal Code “The penalty for insult is imprisonment for a term not exceeding one year or a fine and, if the insult is committed by means of an assault, imprisonment for a term not exceeding two years or a fine.”
Qualification	Insult

Continued on next page

Field	Content
Notes on interpretation and application	The honour of a person is protected. It mainly includes value judgments, i.e., statements of disregard or disrespect. These can concern the ethical value of another person or the social value. Assertions of facts are only factual if they are made to the person concerned (as opposed to § 186 StGB). In principle, the content of the statement must be determined, taking into account all the accompanying circumstances. Satire and caricatures are not insulting, nor are general discourtesies, lack of distance. Examples: TV announcer as 'milked goat', police officers as 'shit cops', comparison of a soldier with 'murderer, torturer, concentration camp supervisor and executioner', 'killer squad' to GSG9, physically handicapped person as 'cripple'. Legal Justification: § 193 Safeguarding legitimate interests The first prerequisite for justification is the exercise of a legitimate interest, i.e., the pursuit of a public or private, non-material or material purpose recognised by law as worthy of protection. The pursuit of the interest must also be proportionate: The pursuit of a legitimate purpose does not in itself justify the violation of honour; rather, it must also be the appropriate means of doing so, taking into account all the circumstances: Only the legitimate pursuit of legally recognised interests is not unlawful. "(1) In the case of a very similar infringement of the obligation to cease and desist under a settlement, there is normally no legal interest in bringing an action for an injunction. (2) If a third party's mail is shared in a social network and marked with a consenting note ('important and correct action'), the disseminator makes the statements contained therein his own. (3) Legal persons governed by private law may not rely on the principles developed in the case law on defamatory criticism."

Table E.25: Code 2-G-1-17

Field	Content
Legal Source	§ 186 German Criminal Code "Whoever asserts or disseminates a fact about another person which is suitable for degrading that person or negatively affecting public opinion about that person, unless this fact can be proved to be true, incurs a penalty of imprisonment for a term not exceeding one year or a fine and, if the offence was committed publicly or by disseminating material (section 11 (3)), a penalty of imprisonment for a term not exceeding two years or a fine." § 188 Towards a person of political life § 193 Safeguarding legitimate interests
Qualification	Malicious gossip (Üble Nachrede)

Continued on next page

Field	Content
Notes on interpretation and application	Attack on the honour of a person through statements of facts against third parties (different from § 185 StGB). The "fact", i.e., the content of the assertion of truth, must be objectively suitable, if it comes to the knowledge of third parties, to disparage the person concerned or degrade him in public opinion. Facts are external events, states and conditions that can be the object of sensory perception, but also internal facts (character traits, motives, goals) as soon as they relate to external phenomena. Facts are everything that can be true or false, i.e., as a statement of truth, the statement of facts must in principle be accessible to verification. Facts must not be proven "true".

Table E.26: Code 2-G-1-18

Field	Content
Legal Source	§ 187 German Criminal Code "Whoever, despite knowing better, asserts or disseminates an untrue fact about another person which is suitable for degrading that person or negatively affecting public opinion about that person or endangering said person's creditworthiness incurs a penalty of imprisonment for a term not exceeding two years or a fine, and, if the act was committed publicly, in a meeting or by disseminating material (section 11 (3)), a penalty of imprisonment for a term not exceeding five years or a fine."
	§ 188 Towards a person of political life § 193 Safeguarding legitimate interests
Qualification	Defamation
Notes on interpretation and application	Perpetrator must knowingly disseminate a proven untrue statement of facts. The untrue fact must be capable of degrading or disparaging a third person. Or the untrue fact must be capable of endangering the credit of the person concerned. Credit is the confidence that someone enjoys in the fulfilment of their pecuniary obligations.

Table E.27: Code 2-G-1-19

Field	Content
Legal Source	§ 201a German Criminal Code "(1) Whoever

Continued on next page

Field	Content
	1. without being authorised to do so creates or transmits photographs or other images of another person in private premises or in a room which is specially protected from view, and thereby violates the intimate privacy of the person depicted, 2. without being authorised to do so produces a photograph or other image exhibiting the helplessness of another person or transmits such image, and thereby violates the intimate privacy of the person depicted, 3. uses a photograph or other image produced by an offence under no. 1 or no. 2 or makes it available to a third party or 4. makes available to a third party, in the awareness of that third party lacking authorisation to do so, a photograph or other image of the type set out in no. 1 or no. 2 which has been produced with authorisation, and thereby violates the intimate privacy of the person depicted, incurs a penalty of imprisonment for a term not exceeding two years or a fine. (2) Whoever, without being authorised to do so, makes available to a third party a photograph or other image of another person which is of such a nature as to significantly damage the reputation of the person depicted incurs the same penalty. (3) Whoever 1. produces or offers to procure for a third party for a consideration or 2. procures for themselves or for a third party for a consideration an image showing the nakedness of another person under 18 years (...)"
Qualification	Violation of intimate privacy by taking photographs or other images
Notes on interpretation and application	Only unauthorized use or access is considered as a criminal act. Use is to play in front of yourself or a third party. Making it accessible is to enable third parties to access the recordings. "Father is not allowed to put pictures of his daughter on the net. A father has been sentenced to a fine for posting photos of his little daughter on the Internet without the permission of the custodial grandmother. The district court of Hanover sentenced him to a fine of 40 daily rates of 40 euros each, as a spokesman for the court said on 03.02.2020 (Ref.: 244 Ds 228/19). Custodial grandmother sued. The grandmother had allowed the 23-year-old, as the custodian of the then nine month old girl, to take photos of herself and the child. However, he published the pictures on Facebook between January and April 2019, despite their prohibition, and thus gave his Facebook friends an insight into the protected space in the grandmother's flat. The grandmother had filed a complaint against the publication of the pictures in social media. The judgment is final."

Table E.28: Code 2-G-1-20

Field	Content
Legal Source	§ 241 German Criminal Code “(1) Whoever threatens a person with the commission of a serious criminal offence against that person or a person close to him or her incurs a penalty of imprisonment for a term not exceeding one year or a fine. (2) Whoever, despite knowing better, pretends to another person that the commission of a serious criminal offence against that person or a person close to him or her is imminent incurs the same penalty.”
Qualification	Threatening commission of serious criminal offence
Notes on interpretation and application	A statement that involves the prospect of committing a crime, with a certain future behaviour that constitutes a crime in the sense of § 12 StGB (i.e., minimum prison sentence of one year). A threat is also considered “an illegal act against the sexual self-determination, the physical integrity, the personal liberty or against a thing of substantial value is to fall. At the same time, the maximum penalty for the threat of a crime will be raised from one to two years’ imprisonment.”

Table E.29: Code 2-G-1-21

Field	Content
Legal Source	§ 269 German Criminal Code “(1) Whoever, for the purposes of deception in legal commerce, stores or modifies data which are of probative value in such a way that a counterfeit or falsified document would be created upon their retrieval, or whoever uses data stored or modified in such a manner incurs a penalty of imprisonment for a term not exceeding five years or a fine. (2) The attempt is punishable. (3) Section 267 (3) and (4) applies accordingly.”
Qualification	Forgery of data of probative value
Notes on interpretation and application	Only the use of incorrect or fake evidence-relevant data can be considered as a criminal act. "According to the findings made in this regard, the defendant, after a dispute with the aggrieved party, praised K. with whom he had previously entered into a relationship, the defendant offered to pay 200 Euros for the killing of the victims on his 'Facebook' page, which can be viewed by all users of this internet platform. He also informed them of her whereabouts and attached several photographs of the victims. The entry did not result in any criminal offences against the injured party. (...) The findings made show, however, that the accused is guilty of falsifying evidence-relevant data (§ 269 (1) StGB) in the variant of the offence of using altered data."

Table E.30: Code 2-G-1-22

Field	Content
Legal Source	§ 14 BtMG labelling and advertising “(…) 5. No advertising shall be made for narcotic drugs referred to in Annex I. The advertising of the stunning substances listed in Annexes II and III shall be restricted to professionals in industry and commerce and to persons and associations of persons operating a pharmacy or a veterinary dispensary, and to doctors, dentists and veterinarians in the case of the stunning substances listed in Annex III.” (§ 29 BtMG Criminal Offences / § 30 BtMG Offences)
Qualification	Labelling and advertising / Criminal Offences
Notes on interpretation and application	Statements related to advertising, offering drugs or commercial trading with drugs. “The advertising ban does not contain any restriction regarding the advertising channel. It therefore covers every medium and thus all advertising channels. Advertising can be done orally, in writing, by telephone, telegram, fax, e-mail, radio, television or internet, by distributing leaflets, brochures, posters, brochures, CDs or DVDs, newspaper or magazine advertisements (KPV/Patzak BtMG § 29 part 18. marginal no. 8). With regard to the restriction of advertising outside professional circles, advertising cannot be carried out on advertising channels that are generally accessible. Legal advertising on the internet for Schedule II and III narcotics is therefore excluded, unless it is a closed user group. In that case, however, it must be ensured that access to it is only possible after verification of the user.”

Table E.31: Code 2-G-1-23

Field	Content
Legal Source	§ 35 WaffG: Advertising, obligations to inform, trade prohibitions “(1) Anyone who offers weapons or ammunition for sale or exchange in advertisements or promotional literature shall draw attention to the requirement of eligibility to purchase the following types of weapons as follows: 1. In the case of firearms requiring a permit and ammunition requiring a permit: Only to holders of an acquisition permit, 2. in the case of firearms and ammunition and other weapons not subject to authorisation: Only to persons of 18 years of age or older, 3. for prohibited weapons: ...to be sold only to holders of a special permit,

Continued on next page

Field	Content
	and to disclose his name, his address and, where applicable, his registered trademark. Advertisements and promotional literature pursuant to sentence 1 may only be published if they contain the name and address of the offerer as well as the information to be provided by him depending on the type of wafer (...)."
Qualification	Advertising, obligations to inform, trade prohibitions Weapons Act / Criminal Offences
Notes on interpretation and application	The provision of § 35 contains both obligations of information for those who offer weapons or ammunition for sale or exchange in advertisements and for those who give weapons subject to authorisation to another in retail trade. Finally, paragraph 3 contains a basic ban on certain forms of distribution or transfer.

Table E.32: Code 2-G-1-24

Field	Content
Legal Source	§ 3, 3a and 11 Heilmittelwerbe-gesetz / Drug Advertising Act (HWG) § 3 HWG "Misleading advertising is not permitted. Misleading advertising exists in particular in this case, 1. if medicinal products, medical devices, processes, treatments, objects or other means are attributed a therapeutic effect or effects which they do not have 2. if the impression is wrongly given that (a) success can be expected with certainty, (b) no harmful effects occur during normal or prolonged use, (c) the advertising is not organised for the purposes of the contest, 3. if untrue information or information suitable for deception (a) on the composition or nature of medicinal products, medical devices, articles or other means, or on the nature of processes or treatments; or (b) are made about the person, education, qualification or success of the manufacturer, inventor or the persons working or having worked for them."
	§ 3a HWG "Advertising for medicinal products which are subject to the obligation to obtain marketing authorisation and which are not authorised or deemed to be authorised under the provisions of the pharmaceutical legislation is prohibited. Sentence 1 shall also apply if the advertising relates to indications or pharmaceutical forms which are not covered by the approval." § 11 HWG

Continued on next page

Field	Content
	“(1) Outside the specialist circles, no advertising may be made for medicinal products, processes, treatments, objects or other means (...) 2. with information or presentations which refer to a recommendation by scientists, health professionals, animal health professionals or other persons who, on the basis of their knowledge, may encourage the use of medicinal products, 3. with the reproduction of patient histories as well as with references to them, if this is done in an abusive, repulsive or misleading manner or if a detailed description or presentation can lead to an incorrect self-diagnosis, 5. with a pictorial representation which makes improper, repulsive or misleading use of changes in the human body as a result of disease or injury or of the action of a medicinal product on the human body or parts of the human body, (...) 7. with advertising statements which suggest that health could be impaired or improved by not using the medicinal product, 8. by means of advertising presentations, which are associated with the offering for sale or acceptance of addresses, 9. with publications whose advertising purpose is misleading or not clearly recognizable, (...) 11. with statements of third parties, in particular with letters of thanks, acknowledgement or recommendation, or with references to such statements, if they are made in an abusive, repulsive or misleading manner, 12. with advertising measures that are exclusively or predominantly aimed at children under 14 years of age, 13. with competitions, draws or other procedures whose outcome depends on chance, if these measures or procedures encourage the improper or excessive use of medicines, 14. by the supply of medicinal products, their samples or specimens or by vouchers for them, 15. by the unsolicited delivery of samples or specimens of other means or objects or by vouchers for them. Sentence 1 No. 7 to 9, 11 and 12 shall apply accordingly to medical devices. Furthermore, the following advertising is not permitted for the operative plastic-surgical procedures mentioned in § 1 paragraph 1 number 2: 1.with the effect of such a treatment by comparative presentation of the body condition or appearance before and after the operation or 2.with advertising measures that are exclusively or predominantly aimed at children and young people. (2) Outside the specialist circles, medicinal products for human use may not be advertised with claims which suggest that the effect of the medicinal product is equivalent or superior to another medicinal product or another treatment.”
Qualification	Drug Advertising Act /Criminal Offences
Notes on interpretation and application	Statements concerning misleading advertising or unlawful advertising with and for medicinal products or medical devices. Advertising is inadmissible if it is made to non-specialist circles or if the impression is made that success will certainly occur, side effects are concealed, or it is concealed that it is advertising etc.
<i>Continued on next page</i>	

Field	Content
	“a mail-order pharmacy established in another Member State is prohibited from advertising a prize draw to attract customers if participation in the prize draw is linked to the submission of a prescription for a medicinal product for human use subject to medical prescription, if the prize offered is not a medicinal product but another article (in this case, an electric bicycle worth EUR 2 500 and electric toothbrushes), and if there is no risk of encouraging the inappropriate or excessive use of medicinal products.”

Table E.33: Code 2-A-1-1

Field	Content
Legal Source	§ 246 Austrian Criminal Code “(1) Whoever establishes a union whose purpose, although not exclusive, is to unlawfully undermine the independence, the form of government laid down in the constitution or a constitutional institution of the Republic of Austria or one of its federal provinces shall be punished by imprisonment from six months to five years. (2) Also to be punished is anyone who acts as a leader in such an association, recruits members for it or supports it with funds or otherwise in a significant way. (3) Anyone who otherwise takes part in such an association or supports it in a manner other than that specified in para. 2 shall be punished by imprisonment for up to one year or a fine of up to 720 daily rates. and § 247a. (1) Whoever founds or leads an anti-state movement shall be punished with imprisonment for up to two years if he or another participant has carried out or contributed to a serious act in which the anti-state orientation is clearly manifest. (2) Whoever takes part in such a movement with the intention of encouraging the commission of anti-state acts or supports it with substantial funds or in any other substantial way, shall be punished under the condition of para. 1 with imprisonment of up to one year or with a fine of up to 720 daily rates. (3) An anti-state movement is a group of many people whose aim is to reject outright the sovereign rights of the Republic of Austria (federal, provincial, municipal or other self-government) or to continue to presume to exercise such or alleged sovereign rights, and whose purpose is to continue to prevent the execution of laws, regulations or other sovereign decisions of the authorities or to enforce the presumed or alleged sovereign rights in a way that clearly manifests the anti-state orientation. (4) The perpetrator shall not be punished in accordance with paras. 1 and 2 if the act is punishable by a more severe penalty under another provision.

Continued on next page

Field	Content
	(5) According to paras. 1 and 2, no punishment shall be imposed on anyone who voluntarily withdraws from the movement, before the authority has learned of his or her fault, in such a way that it is clearly evident that the anti-state orientation is no longer supported."
Qualification	Anti-state activities
Notes on interpretation and application	"The concept of connection corresponds to that of § 279 (armed connection). This is to be understood as a union of a larger number of people. This presupposes, first of all, an agreement of will to form a community. However, a union is only present if a minimum of organisational order is given, even if this is not recorded in writing." "For a connection, a larger number of people must join together. According to the prevailing opinion, a larger number is 10 people. However, this is not an absolute limit, but a guideline. Depending on the degree of organisation and danger, a connection in the sense of § 246 can exist even if the number of persons is just below this guideline value. The greater the danger to the state emanating from the connection, the more likely it is that the facts will be fulfilled even if the number is below 10."

Table E.34: Code 2-A-1-2

Field	Content
Legal Source	§ 248 Austrian Criminal Code "(1) Anyone who hatefully insults or disparages the Republic of Austria or one of its federal provinces in such a way that the act becomes known to a broad public shall be punished with imprisonment of up to one year or a fine of up to 720 daily rates. (2) Anyone who, in the manner described in paragraph 1, in a hateful manner insults, disparages or otherwise degrades a flag of the Republic of Austria or of one of its federal provinces shown on a public occasion or at an event open to the public, a national emblem affixed by an Austrian authority, the federal anthem or a provincial anthem, shall be punished by imprisonment for up to six months or a fine of up to 360 daily rates."
Qualification	Disparagement of the state and its symbols
Notes on interpretation and application	"Insulting means the crude manifestation of disregard. How this manifestation takes place is irrelevant: it can be verbal, through insulting signs, gestures or other actions. Verbal insulting is present, for example, through the exclamation "shit..." or other insulting words. It depends not only on the wording, but also on the way and the circumstances in which the word is used. Insulting by gestures is present, for example, when pointing the outstretched middle finger in the direction of the hoisted flag or when playing the hymn. Spitting out in front of the flag can also fulfil the crime scene."

Continued on next page

Field	Content
Table E.35: Code 2-A-1-3	
Field	Content
Legal Source	§ 242 Austrian Criminal Code “(1) Whoever undertakes to amend the constitution of the Republic of Austria or of one of its federal provinces by force or by threat of force, or to partition off an area belonging to the Republic of Austria, shall be punished with imprisonment from ten to twenty years. (2) An undertaking within the meaning of para. 1 is also already present at an attempt.”
Qualification	High Treason
Notes on interpretation and application	“Violence (margin no. 16 f) or threat of violence (margin no. 18) can be considered an offence. These means of action are aimed at amending the constitution of the Federation (margin no. 21 ff) or of the provinces (margin no. 29 ff) or at partitioning off an area belonging to the Republic of Austria (margin no. 31 ff). The pursuit of the aforementioned objectives by other (unlawful) means is not punishable under § 242. However, a punishability on account of § 246 (connections hostile to the state) is possible. Pursuing these goals by legal means is not punishable at all: For example, holding a strike with highly treasonous aims is not criminal if it is within the scope of what is permissible under labour law.”
Table E.36: Code 2-A-1-4	
Field	Content
Legal Source	§ 244 Austrian Criminal Code “(1) Anyone who conspires with another to commit treason is punishable with imprisonment from one to ten years. (2) Similarly, anyone who prepares a high treason in any other way and thereby causes or significantly increases the danger of a high treasonous enterprise or who prepares a high treason in cooperation with a foreign power is to be punished.”
Qualification	Preparation of a high treason
<i>Continued on next page</i>	

Field	Content
Notes on interpretation and application	“The mere procurement of writing materials for the production of highly treasonous appeals will usually neither cause nor significantly increase the danger of treason. Neither will the distribution of leaflets to propagate highly treasonous aims. Cases must be judged differently in which the perpetrator is already forming teams, procuring money, weapons and the like, trying to get security organs or members of the armed forces on his side or trying to win over persons to take over important positions after taking power. However, it depends on the assessment of the circumstances of the individual case, especially the time horizon.”

Table E.37: Code 2-A-1-5

Field	Content
Legal Source	§ 258 Austrian Criminal Code “(1) Who 1. has a legal relationship between the Republic of Austria or one of its federal provinces and a foreign power or a supranational or international institution or 2. has knowledge of a fact which is important for the relations between the Republic of Austria or one of its federal provinces and a foreign power or a supranational or international institution, produces false evidence or falsifies, destroys, damages or removes genuine evidence and thereby endangers the interests of the Republic of Austria or one of its federal provinces shall be punished by imprisonment from six months to five years. (2) Similarly, anyone who makes use of such false or falsified evidence and thereby endangers the interests of the Republic of Austria or one of its federal provinces shall be punished.”
Qualification	Traitorous counterfeiting and destruction of evidence
Notes on interpretation and application	“The means of evidence concerning the legal relationship (§ 258, Subsection 1, Line 1) include cases where the forgery of a document gives the impression that Austria has concluded this or that agreement with a foreign country. Evidence relating to a fact which is significant for Austria's foreign relations (§ 258.1(2)) exists if the photograph attached to a news item is falsified to the extent that it proves the untruthful presence of a certain public figure at a certain announcement. This could in fact be used to the detriment of Austria to document incorrectly that an Austrian official had participated in a public announcement hostile to a third state.”

Table E.38: Code 2-A-1-6

Field	Content
Legal Source	§ 282 Austrian Criminal Code “(1) Anyone who, in a printed work, on the radio or otherwise in such a way that it becomes accessible to the general public, incites a person to commit an act which is punishable by a penalty shall, if he is not punishable by a more severe penalty as a party to that act (§ 12), be punished by imprisonment for up to two years. (2) Similarly, anyone who, in the manner described in para. 1, approves of an act committed intentionally and punishable by a custodial sentence of more than one year in a manner that is likely to outrage the general sense of justice or to incite people to commit such an act shall be punished.”
Qualification	Incitement to commit punishable acts and approval of punishable acts
Notes on interpretation and application	“An incitement is to be understood as any utterance that is intended to arouse the decision to act in a certain way in others, namely, to commit an act threatened with punishment.” “In contrast to Section 281, which sanctions an incitement to disobey (any) law, a request under Section 282 (1) only relates to criminal laws. According to Section 282 (1), an incitement must involve the commission of an act that is threatened with a judicial penalty; it is irrelevant whether the offense is part of the Criminal Code or ancillary criminal law.” “The appeal “Loot the jewelry stores!” clearly concerns property crimes and fulfils § 282 (58). The same applies to the incitement to violent demonstrations and to destroy business premises, to damage cars (§§ 125 f) or to defend oneself violently against a police force, sections 269, 270 (59). On the other hand, the distribution of leaflets calling for a violent demonstration under the motto “Vienna glöst” or “a hot night” is not sufficient for criminal liability under section 282, Paragraph 1.60.”

Table E.39: Code 2-A-1-7

Field	Content
Legal Source	§ 282a Austrian Criminal Code “(1) Anyone who, in a printed work, on the radio or in any other medium or otherwise publicly in such a way that it becomes accessible to many people, incites the commission of a terrorist offence (§ 278c par. 1 nos. 1 to 9 or 10) shall be punished with imprisonment of up to two years, unless he or she is threatened with a more severe penalty as a participant in this act (§ 12). (2) Also to be punished is anyone who, in the manner described in para 1, condones a terrorist offence (§ 278c para 1 subparas 1 through 9 or 10) in a manner that is likely to result in the commission of one or more such offences.”

Continued on next page

Field	Content
Qualification	Incitement to commit terrorist offences and endorsement of terrorist offences
Notes on interpretation and application	“Mere expressions of sympathy for a terrorist organization do not meet the requirements of approving at least one specific terrorist offense within the meaning of Section 278c Paragraph 1 Z 1 to 9 or 10 StGB.”

Table E.40: Code 2-A-1-8

Field	Content
Legal Source	§ 283 Austrian Criminal Code “(1) Whoever publicly in such a way that it becomes accessible to many people 1. incites to violence or hatred against a church or religious society or any other group of persons defined according to the existing or absent criteria of race, colour, language, religion or belief, nationality, ancestry or national or ethnic origin, gender, disability, age or sexual orientation, or against a member of such a group expressly because of membership of that group, 2. insults one of the groups defined in Z or a person because of the belonging to such a group with the intention to violate the human dignity of the members of the group or the person in a way that is suitable to make the group or person despicable or disparaging in public opinion, or 3. approves, denies, grossly trivialises or justifies crimes in the sense of §§ 321 to 321f as well as § 321k, which have been legally established by a domestic or an international court, where the act is directed against one of the groups referred to in Z 1 or against a member of such a group expressly because of membership of such a group and is committed in a manner likely to incite violence or hatred against such a group or against a member of such a group, shall be punished by imprisonment for a term not exceeding two years. (2) Any person who commits the act referred to in subsection (1) in a printed work, on the radio or otherwise in a manner whereby the acts referred to in subsection (1) become accessible to a wide public shall be liable to a custodial sentence not exceeding three years. (3) Whoever, by an act under subsection (1) or (2), causes other persons to exercise violence against a group referred to in subsection (1)(1) or against a member of such a group because of his membership of that group, shall be punished with imprisonment for a term of six months to five years.

Continued on next page

Field	Content
	(4) Any person who, unless he is punishable by a more severe penalty as a participant (section 12) in an act under subsections (1) to (3), produces written material, pictures or other representations of ideas or theories that incite hatred or violence against a group referred to in subsection (1)(1) or against a member of such a group because of his membership of such a group, in a printed work, on the radio or otherwise in a manner that makes them accessible to a broad public, in an approving or justifying manner, or otherwise makes them publicly available, incurs a penalty of imprisonment for a term not exceeding a year or a fine of up to 720 daily rates."
Qualification	Incitement to violence
Notes on interpretation and application	"Incite means to want to arouse the decision to a certain behaviour in another directly through a certain behaviour, especially through a statement." (...) "The request or incitement to a hostile act only fulfils Section 283 (1) if it was done publicly. According to § 69, an act is committed publicly if it can be directly perceived by a larger group of people. A larger group of people is assumed from a guideline number of around 10 people." "Hate is a tendency to incite hatred and contempt (...)"

Table E.41: Code 2-A-1-9

Field	Content
Legal Source	§ 278 Austrian Criminal Code "(1) Anyone who founds a criminal organisation or participates in such an organisation as a member, is punishable by imprisonment for up to three years. (2) A criminal organisation is a long-term association of more than two persons, the purpose of which is that one or more members of the association commit one or more crimes, other substantial acts of violence against life and limb, not only minor damage to property, theft or fraud, offences under sections 165, 177b, 233 to 239, 241a to 241c, 241e, 241f, 283, 304 or 307, in section 278d subsection 1, other offences or misdemeanours referred to in Article 114(1) or 116 of the Aliens Police Act may be committed. (3) A member participates in a criminal organisation if he or she commits an offence within the framework of its criminal orientation or participates in its activities by providing information or assets or in any other way with the knowledge that he or she is thereby promoting the organisation or its criminal activities (...)"
Qualification	Criminal organization
Notes on interpretation and application	"Organizing" characterizes a process that consists of several expressions of will and is only present if these correspond to one another."

Continued on next page

Field	Content
	“An organisation within the meaning of § 278 only exists if there are corresponding expressions of will from at least three people who create a union of a community.” “The existence of a criminal organization does not depend on whether its members have already committed crimes within the meaning of Section 278 (2) StGB, but only on the orientation of the organization to carry them out.”

Table E.42: Code 2-A-1-10

Field	Content
Legal Source	§ 278a Austrian Criminal Code “Whoever wants to establish a company-like connection of a larger number of employees over a longer period of time or takes part in such a union as a member (Section 278 (3)), 1. which is aimed, although not exclusively, at the recurrent and planned commission of serious criminal offences threatening life, physical integrity, freedom or property, or serious criminal offences in the field of sexual exploitation of human beings, smuggling or illicit trafficking in explosive ordnance, nuclear material and radioactive substances, hazardous waste, counterfeit currency or narcotic drugs, which thereby seeks enrichment on a large scale and seeks to corrupt or intimidate others or to shield himself in a particular way from prosecution, (...)”
Qualification	Criminal organization
Notes on interpretation and application	“If one looks at the various offenses in the StGB in which the establishment of associations or organizations is described, the criminal organization within the meaning of Section 278a has the highest degree of organization.” “According to the Justice Committee, the term “company-like” should “summarize the elements of the division of labour, hierarchical structure and the existence of a certain infrastructure as essential characteristics of organized action” (JAB StRÄG 1996, 10); This feature also makes it clear that the establishment is a successful offense (margin no. 21).”

Table E.43: Code 2-A-1-11

Field	Content
Legal Source	§ 278b Austrian Criminal Code “(1) Whoever leads a terrorist group (paragraph 3) is punishable with imprisonment. (...)”

Continued on next page

Field	Content
	(2) Whoever participates in a terrorist organization as a member (§ 278 para. 3) shall be punished with imprisonment from one to ten years. (3) A terrorist group is a long-term association of more than two persons, the purpose of which is to carry out one or more terrorist offences (Section 278c) or terrorist financing (Section 278d) by one or more of its members.”
Qualification	Terrorist organization
Notes on interpretation and application	“The teaching subsumes as participation in a terrorist association: the establishment of contact with an advertiser and the organized association to form a travel group with the aim of participating in combat operations (116) and recruiting members; (117) the establishment of contact with a person who recruits members for ISIS, as well as the organized association to a travel group with the aim of participating in combat operations; (121) joining of the Asa’ib Ahl al-Haqq militia and the multiple undertaking of supply deliveries in the form of delivery of various goods in connection with battles; (122) the support of members of ISIS by young women through (planned) marriage; (123) the marriage of an ISIS fighter according to the Islamic rite in a ceremony transmitted via Skype and the promise to travel to Syria to support and strengthen him in the fight by starting a family, personal care and housekeeping; (124) the concrete promise of support to a person recruiting fighters in connection with the departure towards the fighting area of the terrorist organization; (125) the reinforcement of the intention of a separately persecuted person to participate as a member of ISIS and to fight for the establishment of a God based state on radical Islamist principles, as well as contacting a liaison officer of the terrorist organization to have their families resettled in homes of displaced Syrians and join the terrorist group as fighters; (126) the departure from Syria in an area controlled by ISIS, the education of under age children according to the propaganda of the terrorist organization, the participation in the development of a radical Islamist-oriented social infrastructure of this terrorist organization, the participation in an education in Islamic law including military training; (139)”

Table E.44: Code 2-A-1-12

Field	Content
Legal Source	§ 207a Austrian Criminal Code “(1) Whoever produces a pornographic depiction of a minor (para. 4) 1. or 2. offers, procures, hands over, demonstrates or otherwise makes available to another person is punishable by imprisonment for up to three years.

Continued on next page

Field	Content
	(2) Any person who produces, imports, transports or exports a pornographic depiction of a minor (par. 4) for the purpose of distribution or who commissions an act referred to in par. 1 for commercial gain shall be punished by imprisonment from six months to five years. Anyone who commits an offence as a member of a criminal organisation or in such a way that it causes a particularly serious disadvantage to the minor shall be punished with a term of imprisonment of between one and ten years; likewise, anyone who produces a pornographic depiction of a minor (paragraph 4) using serious violence or endangers the life of the depicted minor intentionally or through gross negligence (Article 6(3)) shall be punished. (3) Anyone who obtains or possesses a pornographic image of an adult minor (para. 4 nos. 3 and 4) shall be punishable by imprisonment for up to one year or a fine of up to 720 daily rates. Penalties of up to two years imprisonment shall be imposed on anyone who obtains or possesses a pornographic image of a minor (par. 4). (3a) Paragraph 3 also punishes anyone who knowingly accesses a pornographic image of a minor on the Internet. (4) Pornographic depictions of minors are 1. realistic images of a sexual act on a minor or a minor on himself, on another person or with an animal, 2. realistic depictions of an event with a minor, the viewing of which, according to the circumstances, gives the impression that this is a sexual act on the minor or the minor on himself, on another person or with an animal, 3. realistic illustrations a) of a sexual act within the meaning of no. 1 or an event within the meaning of no. 2, but with minors of legal age, or b) the genitals or pubic area of minors, as far as they are luridly distorted images, reduced to themselves and detached from other expressions of life, which serve the sexual arousal of the viewer, 4. pictorial representations, the viewing of which - as a result of alteration of an illustration or without the use of such an illustration - gives the impression according to the circumstances that it is an illustration according to lines 1 to 3. (...)”
Qualification	Pornographic depictions of minors
Notes on interpretation and application	“On the one hand, Section 184b of the German Criminal Code is narrower than Section 207a of the Austrian Criminal Code, because Section 184b of the German Criminal Code requires child pornography as the object of the crime, while Section 207a of the Austrian Criminal Code also includes pornographic depictions of adult minors. In contrast to Austria, where people between 14 and 18 years of age are also included in the protection area, it must therefore be clear from the presentation within the framework of Section 184b of the German Criminal Code that children under 14 years of age are involved. (19) On the other hand, Section 184b of the German Criminal Code is also wider than its Austrian counterpart. In German criminal law, for example, obtaining possession of relevant representations for a third party (Section 184b (2)) as well as obtaining possession (Section 184b (4)) are corporate offenses. In contrast to Austrian law, it is sufficient for the completion of the offense if the perpetrator starts the corresponding acquisition processes, i.e., "undertakes" them (20).”
<i>Continued on next page</i>	

Field	Content
	“The young actors themselves can also be perpetrators of Section 207a Paragraph 1. This applies, for example, if you make the pornographic representation available to third parties or export it. For example, a young person is liable to prosecution according to § 207a Paragraph 1 Z 3, who makes a "luridly distorted" and "reduced to himself" pornographic representation of his genitals (§ 207a Paragraph 4 Z 3 lit b; Rz 38 ff) via webcam or E-Mail available to others. Since minors are not culpable (§ 4 Abs 1 JGG), underage actors can never make themselves punishable because of § 207a Abs 1.”

Table E.45: Code 2-A-1-13

Field	Content
Legal Source	§ 115 Austrian Criminal Code “(1) Anyone who publicly or in front of several people insults, mocks, physically abuses or threatens to physically abuse another person shall be punished by imprisonment for up to three months or a fine of up to 180 daily rates, provided that he is not threatened with a more severe penalty under any other provision. (2) An act is committed in front of several people if it is committed in the presence of more than two people, different from the perpetrator and the assaulted person, who can perceive it. (3) A person who, merely by indignation at the behaviour of another, allows himself to be carried away to insult, mock, maltreat or threaten with maltreatment in a way that is excusable in the circumstances is excused if his indignation is generally understandable, especially in view of the time that has elapsed since the occasion.”
Qualification	Insult
Notes on interpretation and application	“An insult is public when it can be perceived directly by a larger group of people. From a guideline value of about 10 people, such a larger group of people required for the public inspection is to be assumed. Mere dissemination (e.g., by telling others) does not constitute a public disclosure. The acts of crime are committed ex lege (Paragraph 2) in front of several people if at least three persons - different from the perpetrator and the offender - are present at the same time and there is a concrete perceptibility for them. The three unaffected persons can also be close to the attacked person; the extent to which the perpetrator has an insult-free privacy will have to be assessed on a case-by-case basis (above, §§ 111 ff Rz 33).” “Insults can be, for example: ‘idiot’, ‘madman’, ‘shit bull’, ‘bastard’, ‘shit nigger’, ‘asshole’, ‘criminal’, ‘Nazi’, ‘young Nazi’, ‘cretin’, ‘idiot’, ‘donkey’, ‘pig’.” “No insults are, for example, the mere improper addressing of someone informally (unless it is possibly particularly defamatory within a certain group of people), to indicate to sb that they’re crazy by tapping one’s forehead, statements such as ‘I want nothing to do with you’, ‘go sit on a track’, ‘you’re crazy’ or the Götz quote (see margin no. 18). In the context, the statements ‘to suffer from paranoia’ and ‘to live a poor life’ were not viewed as insults.”

Continued on next page

Field	Content
	“Even when describing a certain politician as a ‘hygienically depraved political inferior, it is important whether there is mere disregard or - seen in context - a permissible political expression of opinion. This also applies to the designation of a member of parliament as a member of the red rent mafia.” “While insulting primarily expresses contempt for another, the essence of mockery lies primarily in ridiculing or deriding another. In contrast to § 491 StGB, it is not necessary for the other person to be exposed to public ridicule. Punishable mockery lies, e.g., in imitating the physical or mental infirmities of another (e.g., speech impediment, limp or hump), in calling the attacked person impotent or mentally diminished or in saying that someone is a case for the psychiatrist. Kien- apfel/Schroll aptly describe the mockery as a ‘popular instrument of mobbing at the workplace’, which can also fall in the scope of the offence. As with the insult (Rz 19 ff), the constitutionally guaranteed liberty of the expression of opinion limits the offence. Thus, for example, the criticism of the Vienna Ballet that it “wobbles about the stage inhomogeneously” was not considered mockery, but rather art criticism without punishment.²⁶ Likewise, the judiciary did not see any mockery - probably in connection with a sufficient factual substrate - in the description of a journalist as ‘old-fashioned greasy hand’ or in the phrase of ‘embarrassing foolish act’.²⁷” “1. The accusation that “the government” acts out of a mere will to power in its function and does not want to improve anything is an unpunished assessment. 2. The same is true for the designation of members of the government as “scumbags”, insofar as there is a material basis for this assessment (criticism of the cultural policy of the federal government and the resulting precarious financial situation of cultural workers). 3. If an artist known in Austria presents this designation of the members of the government in an unedited television program and the host of the program distances herself from it, the reason for exclusion under § 6 para. 2 no. 4 MedienG applies.“ “It is therefore necessary to first determine the meaning assigned to the emoji or emoticon by the creator. This is true even if one wants to follow the BMVRDJ in the opinion that the one-time publication of the emojis “strong arm” ☹ without further explanations under an article or a posting could neither fulfil the offence of requesting, provoking or inciting hatred, nor that of insulting in the sense of § 283 StGB.¹⁸ This opinion deals exclusively with the legal evaluation. On the factual level, the court must therefore first determine the meaning.”
	<i>Continued on next page</i>

Field	Content
	“A sign, which has a completely harmless explanatory value due to years of practice, is given a negative meaning by a certain grouping, or it receives this meaning through the way it is used or through the activity of the grouping.³⁴ In this sense the racist "White Power" movement, for example, adopted the emoji "All is well" () - a hand sign, where thumb and index finger form a circle, while the other fingers are stretched. The three stretched fingers should symbolize a "W", the circle formed by thumb and index finger in connection with the forearm a "P"³⁵. Similarly, users deviate from the actual meaning when, for example, the emoji "eggplant" () is used as a phallic symbol,³⁶ with which it could be used, for example, to substantiate a violation of honour (§§ 111 ff StGB) or to reinforce "stalking" (§ 107a StGB) or "cybermobbing" (§ 107c StGB). The Regional Labour Court of Baden-Württemberg had to judge a similar phenomenon of deviating significance. If the person who uttered an insult and a third party who perceives it agree that an emoji has a different meaning than that which the originator has attributed to it, the emoji can be interpreted in this sense. Specifically, an employee used the emoji "monkey face" (), and assumed, like the third party, that it had the meaning "bear head". If "bear's head" meant a certain superior of the employee, who had a pronounced facial expression due to illness, it was grossly offensive to use the nickname "bear's head" with the emoji, even if the emoji used was actually a monkey.” “All this can be illustrated by the example of the judgment of the OLG Innsbruck on 11 Bs 110/13h. The defendant posted a - even from the defence's point of view - "tasteless and also morally reprehensible" xenophobic joke on Facebook, which he combined with the emoticon ;). The OLG Innsbruck took into account the situation in which the specific entry was made, the use of language (tone of voice) during the discussion in question and the fact that the incriminated statement was made in response to xenophobic Facebook entries, including other xenophobic "jokes". Furthermore, the accused was a "simply thinking person". For the OLG Innsbruck, however, the following statement was already decisive: "The meaning of the emoticon used by the defendant follows from the Wikipedia page http://de.wikipedia.org/wiki/Emoticon "The ... used emoticon stands for 'wink, don't take it so seriously!' "The OLG Innsbruck therefore strictly assumed the meaning that the creator attributed to the emoticon, which contributed to the fact that the accused was finally acquitted of the accusation of incitement.”

Table E.46: Code 2-A-1-14

Field	Content
Legal Source	§ 297 StGB “(1) Anyone who exposes another person to the risk of official prosecution by falsely suspecting him of an offence punishable by a penalty to be imposed ex officio or of a breach of an official or professional duty shall, if he knows (§ 5 para. 3) that the suspicion is false, be punished with imprisonment of up to one year or with a fine of up to 720 daily rates, but if the wrongly charged act is punishable with imprisonment of more than one year, with imprisonment of six months to five years.

Continued on next page

Field	Content
	(2) Under subsection 1, no punishment shall be imposed on anyone who voluntarily removes the danger of official prosecution before an authority has taken action to prosecute the suspect."
Qualification	Defamation
Notes on interpretation and application	"Defamation is committed by anyone who falsely suspects another person of having committed a criminal offence or a breach of an official or professional duty and thereby exposes this person to the risk of prosecution by the authorities. The decisive factor is that the perpetrator knows that the suspicion is false. A suspicion is also "false" if essential circumstances exonerating the person are concealed or the person is suspected of a more serious crime than the one actually committed." "There is a risk of official prosecution if this is to be expected immediately due to the circumstances. It is sufficient for the Federal Police to conduct an investigation; criminal or disciplinary proceedings are not necessary." "The perpetrator claimed during his detention hearing that a prison doctor ordered three prison guard officers to beat him, causing a laceration on his lip, and to squeeze his crotch with his hand, causing pain in his intimate area." "The perpetrator claimed to have handed over € 100,000 to an employee (presumably of a bank or insurance company), who, however, did not invest it but used it for other purposes. He thus accused her of the crime of infidelity. He accused the same woman of having committed an insurance fraud by reporting the same water damage in her apartment twice with forged documents. The damage is said to have amounted to more than € 7,000. With this he accused her of a serious fraud." "The perpetrator sent letters to several people under a false name, claiming that a woman had had sexual contact with "his" children, some of whom were minors. In this way he accused this woman of the crime of sexual abuse of minors."

Table E.47: Code 2-A-1-15

Field	Content
Legal Source	§ 116 Austrian Criminal Code "Acts under sections 111 or 115 shall also be punishable if they are directed against the National Council, the Federal Council, the Federal Assembly or a Landtag, against the armed forces, an independent division of the armed forces or against an authority and are committed in public. The provisions of sections 111(3), 112 and 114 shall also apply to such offences."
Qualification	Public insult of a constitutional representative body, the armed forces or an authority

Continued on next page

Field	Content
Notes on interpretation and application	“The National Council, the Federal Council, the Federal Assembly and the state parliaments can be the objects of an insult of honour. The municipal council is not included (in Vienna it also has the functions of a state parliament anyway). However, individually identifiable municipal councils might be protected. However, the municipal council is protected in its entirety if it is acting as an authority and is therefore attacked.” [...] “The Federal President is not a collegiate body and therefore not mentioned here. In the case of punishable violations of honour against his person, even if these are not committed in public, this constitutes, however, an offence prosecutable on complaint according to § 117 Abs. 1.” “Only the protection of honour of the authorities is of greater importance in practice. The criminal law term ‘authority’ includes the organs of a territorial entity that are equipped with decisive and dispositive power and are permanently organized to fulfil state tasks (administration, jurisdiction) within their sphere of activity. This also includes the associated offices (departments with the character of authorities). The federal government, state governments and their offices, and also the city council (see, however, margin no. 5), the district authorities, security directors, federal police departments, state police commands, the Federal Criminal Police Office, or the Federal Asylum Office, including its outposts, are therefore insulatable as ‘authorities’; no authorities are, for example, schools, hospitals, the Austrian Federal Railway Company, and the police inspections [...]” “Furthermore, courts (including the Asylum Court) and public prosecutor’s offices are authorities within the meaning of § 116. The same applies to chambers when they perform sovereign functions, and to collegiate bodies of the universities.”

Table E.48: Code 2-A-1-16

Field	Content
Legal Source	§ 111 Austrian Criminal Code “(1) Anyone who, in a manner perceptible to a third party, accuses another person of a disparaging quality or attitude or accuses him of dishonourable behaviour or behaviour contrary to public morals which is likely to disparage or belittle him in public opinion shall be punished by imprisonment for up to six months or a fine of up to 360 daily rates. (2) Anyone who commits the offence in a printed work, on the radio or in any other way which makes the defamation of character accessible to the general public shall be punished by imprisonment for up to one year or a fine of up to 720 daily rates. (3) The offender is not to be punished if the allegation is proven to be true. In the case of para. 1, the perpetrator shall not be punished even if circumstances are proven which have provided sufficient grounds for the perpetrator to consider the allegation to be true.”
Qualification	Malicious gossip (Üble Nachrede)

Continued on next page

Field	Content
Notes on interpretation and application	“Malicious gossip can be carried out in two ways. The first alternative of malicious gossip is the accusation of a contemptuous characteristic or attitude (invective) and the second is the accusation of dishonourable or immoral behaviour. The boundaries are fluid. An incorrect subordination to one of the two forms of commission does not do any harm, because they are equivalent and form an alternative mixed offence. [...] a) accusing of a contemptuous characteristic or attitude. The first alternative makes it a punishable offense to accuse another person of a contemptuous characteristic or attitude perceptible for a third party. According to the prevailing opinion, the "accusing" is equivalent to the "blaming" of the second alternative. In both cases, the perpetrator makes a verbal, written or dramatic statement. In the first alternative, the person concerned is ascribed an unworthy character. Insignificant character defects are neither contemptuous nor defamatory; accusations such as "unpunctuality", "weakness in decision" or "tactlessness" do not fall in the scope of the offense. This first alternative of para. 1, also called invective, differs from the second in that the accusation does not refer to a certain behaviour, but to the contemptuous character or the general unworthy inner attitude of the person being reviled. The term "crook" or "liar" can be an invective when emphasizing a character flaw, but it can also be an insult under § 115 if the perpetrator wants to express his disregard in general terms. If, however, the accusation refers to a certain behaviour, as for example in the exclamation "You lied in this agreement!", then this statement is normally to be considered as dishonourable conduct under the second alternative of para. 1. The term Nazi, fascist or the belonging to the extreme right, liar, defamer, crook, mafioso, racist or addict can be for example factual invectives, where also accusations of character would meet the standards like intolerable moralities, degeneration, or perfidy. Mere incivilities and disturbances cannot be understood as invectives. The accusations to be a troublemaker, rascal, to be stubborn, to be unpleasant, cunning, senile, mischievous, or jealous does usually not meet the criteria of the fact. If a person is described as belonging to the extreme-left, the criteria will also not be met, because this political opinion is not covered by the law banning National Socialist activities. The claim that the opposing party in an election campaign is lying or deceiving the voters will also generally not meet the threshold of being qualified as an invective. On the one hand, the accusation is targeting a politician who is in a position where he has to bear more than a private person and on the other hand such accusations will also be qualified by the voters themselves as rhetorical exaggeration. Hence a claim that is heading for a § 263 accusation has to be treated differently. The accusation of playing down national socialistic institutions, after all being a Nazi or racist agitator can, as a not punishable judgement of value according to Art 10 (1) EMRK, lead to the exclusion of the offence. Accusation of dishonourable behaviour or violation of the good morals § 111 (1) is fulfilled if someone, observable for a third person, is accusing a person of dishonourable behaviour or violating the good morals in such a way that it is suitable to influence the public opinion contemptuously. The behaviour of a person can be qualified as dishonourable if the dominant social appreciation of the concerned party is damaged. A behaviour that is violating the good morals is a violation of the public decency. The difference to an invective, according to the first case of paragraph 1, can be found in the accusation of a lack of character in general in contrast to an accusation that targets some concrete behaviour.

Continued on next page

Field	Content
	Defamation has to be observable for a third person. It is sufficient for the general case of paragraph 1 that a third person is able to observe the accusation. A person that is also defamed by the accusation will not be qualified as a third person, except the accusation is addressing several persons but is only targeting a concrete one of them. Also, statements that are given within the sphere of privacy and targeting external parties do not meet the criteria of publicity. Related parties generally will be qualified as third parties. The statement that is touching the honour of a person has to be observed by a third person that neither is the offender nor the offended party, where this third person only has to have the opportunity to observe the behaviour. It is sufficient if the third person is within acoustic reach or has the possibility to read the postcard or, as a secretary, opens the office mail."

Table E.49: Code 2-A-1-17

Field	Content
Legal Source	§ 107c Austrian Criminal Code "(1) Anyone who, by means of telecommunication or using a computer system, in a manner likely to unreasonably interfere with a person's conduct of life, 1. commits an offence against the honour of a person in a manner that is perceptible to a large number of people for a longer period of time, or 2. makes a fact or image of a person's most personal sphere of life perceptible to a larger number of people for a longer period of time without that person's consent, shall be punished with imprisonment for not more than one year or with a fine not exceeding 720 daily penalty units. (2) If the act results in the suicide or attempted suicide of the person injured within the meaning of subsection 1, if the perpetrator continues to commit acts within the meaning of subsection 1 directed against the injured person within a period exceeding one year, or if the duration of the perceptibility according to subsection 1 exceeds one year, the perpetrator shall be punished with imprisonment for up to three years."
Qualification	Continued harassment by means of telecommunications or computer systems
Notes on interpretation and application	"After the leading and convincing opinion, a person is infringing another party's 'honour' according to 107 c (1) Z 1 a decrease in standing or reputation of a human beings relevant environment.' This characteristic, therefore, is identical to the category of honour crimes (§§ 111 bis 117) and includes e.g. insult, defamation, mockery or humiliations. Therefore, the (incorrect) claim that somebody has high (gambling) debts is influencing one's sphere of honour. The deceive element is the objective honour of another party or his good reputation."

Continued on next page

Field	Content
	“§ 107 c (1) Z 2 includes the publication of facts or pictures of the personal sphere of the victim without consent. It is sufficient according to the wording to be perceptible for a minimal audience of about 10 persons (Rz 37). Making public is not only the act that includes simultaneous optical or acoustic perception, but also captures gradual public on the side of the addressee. It is neither about the special common element nor if the transmission of the facts or pictures takes place simultaneously, if telecommunication or the use of a computer system include the gradual capture of such a group of persons. Making accessible is sufficient and does not need actual listening or watching.” “The personal sphere includes the sexual preferences, serious illnesses, disabilities or religious beliefs. This includes action that not only addresses the electronic publishing of pictures of the victim, but also recordings of his living space. Other examples include the making public of AIDS-tests of the victim on the internet, the communication of the sheer fact that a person is HIV-positive via Twitter or the documentation of the lived religious practice of a human e.g., a photograph of a person praying in the direction of Mecca or in cases of self-flagellation.”

Table E.50: Code 2-A-1-18

Field	Content
Legal Source	§ 107 Austrian Criminal Code “(1) Anyone who dangerously threatens another in order to make him fear and agitate shall be punished by imprisonment for up to one year or a fine of up to 720 daily rates. (2) Anyone who makes a dangerous threat by threatening death, serious mutilation or striking disfigurement, kidnapping, arson, exposure to nuclear energy, ionising radiation or explosives, or the destruction of economic existence or social position, or by using these means to put the threatened person or another person against whom the violence or dangerous threat is directed into an agonising state for an extended period of time, shall be punished by imprisonment for up to three years. (3) In the cases referred to in section 106(2), the penalty provided for therein shall be imposed.”
Qualification	Dangerous threat
Notes on interpretation and application	“§ 107 (2) is addressing only in the qualified cases, in which there is an act of threat, where the basic offense of § 107 (1) of the dangerous threat in § 74 (1) Z 5 has to be included in the interpretation. This means that the act of the meaning of dangerous threat according to § 107 is only a behaviour in which a personal injury, personal freedom, honour or property is threatened.” “The act of threat that fulfils the criteria of harm according to § 74 (1) Z 5 fulfils the provision when a harm to the personal sphere is committed and can be performed in writing, verbal or through gestures.”

Table E.51: Code 2-A-1-19

Field	Content
Legal Source	§ 225a Austrian Criminal Code “Anyone who intentionally creates false data by entering, altering, deleting or suppressing data or falsifies genuine data with the intention that they are used in legal transactions to prove a right, a legal relationship or a fact shall be punished by imprisonment for up to one year or a fine of up to 720 daily rates.”
Qualification	Data falsification
Notes on interpretation and application	“Only computer data as defined in Art 1 lit b CCC (including data definition § 118a Rz 60 f), which are important for private as well as sovereign legal transactions, are considered to be objects of crime. According to the prevailing doctrine, this data is data stored on data carriers as defined in § 74 (2), which is intended for legal transactions and allows the issuer to be identified (‘electronic documents’). They are of legal significance if they refer to legal transactions in the broadest sense or are equivalent to documents with legal significance. Both personal and non-personal data and programs are protected. The latter do not, however, as a rule fulfil the similarity of documents presupposed by § 225a.” “The unauthorized ‘input’ of correct or incorrect data is comparable to the production of a fake document. Subsequent ‘altering’ (modifying, changing, partially changing), ‘deleting’ (removing data from a data carrier) and ‘suppressing’ (withholding, hiding data) generally correspond to the falsification of a genuine certificate.” “‘Changing’ means any modification of existing data. A reduction of its usability is not assumed, otherwise this variant of the offence would be superfluous. Any manipulation of external data or data contents is covered here. It is sufficient if the creation of a new link makes it possible for the user to recognize different content even without a concrete change to the data output, since the link has changed the content of meaning. In practice, the so-called ‘website spoofing’ in particular causes considerable difficulties in this context. The perpetrator creates a website that looks and functions and has the look and feel of a bank, a serious antique dealer or any other corporate website. Subsequently, unsuspecting Internet users are directed to the fake website via e-mail and asked to provide their credit card number or other bank details or to act as ‘independent sales representatives’ in the antique trade if they have previously made a transfer to a specific account. The data, in this case of the entire foreign, already existing website, is assigned a different information content and thus no longer fulfils its intended purpose. In this respect, a change of data by merely changing the data sequence is also of interest if it changes the context. Ultimately, this is a “virtual identity thief”, which as a rule serves as a predicate offence for large-scale computer fraud. The decisive factor is that this ‘reprogramming’ of the website results in a deception of the identity of the exhibitor, e.g., by simultaneously adding a (slightly) different imprint.” “The ‘deletion’ of data is the equivalent to the destruction of physical objects. The data is thereby destroyed and made unrecognizable, whereby a partial deletion is also covered in the sense of making it unrecognizable.”

Continued on next page

Field	Content
	“The ‘suppression’ of computer data includes any action that prevents or terminates the availability of data to the person who has access to the computer or the data carrier on which it is stored. This act is already affirmative if the data cannot be immediately and completely restored. It is not necessary to destroy the data.” “The protection against falsification starts in any case after the first saving of an electronic document (see margin no. 20). In addition, according to part of the literature, this alternative also covers protection during the transmission process, e.g., by e-mail. Data is therefore falsified if it is changed in the non-volatile state or on the transmission path - if it is stored again, meaning the offender gives the impression that the data in this form no longer originates from the original issuer.”

Table E.52: Code 2-A-1-20

Field	Content
Legal Source	§ 188 StGB “Anyone who publicly disparages or mocks a person or an object of worship of a church or religious community existing in the country, or a doctrine, a legally permissible custom or a legally permissible institution of such a church or religious community in circumstances in which his conduct is likely to cause justified nuisance shall be punished with imprisonment for up to six months or a fine of up to 360 daily rates.”
Qualification	Vilification of religious teachings
Notes on interpretation and application	“This [...] provision penalizes grossly disturbing impairments of the constitutionally guaranteed freedom of religion (Article 14 StGG 1867; Article 63 para. 2 StV v Saint Germain; Article 6 StV v Vienna; Article 9 EMRK). What is protected, however, is not the religion in question as such or the religious sensibilities of the individual, but religious peace (likewise E. Mayer/Tipold, SbgK § 188 Rz 6 mwN).” “The perpetrator claimed that it was no secret that the Prophet Mohammed “also liked to make out a little bit with children”. She expressed this at a seminar of the educational institute of her political party. The Supreme Court of Justice (OGH) described this accusation of paedophilia as a public disparagement or mockery of a person who is revered by a religious community recognized in the country (OGH 11.12.2013, 15 Os 52/12d (“Mohammed-decision“).”

Table E.53: Code 2-A-1-21

Field	Content
Legal Source	§ 120a StGB

Continued on next page

Field	Content
	“(1) Any person who intentionally takes a photograph of the genitals, the pubic region, the buttocks, the female breast or the underwear covering these parts of the body of another person who has protected these areas from view or is in a dwelling or in a room specially protected from view, without that person's consent, shall be liable to a custodial sentence not exceeding six months or to a monetary penalty not exceeding 360 daily penalty units. (2) Any person who makes accessible to a third party or publishes a photograph pursuant to subsection (1) without the consent of the person depicted shall be liable, unless the offence is punishable under another provision by the same or more severe penalty, to a custodial sentence not exceeding twelve months or to a monetary penalty not exceeding 720 daily penalty units. (...)”
Qualification	Unauthorised recording of images
Notes on interpretation and application	“The protected body parts include the genitals, the pubic region, the buttocks and the female breast. The terms genitals and pubic region are to be understood as in Section 207a (4) (3) (b) StGB. Although images of the female breast are not "upskirting" in the narrower sense, following the German model, these images are also to be included in the offence, because these are also supposed to be protected against the views of strangers by the outer clothing and such photos therefore also violate the privacy of the victim. Photographing leggings worn by a person under a dress or skirt, on the other hand, will not constitute an offence because it is not underwear. Whether one of the protected body parts appears on a part of the picture or takes up the whole picture is not relevant for the objective elements of the offence. However, in order to limit liability to conduct particularly worthy of punishment and to counteract problems of application in practice, intentionality is to be required on the subjective side of the offence. The perpetrator must therefore be concerned with the realisation of the elements of the offence.”

Table E.54: Code 2-A-1-22

Field	Content
Legal Source	§ 3g VerbotsG “Anyone who engages in activities in the national socialist sense in a manner other than those described in §§ 3a to 3f shall be punished with imprisonment of one to ten years, or up to 20 years if the offender or the activity is particularly dangerous, unless the act is more severely punishable under another provision.” § 3h VerbotsG “Anyone who denies, grossly trivialises, approves of or attempts to justify the National Socialist genocide or other National Socialist crimes against humanity in a printed work, on the radio or in another medium, or anyone else publicly in such a way that it becomes accessible to many people, will also be punished according to § 3g.”

Continued on next page

Field	Content
Qualification	Verbotsgesetz/Wiederbetätigung
Notes on interpretation and application	A violation of the Verbotsg (VG) is for instance: “Glorification of the person of Adolf Hitler and approval of his mission in life. The propagandistic use of typical National Socialist slogans and symbols, such as “Heil Hitler”, “Sieg Heil”, the Hitler salute or the swastika” § 3h covers the denial of the crimes against humanity committed under the National Socialist regime. “Every conduct not covered by §§ 3a to 3f of the Verbotsg (VG) is covered by § 3g of the Verbotsg (VG), insofar as it is suitable for propagating and updating any objectives of National Socialism in Austria or at least with an effect on the Republic of Austria, the act therefore also having an inherent propagandistic effect which, according to the perpetrator's ideas, develops its effect on Austrian territory. The absence of this domestic reference would, even in the case of the existence of an identical norm under the laws of the scene of the crime covering the conduct in question, exclude punishability under the Verbotsg in Austria.” “§ 3g VG is a catch-all offence and punishes (among other things) any other behaviour (not covered by §§ 3a to 3f VG) - which is not at all accessible to a conclusive legal description - which even in the abstract is suitable to revive one of the specific and manifold objectives of the NSDAP. The combination of several text passages of a single publication (a book) to a factual unit of action in the main questions (in conformity with the lawsuit) requires the examination of the factuality in the sense of § 3g VG on the basis of the overall impression of these text passages; whether each of them - considered in isolation - fulfils the facts, however, is irrelevant.”

Table E.55: Code 2-G-2-1

Field	Content
Legal Source	§7 (4) TMG “If a tele media service has been used by a user in order to infringe the intellectual property rights of another user, and if the holder of this right has no other possibility of remedying the infringement of his right, the holder of the right can demand that the service provider concerned block the use of information in accordance with § 8 (3) TMG in order to prevent the repetition of the infringement. The blocking must be reasonable and proportionate. A claim against the service provider for reimbursement of the pre- and out-of-court costs for asserting and enforcing the claim pursuant to sentence 1 does not exist except in the cases of § 8 (1) TMG sentence 3.”
Qualification	Disturbance liability (Störerhaftung)

Continued on next page

Field	Content
Notes on interpretation and application	The operator of a website is liable for his own content; for the content of a foreign party the operator of a website generally is only liable if he appropriates such content. This criterion is fulfilled if the website operator has knowledge of this content and it is technically possible and reasonable to inhibit the use.

Table E.56: Code 2-G-2-2

Field	Content
Legal Source	§ 1004 I BGB analogous in conjunction with, § 823 BGB
Qualification	Quasi-negatory injunctive relief and tort law
Notes on interpretation and application	"Irrespective of the fact that even in the case of an originally admissible suspicious transaction report, there is now a right to injunctive relief, the Regional Court correctly stated that the requirements for admissible suspicious transaction reports required by case law had not been met. The mere mention of a newspaper article in the "Stuttgarter Zeitung" as proof of the text had rightly not been sufficient for the Regional Court. On the other hand, the Regional Court wrongly stated that the passage "in the context of his television work, he should ... massive intimidation of employees, whom, according to the daily newspaper, he brainwashed in one-on-one interviews lasting several hours. Already with ... employees said that the situation was partly similar to a sect". Correctly, these were factual statements. The statement that he was brainwashing his employees was just as accessible as the statement that sectarian conditions prevailed in his company. The contested passage would give the public concerned the impression that he had exerted considerable psychological pressure on his employees in order to induce them to behave in a certain way. The statements were also accessible to evidence since his employees could easily be questioned as witnesses to this factual claim. Since these defamatory factual claims are untrue, there is also a quasi-legal right to injunctive relief in this respect according to §§ 823 (1), 1004 BGB, Article 1 (2) and Article 2 (1) GG."

Table E.57: Code 2-G-2-3

Field	Content
Legal Source	§ 97 German Copyright Act

Continued on next page

Field	Content
	“(1) Anyone who unlawfully infringes the copyright or any other right protected under this Act may claim from the infringing party the removal of the impairment or, in the event of a risk of repetition, injunctive relief. The right to injunctive relief shall also exist if an infringement is threatened for the first time. (2) Whoever performs the act intentionally or negligently is obliged to compensate the injured party for the resulting damage. In assessing the damages, the profit which the infringer has made from the infringement of the right may also be taken into account. The claim for damages may also be calculated on the basis of the amount which the infringer would have had to pay as reasonable compensation if he had obtained permission to use the infringed right. Authors of scientific publications (section 70), photographers (section 72) and performing artists (section 73) may also claim monetary compensation for damage other than pecuniary loss if and to the extent that this is equitable.”
Qualification	Copyright infringement
Notes on interpretation and application	“Infringement: § 97 only concerns cases of infringement of a third party’s copyright or another right protected under the Copyright Act. Sanctioned is therefore the violation of absolute rights which are effective against everyone (“other rights” as defined in § 823 para. 1 BGB). However, the infringement of authorizations or permits which are only valid for the respective Contract Partner does not trigger the legal consequences of § 97 (amtl. Begr., BT-Drs. IV/270, 103; Schricker/Loewenheim/Leistner nr. 8). Thus, whoever is only the holder of a right of use under the law of obligations, but not of a right of use in rem, cannot take action against the infringer in the event of infringement (only the author himself or the holder of the infringed right of use is legitimated to take action; → marginal no. 18). Absolute rights include, on the one hand, the rights flowing from the author’s moral rights and, on the other hand, the rights in rem arising from the comprehensive exploitation right of the author or the owner of a related right, including the prohibition of exploitation under Art. 96 (BGH ZUM 1986, 199, 202 - GEMA presumption III; Schricker/Loewenheim/Leistner nos. 10 f.). Whether this also includes the violation of the protection against circumvention is disputed. By the bypass protection in accordance with §§ 95a exp. 1, 3 no new utilization or achievement patent right is created. The protection against circumvention merely flanked the copyright powers, which is why the applicability of § 97 can certainly be questioned (left open by BGH GRUR 2008, 996 marginal no. 12 - Clone CD).”

Table E.58: Code 2-G-2-4

Field	Content
Legal Source	§ 14 Trademark Act (MarkG) “(2) A third party shall be prohibited in trade, without the consent of the proprietor of the trademark, from 1. using a sign which is identical to the trademark for goods or services which are identical to those for which it enjoys protection.

Continued on next page

Field	Content
	2. using a sign if the likelihood of confusion exists for the public because of the identity or similarity of the sign to the trademark and the identity or similarity of the goods or services covered by the trademark and the sign, including the likelihood of association with the trademark; or 3. using a sign identical with or similar to the trademark for goods or services which are not similar to those for which the trademark enjoys protection if the trademark is a trademark which has a reputation in Germany and the use of the sign without due cause takes unfair advantage of, or is detrimental to, the distinctive character or the repute of the trademark which has a reputation (...)"
Qualification	Trademark infringement
Notes on interpretation and application	The criteria for a violation against § 14 Trademark Act: - The use of a trademark by a third party - Within commercial practice - Without the permission of the trademark's owner - If the criteria of § 14 (2) Trademark Act is met: Double identity: The usage of an identical trademark for a similar category of goods or services. Danger of confusion case 1: The usage of an identical trademark for similar goods or services. Danger of confusion case 2: The usage of similar trademarks for identical goods or services. Danger of confusion case 3: The usage of similar trademarks for similar goods or services. Protection of prominence: The usage of an identical or similar trademark for unsimilar goods or services, if distinctiveness is met or the appreciation of a well-known trademark is misused or impaired without any justification.

Table E.59: Code 2-G-2-5

Field	Content
Legal Source	§ 241 II BGB, § 280 BGB § 241 "(2) An obligation may also, depending on its contents, oblige each party to take account of the rights, legal interests and other interests of the other party." § 280 "(1) If the obligor breaches a duty arising from the obligation, the obligee may demand compensation for the damage caused thereby. This does not apply if the obligor is not responsible for the breach of duty.

Continued on next page

Field	Content
	(2) Damages for delay in performance may be demanded by the obligee only subject to the additional requirement of section 286. (3) Damages in lieu of performance may be demanded by the obligee only subject to the additional requirements of sections 281, 282 or 283."
Qualification	Contractual obligations
Notes on interpretation and application	"The user of a social media platform, which is intended to enable a general exchange of information and opinions, is entitled to a claim against the operator from the user contract in connection with § 241 Para. 2 BGB and the indirect third-party effect of the basic right to freedom of opinion (Art. 5 Para. 1 Sentence 1 GG) that a permissible expression of opinion posted by him is not removed from the platform and that its posting is not subject to sanctions by the operator. (RN.106) 2) By establishing general rules of conduct, the platform operator can define rights, legal interests and interests in the sense of § 241 (2) BGB (German Civil Code), which the user must take into account. (margin no. 116)"

Table E.60: Code 2-G-2-6

Field	Content
Legal Source	§ 7 Unfair Competition Act in conjunction with § 3a Unfair Competition Act "§ 7 (1) A commercial practice which constitutes an unacceptable nuisance to a market participant shall be illegal. This shall apply to advertising particularly in cases where it is apparent that the solicited market participant does not want this advertising. (2) An unacceptable nuisance shall always be assumed in the case of 1. advertising using a medium of commercial communication not listed under nos. 2 and 3 which is suited to distance marketing and through which a consumer is persistently solicited although it appears that he does not want this. 2. advertising by means of a telephone call, made to a consumer without his prior express consent, or made to another market participant without at least the latter's presumed consent. 3. advertising using an automated calling machine, a fax machine or electronic mail without the addressee's prior express consent; or 4. advertising using a communication a) where the identity of the sender on whose behalf the communication is transmitted is concealed or kept secret or b) which violates section 6 (1) of the Telemedia Act (Telemediengesetz) or in which the recipient is prompted to call up a website which violates said provision or

Continued on next page

Field	Content
Qualification	c) where there is no valid address to which the recipient can send an instruction to terminate transmission of communications of this kind, without costs arising by virtue thereof, other than transmission costs pursuant to the basic rates. (...)" Spam
Notes on interpretation and application	"Spam can reach from messages that contain a virus, offered products, advertised websites, offered prizes or winnings, mediation of business contacts which constitute of an advertising nature and where the receive is unwelcome to the receiving party. The message reaches its advertising nature if a good, service or enterprise is praised. Mails that only suit the purpose of the dissemination of a virus and do not advertise any product cannot be qualified as spam. A message is undesirable if the recipient receives it without his permission or is bothering him in any other way." Unacceptable Harassment: "Harassment: The business act must be perceived as a nuisance by the recipient circle because of the way it reaches them. The annoyance therefore follows from its imposition: The recipient, whether he likes it or not, must first of all deal with it. His opposing will has either been ignored by the actor or could not be formulated at all. The confrontation with the business act interferes with his private life or business processes and can - but need not - lead to a time and/or cost expenditure for the recipient, since he has to check and, if necessary, remove the advertisement, costs for keeping reception equipment are incurred, etc. The content of an advertisement is irrelevant for determining its annoying effect. It is irrelevant whether the addressee feels annoyed by the advertising message transported with a business act, e.g., because it is contrary to his political, religious, or moral views. Irrespective of the fact that a harassing effect cannot already lie in "the fact that the audience is confronted with unpleasant or pitiful realities through images outside the editorial part of the media" or "the advertisement by addressing suffering not in the usual political, charitable or reporting context, but in a commercial context, can be perceived as "disconcerting or deemed improper" and competition law must not amount to censorship of taste, advertising forms such as tasteless advertising or shocking image advertising cannot a priori fall under Section 7 UWG, since the accusation of unfairness is based on their content and not on the way in which the advertising message is communicated. (...)

Continued on next page

Field	Content
	Unacceptability: Not every harassment is unfair within the meaning of § 7 para. 1 sentence 1 UWG because a large part of advertising measures usually reaches the recipient unsolicited. Moreover, competition without extensive and intensive influence on market participants is simply not conceivable and is generally not even expected. This is taken into account by the element of unreasonableness: not every annoyance, no matter how slight, can be attributed the verdict of anti-competitive behaviour; rather, the intervention of the law on unfair competition requires an annoyance that goes beyond the "healthy level". The assessment of unreasonableness does not depend on the viewpoint of the person specifically affected by a business act, but on that of an averagely sensitive market participant who is neither particularly sensitive nor particularly insensitive. The model of the average consumer with a normal understanding, which was developed by the European Court of Justice and the Federal Court of Justice for the assessment of misleading advertising, can also be made fruitful here for the determination of standards. If the advertising is directed at a certain group of people, the view of an averagely sensitive member of this group of people must be taken into account."

Table E.61: Code 2-A-2-1

Field	Content
Legal Source	§ 86 UrhG "(1) Anyone who without authorisation 1. uses a work of literature or art in a manner reserved for the author under §§ 14 to 18a, 2. uses a performance in a manner reserved for the performing artist under § 68, 3. uses a performance in a form of exploitation reserved for the organiser according to § 72, 4. uses a photograph or a sound carrier in a manner reserved for the manufacturer under §§ 74 or 76, 5. uses a radio broadcast in a form of exploitation reserved for the radio broadcaster pursuant to § 76a or 6. uses a database in a manner reserved for the manufacturer under § 76d, must, even if he is not at fault, pay an appropriate compensation to the injured party whose consent would have been required. (2) However, there shall be no claim to such remuneration if a radio broadcast, a communication to the public or a making available to the public has been inadmissible only because it has been made with the aid of image or sound carriers or radio broadcasts, which were made in accordance with § 50 para. 2, § 53(2), § 56(3), § 56b(2), § 56c(3)(2), § 56d(1)(2), §§ 68, 72, 74, 76 or 76a(2) and (3), and if this characteristic of the image or sound carriers or broadcasts was unknown to their user without his fault. (3) Anyone who uses a press report in contravention of § 79 shall pay the news collector an appropriate fee, even if he is not at fault."

Continued on next page

Field	Content
	§ 87 UrhG “(1) Any person who culpably damages another person by violating this Act shall also compensate the injured party for loss of profit, irrespective of the degree of fault. (2) In such a case, the injured party may also claim reasonable compensation for the disadvantages he has suffered as a result of the act, which do not consist in any financial loss. (3) The injured party, whose consent would have had to be obtained, may claim double the remuneration due to him under § 86 as compensation for the pecuniary loss culpably inflicted on him (para. 1), unless higher losses are proven. (4) If a work of literature or art is reproduced or distributed without authorisation, the injured party, whose consent would have had to be obtained, may also demand the surrender of the profit which the injuring party has made through the culpable intervention. The same shall apply if a performance contravenes section 68 subsection (1) or a radio broadcast is exploited in contravention of section 76a on a picture or sound carrier or if a photograph is reproduced or distributed in contravention of section 74 or a sound carrier in contravention of section 76. Finally, the same shall apply if the right of making available (§ 18a) is infringed. (5) In addition to an appropriate remuneration (§ 86) or the surrender of the profit (paragraph 4), compensation for financial loss can only be claimed if it exceeds the remuneration or the profit to be surrendered.”
Qualification	Copyright infringement
Notes on interpretation and application	“The claim to adequate remuneration according to § 86 UrhG is dogmatically - like the comparable claims under § 53 (1) MSchG and § 150 (1) PatG - a special form of the use claim under the law of enrichment in the sense of § 1041 ABGB. It concerns a statutory claim for remuneration for the unjustified use of a copyright protected work or a creative result subject to ancillary copyright.” “The claim exists independently of a damage of the injured party and of the fault of the infringer.” “§ 87 UrhG contains some regulations that deviate from the general law of damages or supplement it, which take into account the special features of copyright and related rights. The provision deals with the extent of the damages to be compensated (para. 1), the compensation of non-material damage for a special impairment of the infringed party connected with the infringement (para. 2), the lump sum compensation for damages - which serves to avoid difficult damage calculations in practice and to facilitate evidence - (para. 3), the right to the distribution of profits (para. 4) as well as questions of delimitation (para. 5). [...]” A violation of copyright regulations in general also only obligates under the general conditions to the compensation. The damage incurred by the author or the person entitled to an ancillary copyright must therefore have been caused unlawfully and culpably by the infringer, whereby slight negligence is sufficient.”

Table E.62: Code 2-A-2-2

Field	Content
Legal Source	§ 10 MarkenSchG “(1) Subject to the safeguarding of earlier rights, a registered trademark shall confer on the proprietor the exclusive right to prohibit all third parties, in the course of trade and without his consent 1. to use a sign which is identical with the trademark in relation to goods or services which are identical with those for which the trademark is registered (Paragraph 10a) 2. to use a sign which is identical or similar to the trademark in relation to identical or similar goods or services (Section 10a) if there exists a likelihood of confusion on the part of the public, which includes the likelihood of association between the sign and the trademark. (2) The proprietor of a registered trade mark shall also be entitled to prevent all third parties not having his consent from using, in the course of trade, any sign which is identical with, or similar to, the trade mark in relation to goods or services which are identical with or similar or dissimilar to those for which the trade mark is registered (Section 10a), where the trade mark has a reputation in the country and where use of that sign without due cause takes unfair advantage of, or is detrimental to, the distinctive character or the repute of the trade mark. The reputation of the earlier trademark must be established at the latest on the date of application for registration of the later trademark, where applicable on the date on which priority or seniority is claimed, or on the date on which the later other sign right comes into existence. 2a. Without prejudice to earlier rights, the proprietor of a registered trade mark shall also be entitled to prohibit a third party from introducing goods into his territory in the course of trade without putting them into free circulation, where those goods, including their packaging, originate in third countries and bear without authorization a trade mark which is identical to the trade mark registered for such goods or which cannot be distinguished in its essential aspects from such trade mark. This entitlement of the trademark owner expires if, during proceedings aimed at determining whether a registered trademark has been infringed and initiated pursuant to Regulation (EU) No. 608/2013 on the enforcement of intellectual property rights by the customs authorities, OJ No. L 181, 29.06.2013, p. 15, the customs applicant or the owner of the goods proves that the owner of the registered trademark is not entitled to prohibit the placing of the goods on the market in the country of final destination. (2b) If there is a risk that the packaging, labels, tags, safety or authenticity notices or proofs or other means of identification to which the trademark is affixed are used for goods or services and that such use constitutes an infringement of the trademark proprietor's rights under subsection 1 or 2, the trademark proprietor shall be entitled to prohibit the following acts if they are carried out in the course of trade: 1. the affixing of a sign identical with, or similar to, the trademark to those means of identification 2. offering, putting on the market or stocking for these purposes or importing or exporting such means of identification to which the trademark or a sign similar to it is affixed.

Continued on next page

Field	Content
	(3) A registered trademark shall not entitle the proprietor to prohibit a third party from using. 1. the name or address of the third party, if the third party is a natural person 2. signs or indications which are devoid of any distinctive character or which may serve to designate the kind, quality, quantity, intended purpose, value, geographical origin, or the time of production of the goods or of rendering of the service, or other characteristics of the goods or service 3. the trademark for the purposes of identifying or referring to goods or services as those of the proprietor of the trademark, in particular where the use of the trademark is necessary to indicate the intended purpose of a product or service, for example as accessories or spare parts in the course of trade, provided that such use is in accordance with honest practices in industrial or commercial matters.”
Qualification	Trademark infringement
Notes on interpretation and application	§ 10 Abs 1 Z 1: Identity protection/ double protection – if the trademark is used in an identical form § 10 Abs 1 Z 2: Protection against a way of usage that is likely to be confused by third parties if there is a similarity to the brand in a degree that makes the party believe the good or service offered comes from the same enterprise. § 10 Abs 2: Protection against exploitation of reputation and impairment: special protection of well-known brands. In the case of a trademark infringement the injured party has the remedy of the injunctive relief according to § 51 MarkenSchG. § 53 MarkenSchG entitles, in the case of unauthorized use of a trademark, the injured party to adequate remuneration, in the case of culpable trademark infringement the injured party also has the claim to damages including the lost profit or the issuance of profit (Summary of MarkenSchG, translation by the authors).

Table E.63: Code 2-A-2-3

Field	Content
Legal Source	§ 107 TKG “(1) Calls - including the sending of faxes - for advertising purposes without prior consent of the subscriber are not permitted. The consent of the subscriber shall be equivalent to the consent of a person authorised by the subscriber to use his line. The consent given can be revoked at any time; revocation of consent has no effect on a contractual relationship with the addressee of the consent. 1a. In the case of telephone calls made for advertising purposes, the presentation of calling line identification must not be suppressed or falsified by the caller and the service provider must not be induced to suppress or falsify it.

Continued on next page

Field	Content
	The sending of electronic mail, including SMS messages, shall not be allowed without the prior consent of the recipient where the sending is for the purposes of direct marketing. (3) Prior consent for the sending of electronic mail pursuant to paragraph 2 shall not be required if 1. the sender has received the contact information for the message in connection with the sale or service to his customers and 2. this message is for direct advertising of own similar products or services and 3. the recipient has been clearly and unambiguously given the opportunity to refuse such use of the electronic contact information at the time of its collection and, in addition, at each transmission, free of charge and without difficulty; and 4. the recipient has not refused the mailing from the outset, in particular by being entered on the list referred to in § 7 para. 2 of the E-Commerce Act."
Qualification	Spam
Notes on interpretation and application	"The plaintiff explains in his remedy that the Facebook or WhatsApp advertising in question addresses a commercially inexperienced audience and therefore an audience that is especially worthy of protection. The messages sent to students without given consent via Facebook violate § 107 (2) TKG." (translation by the author) "Without the consent for the messages sent via Facebook to students the act violated § 107 (2) TKG. Regardless of if the addressee has told the advertiser that he does not desire any more consignment." "Relating to the messages received via Facebook or WhatsApp, the reference of the plaintiff regarding the legal provisions concerning undesired (unsolicited or unwelcome) advertising for distance selling or sales via electronic means (§ 107 Abs 2 TKG; § 7 ECG) per se are applicable. According to those provisions direct advertising using electronic means or messages is in the need of consent of the addressee that is given prior to the advertising act (and has to be revocable). For that, undesired electronic advertising per se is impermissible in Austria (Opt-in-System)". A consignment is 'Unsolicited' (undesired or unwelcome) if in coordination with the transmission the advertiser has no prior given consent of the addressee (see Brenn, ECG 218), therefore, the addressee is not in need of an active contradiction."

Table E.64: Code 2-A-2-4

Field	Content
Legal Source	§ 6 MedienG

Continued on next page

Field	Content
	“(1) If the objective facts of defamation, insult, ridicule or slander are established in a medium, the person concerned shall be entitled to compensation from the media owner for the personal harm suffered (section 8(1)). (2) The claim under subsection 1 shall not exist if 1. it concerns a truthful report on a hearing in a public session of the National Council, the Federal Council, the Federal Assembly, a Land parliament or a committee of one of these general representative bodies, 2. in the case of defamation a) the publication is true; or b) there was an overriding public interest in the publication and there were sufficient reasons to believe the allegation to be true, even if due journalistic care had been exercised, 3. it is a direct broadcast on the radio (live broadcast) without an employee or agent of the radio having disregarded the due journalistic care, 3a. it is a matter of retrievability on a website without the media owner or one of its employees or agents having disregarded due diligence, or 4. it is a truthful reproduction of the statement of a third party and there was an overriding interest of the public in knowing the statement quoted. (3) If the publication relates to the highly personal sphere of life, the claim under subsection 1 shall only be excluded on the grounds of subsection 2 (1), subsection 2 (2) (a), subsection 2 (3) or subsection 2 (3a), but in the case of subsection 2 (2) (a) only if the facts published are directly related to public life.”
Qualification	Protection of Personality – Malicious gossip, insult, mockery, and defamation
Notes on interpretation and application	§ 6 MedienG statutorily guaranteed, in the case of the fulfilment of the physical elements of the offence for defamation, insult, mockery or traducement, damages for the insult suffered against the media owner. Media owner is according to § 1 Z 8 MedienG, who a) operates a media business or service, or b) in other ways is giving shape to the media work or its manufacture or dissemination or, c) in the case of electronic means, the person concerning or arranging for influencing the design, availability, broadcast or dissemination, or d) is concerning the content design of a medium for the purpose of broadcast, availability or dissemination.

Table E.65: Code 2-A-2-5

Field	Content
Legal Source	§ 549 (1) ZPO “In legal disputes concerning lawsuits in which only claims for injunctive relief are asserted due to a substantial violation of personal rights in an electronic communications network affecting a natural person's human dignity, the court shall, at the request of the plaintiff, issue an injunction without a prior oral hearing and without hearing the defendant if the alleged claim can be conclusively derived from the information in the complaint. The complaint shall be accompanied by evidence from the electronic communications network showing or making apparent the infringing content.”
Qualification	Proceedings for substantial violation of personality rights in an electronic communications network
Notes on interpretation and application	“The proposed provision is intended to establish a special procedure in the Code of Civil Procedure that is to be available for particularly massive cases of violations of personality rights. This procedure is to be used exclusively in legal disputes on actions in which claims for injunctive relief are filed due to a significant violation of personality rights in an electronic communications network that impairs the human dignity of a natural person. Following suggestions in the review procedure, the wording of the offence “significant violation of personality rights affecting the human dignity of a natural person” is intended to clarify that a personality right must be violated, regardless of whether it is a violation of honour and reputation, privacy, the right to one's own image or another personality right. However, since it must be a question of violations by means of the representation of thought content (words and images) in an electronic communications network, a violation of the right of personality to bodily integrity, for example, will not come into consideration [...]” Human dignity is violated if the right to be a human being is denied directly or indirectly, for example by denying the right to life as an equal citizen or by portraying the person as an inferior or worthless part of the population as a whole, or if the person is otherwise subjected to inhuman or degrading treatment (cf. Plöchl in Höpfel/Ratz, WK² StGB § 283 Rz 23). In the given context, the aim is to deal with violations that cannot be justified by freedom of expression, such as when a statement no longer has any comprehensible connection to a factual dispute and is essentially only about the groundless disparagement of the person concerned as such, or in the case of swear words used with premeditation and not only in the heat of a dispute, which are, according to general opinion, particularly blatant and disparaging in themselves (such as from faecal language).”

Table E.66: Code 3-1

Field	Content
Source	FB Community Standards Section I.

Continued on next page

Field	Content
	Gutefrage.net, Community Guidelines
Qualification	Violence and incitement
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	Facebook Inc.: “We remove content, disable accounts, and work with law enforcement when we believe there is a genuine risk of physical harm or direct threats to public safety. We also try to consider the language and context in order to distinguish casual statements from content that constitutes a credible threat to public or personal safety. In determining whether a threat is credible, we may also consider additional information like a person's public visibility and the risks to their physical safety. In some cases, we see aspirational or conditional threats directed at terrorists and other violent actors (e.g., Terrorists deserve to be killed), and we deem those non credible absent specific evidence to the contrary.” Gutefrage.net: “Threat, trivialization or spreading of violence good question protects you and all users. Therefore, we expect you neither to threaten violence nor to use violence. Neither against yourself nor against others. This also includes statements, writings and media that glorify and trivialize violence. Examples: “Threat of violence unnecessarily detailed descriptions of extreme violence and fantasies of violence representation of extreme violence contributions glorifying violence harassing contributions, user stalking and mobbing trivialization of sexual, psychological or physical abuse, especially of children (this includes statements such as ‘You only convince the victims that rape is something bad’, ‘In the past it was quite normal to have sex with children’, ‘Give your child a slap in the face more often, then it will learn respect’ or similar). It is also not allowed to link to pages with corresponding content. When writing your contributions, always remember that younger users can see them.” Facebook Inc.: “Do not post: Threats that could lead to death (and other forms of high-severity violence) of any target(s) where threat is defined as any of the following: Statements of intent to commit high-severity violence; or

Continued on next page

Field	Content
	Calls for high-severity violence including content where no target is specified but a symbol represents the target and/or includes a visual of an armament to represent violence; or Statements advocating for high-severity violence; or Aspirational or conditional statements to commit high-severity violence Content that asks or offers services for hire to kill others (for example, hitmen, mercenaries, assassins) or advocates for the use of a hitman, mercenary or assassin against a target. Admissions, statements of intent or advocacy, calls to action, or aspirational or conditional statements to kidnap a target. Threats that lead to serious injury (mid-severity violence) towards private individuals, minor public figures, high risk persons, or high risk groups where threat is defined as any of the following: Statements of intent to commit violence; or Statements advocating violence; or Calls for mid-severity violence including content where no target is specified but a symbol represents the target; or Aspirational or conditional statements to commit violence; or Content about other target(s) apart from private individuals, minor public figures, high risk persons, or high risk groups and any credible: Statements of intent to commit violence; or Calls for action of violence; or Statements advocating for violence; or Aspirational or conditional statements to commit violence Threats that lead to physical harm (or other forms of lower-severity violence) towards private individuals (self-reporting required) or minor public figures where threat is defined as any of the following: Private individuals (name and/or face match are required) or minor public figures that includes: Statements of intent, calls for action, advocating, aspirational or conditional statements to commit low-severity violence Imagery of private individuals or minor public figures that has been manipulated to include threats of violence either in text or pictorial (adding bulls eye, dart, gun to head, etc.) Any content created for the express purpose of outing an individual as a member of a designated and recognizable at-risk group Instructions on how to make or use weapons if there's evidence of a goal to seriously injure or kill people, through: Language explicitly stating that goal, or Photos or videos that show or simulate the end result (serious injury or death) as part of the instruction Unless the aforementioned content is shared as part of recreational self-defence, for military training purposes, commercial video games, or news coverage (posted by Page or with news logo)
<i>Continued on next page</i>	

Field	Content
	Providing instructions on how to make or use explosives: Unless there is clear context that the content is for a non-violent purpose (for example part of commercial video games, clear scientific/educational purpose, fireworks, or specifically for fishing) Any content containing statements of intent, calls for action, or advocating for high or mid-severity violence due to voting, voter registration, or the outcome of an election Misinformation that contributes to the risk of imminent violence or physical harm. Statements of intent or advocacy, calls to action, or aspirational or conditional statements to bring weapons to locations, including but not limited to places of worship, educational facilities or polling places (or encouraging others to do the same)."

Table E.67: Code 3-2

Field	Content
Source	FB Community Standards Section I. Gutefrage.net, Community Guidelines
Qualification	Content of dangerous individuals and organizations
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	"We do not allow any organizations or individuals that proclaim a violent mission or are engaged in violence to have a presence on Facebook. This includes organizations or individuals involved in the following: Terrorist activity Organized hate Mass murder (including attempts) or multiple murder Human trafficking Organized violence or criminal activity We also remove content that expresses support or praise for groups, leaders, or individuals involved in these activities." Facebook: "We do not allow the following people (living or deceased) or groups to maintain a presence (for example, have an account, Page, Group) on our platform: Terrorist organizations and terrorists, which include: Any non-state actor that: Engages in, advocates, or lends substantial support to purposive and planned acts of violence,

Continued on next page

Field	Content
	Which causes or attempts to cause death, injury or serious harm to civilians, or any other person not taking direct part in the hostilities in a situation of armed conflict, and/or significant damage to property linked to death, serious injury or serious harm to civilians With the intent to coerce, intimidate and/or influence a civilian population, government, or international organization In order to achieve a political, religious, or ideological aim. Hate organizations and their leaders and prominent members A hate organization is defined as: Any association of three or more people that is organized under a name, sign, or symbol and that has an ideology, statements, or physical actions that attack individuals based on characteristics, including race, religious affiliation, nationality, ethnicity, gender, sex, sexual orientation, serious disease or disability. Mass and multiple murderers (including attempts) We consider a homicide to be a mass murder if it results in three or more deaths in one incident We consider an attempted mass murder to be one where an individual uses a weapon or vehicle to attempt mass harm in a public space or against more than one person We consider any individual who has committed two or more murders over multiple incidents or locations a multiple murderer Human trafficking groups and their leaders Human trafficking groups are organizations responsible for any of the following: Prostitution of others, forced/bonded labour, slavery, or the removal of organs Recruiting, transporting, transferring, detaining, providing, harbouring, or receiving a minor, or an adult against their will Criminal organizations and their leaders and prominent members A criminal organization is defined as: Any association of three or more people that is united under a name, colour(s), hand gesture(s) or recognized indicia, that has engaged in or threatens to engage in criminal activity, including (but not limited to) Homicide, Drug trafficking, Arms trafficking, Identity theft, Money laundering. Extortion or trafficking, Assault, Kidnapping, Sexual exploitation (covered in section 7 and section 8) We do not allow symbols that represent any of the above organizations or individuals to be shared on our platform without context that condemns or neutrally discusses the content. We do not allow content that praises any of the above organizations or individuals or any acts committed by them. We do not allow coordination of support for any of the above organizations or individuals or any acts committed by them. We do not allow content that praises, supports, or represents events that Facebook designates as terrorist attacks, hate crimes or mass shootings.”

Table E.68: Code 3-3

Field	Content
Source	FB Community Standards Section I. Gutefrage.net, Community Guidelines
Qualification	Coordinated harm and publicizing crime
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	“Prevent and disrupt offline harm and copycat behaviour, we prohibit people from facilitating, organizing, promoting, or admitting to certain criminal or harmful activities targeted at people, businesses, property or animals. We allow people to debate and advocate for the legality of criminal and harmful activities, as well as draw attention to harmful or criminal activity that they may witness or experience as long as they do not advocate for or coordinate harm.” Facebook: “Do not post content that falls into the following categories: Harm against people Statements of intent, calls to action, or advocating for harm against people that depicts, admits to, or promotes the following acts committed by you or your associates: Acts of physical harm against humans, including acts of domestic violence, except when shared in a context of redemption or defence of self or another person Swatting Encouraging participation in a high risk viral challenge Harm against animals Statements of intent, calls to action, or advocating for harm against animals that depicts, admits to, or promotes: Acts of physical harm against animals committed by you or your associates except in cases of hunting, fishing, religious sacrifice, food preparation or processing, pest or vermin, self-defence or redemption Staged animal vs. animal fights, including acts committed by a third party Harm against property Statements of intent, calls to action, or advocating for harm against property that depicts, admits to, or promotes the following acts committed by you or your associates: Vandalism Hacking when the intent is to hijack a domain, corrupt or disrupt cyber systems, seek ransoms, or gain unauthorized access to data systems Theft when committed by you or your associates, as well as positive statements about theft when committed by a third party Voter and/or Census Fraud

Continued on next page

Field	Content
	Offers to buy or sell votes with cash or gifts Statements that advocate, provide instructions, or show explicit intent to illegally participate in a voting or census process Voter and/or Census Interference Misrepresentation of the dates, locations, and times, and methods for voting or voter registration or census participation Misrepresentation of who can vote, qualifications for voting, whether a vote will be counted, and what information and/or materials must be provided in order to vote. Misrepresentation of whether a candidate is running or not Misrepresentation of who can participate in the census and what information or materials must be provided in order to participate Calls for coordinated interference that would affect an individual's ability to participate in the census Content stating that census or voting participation may or will result in law enforcement consequences (e.g., arrest, deportation, imprisonment) Misrepresentation of government involvement in the census, including that an individual's census information will be shared with another government agency. Other misrepresentations related to voting in an official election or census participation may be subject to false news standards, as referenced in section 20 For the following content, we include a label so that people are aware the content may be sensitive: Imagery depicting a high risk viral challenge if shared with a caption that condemns or raise awareness of the associated risks"

Table E.69: Code 3-4

Field	Content
Source	FB Community Standards Section I. Gutefrage.net, Community Guidelines
Qualification	Content on regulated goods

Continued on next page

Field	Content
Example and detailed descriptions by Facebook Inc. and Gutfrege.net, Community Guidelines	“To encourage safety and compliance with common legal restrictions, we prohibit attempts by individuals, manufacturers, and retailers to purchase, sell, or trade non-medical drugs, pharmaceutical drugs, and marijuana. We also prohibit the purchase, sale, gifting, exchange, and transfer of firearms, including firearm parts or ammunition, between private individuals on Facebook. Some of these items are not regulated everywhere; however, because of the borderless nature of our community, we try to enforce our policies as consistently as possible. Firearm stores and online retailers may promote items available for sale off of our services as long as those retailers comply with all applicable laws and regulations. We allow discussions about sales of firearms and firearm parts in stores or by online retailers and advocating for changes to firearm regulation. Regulated goods that are not prohibited by our Community Standards may be subject to our more stringent Commerce Policies. e.g., non-medical drugs, content that depicts the sale or attempt to purchase marijuana and pharmaceutical drugs, firearms, firearms parts, ammunition, explosives, trade of human blood, sale of live animals, sale of endangered species, trade of alcohol or tobacco between private individuals etc.” “Sell / request services or products gutfrege is not a sales platform. Therefore we do not allow any contributions that aim to sell or receive products, services or other things. It is the same for us whether for commercial reasons or free of charge. Examples: Who can create a logo for my YouTube channel? Have to give away winter tires for 10 Euro. Who wants to buy them? Who wants to take over my Minecraft account? Still have a voucher, who wants it?” And more general: “Illegal or unlawful contents gutfrege is open and versatile, but not a legal vacuum. Therefore we do not allow illegal content to be distributed on gutfrege. Also contributions that are intended to deliberately harm other users or companies will be deleted by our moderators. Examples: Where can I spray well? Where can I download the full version of Photoshop for free? Where can I get a weapon (without a license)? How do I get the anti-theft device off my new trousers? What do I need to make Speed? At the company XY only rip-off artists work, who have nothing in mind except cheating! How can I stop them?!

Continued on next page

Field	Content
	Please note: The violation of copyrights (e.g. by uploading a picture from the Google image search or copying texts from other websites) can have serious legal consequences. Please contact us if you are the copyright holder of illegal media.” Facebook: “Do not post: Content about non-medical drugs that Coordinates or encourages others to sell non-medical drugs Depicts, admits to, attempts purchase, or promotes sales of non-medical drugs by the poster of the content or their associates Promotes, encourages, coordinates, or provides instructions for use or make of non-medical drugs Admits, either in writing or verbally, to personal use of non-medical drugs unless posted in a recovery context Content that depicts the sale or attempt to purchase marijuana and pharmaceutical drugs. This includes content that Mentions or depicts marijuana or pharmaceutical drugs Makes an attempt to sell or trade, by which we mean any of the following: Explicitly mentioning the product is for sale or trade or delivery Asking the audience to buy Listing the price Encouraging contact about the product either by explicitly asking to be contacted or including any type of contact information Attempting to solicit the product, defined as: Stating interest in buying the product, or Asking if anyone has the product for sale/trade This applies to both individual pieces of content and Pages and Groups primarily dedicated to the sale of marijuana or pharmaceutical drugs Content that attempts to offer, sell, gift, exchange, or transfer firearms, firearm parts, ammunition, or explosives between private individuals, unless posted by a Page representing a real brick-and-mortar store, legitimate website, brand or government agency (e.g. police department, fire department). This includes content that: Mentions or depicts firearms, firearm parts, ammunition, explosives, or 3D gun printing files of any firearm or its parts*, and Makes an attempt to sell or trade, by which we mean: Explicitly mentioning the product is for sale or trade, or Asking the audience to buy, or Listing the price or noting that the product is free

Continued on next page

Field	Content
	Encouraging contact about the product either by Explicitly asking to be contacted Including any type of contact information Making an attempt to solicit the item for sale, defined as Stating that they are interested in buying the good, or Asking if anyone else has the good for sale/trade * 3D gun printing files or instructions to manufacture firearms using 3D printers or CNC milling machines, including links to websites where such files or instructions are made available, may not be shared by anyone. A Page representing a real brick-and-mortar store, legitimate website or brand can otherwise post content that may promote these items only for sale off of our services and as long as those retailers comply with all applicable laws and regulations. Content that depicts the trade (buying or selling) of human blood where trade is defined as: Mentioning or depicting the human organs and/or blood, and Indicating that human organs and/or blood are available for selling or buying, or Listing a price or expressing willingness to discuss price Content that encourages contact to facilitate the trade of human blood Content that attempts to sell live animals between private individuals. Content that coordinates or supports the poaching, selling, buying or trade/trading of endangered species and their parts Content that depicts poaching or selling of endangered species and their parts committed by you or your associates Content about endangered species: Promoting or speaking positively of the poaching or selling of endangered species and their parts. Depicts poaching or selling of endangered species and their parts committed by the poster of the content or their associates. Content that attempts to sell, poach, buy, giveaway endangered species and their parts. Content that attempts to offer, sell, gift, exchange, or transfer alcohol or tobacco products between private individuals. This includes content that: Mentions or depicts alcohol or tobacco, and Makes an attempt to sell or trade, by which we mean: Explicitly mentions the product is for sale or trade, or Asks the audience to buy, or Lists price or notes product is free And is not posted by a Page representing a real brick-and-mortar store, legitimate website or brand
	<i>Continued on next page</i>

Field	Content
	Content promoting weight loss products that include a miracle claim with a commercial offer For the following content, we restrict visibility to adults twenty-one years of age and older: Content posted by or promoting a brick-and-mortar store, legitimate website, or brand, which coordinates or promotes the sale or transfer of firearms, firearm parts, ammunition, or explosives. This includes content that Explicitly mentions the product is for sale or transfer and Asks the audience to buy the product, or Lists the price or notes that the product is free, or Encourages contact about the product either by explicitly asking to be contacted or including any type of contact information Content that depicts the sale or attempt to purchase knives, which includes content that Mentions or depicts a knife or bladed weapon, and Explicitly mentions the product is for sale or trade or delivery, or Asks the audience to buy, or Lists the price, or Asks or gives away the product for free between private individuals, or Encourages contact about the product either by: Explicitly asking to be contacted, or Including any type of contact information For the following content, we restrict visibility to adults eighteen years of age and older: Content posted by or promoting a brick and mortar store, legitimate website or brand, which coordinates or promotes the sale or transfer of alcohol or tobacco products off of Facebook's services. This includes content that: Mentions or depicts alcohol or tobacco, and Explicitly mentions the product is for sale or trade, or Asks the audience to buy, or Lists the price or notes product is free or Encourages contact about the product either by: Explicitly asking to be contacted, or Including any type of contact information Content that promotes the use or depicts the sale of weight loss products Mentions or depicts a weight loss product or supplement, and Promotes, encourages, coordinates, or provides instructions for use, or Explicitly mentions the product is for donation or sale or trade or delivery, or
<i>Continued on next page</i>	

Field	Content
	Asks the audience to buy, or Lists the price, or Gives away the product for free Promotes the use or depicts the sale of certain cosmetic procedures such as those involving body enhancement: Mentions or depicts the cosmetic procedure, and Promotes, encourages, coordinates, or provides instructions for the procedure, or Explicitly mentions the procedure is for sale, or Asks the audience to purchase the procedure, or Lists the price of the procedure, or Gives away the procedure for free.”

Table E.70: Code 3-5

Field	Content
Source	FB Community Standards Section I. Gutefrage.net, Community Guidelines
Qualification	Fraud and deception
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	“In an effort to prevent and disrupt harmful or fraudulent activity, we remove content aimed at deliberately deceiving people to gain an unfair advantage or deprive another of money, property, or legal right. However, we allow people to raise awareness and educate others as well as condemn these activities using our platform.” “Illegal or unlawful contents gutefrage is open and versatile, but not a legal vacuum. Therefore we do not allow illegal content to be distributed on gutefrage. Also contributions that are intended to deliberately harm other users or companies will be deleted by our moderators. Examples: “Where can I spray well? Where can I download the full version of Photoshop for free? Where can I get a weapon (without a license)? How do I get the anti-theft device off my new trousers? What do I need to make Speed? At the company XY only rip-off artists work, who have nothing in mind except cheating! How can I stop them?!

Continued on next page

Field	Content
	Please note: The violation of copyrights (e.g. by uploading a picture from the Google image search or copying texts from other websites) can have serious legal consequences. Please contact us if you are the copyright holder of illegal media.” Facebook: “Do not post: Content that engages in, promotes, encourages, facilitates, or admits to the following activities: Offering, solicitation and trade of: Goods/property purchased with stolen financial information Future exam papers or answer sheets where answers are not meant to be widely shared Educational and professional certificates Credentials for paid subscription services Fake user reviews Fake or manipulated documents, e.g. coupons, medical prescriptions Fake or counterfeit currency Offering or selling of personally identifiable information Coordination or facilitation of betting manipulation, e.g. match fixing Recruitment of workforce for the purpose of fraudulent activities, e.g. Social Security Number (SSN) scams Other forms of deception aimed at depriving people of money or property such as: Confidence schemes, e.g. romance/military impersonation scams Ponzi schemes or Pyramid schemes Setting up false businesses or entities Investment schemes with promise of high rates of return Bribery Embezzlement Money laundering (concealment of the origins of criminally obtained money) Charity scams or fake fundraising campaigns Debt Relief or Credit Repair Scam.”

Table E.71: Code 3-6

Field	Content
Source	Facebook Community Standards Section II and

Continued on next page

Field	Content
	Gutefrage.net, Community Guidelines
Qualification	Suicide and Self-Injury
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	“Self-injury is defined as the intentional and direct injuring of the body, including self-mutilation and eating disorders. We want Facebook to be a space where people can share their experiences, raise awareness about these issues, and seek support from one another, which is why we allow people to discuss suicide and self-injury. We have been advised by experts that we should not remove live videos of self-injury while there is an opportunity for loved ones and authorities to provide help or resources. In contrast, we remove any content that identifies and negatively targets victims or survivors of self-injury or suicide seriously, humorously, or rhetorically.” “Contributions hazardous to health The protection of our users is very important to us. Contributions that could endanger your own health or the health of other users (e.g. ‘20 paracetamol can be taken without any problems’) are therefore not allowed. Should you find yourself in a difficult life situation and are looking for support, we have some offers of help for you here: For depression and suicidal thoughts For self-harming behaviour With negative body perception & eating disorders With mobbing” Facebook: “Do not post: Content that promotes, encourages, coordinates, or provides instructions for Suicide Self-injury Eating disorders Content that depicts graphic self-injury imagery Except in limited situations of newsworthiness, it is against our policies to post content depicting a person who engaged in a suicide attempt or death by suicide Content that focuses on depiction of ribs, collar bones, thigh gaps, hips, concave stomach, or protruding spine or scapula when shared together with terms associated with eating disorders. Content that contains instructions for drastic and unhealthy weight loss when shared together with terms associated with eating disorders. For the following content, we restrict content to adults over the age of 18, and include a sensitivity screen so that people are aware the content may be upsetting: Photos or videos depicting a person's death by suicide that are determined to be newsworthy
<i>Continued on next page</i>	

Field	Content
	Photos or videos depicting a person who engaged in euthanasia/assisted suicide in a medical setting For the following content, we include a sensitivity screen so that people are aware the content may be upsetting to some: Content that depicts older instances of self-harm such as healed cuts or other non-graphic self-injury imagery in a context of recovery Content that depicts ribs, collar bones, thigh gaps, hips, concave stomach, or protruding spine or scapula in a recovery context. We provide resources to people who post written or verbal admissions of engagement in self injury, including: Suicide, Euthanasia/assisted suicide, Self-harm Eating disorders”

Table E.72: Code 3-7

Field	Content
Source	Facebook Community Standards Section II and Gutefrage.net, Community Guidelines
Qualification	Child Nudity and sexual exploitation of children
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	“We do not allow content that sexually exploits or endangers children. When we become aware of apparent child exploitation, we report it to the National Center for Missing and Exploited Children (NCMEC), in compliance with applicable law. We know that sometimes people share nude images of their own children with good intentions; however, we generally remove these images because of the potential for abuse by others and to help avoid the possibility of other people reusing or misappropriating the images. We also work with external experts, including the Facebook Safety Advisory Board, to discuss and improve our policies and enforcement around online safety issues, especially with regard to children. Learn more about the new technology we’re using to fight against child exploitation.” “Strongly offensive, obscene or vulgar posts Objectionable, vulgar or obscene contributions may not be posted on our site. Likewise prohibited are contributions that are solely aimed at the sexualization of a certain group of people (especially children), or to live out his own sexual inclinations. Not allowed are for example: Contributions with photos of genitals like: Would this penis satisfy you? Questions about porn links: Where can I find the link to the latest porn by porn star Julia?

Continued on next page

Field	Content
	Contributions with strongly offensive language
	Articles like "What makes children so attractive/sexy in your eyes?"
	Fetish questions with the intention to live out this fetish with the help of the given answers / to inspire his fantasy"
	Facebook:
	"Do not post:
	Content that depicts participation in or advocates for the sexual exploitation of children, including (but not limited to)
	Engaging in any sexual activity involving minors
	Adults soliciting minors
	Minors soliciting minors
	Minors soliciting adults
	Using our products and site functionality with the intention of sexualizing minors
	Content that constitutes or facilitates inappropriate interactions with children, such as
	Initiating unsolicited contact with minors (for example, private messages between stranger adults and minors)
	Soliciting, displaying, sharing, or viewing imagery of nude, sexualized, or sexual activity with minors
	Arranging real-world sexual encounters or obtaining sexual material from a minor directly
	Displaying nudity to minors
	Content (including photos, videos, real-world art, digital content, and text) that depicts
	Any sexual activity involving minors
	Minors in a sexual fetish context
	Minors with sexual elements, including (but not limited to):
	Restraints
	Focus on genitals
	Presence of aroused adult
	Presence of sex toys
	Sexualized costume
	Stripping
	Staged environment (for example, on a bed) or professionally shot (quality/focus/angles)
	Open-mouth kissing with minor or adult
	Content (including photos, videos, real-world art, digital content, and verbal depictions) that shows minors in a sexualized context
	Content that depicts child nudity where nudity is defined as

Continued on next page

Field	Content
	Visible genitalia (even when covered or obscured by transparent clothing) Visible anus and/or fully nude close-up of buttocks Uncovered female nipples for children older than toddler-age No clothes present from neck to knee for children older than toddler-age digitally-created depictions of nude minors, unless the image is for health or educational purposes.”

Table E.73: Code 3-8

Field	Content
Source	Facebook Community Standards Section II and Gutefrage.net, Community Guidelines
Qualification	Sexual exploitation of adults
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	“We remove content that depicts, threatens or promotes sexual violence, sexual assault, or sexual exploitation, while also allowing space for victims to share their experiences. We remove content that displays, advocates for, or coordinates sexual acts with non-consenting parties or commercial sexual services, such as prostitution and escort services. We do this to avoid facilitating transactions that may involve trafficking, coercion, and non-consensual sexual acts. To protect victims and survivors, we also remove images that depict incidents of sexual violence and intimate images shared without permission from the people pictured. We’ve written about the technology we use to protect against intimate images and the research that has informed our work. We’ve also put together a guide to reporting and removing intimate images shared without your consent.” General: “Illegal or unlawful contents gutefrage is open and versatile, but not a legal vacuum. Therefore we do not allow illegal content to be distributed on gutefrage. Also contributions that are intended to deliberately harm other users or companies will be deleted by our moderators.” Facebook: “Do not post: In instances where content consists of any form of non-consensual sexual touching, crushing, necrophilia or bestiality, or forced stripping, including: Depictions (including real photos/videos), or Advocacy (including aspirational and conditional statements), or

Continued on next page

Field	Content
	Statements of intent, or Calls for action, or Threatening, soliciting, or stating an intent to share imagery, or Admitting participation, or Mocking victims of any of the above Content that attempts to exploit people by any of the following: Coercing money, favours, or images from people by threats of exposure of their naked or semi-naked photos/videos Sharing, threatening, stating an intent to share, offering or asking for imagery known as "Revenge Porn" or "Non-Consensual Intimate Images" that fulfills all three of the following conditions: Image is non-commercial or produced in a private setting Person in the image is (near) nude, engaged in sexual activity, or in a sexual pose Lack of consent to share the image is indicated by Vengeful context (for example, caption, comments, or page title) Independent sources (for example, law enforcement record) including entertainment media (for example, leak of images confirmed by media) A visible match between the person depicted in the image and the person who has reported the content to us The person who reported the content to us shares the same name as the person depicted in the image Secretly taken non-commercial imagery of a real person's commonly sexualized body parts (breasts, groin, buttocks, or thighs) or of a real person engaged in sexual activity. This imagery is commonly known as "creepshots" or "upskirts" and includes photos or videos that mocks, sexualizes or exposes the person depicted in the imagery. Threatening or stating an intent to share private sexual conversations that meets the following criteria: Lack of consent is indicated by: Vengeful context and/or threatening context, or A visible match between the person depicted in the image and the person who has reported the content to us The person who reported the content to us shares the same name as the person depicted in the image Attempting to coordinate adult commercial sexual services or prostitution activities, such as requesting or offering or asking for rates for escort services and paid sexual fetish or domination services. For the following content, we include a warning screen so that people are aware the content may be disturbing: Narratives and statements that contain a depiction of non-consensual sexual touching (written or verbal) that includes details beyond mere naming or mentioning the act if: Shared by the victim, or

Continued on next page

Field	Content
	Shared by a third party (other than the victim) in support of victim or condemnation of act or for general awareness to be determined by context/caption Content mocking the concept of non-consensual sexual touching."

Table E.74: Code 3-9

Field	Content
Source	Facebook Community Standards Section II and Gutefrage.net, Community Guidelines
Qualification	Bullying and Harassment
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	"Bullying and harassment happen in many places and come in many different forms, from making threats to releasing personally identifiable information, to sending threatening messages, and making unwanted malicious contact. We do not tolerate this kind of behaviour because it prevents people from feeling safe and respected on Facebook. We distinguish between public figures and private individuals because we want to allow discussion, which often includes critical commentary of people who are featured in the news or who have a large public audience. For public figures, we remove attacks that are severe as well as certain attacks where the public figure is directly tagged in the post or comment. For private individuals, our protection goes further: we remove content that's meant to degrade or shame, including, for example, claims about someone's sexual activity. We recognize that bullying and harassment can have more of an emotional impact on minors, which is why our policies provide heightened protection for users between the ages of 13 and 18. Context and intent matter, and we allow people to share and re-share posts if it is clear that something was shared in order to condemn or draw attention to bullying and harassment. In certain instances, we require self-reporting because it helps us understand that the person targeted feels bullied or harassed. In addition to reporting such behaviour and content, we encourage people to use tools available on Facebook to help protect against it. We also have a Bullying Prevention Hub, which is a resource for teens, parents, and educators seeking support for issues related to bullying and other conflicts. It offers step-by-step guidance, including information on how to start important conversations about bullying. Learn more about what we're doing to protect people from bullying and harassment here." "Insults & user demonstration So that everyone can feel comfortable in our community, insults as well as mutual presentation / ridiculing are not allowed (no matter whether publicly or e.g. by private message). This includes among others: Insinuations

Continued on next page

Field	Content
	accusations (e.g. "user xy is a troll") Tracking" individual users & their contributions with negative intent (for example, with the intention of exposing the user)" "Threat, trivialization or spreading of violence good question protects you and all users. Therefore we expect you neither to threaten violence nor to use violence. Neither against yourself nor against others. This also includes statements, writings and media that glorify and trivialize violence. Examples: Threat of violence unnecessarily detailed descriptions of extreme violence and fantasies of violence Representation of extreme violence contributions glorifying violence harassing contributions, user stalking and mobbing trivialization of sexual, psychological or physical abuse, especially of children (this includes statements such as "You only convince the victims that rape is something bad") "In the past it was quite normal to have sex with children" Give your child a slap in the face more often, then it will learn respect" or similar). It is also not allowed to link to pages with corresponding content. When writing your contributions, always remember that younger users can see them." Facebook: "Do not: Repeatedly contact someone in a manner that is: Unwanted or Sexually harassing or Directed at a large number of individuals with no prior solicitation Target anyone maliciously by: Attacking them based on their status as a victim of sexual assault, sexual exploitation, or domestic abuse Calling for self-injury or suicide of a specific person, or group of people Attacking them through derogatory terms related to sexual activity (e.g. whore, slut) Posting content about a violent tragedy, or victims of violent tragedies that include claims that a violent tragedy did not occur Posting content about victims or survivors of violent tragedies by name or by image, with claims that they are Acting/pretending to be a victim of an event

Continued on next page

Field	Content
	Otherwise paid or employed to mislead people about their role in the event Threaten to release an individual's private phone number, residential address, or email address Creating or posting content to Pages or Groups that are dedicated to attacking individual(s) by, for example Calling for death or serious disease or disability Making statements of intent of advocating to engage in sexual activity Claims about sexually transmitted diseases Sending messages that contain the following attacks when aimed at an individual or group of individuals in the thread Targeted cursing Calls for death, serious disease or disability, or physical harm Target another private individual, or public figure who is a minor, with: Calls for death or serious disease or disability Female-gendered cursing terms when used in a derogatory way Claims about sexual activity or sexually transmitted diseases Pages or Groups created to attack through: Targeted cursing High-severity physical descriptions Claims about religious identity or blasphemy Expressions of contempt or disgust Female-gendered cursing terms when used in a derogatory way Target public figures by purposefully exposing them to: For adults Calls for death or serious disease or disability Statements of intent to engage in a sexual activity or advocating to engage in a sexual activity Claims about sexually transmitted disease Female-gendered cursing terms when used in a derogatory way Content that praises, celebrates, or mocks their death For minors Comparisons to animals or insects that are culturally perceived as intellectually or physically inferior or to an inanimate object ("cow," "monkey" "potato") Content manipulated to highlight, circle, or otherwise negatively draw attention to specific physical characteristics (nose, ear, etc.)
<i>Continued on next page</i>	

Field	Content
	Target private individuals or involuntary minor public figures with: Comparisons to animals or insects that are culturally perceived as intellectually or physically inferior or to an inanimate object ("cow," "monkey" "potato") Content manipulated to highlight, circle, or otherwise negatively draw attention to specific physical characteristics (nose, ear, etc.) Attacks through negative physical descriptions Content that ranks individuals on physical appearance or personality Content that depicts others in the process of, or right after menstruating, urinating, vomiting, or defecating where context further degrades the individual or contains an expression of disgust Physical bullying where the context further degrades the individual In addition to the above, attacks through Pages or Groups: Negative character or ability claims Self-referential bullying; Only if the Object targets more than 1 private individual Content that praises, celebrates, or mocks their death Target private adults (who must self-report) or any private minors or involuntary minor public figures with: Targeted cursing Claims about romantic involvement, sexual orientation or gender identity Coordination, advocacy, or promotion of exclusion Negative character or ability claims, except in the context of criminal allegations against adults Expressions of contempt or disgust, except in the context of criminal allegations against adults Target private individuals who are minors with: Allegations about criminal or illegal behaviour Videos of physical bullying or violence against minors in a fight context shared in a non-condemning context Target private individuals (who must self-report) with: Self-referential bullying Unwanted manipulated imagery Comparison to other public, fictional or private individuals on the basis of physical appearance Claims about religious identity or blasphemy, except in at-risk countries where the Violence and Incitement policy should be applied Comparisons to animals or insects that are not culturally perceived as intellectually or physically inferior ("tiger," "lion") Neutral or positive physical descriptions Non-negative character or ability claims
<i>Continued on next page</i>	

Field	Content
	Any bullying or harassment violation, when shared in an endearing context Attacks through derogatory terms related to a lack of sexual activity For the following content, we include a warning screen so that people are aware the content may be disturbing: Videos of physical bullying or violence against minors in a fight context shared with a condemning caption.”

Table E.75: Code 3-10

Field	Content
Source	Facebook Community Standards Section II and Gutefrage.net, Community Guidelines
Qualification	Human exploitation
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	“In an effort to disrupt and prevent harm, we remove content that facilitates or coordinates the exploitation of humans, including human trafficking. We define human trafficking as the business of depriving someone of liberty for profit. It is the exploitation of humans in order to force them to engage in commercial sex, labour, or other activities against their will. It relies on deception, force and coercion, and degrades humans by depriving them of their freedom while economically or materially benefiting others. Human trafficking is multi-faceted and global; it can affect anyone regardless of age, socioeconomic background, ethnicity, gender, or location. It takes many forms, and any given trafficking situation can involve various stages of development. By the coercive nature of this abuse, victims cannot consent. While we need to be careful not to conflate human trafficking and smuggling, the two can be related and exhibit overlap. The United Nations defines human smuggling as the procurement or facilitation of illegal entry into a state across international borders. Without necessity for coercion or force, it may still result in the exploitation of vulnerable individuals who are trying to leave their country of origin, often in pursuit of a better life. Human smuggling is a crime against a state, relying on movement, and human trafficking is a crime against a person, relying on exploitation.” “Illegal or illegal contents gutefrage is open and versatile, but not a legal vacuum. Therefore we do not allow illegal content to be distributed on gutefrage. Also contributions that are intended to deliberately harm other users or companies will be deleted by our moderators.” Facebook: “Do not post: Content or exhibit behaviour that may lead to human exploitation, which includes any of the following:

Continued on next page

Field	Content
	Sex Trafficking, covering minors and adults The sale of children for illegal adoption Orphanage Trafficking and Orphanage Voluntourism Forced Marriages Labour Exploitation (incl. bonded labour) Domestic Servitude Non-regenerative Organ Trafficking Forced Criminal Activity (e.g. forced begging, forced drug trafficking) Recruitment of Child Soldiers Content geared towards the: Recruitment of potential victims through force, fraud, coercion, enticement, deception, blackmail or other non-consensual acts. Facilitation of human exploitation by coordinating, transporting, transferring, harbouring or brokering of victims prior or during the exploitation. Exploitation of humans by promoting, depicting, or advocating for it. Content that offers or assists in smuggling of humans.”

Table E.76: Code 3-11

Field	Content
Source	Facebook Community Standards Section II and Gutefrage.net, Community Guidelines
Qualification	Privacy violations and image privacy rights
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	Privacy and the protection of personal information are fundamentally important values for Facebook. We work hard to keep your account secure and safeguard your personal information in order to protect you from potential physical or financial harm. You should not post personal or confidential information about others without first getting their consent. We also provide people ways to report imagery that they believe to be in violation of their privacy rights. “Illegal or illegal contents

Continued on next page

Field	Content
	gutefrage is open and versatile, but not a legal vacuum. Therefore we do not allow illegal content to be distributed on gutefrage. Also contributions that are intended to deliberately harm other users or companies will be deleted by our moderators.” Facebook: “Do not post: Content that facilitates identity theft by posting or soliciting personally identifiable information, including (but not limited to) National identification numbers, Social Security numbers, passport numbers, or exam numbers Government IDs School and education IDs featuring two of the following: (1) name, (2) photo, or (3) ID number Digital identities, including passwords Content that contains medical/psychological, biometric, or genetic records or official documentation of others Content that facilitates identity theft by sharing personally identifiable information via an external link Content that facilitates identity theft by sharing private financial information of an organization or business Content that facilitates identity theft by disclosing the following personal financial information (of either the self or others) Bank account and/or card information Financial records paired with account information Content that facilitates identity theft by sharing the private contact information of others defined as Private phone numbers or addresses Email, Messenger, and chat identities The above information may be shared to promote charitable causes, non-violating services, or to facilitate finding missing people or animals Except in limited cases of newsworthiness, content claimed or confirmed to come from a hacked source, regardless of whether the affected person is a public figure or a private individual. Content that identifies individuals by name and depicts their personal information, including: Driver’s licenses, Government IDs other than driver’s licenses, Green Cards, or immigration papers Marriage, birth, and name change certificates Digital identities, including passwords License plates Content that includes photographs that display the external view of private residences if the following conditions apply: The residence is a single-family home, or the resident’s unit number is identified in the image/caption
	<i>Continued on next page</i>

Field	Content
	The city or neighbourhood is identified A resident is mentioned or depicted That same resident objects to the exposure of their private residence Content that exposes the undercover status of law enforcement personnel if The content contains the agent's full name or other explicit identification and explicitly mentions their undercover status, or The content contains images identifying the faces of the law enforcement personnel and explicitly mentions their undercover status Content that exposes information about safe houses by sharing any of the below, unless the safe house is actively promoting its location, contact information, or the type of service and protection it offers through comments, posts, Pages or Groups: Actual address of the safe house (post box only is allowed), Images of the safe house, Identifiable city/neighbourhood of the safe house, or Information outing residents of the safe house The following content also may be removed A reported photo or video of people where the person depicted in the image is: A minor under thirteen years old, and the content was reported by the minor or a parent or legal guardian, A minor between thirteen and eighteen years old, and the content was reported by the minor, An adult, where the content was reported by the adult from outside the United States and applicable law may provide rights to removal Any person who is incapacitated and unable to report the content on their own."

Table E.77: Code 3-12

Field	Content
Source	Facebook Community Standards Section III Gutefrage.net Community Guidelines
Qualification	Hate Speech

Continued on next page

Field	Content
Example and detailed descriptions by Facebook Inc. and Gutfreage.net, Community Guidelines	"We define hate speech as a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We also provide some protections for immigration status. We define attack as violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation. We separate attacks into three tiers of severity, as described below. Sometimes people share content containing someone else's hate speech for the purpose of raising awareness or educating others. In some cases, words or terms that might otherwise violate our standards are used self-referentially or in an empowering way. People sometimes express contempt in the context of a romantic break-up. Other times, they use gender-exclusive language to control membership in a health or positive support group, such as a breastfeeding group for women only. In all of these cases, we allow the content but expect people to clearly indicate their intent, which helps us better understand why they shared it. Where the intention is unclear, we may remove the content. We allow humour and social commentary related to these topics. In addition, we believe that people are more responsible when they share this kind of commentary using their authentic identity." "Extremism, Xenophobia and Hate-Speech good question is cosmopolitan. We therefore tolerate neither discrimination, defamation nor extremism in any form. gutfreage is not a place for spreading hate and incitement to hatred Examples of hate contribution: Conscious dissemination of false statements "Refugees all get expensive cell phones for free!" "Refugees may steal from supermarkets, they won't be punished anyway." Camouflage as humour / irony "In my next life I'll be an asylum seeker, then I'll get a new smartphone!" Degrading and denigrating terms / sexist and racist insults "Kan**ake", "Faggot.", "Bitch.", "Headscarf Girl", "gold piece", "cultural enricher", "knife" in relation to (male) refugees Striking use of stereotypes / use of certain terms and language patterns "The pharmaceutical lobby wants to vaccinate us all sick.", "Foreigners out!", "Islamization of the German Occident", "The state wants to raise our children as homos!", "Sacrifice subscription", Devaluating / Negative Generalization, "Jews are profit hungry", "Carnivores are murderers" Striking visual language Profile pictures (photos, cartoons, etc.) with racist, sexist or other degrading images Images that reproduce stereotypes, for example by associating Muslim men with bestiality Criminal equations
<i>Continued on next page</i>	

Field	Content
	"It's a known fact that most homosexuals also molest children.", "No wonder that so many cars are stolen in Brandenburg - it's close to the Polish border." Advocacy / trivialization or threat of sexualized violence "You should give them a real hard time!", "If they touch my wife, I'll cut their balls off myself!", "Women shouldn't complain about sexism when they're so slutty!" advocacy of or call for acts of violence "They should all be shot/burned/gasified.", "Take them to the gallows!", "I vote we open the gas chambers again and put the whole brood in there." We therefore do not allow: to attack persons or groups of persons verbally or visually, e.g. through videos and pictures - on the basis of their ethnic and religious affiliation, national origin, sexual orientation, gender (identity), social status, world view, lifestyle, physical or mental disability, as well as illness. An attack also includes the use of derogatory or violent language, defaming users in any way, and deliberately spreading prejudice and misinformation. To spread or quote hate contributions of other people. Use certain language patterns that use stereotypes to devalue others. To write tendentious contributions whose only aim is to convince other users of the truth of their own views. What is good question about questions on political issues? Please always remember that your anonymity on the net does not protect you from punishment. Spreading hate is not an expression in the sense of freedom of opinion. There can be criminal offences; for this purpose, StGB § 111 Public incitement to commit criminal offences, § 130 Sedition, § 185 Insult, § 186 Defamation and slander and § 187 Defamation apply. We reserve the right to take legal action or to pass on such incidents to the authorities". Facebook: "Do not post: Tier 1 Content targeting a person or group of people (including all subsets except those described as having carried out violent crimes or sexual offenses) on the basis of their aforementioned protected characteristic(s) or immigration status with: Violent speech or support in written or visual form Dehumanizing speech or imagery in the form of comparisons, generalizations, or unqualified behavioural statements to or about: Insects Animals that are culturally perceived as intellectually or physically inferior Filth, bacteria, disease and feces

Continued on next page

Field	Content
	Sexual predator Subhumanity Violent and sexual criminals Other criminals (including but not limited to “thieves,” “bank robbers,” or saying “all [protected characteristic or quasi-protected characteristic] are ‘criminals’”) Statements denying existence Mocking the concept, events or victims of hate crimes even if no real person is depicted in an image Designated dehumanizing comparisons, generalizations, or behavioural statements (in written or visual form)- that include: Black people and apes or ape-like creatures Black people and farm equipment Jewish people and rats Muslim people and pigs Muslim person and sexual relations with goats or pigs Mexican people and worm like creatures Women as household objects or referring to women as property or ‘objects’ Transgender or non-binary people referred to as ‘it’ Tier 2 Content targeting a person or group of people on the basis of their protected characteristic(s) with: Generalizations that state inferiority (in written or visual form) in the following ways: Physical deficiencies are defined as those about: Hygiene, including but not limited to: filthy, dirty, smelly Physical appearance, including but not limited to: ugly, hideous Mental deficiencies are defined as those about: Intellectual capacity, including but not limited to: dumb, stupid, idiots Education, including but not limited to: illiterate, uneducated Mental health, including but not limited to: mentally ill, retarded, crazy, insane Moral deficiencies are defined as those about: Culturally perceived negative character trait, including but not limited to: coward, liar, arrogant, ignorant Derogatory terms related to sexual activity, including but not limited to: whore, slut, perverts Other statements of inferiority, which we define as:
<i>Continued on next page</i>	

Field	Content
	Expressions about being less than adequate, including but not limited to: worthless, useless Expressions about being better/worse than another protected characteristic, including but not limited to: "I believe that males are superior to females." Expressions about deviating from the norm, including but not limited to: freaks, abnormal Expressions of contempt or their visual equivalent, which we define as: Self-admission to intolerance on the basis of a protected characteristics, including but not limited to: homophobic, islamophobic, racist Expressions that a protected characteristic shouldn't exist Expressions of hate, including but not limited to: despise, hate Expressions of dismissal, including but not limited to: don't respect, don't like, don't care for Expressions of disgust or their visual equivalent, which we define as: Expressions that suggest the target causes sickness, including but not limited to: vomit, throw up Expressions of repulsion or distaste, including but not limited to: vile, disgusting, yuck Cursing, defined as: Referring to the target as genitalia or anus, including but not limited to: cunt, dick, asshole Profane terms or phrases with the intent to insult, including but not limited to: fuck, bitch, motherfucker Terms or phrases calling for engagement in sexual activity, or contact with genitalia or anus, or with feces or urine, including but not limited to: suck my dick, kiss my ass, eat shit Tier 3 Content targeting a person or group of people on the basis of their protected characteristic(s) with any of the following: Calls for segregation Explicit Exclusion which includes but is not limited to "expel" or "not allowed". Political Exclusion defined as denial of right to political participation. Economic Exclusion defined as denial of access to economic entitlements and limiting participation in the labour market, Social Exclusion defined as including but not limited to denial of opportunity to gain access to spaces (incl. online) and social services. We do allow criticism of immigration policies and arguments for restricting those policies. Content that describes or negatively targets people with slurs, where slurs are defined as words commonly used as insulting labels for the above-listed characteristics."

Table E.78: Code 3-13

Field	Content
Source	Facebook Community Standards Section III Gutefrage.net Community Guidelines
Qualification	Violent Graphic Content
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	“We remove content that glorifies violence or celebrates the suffering or humiliation of others because it may create an environment that discourages participation. We allow graphic content (with some limitations) to help people raise awareness about issues. We know that people value the ability to discuss important issues like human rights abuses or acts of terrorism. We also know that people have different sensitivities with regard to graphic and violent content. For that reason, we add a warning label to especially graphic or violent content so that it is not available to people under the age of eighteen and so that people are aware of the graphic or violent nature before they click to see it.” “Strongly offensive, obscene or vulgar posts Objectionable, vulgar or obscene contributions may not be posted on our site. Likewise prohibited are contributions that are solely aimed at the sexualization of a certain group of people (especially children), or to live out his own sexual inclinations. Not allowed are for example: Contributions with photos of genitals like: Would this penis satisfy you? Questions about porn links: Where can I find the link to the latest porn by porn star Julia? Contributions with strongly offensive language Articles like "What makes children so attractive/sexy in your eyes? Fetish questions with the intention to live out this fetish with the help of the given answers / to inspire his fantasy” “Threat, trivialization or spreading of violence good question protects you and all users. Therefore we expect you neither to threaten violence nor to use violence. Neither against yourself nor against others. This also includes statements, writings and media that glorify and trivialize violence. Examples: Threat of violence unnecessarily detailed descriptions of extreme violence and fantasies of violence Representation of extreme violence contributions glorifying violence harassing contributions, user stalking and mobbing

Continued on next page

Field	Content
	trivialization of sexual, psychological or physical abuse, especially of children (this includes statements such as "You only convince the victims that rape is something bad") "In the past it was quite normal to have sex with children" Give your child a slap in the face more often, then it will learn respect" or similar). It is also not allowed to link to pages with corresponding content. When writing your contributions, always remember that younger users can see them." Facebook: "Do not post: Videos of people or dead bodies in non-medical settings if they depict Dismemberment Visible internal organs; partially decomposed bodies Charred or burning people unless in the context of cremation or self-immolation when that action is a form of political speech or newsworthy Victims of cannibalism Throat-slitting Live streams of capital punishment of a person Videos or photos that depict non-sexual child abuse, defined as Repeated kicking, beating, slapping, or stepping on by an adult or animal Strangling or suffocating by an adult or animal Drowning by an adult or animal Biting through skin by an adult or animal Poisoning by an adult Forcible restraint by an adult Inflicting of burn or cut wounds by an adult Forcible smoking Tossing, rotating, or shaking of an infant (too young to stand) by their wrists/ankles, arms/legs, or neck."

Table E.79: Code 3-14

Field	Content
Source	Facebook Community Standards Section III Gutefrage.net Community Guidelines

Continued on next page

Field	Content
Qualification	Adult Nudity and Sexual Activity
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	We restrict the display of nudity or sexual activity because some people in our community may be sensitive to this type of content. Additionally, we default to removing sexual imagery to prevent the sharing of non-consensual or underage content. Restrictions on the display of sexual activity also apply to digitally created content unless it is posted for educational, humorous, or satirical purposes. “Strongly offensive, obscene or vulgar posts Objectionable, vulgar or obscene contributions may not be posted on our site. Likewise prohibited are contributions that are solely aimed at the sexualization of a certain group of people (especially children), or to live out his own sexual inclinations. Not allowed are for example: Contributions with photos of genitals like: Would this penis satisfy you? Questions about porn links: Where can I find the link to the latest porn by porn star Julia? Contributions with strongly offensive language Articles like "What makes children so attractive/sexy in your eyes? Fetish questions with the intention to live out this fetish with the help of the given answers / to inspire his fantasy” Facebook: “For example, while we restrict some images of female breasts that include the nipple, we allow other images, including those depicting acts of protest, women actively engaged in breast-feeding, and photos of post-mastectomy scarring. We also allow photographs of paintings, sculptures, and other art that depicts nude figures.”

Table E.80: Code 3-15

Field	Content
Source	Facebook Community Standards Section III Gutefrage.net Community Guidelines
Qualification	Sexual Solicitation

Continued on next page

Field	Content
Example and detailed descriptions by Facebook Inc. and Gutterfringe.net, Community Guidelines	“As noted in Section 8 of our Community Standards (Sexual Exploitation of Adults), people use Facebook to discuss and draw attention to sexual violence and exploitation. We recognize the importance of and want to allow for this discussion. We draw the line, however, when content facilitates, encourages or coordinates sexual encounters between adults. We also restrict sexually explicit language that may lead to solicitation because some audiences within our global community may be sensitive to this type of content and it may impede the ability for people to connect with their friends and the broader community.” “Strongly offensive, obscene or vulgar posts Objectionable, vulgar or obscene contributions may not be posted on our site. Likewise prohibited are contributions that are solely aimed at the sexualization of a certain group of people (especially children), or to live out his own sexual inclinations. Not allowed are for example: Contributions with photos of genitals like: Would this penis satisfy you? Questions about porn links: Where can I find the link to the latest porn by porn star Julia? Contributions with strongly offensive language Articles like "What makes children so attractive/sexy in your eyes? Fetish questions with the intention to live out this fetish with the help of the given answers / to inspire his fantasy” Facebook: “Do not post: Attempted coordination of or recruitment for adult sexual activities, including but not limited to: Filmed sexual activities Pornographic activities, strip club shows, live sex performances, erotic dances Sexual, erotic, or tantric massages Explicit sexual solicitation by, including but not limited to the following, offering or asking for: Sex or sexual partners Sex chat or conversations Nude photos/videos/imagery Content that meets both of the following criteria: Criteria 1: Offer or Ask Content implicitly or indirectly* (typically through providing a method of contact) offers or asks for: Nude imagery, or Sex or sexual partners, or Sex chat conversations

Continued on next page

Field	Content
	Criteria 2: Suggestive Elements Content makes the aforementioned offer or ask using one of the following sexually suggestive elements: Contextually specific and commonly sexual emojis or emoji strings, or Regional sexualized slang, or Mentions or depictions of sexual activity (including hand drawn, digital, or real world art) such as: sexual roles, sex positions, fetish scenarios, state of arousal, act of sexual intercourse or activity (sexual penetration or self-pleasuring), or Imagery of real individuals with nudity covered by human parts, objects, or digital obstruction, including long shots of fully nude butts Content must meet Criteria 1 (offer or ask) and be implicitly or indirectly offering or asking for sexual solicitation in order to be deemed violating. For example, if content is a hand-drawn image depicting sexual activity but does not ask or offer sexual solicitation, it is not violating. An offer or ask for pornographic material (including, but not limited to, sharing of links to external pornographic websites) Sexually explicit language that goes into graphic detail beyond mere reference to: A state of sexual arousal (wetness or erection) or An act of sexual intercourse (sexual penetration, self-pleasuring or exercising fetish scenarios)".

Table E.81: Code 3-16

Field	Content
Source	Facebook Community Standards Section III Gutefrage.net Community Guidelines
Qualification	Cruel and insensitive content
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	"We believe that people share and connect more freely when they do not feel targeted based on their vulnerabilities. As such, we have higher expectations for content that we call cruel and insensitive, which we define as content that targets victims of serious physical or emotional harm. We remove explicit attempts to mock victims and mark as cruel implicit attempts, many of which take the form of memes and GIFs." "Strongly offensive, obscene or vulgar posts

Continued on next page

Field	Content
	Objectionable, vulgar or obscene contributions may not be posted on our site. Likewise prohibited are contributions that are solely aimed at the sexualization of a certain group of people (especially children), or to live out his own sexual inclinations. Not allowed are for example: Contributions with photos of genitals like: Would this penis satisfy you? Questions about porn links: Where can I find the link to the latest porn by porn star Julia? Contributions with strongly offensive language Articles like "What makes children so attractive/sexy in your eyes?" Fetish questions with the intention to live out this fetish with the help of the given answers / to inspire his fantasy" Facebook: "Do not post: Content that depicts real people and laughs at or makes fun of their premature death, serious or fatal disease or disability, starvation, physical violence, domestic violence, or serious physical injury Content that contains sadistic remarks and any visual or written depiction of real people experiencing premature death, serious physical injury, physical violence or domestic violence. Imagery that depicts real animals that are visibly experiencing and being laughed at, made fun of, or contain sadistic remarks for any of the following (except staged animal vs. animal fights or animal fights in the wild): premature death serious physical injury (including mutilation) physical violence from a human."

Table E.82: Code 3-17

Field	Content
Source	Facebook Community Standards Section IV Gutefrage.net Community Guidelines
Qualification	Misrepresentation

Continued on next page

Field	Content
Example and detailed descriptions by Facebook Inc. and Gutsyfringe.net, Community Guidelines	“Authenticity is the cornerstone of our community. We believe that people are more accountable for their statements and actions when they use their authentic identities. That’s why we require people to connect on Facebook using the name they go by in everyday life. Our authenticity policies are intended to create a safe environment where people can trust and hold one another accountable.” “Contributions that are not meant seriously (troll contributions), deliberately contain false information or mislead users are not allowed.” Facebook: “Do not: Misrepresent your identity by Using a name that does not abide by our name policies Providing a false date of birth Misuse our products by Creating a profile for someone under thirteen years old Maintaining multiple accounts Creating inauthentic profiles Sharing an account with any other person Creating another Facebook account after being banned from the site Creating or Managing a Page, Group, Event or Instagram Profile because the previous Page, Group, Event or Instagram Profile was removed from the site Evading the registration requirements outlined in our Terms of Service Impersonate others by Using their images with the explicit aim to deceive people Creating a profile assuming the persona of or speaking for another person or entity Creating a Page assuming to be or speak for another person or entity for whom the user is not authorized to do so. Posting imagery that is likely to deceive the public as to the content’s origin, if: The entity or an authorized representative object to the content, and Can establish a risk of harm to members of the public.”

Table E.83: Code 3-18

Field	Content
Source	Facebook Community Standards Section IV

Continued on next page

Field	Content
	Gutefrage.net Community Guidelines
Qualification	Spam
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	“We work hard to limit the spread of spam because we do not want to allow content that is designed to deceive, or that attempts to mislead users to increase viewership. This content creates a negative user experience and detracts from people’s ability to engage authentically in online communities. We also aim to prevent people from abusing our platform, products, or features to artificially increase viewership or distribute content en masse for commercial gain.” “Publish spam articles We consider contributions that advertise products, offers or services as unauthorized advertising (spam) and delete them. By the way, this also applies to external surveys, petitions or, for example, advertising a private YouTube channel. Other, repetitive content can also be considered spam. For example, multiple posting of the same answer or fake / manipulated posts. It does not matter whether these contributions are posted publicly or, for example, by private message.” Facebook: “Do not: Post, share, engage with content, or create accounts, groups, pages, events or other assets, either manually or automatically, at very high frequencies. Attempt to or successfully sell, buy, or exchange site privileges or product features, such as accounts, admin roles, permission to post, pages, groups, likes, etc., except in the case of clearly identified branded content, as defined by our Branded Content Policy. Require or claim that users are required to engage with content (e.g. liking, sharing) before they are able to view or interact with promised content. Encourage likes, shares, follows, clicks, or the use of apps or websites under false pretences, such as: Offering false or non-existent services or functionality(e.g., “Get a ‘Dislike’ button!”) Failing to direct to promised content (e.g., “Click here for a discount code at Nordstrom”; false play buttons) Directing users off of Facebook through the deceptive or misleading use of URLs, defined as: Cloaking: Presenting different content to Facebook users and Facebook crawlers or tools. Misleading content: Content contains a link that promises one type of content but delivers something substantially different. Deceptive Pop-up websites: Websites that require an action (captcha, watch ad, click here) in order to view the expected landing page content and the domain name of the URL changes after the required action is complete. Like / Share-Gating: Landing pages that require users to like, share, or otherwise engage with content before gaining access to off-Facebook content.
<i>Continued on next page</i>	

Field	Content
	Deceptive Landing Page Functionality: Websites that have a misleading user interface, which results in accidental traffic being generated (pop-ups/unders, clickjacking, etc.) Impersonation: The website pretends to be a reputable brand or service by using a name, domain, or content featuring typos, misspellings, or other means to impersonate well-known brands www.facebook.com, www.facebook.com, using a landing page similar to another, trusted site to mislead visitors). And other behaviours that are substantially similar to those listed above.”

Table E.84: Code 3-19

Field	Content
Source	Facebook Community Standards Section IV Gutefrage.net Community Guidelines
Qualification	Cybersecurity
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	“We recognize that the safety of our users extends to the security of their personal information. Attempts to gather sensitive personal information by deceptive or invasive methods are harmful to the authentic, open, and safe atmosphere that we want to foster. Therefore, we do not allow attempts to gather sensitive user information through the abuse of our platform and products.” No matching stipulation in the community guidelines of Gutefrage.net. Facebook: “Do not: Attempt to compromise user accounts or gather sensitive information through unauthorized means, defined as: Gaining access to any account or user data other than your own without explicit permission from the account owner. Encouraging or deceiving users and entities to download or run files or programs that will compromise a user's online or data security. Such files and programs will be deemed malicious software or "malware" if they harm or gain unauthorized access to a computer, device, or network. Acquiring or requesting another user's login credentials, whether explicitly or through deceptive means such as phishing (e.g. fake surveys designed to capture log-in info or links to fake login pages or impostor websites). Attempting to obtain the sensitive information of others such as usernames, passwords, or other personal information, through deceptive means or the use of malicious software or websites. Publicly sharing your own or others' login information on platform or through a third party service.

Continued on next page

Field	Content
	Creating, sharing, or hosting malicious software or malicious browser extensions to compromise accounts or gain access to user data. Providing online infrastructure, including Web Hosting Services, Domain Name System servers, and ad networks that enables abusive links such that a majority of those links on Facebook or Instagram violate the Spam or Cybersecurity sections of the Community Standards.”

Table E.85: Code 3-20

Field	Content
Source	Facebook Community Standards Section IV Gutefrage.net Community Guidelines
Qualification	Inauthentic Behaviour
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	“In line with our commitment to authenticity, we don't allow people to misrepresent themselves on Facebook, use fake accounts, artificially boost the popularity of content, or engage in behaviours designed to enable other violations under our Community Standards. This policy is intended to create a space where people can trust the people and communities they interact with.” “Contributions that are not meant seriously (troll contributions), deliberately contain false information or mislead users are not allowed.” Facebook: “Do not: Use multiple Facebook accounts or share accounts between multiple people, Misuse Facebook or Instagram reporting systems to harass others, Conceal a Page's purpose by misleading users about the ownership or control of that Page, Engage in or claim to engage in Inauthentic Behaviour, which is defined as the use of Facebook or Instagram assets (accounts, pages, groups, or events), to mislead people or Facebook: - about the identity, purpose, or origin of the entity that they represent - about the popularity of Facebook or Instagram content or assets - about the purpose of an audience or community - about the source or origin of content to evade enforcement under our Community Standards Engage in, or claim to engage in Coordinated Inauthentic Behaviour, defined as the use of multiple Facebook or Instagram assets, working in concert to engage in Inauthentic Behaviour (as defined above), where the use of fake accounts is central to the operation

Continued on next page

Field	Content
	Engage in or claim to engage in Foreign or Government Interference, which is Coordinated Inauthentic Behaviour conducted on behalf of a foreign or government actor.

Table E.86: Code 3-21

Field	Content
Source	Facebook Community Standards Section IV Gutefrage.net Community Guidelines
Qualification	False News
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	“Reducing the spread of false news on Facebook is a responsibility that we take seriously. We also recognize that this is a challenging and sensitive issue. We want to help people stay informed without stifling productive public discourse. There is also a fine line between false news and satire or opinion. For these reasons, we don't remove false news from Facebook but instead, significantly reduce its distribution by showing it lower in the News Feed.” “We are working to build a more informed community and reduce the spread of false news in a number of different ways, namely by - Disrupting economic incentives for people, Pages, and domains that propagate misinformation - Using various signals, including feedback from our community, to inform a machine learning model that predicts which stories may be false - Reducing the distribution of content rated as false by independent third-party fact-checkers - Empowering people to decide for themselves what to read, trust, and share by informing them with more context and promoting news literacy - Collaborating with academics and other organizations to help solve this challenging issue.”

Table E.87: Code 3-22

Field	Content
Source	Facebook Community Standards Section IV Gutefrage.net Community Guidelines
Qualification	Manipulated Media

Continued on next page

Field	Content
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	“Media, including image, audio, or video, can be edited in a variety of ways. In many cases, these changes are benign, like a filter effect on a photo. In other cases, the manipulation isn’t apparent and could mislead, particularly in the case of video content. We aim to remove this category of manipulated media when the criteria laid out below have been met.” No matching stipulation in the community guidelines of Gutefrage.net. Facebook: “Do not post: Video that has been edited or synthesized, beyond adjustments for clarity or quality, in ways that are not apparent to an average person, and would likely mislead an average person to believe that a subject of the video said words that they did not say and is the product of artificial intelligence or machine learning, including deep learning techniques (e.g., a technical deepfake), that merges, combines, replaces, and/or superimposes content onto a video, creating a video that appears authentic. This policy does not extend to content that is parody or satire or is edited to omit words that were said or change the order of words that were said.

Table E.88: Code 3-23

Field	Content
Source	Facebook Community Standards Section IV Gutefrage.net Community Guidelines
Qualification	Memorialization
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	“When someone passes away, friends and family can request that we memorialize the Facebook account. Once memorialized, the word “Remembering” appears above the name on the person’s profile to help make it that the account is now a memorial site and protects against attempted logins and fraudulent activity. To respect the choices someone made while alive, we aim to preserve their account after they pass away. We have also made it possible for people to identify a legacy contact to look after their account after they pass away. To support the bereaved, in some instances we may remove or change certain content when the legacy contact or family members request it.” Facebook: “For victims of murder and suicide we will remove the following content if it appears on the deceased’s profile photo, cover photo, or among recent timeline posts when requested by a legacy contact or family member of the deceased: Visual depiction of the object used in the deceased’s death Imagery of the convicted or alleged murderer of the deceased

Continued on next page

Field	Content
	Content related to the deceased's death For victims of murder, we will also remove the convicted or alleged murderer from the deceased's profile if referenced in relationship status or among friends. When we have additional context provided by the legacy contact or a family member of the deceased, we may: Remove bullying or harassment violations on a memorialized profile, which normally require the individual to self-report Remove praise or support for the death, disease, or harm of the deceased person on a memorialized profile Change the deceased's individual's privacy settings from public to friends-only when there is harmful content on the profile Change a violating account name on the profile of the deceased individual."

Table E.89: Code 3-24

Field	Content
Source	Facebook Community Standards Section V Gutefrage.net Community Guidelines
Qualification	Intellectual Property
Example and detailed descriptions by Facebook Inc. and Gutefrage.net, Community Guidelines	Copyright infringement Trademark infringement Facebook: "Upon receipt of a report from a rights holder or an authorized representative, we will remove or restrict content that engages in: Copyright infringement Trademark infringement"

Table E.90: Code 4-1

Field	Content
Legal Source	-
Qualification	Does not meet the threshold of a legally relevant behaviour but can affect people's feelings of comfort.

Continued on next page

Field	Content
	May lead to uncomfortable feelings.
Notes on interpretation and application	Targeting to the person individual insecurities (calling out weaknesses)

Table E.91: Code 4-2

Field	Content
Legal Source	Analogy to legally relevant category.
Qualification	Does not fit into a legally relevant category but can be considered to be equally relevant/invasive. Because of the intensity or extent. And because of the fact that there have not yet been cases decided by courts regarding these categories and new legislation has not been passed, yet.
Notes on interpretation and application	Deadnaming (Calling a person by the sex they do not longer identify with)

F

APPENDIX FOR CHAPTER 10

F.0.1. SEMI-STRUCTURED INTERVIEW GUIDELINES

INTRODUCTION:

1. Introduce the purpose of the interview and the focus on flagging mechanisms.
2. Introduce the flagging mechanisms [*Recital 50, Article 16 DSA*] (Terms of Service (ToS) [*Article 14 DSA*] and legal reasons [*Article 3(h) DSA*]) used by platforms like Facebook and TikTok etc. [We asked this question to regulators].
3. Can you explain your position and your work briefly?
4. How does your work relate to users? And to which users does it relate?
5. Which user groups or types do you represent? [Not asked for regulators].

RQ1: HOW WOULD USER CIRCUMSTANCES/SITUATIONS ALIGN WITH THE CONTENT-FLAGGING PROCESS?

1. The DSA refers to “notice-and-action mechanisms” [*Article 16 DSA*]. Reporting or flagging (terms used interchangeably here) describes the process by which a user employs an interface to inform a platform about perceived problematic content. How does the flagging process start, and how does it end? When is reporting “over” or closed? [Not asked for regulators].
2. What is a ‘normal’ flagging process for a user?
3. Describe situations and circumstances of users that motivates / incentivizes users to report? Can you describe situations that would motivate a user to report? (Don’t walk on the “bad street” of platform content. You might be associated with problematic/illegal content).
4. Under which situations do users, in your opinion, flag content?

5. How do these situations/ circumstances differentiate from/ between user groups normal users, flagging entities, Trusted Flaggers [*Article 22 DSA*]? [We asked this question to regulators].
6. How many users have to flag a piece of content before consensus is reached to enforce platforms rules (Terms of Service and legal reasons [*Articles 14, 3(h) and 16 DSA*])? When is user representation solidified?
7. How do these situations/ circumstances differentiate from your own experience when flagging content? [Not asked for regulators].
8. Are you aware of other alternatives for users than flagging to deal with problematic/upsetting content?
9. How do professional flagging entities become aware of problematic content [*Articles 9-10 DSA, Article 22 DSA*]?
10. Are platforms actively approaching flagging entities?
11. Do users actively approach flagging entities?
12. What do you expect from Trusted Flaggers [*Recital 61, Article 22 DSA*]?
13. What do you expect flagging entities and Trusted Flaggers not to catch [*Article 22 DSA*]? [Not asked for regulators].
14. Introduce the flagging mechanisms (Terms of Service and legal reasons) used by platforms like Facebook and TikTok etc. [Not asked for regulators] [*Article 16(1) DSA*].
15. How does flagging under the Terms of Service and the law differentiate [*Articles 16(2) and (3) DSA*]?
16. How does the flagging process start, and how does it end [*Articles 16, 20, 21 DSA*]? [We asked this question to regulators].

RQ2: WHAT ARE THE EXPECTATIONS ON USERS TO FLAG CONTENT?

1. What are the necessities for a user to flag content [*Recitals 52-53, 58 DSA*]? What has to be clear to the user [The second part of the question was not asked for regulators]?
2. Do you expect that flagging for users can have negative consequences [*Recital 58 DSA*]?
3. Are there differences in how to flag under the Terms of Service and the law [*Articles 3(h) and 14 DSA*]?
4. Are users aware of these differences?
5. Are users able [for regulators "equipped"] to make informed flagging decisions under either alternative (Terms of Service and the law [*Articles 3(h) and 14 DSA*])?

6. How are professional flagging entities navigating flagging under the Terms of Service and the law?
7. Are there different expectations for Trusted Flaggers [*Article 22 DSA*] under the DSA?

PROMPTS FOR USER-FLAGGING CHALLENGES

1. Online platforms have ‘notice action mechanisms’ or flagging mechanisms that provide options to report content through the Terms of Service and legal reasons [*Recital 50 DSA*]. How does flagging under the Terms of Service and the law differentiate through those mechanisms? And are they differently designed? [*Recital 53 DSA*]
2. Can you comment on the high level view of flagging mechanisms illustrated for Facebook and TikTok? How do they differentiate?
3. What are the essential requirements for a user to flag content? What information must be clear to the user before flagging content? [*Recital 53 and Article 16 DSA*].
4. Are items flagged as ‘illegal content’ treated differently than Terms and Conditions flags [*Recital 61, Articles 3(h) and 14 DSA*]?
5. Can you comment on the design of Terms of Service and illegal content flagging of the two platforms?
6. Are there potential negative consequences for users who flag content? If so, what are they?
7. Can you articulate your thoughts on the different designs for flagging illegal content for the four Member States for Facebook (France, Austria, Germany, and other) and related regulations in those countries [*Recitals 52-53, Article 3(h) DSA*]?
8. On point five (referring to the Miro board) you see hiding content or don’t show me this again, can you comment on this alternative to flagging under the Terms and Conditions for the user and flagging because of a legal reason [*Recitals 50, 52-53, Article 3(h), 14, 16 DSA*]?

RQ3: HOW CAN CONTENT-FLAGGING INTERFACES BE DESIGNED TO BE USABLE AND SATISFY LEGAL REQUIREMENTS?

1. The DSA requires “notice and action mechanisms” to be “easy to access” and “user-friendly” [*Recitals 50, 52-53, 61, and Article 16 DSA*]. In your opinion, what restrictions or expectations do you identify for compliance with the law?
2. What requirements are necessary from a user’s point of view, in your opinion, to design flagging interfaces in a DSA-compliant way [*Recitals 50, 52-53, 61, and Article 16 DSA*]?
3. What tensions do you see in user flagging, what obstacles must users overcome to flag content, and how should flagging be structured around user needs? [Not asked of regulators].

F.0.3. CODEBOOK AND THEMATIC FINDINGS IN THEMES AND SUBTHEMES

Table F.2: Codebook: Code Families and Associated Codes

Code Family	Associated Codes
Participants	Expert's position, years of experience, other regulation, GDPR, CDP, NetzDG, Loi n° 2004-575
Stakeholders	User representation, Own tool, NGO representation through the "Right to lodge a complaint," Concerns about user privacy, Whistleblower protection for users, User as a witness, Unreportable content, Insufficient user reporting bootstrapping, User burden, Dark pattern, Offended user, Missing information, Revictimization through reporting, Stakeholder concern about intentional online platform non-compliance
DSA	User right, Information about user rights, Cooperation, Platform obligation, Legal uncertainty
Content	Awareness of problematic content, Awareness of user content screenshots, Comments, Public authority information, Illegal content, Different levels of ease to identify 'manifestly illegal content,' Overlaps in reporting, Fragmentation of definition of illegal content, Terms of service, Contractual freedom
Reporting	Reporting process, Complexity of the reporting process, User follow-up, Information about the reporting process, Information about content moderation decision, Uncertainty in reporting, Uncertainty about the handling of flags, Different treatment of illegal content or ToS flags, Uncertainty about public authorities in the process, User motivation, Social role of reporting, User perception about the impact of own reporting, Professional reporting, Prioritization of trusted flaggers, Accreditation of trusted flaggers, Trust systems in user reporting, Enforcing cross-member state content reporting, Malicious reporting
Transparency	Transparency, Transparency report, Statement of Reason Database, Transparency of the reporting process
Design	Design of reporting mechanism, Symmetry/Relation between ToS and Law reporting, Artificial Intelligence (AI) tools for reporting, Labeling, Design of file uploading options, Single click solution, Progress bar, User friendly
Content Moderation	Moderation decision, Rapid response system, Recommendation system, Unbalanced decision consensus

Code Family	Associated Codes
Hiding Content	Hide/don't show me this again, Information about the effect of hiding

Table E.3: Thematic Findings in Themes and Subthemes

No.	Theme	Subtheme	Summary	Related Codes
1	Content reporting – a multi-actor process	Connection to users	Reporting is dependent on the user's motivation. Participants represent users in different positions against various threats of online content by offering legal expertise.	User motivation, user representation, NGO representation through the "right to lodge a complaint", insufficient user reporting bootstrapping, offended user, user burden, professional reporting, design of reporting mechanism, user friendly
2	Content reporting – a multi-actor process	Legal representation for users	DSA reporting empowers the user. Users are not expected to have legal knowledge. Legal understanding is needed in DSA reporting. Regulatory bodies offer enforcement and reporting support to users.	Insufficient user reporting bootstrapping, user burden, illegal content, user right, user follow-up, user representation, NGO representation to lodge a complaint, complexity in the reporting process
3	Content reporting	Upholding representation of users	The new DSA Trusted Flaggers specialize in educating and empowering user representation.	User representation, illegal content, professional representation, NGO, reporting process, complexity of the reporting process
4	Content reporting	Reporting that supports the user	User representatives emphasize creating tools that support user reporting and representation.	Information about the reporting process, user right, information about user rights, own tool, user representation, NGO representation
5	Design of RMs	The need for building tools	User representatives emphasize creating tools that support user reporting and representation.	Own tool, user representation, NGO representation
6	Design of RMs	Accessibility of reporting mechanisms	Reporting mechanisms need to be easy to find and accessible to all users.	User accessibility, simple UI design, user-friendly reporting mechanisms
7	Content moderation	Moderation decision	Users often feel left in the dark about the results of their reporting. Transparency is critical. Reports are sometimes handled inconsistently, leading to perceptions of bias.	User follow-up, transparency, clarity in reporting processes, fairness, report consistency, bias mitigation.
8	Walkthrough exercise	ToS/law reporting path	Platforms choose different layouts and paths for how users can report. These can be complex for users to navigate.	Overlaps in reporting, illegal content, terms of service, user feedback, iterative improvements, user-centric design, symmetry/relation between ToS and Law reporting.
9	Walkthrough exercise	Differences in member state design	Reporting mechanisms show different designs for member states. This makes reporting across national jurisdictions fragmented and complicated.	Insufficient user reporting bootstrapping, User burden, Stakeholder concern about intentional online platform non-compliance

No.	Theme		Subtheme	Summary	Related Codes
10	Design of mechanism	reporting	The act of reporting: motivation, emotion, and duty	Users face shocking content, which influences their motivation and emotional burden while reporting.	User burden, user motivation, uncertainty in reporting
11	Design of mechanism	reporting	Easy to access interface design	Reporting mechanisms need to be easy to find and accessible to all users.	User accessibility, simple UI design, user-friendly reporting mechanisms
12	Design of mechanism	reporting	Reporting labyrinth – a form of dark patterns?	Reporting processes should empower users without overly burdening them.	User empowerment, reporting burden, process simplicity
13	Design of mechanism	reporting	Support of different user needs	External actors, such as NGOs, play a crucial role in supporting effective reporting.	NGO oversight, reporting audits, external expertise
14	Design of mechanism	reporting	Hiding content as the user-friendly default	Hiding content can be seen as user control that burdens the user less	Hide/don't show me this again, Information about the effect of hiding
15	Stakeholder relation	cooperation	Collaboration between users	Collaboration between users, platforms, and regulators is key to effective enforcement.	Collaboration, enforcement, stakeholder involvement

G

APPENDIX FOR CHAPTER 11

G.1. SUMMARY OF HYPOTHESIS

Table G.1: Hypotheses overview.

Hypothesis	Statement
H1	Under AI error, Evaluative AI yields higher provision-level accuracy than conventional XAI (less overreliance on incorrect outputs).
H2a H2b	Evaluative AI reduces misclassification distance relative to XAI. The H2a reduction varies by error type; the Evaluative effect is attenuated for Out-of-Scope errors relative to Near-Miss/Overbreadth.
H3a H3b	When the AI is correct, Evaluative and XAI have comparable accuracy. When the AI is correct, XAI decisions are faster than Evaluative.
H4	Evaluative AI elicits higher-quality explanations than Manual and XAI.
H5a H5b	AI error types differ in the magnitude of their impact on reporting performance. Evaluative AI differentially attenuates error-type-specific performance drops relative to XAI.
H6	Versus XAI, Evaluative better aligns with DSA goals: (i) lower over-removal under Overbreadth; (ii) lower misrouting under OS.

**G.2. H3: DECISION TIME ON NO-ERROR TRIALS BY ASSISTANCE
CONDITION**

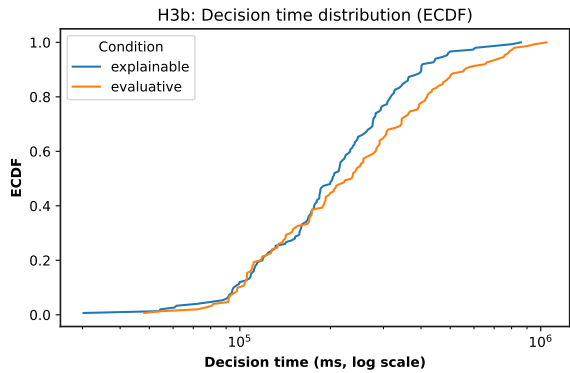


Figure G.1: Decision time on no-error trials by assistance condition.

G.3. ERROR TYPES

Table G.5: Error types (Near-miss, Overbreadth, Out-of-scope) for selected legal content categories.

Provision	Near-miss	Overbreadth	Out-of-scope
Incitement to crime (§ 111)	Prohibited symbols (§ 86a), Breach of peace (§ 126)	Incitement to hatred (§ 130), Criminal organizations (§ 129)	Privacy images (§ 201a), Defamation (§ 186–188)
Incitement to hatred (§ 130)	Incitement to crime (§ 111), Encouraging violence (§ 91)	Insult (§ 185–188), Religious defamation (§ 166)	Privacy images (§ 201a), Data forgery (§ 269)
Use of prohibited symbols (§ 86a)	Propaganda (§ 86), Incitement to crime (§ 111)	Incitement to hatred (§ 130), Encouraging violence (§ 91)	Privacy images (§ 201a), Child pornography (§ 184b)
“Something else” (placeholder)	Incitement to crime (§ 111), Data forgery (§ 269)	Incitement to hatred (§ 130), Privacy images (§ 201a)	Prohibited symbols (§ 86a), Child pornography (§ 184b)

G.4. PRE-STUDY QUESTIONNAIRE

Please answer each question to the best of your ability. This helps us understand your background and experience.

G.4.1. NEED FOR COGNITION (5-POINT LIKERT)

Answer options: 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree.

1. I would prefer complex to simple problems. (*Need for Cognition — Question 1*)
2. I like to have the responsibility of handling a situation that requires a lot of thinking. (*Need for Cognition — Question 2*)
3. Thinking is not my idea of fun. (*Need for Cognition — Question 3*)
4. This is an attention check question, please select "Strongly Agree". (*Attention Check — Question 4*)
5. I would rather do something that requires little thought than something that is sure to challenge my thinking abilities. (*Need for Cognition — Question 5*)
6. I really enjoy a task that involves coming up with new solutions to problems. (*Need for Cognition — Question 6*)
7. I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought. (*Need for Cognition — Question 7*)

G.4.2. USE & APPLY AI (0–10 SCALE)

Select one: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. (0 = Not at all; 10 = Extremely well.)

1. I can operate AI applications in everyday life. (*Use & Apply AI — Question 1*)
2. I can use AI applications to make my everyday life easier. (*Use & Apply AI — Question 2*)
3. I can use artificial intelligence meaningfully to achieve my everyday goals. (*Use & Apply AI — Question 3*)
4. To ensure data quality, please rate this item as 7 out of 10. (*Attention Check — Question 4*)
5. In everyday life, I can interact with AI in a way that makes my tasks easier. (*Use & Apply AI — Question 5*)
6. In everyday life, I can work together gainfully with an artificial intelligence. (*Use & Apply AI — Question 6*)
7. I can communicate gainfully with artificial intelligence in everyday life. (*Use & Apply AI — Question 7*)

G.4.3. KNOW & UNDERSTAND AI (0–10 SCALE)

Select one: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. (0 = Not at all; 10 = Extremely well.)

1. I know the most important concepts of the topic “artificial intelligence”. (*Know & Understand AI — Question 1*)
2. I know definitions of artificial intelligence. (*Know & Understand AI — Question 2*)
3. I can assess what the limitations and opportunities of using an AI are. (*Know & Understand AI — Question 3*)
4. I can assess what advantages and disadvantages the use of an artificial intelligence entails. (*Know & Understand AI — Question 4*)
5. I can think of new uses for AI. (*Know & Understand AI — Question 5*)
6. I can imagine possible future uses of AI. (*Know & Understand AI — Question 6*)

G.4.4. DETECT AI (0–10 SCALE)

Select one: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. (0 = Not at all; 10 = Extremely well.)

1. I can tell if I am dealing with an application based on artificial intelligence. (*Detect AI — Question 1*)
2. I can distinguish devices that use AI from devices that do not. (*Detect AI — Question 2*)
3. I can distinguish if I interact with an AI or a “real human”. (*Detect AI — Question 3*)

G.4.5. AI ETHICS (0–10 SCALE)

Select one: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. (0 = Not at all; 10 = Extremely well.)

1. I can weigh the consequences of using AI for society. (*AI Ethics — Question 1*)
2. I can incorporate ethical considerations when deciding whether to use data provided by an AI. (*AI Ethics — Question 2*)
3. I can analyze AI-based applications for their ethical implications. (*AI Ethics — Question 3*)

G.4.6. AI SELF-EFFICACY — PROBLEM SOLVING (0–10 SCALE)

Select one: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. (0 = Not at all; 10 = Extremely well.)

1. I can rely on my skills in difficult situations when using AI. (*AI Self-Efficacy — Problem Solving — Question 1*)
2. I can handle most problems in dealing with artificial intelligence well on my own. (*AI Self-Efficacy — Problem Solving — Question 2*)
3. I can also usually solve strenuous and complicated tasks when working with artificial intelligence well. (*AI Self-Efficacy — Problem Solving — Question 3*)

G.4.7. AI SELF-EFFICACY — LEARNING (0–10 SCALE)

Select one: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. (0 = Not at all; 10 = Extremely well.)

1. I can keep up with the latest innovations in AI applications. (*AI Self-Efficacy — Learning — Question 1*)
2. Despite the rapid changes in the field of artificial intelligence, I can always keep up to date. (*AI Self-Efficacy — Learning — Question 2*)
3. Although there are often new AI applications, I manage to always be “up-to-date”. (*AI Self-Efficacy — Learning — Question 3*)

G.4.8. AI SELF-COMPETENCY — PERSUASION LITERACY (0–10 SCALE)

Select one: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. (0 = Not at all; 10 = Extremely well.)

1. I don't let AI influence me in my everyday decisions. (*AI Self-Competency — Persuasion Literacy — Question 1*)
2. I can prevent an AI from influencing me in my everyday decisions. (*AI Self-Competency — Persuasion Literacy — Question 2*)
3. I realise if artificial intelligence is influencing me in my everyday decisions. (*AI Self-Competency — Persuasion Literacy — Question 3*)

G.4.9. AI SELF-COMPETENCY — EMOTION REGULATION (0–10 SCALE)

Select one: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. (0 = Not at all; 10 = Extremely well.)

1. I keep control over feelings like frustration and anxiety while doing everyday things with AI. (*AI Self-Competency — Emotion Regulation — Question 1*)
2. I can handle it when everyday interactions with AI frustrate or frighten me. (*AI Self-Competency — Emotion Regulation — Question 2*)
3. I can control my euphoria that arises when I use artificial intelligence for everyday purposes. (*AI Self-Competency — Emotion Regulation — Question 3*)

G.4.10. TRUST IN AUTOMATION (5-POINT LIKERT)

Answer options: 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree.

1. One should be careful with unfamiliar automated systems. (*Trust in Automation — Question 1*)
2. I rather trust a system than I mistrust it. (*Trust in Automation — Question 2*)
3. For quality control purposes, please select "Disagree" for this statement. (*Attention Check — Question 3*)
4. Automated systems generally work well. (*Trust in Automation — Question 4*)

G.4.11. LEGAL KNOWLEDGE (5-POINT LIKERT)

Answer options: 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree.

1. I am familiar with basic legal restrictions on online content in my country. (*Legal Knowledge — Question 1*)
2. I can identify when a post likely violates the law. (*Legal Knowledge — Question 2*)
3. I know where to look up legal rules for online speech. (*Legal Knowledge — Question 3*)

G.5. POST-STUDY QUESTIONNAIRE**G.5.1. USEFULNESS AND EFFECTIVENESS (7-POINT LIKERT)**

Answer options: 1 = Strongly Disagree; 2 = Disagree; 3 = Somewhat Disagree; 4 = Neutral; 5 = Somewhat Agree; 6 = Agree; 7 = Strongly Agree.

1. Using the AI assistance enabled me to accomplish the reporting task more quickly. (*Usefulness — Effectiveness — Question 1*)
2. Using the AI assistance improved the quality of my reporting decisions. (*Usefulness — Effectiveness — Question 2*)

3. Using the AI assistance increased my productivity in completing the task. (*Usefulness — Productivity — Question 3*)
4. Using the AI assistance enhanced my effectiveness in the reporting task. (*Usefulness — Task Effectiveness — Question 4*)
5. Using the AI assistance made the reporting task easier to complete. (*Usefulness — Ease — Question 5*)
6. I found the AI assistance useful for the reporting task. (*Usefulness — Overall — Question 6*)

G.5.2. WORKLOAD (7-POINT LIKERT)

Answer options: 1 = Very Low; 2 = Low; 3 = Somewhat Low; 4 = Neutral; 5 = Somewhat High; 6 = High; 7 = Very High.

1. How mentally demanding was the reporting task with AI assistance? (*Workload — Mental Demand — Question 1*)
2. How much effort did you have to expend to accomplish the task with AI assistance? (*Workload — Effort — Question 2*)
3. How frustrated did you feel during the task with AI assistance? (*Workload — Frustration — Question 3*)

G.5.3. TRUST, CONTROL, AND RESPONSIBILITY (7-POINT LIKERT)

Answer options: 1 = Strongly Disagree; ... 7 = Strongly Agree.

1. I felt in control of my reporting decisions when using the AI assistance. (*Trust/Control — Question 1*)
2. I felt responsible for the outcomes of my reporting decisions when using AI assistance. (*Responsibility — Question 2*)
3. I felt like I was just following the AI suggestions rather than making my own decisions. (*Reliance — Question 3*)
4. I could easily override or ignore the AI's suggestions when I wanted to. (*Trust/Control — Override — Question 4*)

G.5.4. DEMOGRAPHICS AND FINAL COMMENTS

1. Please indicate your age. (*Demographics — Question 1*)
2. Please indicate your gender. (*Demographics — Question 2*)
3. Any final comments about your experience? (*Open Feedback — Question 1*)
1. Age. (*Demographic Data Collection — Question 1*)
2. Gender. (*Demographic Data Collection — Question 2*)
3. Any final comments? (*Optional Feedback — Question 3*)

G.6. DESCRIPTIVES FOR PARTICIPANTS, ERROR TYPES, AND LEGAL CONDITIONS

Table G.6: Descriptives for participants, error types (malicious trials), and case law stimuli.

	Evaluative	Explainable	Manual
Participants (valid)	150	150	150
<i>Error type counts (malicious trials; AI arms only)</i>			
Near-Miss (NM)	52	52	–
Overbreadth (OB)	48	54	–
Out-of-Scope (OS)	50	44	–
<i>Demographics (valid N=450)</i>			
Gender (m / f / nb)	293 / 152 / 5		
Age	M=31.32, SD=8.7, range 18–71		
<i>Case law exposure (number of participants; AI arms)</i>			
§ 86a (symbols)	92		
§ 111 (incitement to crime)	76		
§ 130 (incitement to hatred)	71		
Other (e.g., § 30a BtMG)	61		

G.7. CODEBOOK FOR EXPLANATION OF THE REASON QUALITY AND RELATED MEASURES

1. Element Coverage

- *Definition:* Extent to which participants include all required legal elements in their explanation.
- *Item:* “The explanation covered all legally relevant elements required for this category.”
- *Scale:*
 - 1 None of the elements covered
 - 2 Few elements covered (1–2)
 - 3 About half covered
 - 4 Most covered (more than half)
 - 5 All covered comprehensively

2. Proportionality Reasoning

- *Definition:* Depth, proportionality, and correctness of reasoning.
- *Item:* “The explanation was clear, accurate, and legally proportional.”
- *Scale:*
 - 1 Very poor (unclear, irrelevant, incorrect)
 - 2 Weak (minimal detail, vague, partially incorrect)
 - 3 Moderate (some detail, limited proportionality)
 - 4 Good (clear, mostly correct, proportionate)
 - 5 Excellent (comprehensive, precise, legally sound)

3. Reasoning Depth

- *Definition:* Extent of elaboration and integration of legal/contextual reasoning.
- *Item:* “The explanation demonstrates deep reasoning and elaboration.”
- *Scale:*
 - 1 Very shallow (minimal or no reasoning, single-word/phrase)
 - 2 Shallow (basic, limited detail, surface-level)
 - 3 Moderate (some elaboration, partial justification)
 - 4 Deep (multiple relevant points, good detail)
 - 5 Very deep (comprehensive, nuanced arguments)

4. Perceived Overall Quality

- *Definition:* Rater’s holistic judgment of clarity, persuasiveness, and adequacy.
- *Item:* “Overall, how good was the explanation of the reason?”

- *Scale:*
 - 1 Very poor (unclear, incoherent)
 - 2 Poor (vague, with errors)
 - 3 Moderate (adequate but incomplete)
 - 4 Good (clear, coherent, minor gaps)
 - 5 Very good (comprehensive, persuasive, legally sound)

5. Moderation Decision

- *Definition:* The action recommended by the participant.
- *Options:* Remove, Escalate, Warn, Suspend, or No action.

6. Notes (Optional)

- *Definition:* Free-text space for coders to note ambiguities, uncertainties, or context.
- *Examples:* “Participant confused \$111 with \$130 but reasoning was detailed.”

G.8. PROMPTING TEMPLATE FOR AI GENERATED SUGGESTIONS FOR CONDITION 2

- **Be explicit about the model's role and instructions.** A well-crafted prompt should specify the task, the context and the criteria the model must follow [217, 582]. The literature recommends a three-part prompt with a system message, task description and formatting examples [217].
- **Provide structured and concise instructions.** Structured prompts using bullet points and clear sections improve moderation accuracy; verbose prompts reduce it [582]. The template therefore uses numbered steps and bullet points.
- **Include examples to guide output.** In-context examples of compliant and non-compliant content help the model contextualise the task[582] and avoid misclassification. The template asks the experimenter to insert examples showing both reportable and non-reportable content.
- **Generate arguments for and against the decision.** ArgLLMs provide arguments for and against particular outputs, yielding explainable and contestable decisions [273]. The template requires each agent to provide at least one argument supporting and one opposing the reporting decision, including references to the relevant ToS and legal rules.
- **Use multiple LLM agents or perspectives.** Multi-agent reasoning frameworks (e.g., SocraSynth) use two LLM agents representing opposing viewpoints, moderated by a human [125]. Including in our paper both a platform-leaning and a regulator-leaning agent helps surface biases and promotes value pluralism.

- **Avoid randomness and ensure reproducibility.** To reduce variability in responses, Elaraby et al set the generation temperature to 0 [217]. The template instructs the experimenter to set a low temperature or deterministic sampling.
- **Make the output accessible and user-friendly.** The Digital Services Act requires that terms and conditions be presented in clear, plain, intelligible language and reporting mechanisms user-friendly and easy to access. The template instructs the LLM to use simple language and avoid legal jargon where possible, while still citing the relevant rule.
- **Keep response lengths consistent across scenarios.** To avoid length-based bias and to ensure comparability, responses must have roughly the same number of words [582]. The template specifies word-count targets for each section.

G.9. SELECTION OF AND CREATION OF REPORTING SCENARIOS AND LEGAL CASES RELEVANT TO ONLINE CONTENT REPORTING IN GERMANY

Table G.7: Examples of illegal content categories, sources, case law, and posts.

Num.	Category	Level	Case law	Post on SocialNet
L_6	Using symbols of unconstitutional and terrorist organizations	Easy	§ 86a German Criminal Code	<i>Finally got this retro banner up in my room. Looks awesome! [photo shows a flag with a prohibited emblem]. New profile pic! Proud of my heritage — some symbols carry power. [prohibited symbol image in avatar].</i>
L_10	Public incitement to crime	Hard	§ 111 German Criminal Code	<i>Share this: at dawn we smash their labs and shut them down for good. They won't listen unless we burn down their offices tomorrow. Who's in?</i>
L_15	Incitement to hatred	Hard	§ 130 German Criminal Code	<i>Foreigners are parasites ruining everything — they should be driven out! Refugees don't belong here — they're an infestation. Get them out.</i>
L_27	Something else	Easy	(e.g., § 30a Abs. 1 BtMG)	<i>Selling unlicensed prescription antibiotics — message me for prices and pickup. Got extra opioid pain meds, no prescription needed. DM to buy discreetly.</i>

G.9.1. SCREENSHOT OF FACEBOOK'S REPORTING UI FOR ILLEGAL CONTENT

Legal removal request

This form is intended for reporting content posted on Facebook that you believe violates your personal legal rights or applicable local laws. If you are trying to report a legal issue related to Defamation, Privacy or Intellectual Property, please use the drop-down reporting forms provided below. Before making this request, you may want to seek guidance from a qualified lawyer.

Please note that the content you'd like to report may violate our:

- Community Standards (e.g. bullying, harassment, hate speech),
- Advertising Policies
- Commerce Policies

If you would like to report the content under our Community Standards, please visit the [Help Centre](#).

Select the country where you are claiming legal rights.

This should be the country whose laws you claim are violated by the relevant content.

Germany

What legal issue do you wish to report?

- ☐ Intellectual property (e.g. copyright, trademark)
- ☐ Defamation
- ☐ Privacy/erasure under GDPR
- ☒ I believe that the content violates a law not mentioned above (e.g. hate speech)

Do you wish to report a violation of any of the following provisions of German law:

- ☐ Dissemination of propaganda material of unconstitutional and terrorist organisations (§ 86 German Criminal Code)
- ☐ Using symbols of unconstitutional and terrorist organisations (§ 86a German Criminal Code)
- ☐ Preparation of a serious violent offence endangering the state (§ 89a German Criminal Code)
- ☐ Encouraging the commission of a serious violent offence endangering the state (§ 91 German Criminal Code)
- ☐ Treasonous forgery (§ 100a German Criminal Code)
- ☐ Public incitement to crime (§ 111 German Criminal Code)
- ☐ Breach of the public peace by threatening to commit offences (§ 126 German Criminal Code)
- ☐ Forming criminal organisations (§ 129 German Criminal Code)
- ☐ Forming terrorist organisations (§ 129a German Criminal Code)
- ☐ Forming criminal and/or terrorist organisations abroad (§ 129b German Criminal Code)
- ☐ Incitement to hatred (§ 130 German Criminal Code)
- ☐ Dissemination of depictions of violence (§ 131 German Criminal Code)
- ☐ Rewarding and approving of offences (§ 140 German Criminal Code)
- ☐ Defamation of religions, religious and ideological associations (§ 166 German Criminal Code)
- ☐ Distribution, acquisition and possession of child pornography content (§ 184b German Criminal Code)
- ☐ Insult (§ 185/188 German Criminal Code)
- ☐ Defamation (§ 186/187/188 German Criminal Code)
- ☐ Disparagement of the memory of the deceased (§ 189 German Criminal Code)
- ☐ Violation of intimate privacy or personality rights by taking photographs or other images (§ 201a German Criminal Code)
- ☐ Threatening the commission of a felony (§ 241 German Criminal Code)
- ☐ Forgery of data intended to provide proof (§ 269 German Criminal Code)
- ☐ Image privacy rights pertaining to German Artistic Copyright Act (§ 22 KUG)
- ☐ Something else

Send

Figure G.2: Example of Facebook's illegal content reporting UI for Germany

G.9.2. PROMPTING AND EVALUATION STRATEGY

PROMPTING

A robust prompt template is essential, because the quality of an LLM's answers depends strongly on the phrasing, specificity and structure of the prompt. Empirical work on *policy-as-prompt* shows that clear, organized prompts, with bullet points and sections rather than long paragraphs, yield higher accuracy than wordy or verbose prompts and that even small formatting changes can cause different outputs (predictive multiplicity)

[582]. Prompting as an application of reasoning in the use of LLMs is proposed in utilizing examples or descriptions of desired output in a structured way. Research on *argumentative LLMs* provides an explicit method for generating *arguments for and against* a decision so that outputs are explainable and contestable [273].

Recent work on ranking arguments recommends structuring prompts into three parts:

- A system message defining the model's role,
- A task description with clear criteria and numbering, and
- Formatting examples that illustrate the desired output [217].

The template is designed to generate a *stable set of responses* for each piece of content in the experiment. It supports two LLM agents—a **platform-learning agent** that interprets content through the lens of legal definitions (e.g., the Digital Services Act and other relevant national provisions).

We generated the *Case Descriptions* with ChatGPT 4.1 based on the following prompting template and automatically evaluated the outcome by a judge model (ChatGPT 4.5). This automatic evaluation was in a next step evaluated by two of the authors who have been working in legal adjacent fields for more than four years and have knowledge about German national law:

- ChatGPT4.1: Make a short summary of the cases provided in the excel file.

SCENARIO: CHATGPT 4.1.:

System Message: You are generating detailed, realistic scenarios for a user study on illegal content flagging in online platforms. For each case, use all available information (summary and full case via link) to write a single, clear scenario of user-generated content or action that would qualify as potentially illegal for the specific category. The scenario should be plausible, anonymized, understandable to a layperson, and sufficiently detailed for use in a user experiment.

INSTRUCTIONS:

Read the full case and summary. Write a stand-alone scenario describing a post, comment, message, or online action. The scenario should: Clearly illustrate the type of illegal content (e.g., hate speech, defamation, drug promotion, threats, etc.) Be specific and detailed enough for users to understand the situation. Be anonymized (no real names). Use plain language, without legal or procedural references. Provide context and any relevant reactions, if they help clarify the situation. Be up to 200 words if needed for clarity. The output should be only the scenario text—do not include the case summary, statute, link, or any other information.

EXAMPLE OUTPUT:

A user posts in a local Facebook group: “Don’t trust the owner of the bakery on Main Street. He poisons his customers and cheats people out of money. Spread the word so nobody else gets scammed!” Several group members reply with supportive comments and share the post. Others ask for evidence, but the original poster does not respond.

Prompt for LLM or Human Writer (per case): Write a detailed, realistic, and clear scenario describing user-generated content or online behavior that would fit the category of illegal content in a social media context. Use the full case and summary for context and details. Do not reference the case, statute, or any legal terms. Only output the scenario.

- Judge: Chat GPT 4.5: Take the summary information and check with the case PDF in batches (1-4) if the summaries are correct.

G.9.3. REPORTING SCENARIOS AND LEGAL CASES FOR REPORTING ILLEGAL CONTENT IN GERMANY

G.10. ADDITIONAL ANALYSES FOR H4: EXPLANATION-OF-REASONS (EOR) QUALITY

G.10.1. INTER-RATER RELIABILITY

Each explanation was independently rated by two trained annotators on four dimensions: *element coverage*, *proportionality reasoning*, *reasoning depth*, and *perceived overall quality*, using 1–5 Likert scales.

Inter-rater reliability was acceptable to high across all dimensions. Krippendorff's α ranged from 0.67 to 0.82, and quadratic-weighted Cohen's κ ranged from 0.69 to 0.82. Agreement within ± 1 scale point exceeded 79% for all dimensions. These values meet or exceed commonly accepted thresholds for subjective content analysis, supporting aggregation of coder ratings via the mean. Full reliability statistics are reported in Table G.11.

Table G.11: Inter-rater reliability for EoR quality dimensions.

Dimension	α	κ_q	Exact Agree.	Agree. ± 1
Element coverage	0.76	0.75	0.82	0.86
Proportionality reasoning	0.82	0.82	0.86	0.89
Reasoning depth	0.67	0.69	0.75	0.79
Perceived overall quality	0.69	0.70	0.80	0.83

G.10.2. MODEL SPECIFICATION AND ROBUSTNESS

In the main analysis, we treated coder-mean ratings as approximately continuous and estimated ordinary least squares (OLS) regression models of the form:

$$Y_{ij} = \beta_0 + \beta_1 \text{Manual}_i + \beta_2 \text{Evaluative}_i + \varepsilon_{ij},$$

with Conventional XAI as the reference condition.

Because participants could contribute more than one rated report, we used participant-clustered robust standard errors whenever repeated observations were present; otherwise, HC1 heteroskedasticity-robust standard errors were applied. All reported inference is two-sided.

As a robustness check, we additionally estimated cumulative logit (ordered logit) models treating the outcomes as ordinal rather than continuous. Across all four dimensions, the ordered-logit results were directionally consistent with the OLS estimates and led to

identical substantive conclusions: no reliable advantage of Evaluative AI over Conventional XAI or Manual reporting. Ordered-logit model summaries and Wald contrasts are reported in Appendix Tables G.12–G.13.

Table G.12: OLS regression results for Explanation-of-Reasons (EoR) quality (coder-mean ratings). Conventional XAI is the reference condition. Participant-clustered robust standard errors.

Dimension	Term	Coef.	SE	95% CI	<i>p</i>
Element coverage	Evaluative vs. XAI	0.01	0.11	[−0.22, 0.23]	.95
	Manual vs. XAI	−0.13	0.11	[−0.36, 0.09]	.24
Proportionality reasoning	Evaluative vs. XAI	0.03	0.13	[−0.23, 0.29]	.82
	Manual vs. XAI	−0.05	0.13	[−0.30, 0.20]	.69
Reasoning depth	Evaluative vs. XAI	0.07	0.12	[−0.16, 0.30]	.56
	Manual vs. XAI	−0.10	0.12	[−0.33, 0.13]	.39
Perceived overall quality	Evaluative vs. XAI	−0.01	0.12	[−0.23, 0.22]	.95
	Manual vs. XAI	−0.15	0.12	[−0.39, 0.09]	.22

Table G.13: Pairwise contrasts for EoR quality dimensions (coder-mean). Participant-clustered robust standard errors.

Dimension	Contrast	Diff.	SE	<i>p</i>
Element coverage	Evaluative vs. XAI	0.01	0.11	.95
	Evaluative vs. Manual	0.14	0.11	.22
Proportionality reasoning	Evaluative vs. XAI	0.03	0.13	.82
	Evaluative vs. Manual	0.08	0.13	.53
Reasoning depth	Evaluative vs. XAI	0.07	0.12	.56
	Evaluative vs. Manual	0.17	0.12	.16
Perceived overall quality	Evaluative vs. XAI	−0.01	0.12	.95
	Evaluative vs. Manual	0.06	0.12	.61

G.10.3. CODEBOOK

G.11. APPENDIX: DETAILED ANALYSIS PLAN

G.11.1. OVERVIEW

This appendix provides the full analysis specifications corresponding to the compressed methodology reported in the main text. It documents outcome definitions, model equations, robustness checks, and supplementary analyses for H1–H6. Unless otherwise noted, all hypothesis tests are two-sided and all confidence intervals are 95%.

G.11.2. OUTCOME DEFINITIONS AND CODING

Provision accuracy. Provision accuracy $Y_{ij}^{acc} \in \{0, 1\}$ indicates whether participant i selected the ground-truth legal provision for case j .

Misclassification distance. Misclassification distance $D_{ij} \in \{0, 1, 2, 3\}$ captures how far the selected provision deviated from the ground truth, where 0 = correct, 1 = near-miss, 2 = related-but-wrong, and 3 = out-of-scope. This ordinal outcome applies to the

participant's selected provision; it should not be confused with the AI error condition itself, which is modeled categorically.

Explanation-of-Reasons outcomes. Explanation-of-Reasons (EoR) quality was rated independently by two trained coders on four 1–5 dimensions: (i) element coverage, (ii) proportionality reasoning, (iii) reasoning depth, and (iv) perceived overall quality. For the primary analyses, we use the mean of the two coder ratings for each dimension. Reliability statistics are reported separately in the appendix.

Downstream risk outcomes. For Overbreadth error trials, *over-removal* is a binary indicator equal to 1 if the participant selected a severe enforcement action (e.g., content removal, account suspension, or deplatforming), and 0 otherwise. For Out-of-Scope error trials, *misrouting* is a binary indicator equal to 1 if the participant failed to route the case as out-of-scope, and 0 otherwise.

Quality-control exclusions. Analyses exclude participants removed for technical issues and, where specified, participants who reported benign filler posts. Excluding participants with false-positive reports on benign filler posts is treated as a robustness check for several hypotheses.

G.11.3. H1: PROVISION ACCURACY UNDER AI ERROR

Estimand. Among AI-assisted participants (Evaluative AI vs. Conventional XAI), we estimate the effect of Evaluative AI on provision-level accuracy on illegal-content error trials (one error trial per participant by design). Benign posts are excluded; excluding participants who reported benign filler posts is a sensitivity check.

Outcome. Provision accuracy $Y_{ij}^{\text{acc}} \in \{0, 1\}$ indicates whether participant i selected the ground-truth provision for case j .

Model. We fit a logistic GLM with AI error condition and case fixed effects:

$$\Pr(Y_{ij}^{\text{acc}} = 1) = \text{logit}^{-1}\left(\beta_0 + \beta_1 \text{Evaluative}_i + \boldsymbol{\gamma}^\top \mathbf{E}_j + \boldsymbol{\eta}^\top (\text{Evaluative}_i \times \mathbf{E}_j) + \boldsymbol{\delta}^\top \mathbf{C}_j\right),$$

where $\mathbf{E}_j$ are AI error condition indicators (NM/OB/OS) and $\mathbf{C}_j$ are case fixed effects. Inference uses HC1 robust standard errors; participant-clustered standard errors are reported as a robustness check (numerically similar given one observation per participant). We report odds ratios with 95% confidence intervals and two-sided p -values.

Robustness. We re-estimate the model with pre-task covariates (MAILS, NCS-6, legal knowledge, propensity to trust automation) and under exclusion of participants who reported benign filler posts. Balance checks by condition and AI error condition accompany Fig. 11.3.

G.11.4. H2: MISCLASSIFICATION DISTANCE UNDER AI ERROR

Estimands and sample. On illegal-content error trials in the AI-assisted conditions, H2a tests whether Evaluative AI reduces misclassification distance relative to Conventional XAI, using an ordinal distance $D \in \{0, 1, 2, 3\}$ (0 = correct; 3 = farthest misclassification). H2b tests whether this effect varies by AI error condition, with particular focus on Out-of-Scope errors.

Model. We estimate proportional-odds (ordered logit) models with participant-clustered

robust standard errors:

$$\Pr(D \leq k) = \text{logit}^{-1}(\tau_k - \eta), \quad \eta = \beta_1 \text{Evaluative} + \boldsymbol{\gamma}^\top \mathbf{E} + \boldsymbol{\theta}^\top (\text{Evaluative} \times \mathbf{E}),$$

where $\mathbf{E}$ are AI error condition indicators (Overbreadth and Out-of-Scope, with Near-Miss as the reference). Negative coefficients indicate a shift toward smaller misclassification distances. The primary specification excludes case fixed effects; case fixed effects are added only as a design-control robustness check when available.

Robustness. We re-estimate models with pre-task covariates (MAILS, NCS-6, legal knowledge, propensity to trust automation), exclude participants who reported benign filler posts, and conduct non-parametric Mann–Whitney tests (Evaluative < XAI). To isolate Near-Miss vs. Overbreadth moderation, we additionally estimate a restricted model with an Evaluative $\times$ Overbreadth interaction. Results are summarized in Fig. 11.4.

G.11.5. H3: NO-ERROR TRIALS (CORRECT AI ASSISTANCE)

Panel. We analyze no-error trials in the AI-assisted conditions only (Evaluative AI and Conventional XAI) after standard quality control. The panel contains one no-error trial per participant in the AI-assisted conditions ($N = 300$, 150 per condition); the Manual condition did not include no-error trials by design.

Outcomes and models. H3a (accuracy): Provision accuracy is modeled using a binomial GLM with HC1 robust standard errors. Case fixed effects are excluded in the primary specification and added only as a design-control robustness check when available. We additionally conduct a two one-sided tests (TOST) equivalence test on the difference in proportions with margin $\delta = 0.05$.

H3b (decision time): Decision time (ms) is analyzed on the log scale using OLS with HC1 robust standard errors. Times are winsorized at the 1st/99th percentiles prior to transformation. As a non-parametric check, we report a one-tailed Mann–Whitney U test on raw times testing whether Conventional XAI is faster than Evaluative AI.

Robustness. Results are robust to inclusion of case fixed effects (when available) and to non-parametric decision-time tests.

G.11.6. H4: EXPLANATION-OF-REASONS QUALITY

Panel. We analyze illegal-content reports with completed coder ratings across all three assistance conditions (Manual, Conventional XAI, Evaluative AI). Each report was independently rated by two trained annotators. Participants contributed up to two rated reports, resulting in repeated observations at the participant level.

Outcomes. EoR quality was assessed on four 1–5 rating scales: (i) element coverage, (ii) proportionality reasoning, (iii) reasoning depth, and (iv) perceived overall quality. For each dimension, we use the mean of the two coder ratings as the dependent variable. Inter-rater reliability was acceptable to high across dimensions (Krippendorff's $\alpha \approx 0.67$ – 0.82 ; quadratic-weighted Cohen's $\kappa \approx 0.69$ – 0.82 ; Appendix), supporting aggregation.

Model. For each outcome, we estimate linear regression models on coder-mean ratings, treating aggregated scores as approximately continuous. Assistance condition is entered as a categorical predictor with Conventional XAI as the reference. Because participants may contribute multiple reports, we use participant-clustered robust standard errors. We report coefficient estimates with 95% confidence intervals and two-sided

p -values. Pairwise contrasts (Evaluative vs. XAI, Manual vs. XAI, Evaluative vs. Manual) are evaluated via linear hypothesis tests.

G.11.7. H5: ERROR-TYPE IMPACT AND ATTENUATION

Panel. We restrict analysis to illegal-content trials in the AI-assisted conditions (Conventional XAI and Evaluative AI) and construct a paired difference-in-differences (DiD) panel. Each participant contributes exactly two trials: (i) one illegal-content no-error trial ($post = 0$) and (ii) one illegal-content error trial ($post = 1$), where the AI error condition is either Near-Miss, Overbreadth, or Out-of-Scope. AI error condition is therefore assigned between participants, while error presence (correct AI assistance vs. AI error condition) varies within participants. We define $pair_errorCondition_i$ as the AI error condition of participant i 's error trial and assign it to both rows of the participant's pair, yielding a constant pair-level factor used for DiD estimation. Benign content is excluded upstream through quality-control filtering.

Outcomes. We analyze two classes of dependent variables. First, provision accuracy is a binary indicator of whether the reported legal provision matches the ground truth. Second, we analyze four coder-rated Explanation-of-Reasons dimensions on 1–5 scales: (i) element coverage, (ii) proportionality reasoning, (iii) reasoning depth, and (iv) perceived overall quality. For each dimension, we compute the mean rating across the two coders. Inter-rater reliability was acceptable across dimensions (Appendix), supporting use of the coder-mean as a stable estimate of explanation quality.

Models. *Provision accuracy (paired DiD GLM).* We estimate a generalized linear model with a binomial family and logit link on the long-format paired panel, reporting participant-clustered robust standard errors:

$$\begin{aligned} \Pr(Y_{it}^{acc} = 1) = \text{logit}^{-1} & \left(\beta_0 + \beta_1 \text{Evaluative}_i + \beta_2 \text{post}_{it} \right. \\ & + \boldsymbol{\gamma}^\top \mathbf{E}_i + \beta_3 (\text{Evaluative}_i \times \text{post}_{it}) \\ & \left. + \boldsymbol{\eta}^\top (\text{post}_{it} \times \mathbf{E}_i) + \boldsymbol{\phi}^\top (\text{Evaluative}_i \times \text{post}_{it} \times \mathbf{E}_i) \right). \end{aligned} \quad (\text{G.1})$$

where Evaluative_i indicates assistance condition (Evaluative vs. Conventional XAI), post_{it} indicates the error trial (1 for error, 0 for no-error), and $\mathbf{E}_i$ encodes $pair_errorCondition_i$ (Overbreadth and Out-of-Scope, with Near-Miss as the reference category). This paired DiD specification identifies (i) the within-participant accuracy change induced by AI error and (ii) whether that change differs by AI error condition and is attenuated under Evaluative AI.

Explanation-of-Reasons quality (primary: first-difference OLS). Because EoR outcomes are coder-means on 1–5 scales and participants contribute paired observations, our primary analysis uses a participant-level first-difference specification on continuous coder-mean scores. For each participant i and dimension d , we compute

$$\Delta Y_i^{(d)} = Y_{i,error}^{(d)} - Y_{i,noerror}^{(d)}.$$

We then estimate OLS models with heteroskedasticity-robust (HC1) standard errors (one row per participant):

$$\Delta Y_i^{(d)} = \alpha_0 + \alpha_1 \text{Evaluative}_i + \boldsymbol{\kappa}^\top \mathbf{E}_i + \boldsymbol{\lambda}^\top (\text{Evaluative}_i \times \mathbf{E}_i) + \varepsilon_i.$$

This removes participant fixed effects by construction and directly captures error-induced changes in explanation quality, while allowing these changes to vary by AI error condition and assistance condition.

Explanation-of-Reasons quality (sensitivity: ordered logit DiD). As a sensitivity analysis, we discretize coder-mean scores back to the original 1–5 categories (rounded to the nearest integer and clipped to [1, 5]) and estimate proportional-odds (ordered logit) models on the long-format paired panel with participant-clustered robust standard errors. These ordinal models include the same DiD interaction structure and yield substantively consistent conclusions (Appendix).

Hypothesis tests. H5a tests whether the *error penalty* differs by AI error condition. For provision accuracy, we test whether the within-participant error effect (the post term) varies across *pair_errorCondition* levels in the Conventional XAI baseline using Wald contrasts (e.g., Out-of-Scope vs. Near-Miss; Out-of-Scope vs. Overbreadth). For Explanation-of-Reasons quality, H5a is evaluated analogously using Wald contrasts on the AI error condition coefficients in the first-difference models, comparing mean ΔY across AI error conditions in the Conventional XAI baseline.

H5b tests attenuation by assistance condition. For provision accuracy, attenuation is assessed via the three-way interaction terms *Evaluative* $\times$ *post* $\times$ *pair_errorCondition*, with primary focus on the Out-of-Scope attenuation contrast. For Explanation-of-Reasons quality, attenuation is assessed via *Evaluative* $\times$ *pair_errorCondition* interactions in the first-difference models, again with primary focus on Out-of-Scope. We report odds ratios for accuracy models and mean first-difference estimates for Explanation-of-Reasons models, alongside robust confidence intervals and *p*-values. Model-implied margins and descriptive plots are provided to aid interpretation.

G.11.8. H6: DSA-RELEVANT DOWNSTREAM RISK OUTCOMES

Estimand. H6 evaluates whether, relative to Conventional XAI, Evaluative AI leads to user reporting behaviors that are more consistent with downstream risk-relevant outcomes in DSA-aligned reporting, focusing on proportionality and appropriate routing under known AI error conditions.

Panels and outcomes. Analyses are restricted to AI-assisted conditions (Evaluative AI and Conventional XAI) and to trials in which the AI output was erroneous. We estimate separate models by AI error condition. For Overbreadth errors, we operationalize *over-removal* as a binary indicator equal to 1 if the participant selected a severe enforcement action (e.g., content removal, account suspension, or deplatforming), and 0 otherwise. This outcome captures disproportionate intervention in cases where the AI's error overextends policy scope. For Out-of-Scope (OS) errors, we operationalize *misrouting* as a binary indicator equal to 1 if the participant failed to route the case as out-of-scope, and 0 otherwise. This outcome reflects potential misallocation of moderation effort when content does not fall under the platform's policy jurisdiction.

Models. For each AI error condition panel, we estimate a binomial generalized linear model with a logit link and participant-clustered standard errors:

$$\Pr(Y = 1) = \text{logit}^{-1}(\alpha + \beta \text{Evaluative}),$$

where Evaluative is an indicator for Evaluative AI (Conventional XAI is the reference

category) and Y denotes either *over-removal* (Overbreadth panel) or *misrouting* (Out-of-Scope panel). We report odds ratios with cluster-robust confidence intervals and present descriptive outcome rates by condition. When outcome variation is insufficient for model estimation, we report descriptive statistics only and treat the corresponding hypothesis test as not identifiable.

G.11.9. PARTICIPANTS

We recruited Prolific participants region-locked to Germany (fluent English; approval rate $\geq 90\%$). Compensation was £1.60 base (10–15 min) plus up to £2.00 bonus (maximum £3.60). We enrolled 489 participants; after exclusions for technical issues or false-positive reports on benign filler posts, $N=450$ remained, with 150 participants per condition (**EVALAI**, **XAI**, and **BASELINE**) (Appendix Table G.6). Demographics were 293 men, 152 women, and 5 non-binary participants; mean age was $M=31.32$ years ($SD=8.7$, range 18–71).

Dimension	Definition, item, and scale
Element coverage	Definition: Extent to which participants include all required legal elements in their explanation. Item: “The explanation covered all legally relevant elements required for this category.” Scale: 1 = None of the elements covered 2 = Few elements covered (1–2 elements) 3 = About half of the elements covered 4 = Most elements covered (more than half) 5 = All elements covered comprehensively
Proportionality reasoning	Definition: Depth, proportionality, and correctness of reasoning. Item: “The explanation was clear, accurate, and legally proportional.” Scale: 1 = Very poor (unclear, irrelevant, incorrect) 2 = Weak (minimal detail, vague, partially incorrect) 3 = Moderate (some relevant detail, but limited accuracy/proportionality) 4 = Good (clear, relevant, mostly correct, proportionate) 5 = Excellent (comprehensive, precise, legally sound, proportionate)
Reasoning depth	Definition: Extent to which a participant’s explanation demonstrates depth of thought, elaboration, and integration of legal or contextual reasoning. Item: “The explanation demonstrates deep reasoning and elaboration.” Scale: 1 = Very shallow (minimal or no reasoning, single-word/phrase) 2 = Shallow (basic reasoning, limited detail, mostly surface-level) 3 = Moderate (some elaboration, partial justification, somewhat developed) 4 = Deep (clear justification with multiple relevant points, good detail) 5 = Very deep (comprehensive, well-developed reasoning with nuanced legal/contextual arguments)
Moderation decision	Definition: Rating of the decision outcome assigned by the participant. Item: “The participant selected the appropriate moderation decision for the case.” Scale: Use the task’s decision labels/codes (report separately if nominal; do not force 1–5 unless your instrument does so).
Perceived overall quality	Definition: Rater’s perceived overall rating of the explanation. Item: “Overall, the explanation was of high quality.” Scale: 1–5, where 5 = very good.

Table G.15: Codebook used for evaluating participants’ explanations of the reason.

H

APPENDIX FOR CHAPTER 12

H.1. INTERVIEW GUIDELINE

Introduction

1. Can you explain your position and your work briefly?
2. How does your work relate to users ['Recipients of the service' *Art. 3(b) DSA*]? Which kinds of users or communities does your organization represent or support?
3. Which of the work activities you have mentioned would you say fall within your Trusted Flagger role [*Art. 22 DSA*]?

Motivation / Process to Become a Trusted Flagger

1. Have you been a Trusted Flagger before the DSA?
2. Can you elaborate on the reasons to become a Trusted Flagger [*Art. 22 DSA*]? (As opposed to conducting your activities without it; why did your organization decide to apply? [*Rec. 62 DSA*])
3. What specific qualifications did you provide in the process to become a Trusted Flagger [*Art. 22(1) DSA*]?

Function of Trusted Flaggers

1. What do you see as the main objectives for Trusted Flaggers [*Art. 22(1)* in conjunction with *Art. 16 DSA*]?
2. How do organizations value or perceive the work of Trusted Flaggers?

Reporting Process

1. What does your organization do as a Trusted Flagger [*Art. 22(1)* and (3) *DSA*]?

2. What kind of content do you report? Is it largely within your area of expertise, or do you go beyond it [Art. 22(1), Art. 22(2)(a) DSA]?
3. What do you conceive as the role of average users in the Trusted Flagger system [Art. 3(b) and Art. 22 DSA]? How do users reach your organization, and what does 'good' user support look like from your perspective?
4. Can you describe the tools you use to fulfill the requirements for Trusted Flaggers [Art. 22(3) DSA]?
5. Can you describe the reporting process for Trusted Flaggers in your experience [Art. 22(1), Art. 16 DSA]?
6. Who within your organization is responsible for carrying out Trusted Flagger tasks?
7. What kinds of access controls or safeguards are in place for staff performing Trusted Flagger activities?
8. Are you in contact with public authorities, and are you sharing your flagging activities with them [Art. 9, 10, 22 DSA]?

Interaction with Platforms

1. For which platforms do you conduct Trusted Flagger reporting tasks?
2. In your opinion, how do platforms react to your flagging notifications [Art. 22, 35(1)(g) DSA] ?
3. In your experience, how does prioritization play out in content moderation for your flags [Rec. 61]?
4. What kinds of interactions do you have with platforms, and how regular are they [Rec. 87]?
5. Are you reporting only for VLOPs and VLOSEs, or also for smaller online platforms? [Art. 16 and 22(1) DSA]

Legal vs Terms of Service

1. Do you only flag content under illegal content [Art. 3(h) DSA] (legally relevant) reporting categories, or do you also flag [Art. 16 DSA] under the Terms of Service of platforms [Art. 14 DSA]?

Statement of Reasons Data

1. Would you agree that the numbers indicated in the Statement of Reasons Database about Trusted Flagger notices accurately reflect your experience [Art. 17, Art. 22 DSA]?
2. (Optional) When looking at the Statement of Reasons data, how do you expect your organization's activity—or your cause/community more broadly—to appear in those reports over time [Art. 17 DSA]?

Resources

1. How do you support the Trusted Flagger activity financially or in terms of resources [Art. 22 DSA]?
2. How many people are conducting Trusted Flagger tasks in your organization [Art. 22 DSA]?
3. Can you describe the activities and amount of work that goes into completing a report [Art. 22(3) DSA]?
4. What happens to your records or stored data when a post you report is taken down by the platform [Art. 15(1)(b) DSA]?

Website and Visibility

1. Could you describe how your website communicates your Trusted Flagger status and activities [Art. 22]?
2. Are there any specific features or elements on your website related to your Trusted Flagger work [Art. 22]?
3. Is there anything else you would like to mention or add that we haven't covered?

H.2. TRUSTED FLAGGER ACTIVITY IN THE STATEMENT OF REASONS DATABASE

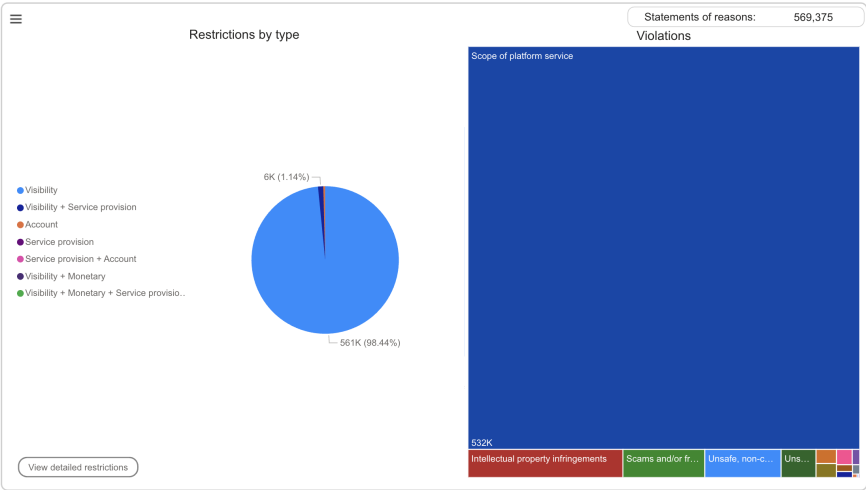


Figure H.1: Statement of Reasons Database Capturing Information About TF Notices About Restriction Types and Violations by Statement of Reason Category Reason taken on the 11/5/2025 [1, 166].

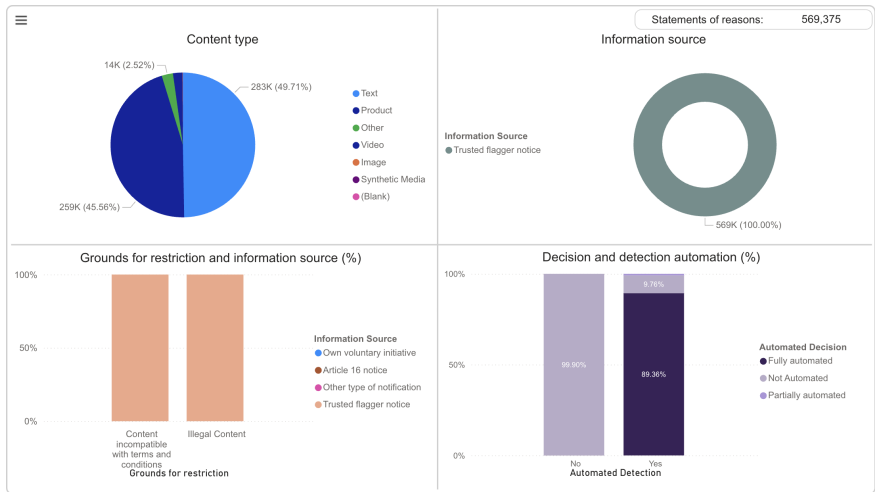


Figure H.2: Statement of Reasons Database Capturing Information About Content Type and Content Moderation Measures of TF Notices taken on the 11/5/2025 [1, 166].

H.3. MAIN THEMES PRESENTED TO WORKSHOP PARTICIPANTS

- Platforms obstruct or fragment Trusted Flagger work
- Capacity, resourcing, and training constraints shape what TFs can do
- Regulatory coordination is uneven and (sometimes) difficult
- Platforms' incentives and response patterns shape outcomes
- Illegality is operationally hard to pin down in practice
- Transparency and accountability artifacts can misrepresent TF labor and role

I

APPENDIX FOR CHAPTER 13

I.1. FACTS-AS-GIVEN SHEET (EVALUATION INPUT PROTOCOL)

For the author-led walkthrough evaluation, we fixed a “facts-as-given” sheet prior to applying the decision tree. The purpose is to (i) document the inputs to the assessment in a reproducible form, (ii) separate stipulated facts from assumptions, and (iii) prevent post-hoc inference when the decision tree requires unavailable information. The facts-as-given sheet should help to use the toolkit and input information in a structured form.

G1. Stipulated facts (used as inputs). Record only information explicitly provided by the use case description or accompanying materials:

- **Platform/service context:** service type; EU availability; relevant jurisdictional scope.
- **Feature/change under assessment:** what is new/changed; where it operates (surfaces/channels).
- **Relevant content/behaviour:** content type(s); target(s); creation method; distribution mechanisms.
- **Observed harm/risk signals:** what is happening; plausible scale indicators (qualitative if necessary).
- **Platform response (if any):** mitigation(s) introduced; timing; stated rationale; external critiques (if part of the case materials).
- **Triggering events:** notices/complaints; enforcement requests; escalation events; deadlines.

G2. Minimal assumptions (kept explicit). If the use case omits operational details required to run the tree, state the minimal assumption needed to proceed, and keep it stable across cases:

- **System structure:** assume the presence of relevant subsystems only where necessary (e.g., moderation, ranking, reporting).
- **Time horizon:** define the assessment window (e.g., first week, first month) if not specified.
- **Scope boundaries:** define what is considered “in scope” (e.g., on-platform surfaces only; excluding external websites).

G3. Explicit unknowns (must not be inferred). List critical elements that are unknown and therefore cannot be used to justify determinate conclusions:

- **Process unknowns:** internal workflows (moderation, escalation, SLAs), governance arrangements, training procedures.
- **Technical unknowns:** detection tools, ranking signals, recommender settings, similarity matching, watermarking, model details.
- **Measurement unknowns:** prevalence, reach, duration, geographic/language distribution, affected populations, error rates.
- **Legal unknowns:** applicable national law determinations; court orders; lawfulness findings (unless explicitly provided).
- **Actor unknowns:** coordination, intent, identity verification, repeat-offender behaviour (unless explicitly provided).

I.2. USE CASE TESTING WALKTHROUGH

I.2.1. DSA APPLICABILITY

This subsection documents the *walkthrough inputs* used to run the toolkit on the worked case. We first record the DSA scoping assumptions that determine whether the decision-tree paths for systemic risk assessment and auditability are applicable: (i) the service is treated, for evaluation purposes, as a Very Large Online Platform (VLOP) meeting the Art. 33 DSA user-threshold, and (ii) the service is classified as an online platform (rather than a mere hosting provider or search engine).

Does the platform host 45 million active users in the EU?

Yes (assumed for evaluation). For this case, the service is treated as a Very Large Online Platform (VLOP) within the meaning of Art. 33 DSA, i.e., it reaches an average of at least 45 million active monthly service recipients in the EU. This assumption is necessary because the decision-tree toolkit targets the DSA’s systemic risk and audit obligations (Arts. 34–35 and 37 DSA).

How would you classify the system?

Online platform. The system is a (see above – very large) social media platform that hosts and disseminates user-generated content and enables user interaction (posting, replies, reposts, profile viewing, and search/discovery).

I.2.2. CONTEXT INFORMATION ANSWERS FOR THE WORKED CASE**Briefly describe the purpose of the system.**

The purpose is to provide a large-scale social communication and content distribution infrastructure in the EU, supporting user engagement and content creation (including AI-assisted editing) and associated monetization (e.g., advertising and subscriptions). In the context of this paper, the system's purpose also implies that risk governance must be operationalized in a manner compatible with the DSA accountability architecture for VLOPs (systemic risk assessment, mitigation, and auditability).

Who is deploying the system?

A private platform provider (the VLOP operator) deploys the system and the new feature set (in-product AI image editor in replies and a companion creator app). Deployment decisions include feature design, rollout, access controls (e.g., subscription gating), content governance policies, and the technical and organizational measures used for detection and enforcement.

In which domain is the system used?

Online social networking and user-generated content distribution, including public communication (posts/replies), content discovery (feeds/search), and creator tooling for media generation and editing. The relevant regulatory domain is EU platform governance under the DSA, with adjacent relevance to e.g. data protection, copyright regulation, terrorist content, and media/or political content governance depending on the assessment focus.

What are the core technical capabilities of the system?

At a minimum, the system provides:

- Content hosting and dissemination:** user accounts, posting, replies, reposts/shares, public/private visibility, and content discovery via feed and search.
- Algorithmic curation:** ranking/recommendation mechanisms that determine order and visibility in feeds and discovery surfaces (e.g. keyword search, or hashtags).
- AI-assisted image editing:** an integrated image editor capable of transforming and generating images (including photorealistic edits) within the reply composer, plus export/creation via a companion creator app for paying users.
- Governance and enforcement surfaces:** terms and conditions and content policies, reporting/complaint content, moderation actions (e.g., removal, demotion, account measures such as demonetisation or deplatforming), and user-facing enforcement communication (e.g., internal complaint-handling mechanisms) at least at the level necessary to operate a very large online platform.
- Content persistence and re-upload handling:** storage and re-distribution of user-generated media (if user upload similar content), including the practical challenge

of dealing with copies and near-duplicates (byte-identical and non-byte-identical variants, hashing).

I.2.3. SUB-SYSTEM IDENTIFICATION AND STAKEHOLDER SPECIFICATION

Building on the context information (Section I.2.2), we decompose the platform as a system-of-systems and select relevant subsystems for the worked case. For each selected subsystem, we then specify key *direct* stakeholders (interacting with the subsystem) and *indirect* stakeholders (affected by subsystem operation).

What subsystems does the system consist of?

For this use case walkthrough, we treat the VLOP as a system-of-systems and select the following subsystems as in-scope for subsequent steps:

- **Interaction layer (front-end):** posting and reply composer, media upload/display, LLM-based image editing UI, profile related UI, and user reporting interfaces, internal complaint handling mechanisms.
- **Data layer (back-end):** media storage, metadata management, logs relevant to (re)distribution and enforcement (e.g., impressions/shares/takedowns), and any similarity/near-duplicate indexing (where implemented).
- **Algorithmic curation subsystem:** ranking/recommender logic for feeds and discovery (including reply visibility of comments and amplification of content).
- **Content moderation subsystem:** automated and/or human review processes, enforcement tooling, escalation pathways, and appeal/complaint handling.
- **Policy/Terms of Service enforcement subsystem:** policy definition and operationalization, including enforcement thresholds and consistency mechanisms.
- **Monetization/access-control subsystem:** subscription gating (paid access) and related feature entitlement checks used as a mitigation measure.
- **Creator tooling subsystem:** LLM creator app and its integration points (export/re-upload pathways, interoperability with posting surfaces, and links to the paywall).

STAKEHOLDER ROLE IDENTIFICATION, VULNERABILITY LENS, AND RISK ASSESSMENT

This subsection instantiates the decision-tree for the walkthrough of the use case, focusing on the subsystems most causally implicated in the generation, dissemination, and remediation of non-consensual sexualized imagery. In line with the protocol in Appendix I.1, where required information is not available from the vignette/ use case (e.g., internal metrics, model specifications, enforcement), we record *insufficient evidence* and log the minimum evidence required.

Stakeholder role identification (per relevant subsystem). We specify stakeholders as *roles* (direct/indirect) and select those most relevant to the rights impacts in this use case.

S1: Interaction layer (front-end surfaces)

Scope: reply composer, AI image editor UI, media upload/display, profile/viewing surfaces, reporting and complaint intake UI.

Direct stakeholders: (D1) users creating edits within the in-product editor; (D2) users uploading/posting/reposting images; (D3) users viewing/engaging with posts and replies; (D4) reporters/complainants submitting notices (incl. Trusted Flaggers where applicable); (D5) advertisers on the platform; (D6) depicted/targeted persons, user the deepfake was generated on (Austrian female politician).

Indirect stakeholders: (I1) other depicted/targeted persons (incl. non-users); (I2) groups disproportionately targeted by sexualized abuse (e.g. women and girls with a specific racial background or exposure to the public, like female politicians); (I3) broader audiences exposed to the content; (I4) users whose experience is affected by UI friction or reporting burdens (e.g. parents).

S2: Algorithmic curation subsystem (ranking/recommender + discovery)

Scope: feed/reply ranking, recommendation and amplification logic, discovery surfaces (incl. search results ordering where relevant), demotion/downranking interventions.

Direct stakeholders: (D5) product/ML teams configuring the LLM tool (e.g. add guardrails) ranking and recommender parameters; (D6) trust & safety teams applying distribution and other content moderation controls (downranking, visibility limits); (D7) internal reviewers/auditors assessing curation-related controls.

Indirect stakeholders: (I1–I3) insofar as amplification increases reach and persistence; (I5) lawful creators/users whose expression or reach is affected by demotion or false positives.

S3: Content moderation subsystem (detection, review, enforcement, remedies)

Scope: automated detection and triage, human review (platform employed/ outsourced), enforcement actions (removal/visibility restriction/account measures), escalations, appeals/complaints handling, user communications (e.g., reasons, outcomes in internal complaint handling).

Direct stakeholders: (D8) moderators (in-house/contracted); (D9) detection/abuse-response operators; (D10) escalation and case-handling teams (incl. high-priority cases or Trusted Flagger prioritisation); (D11) appeals/complaints handlers.

Indirect stakeholders: (I1) targets seeking swift remedy (e.g. traumatized users); (I6) users subject to enforcement actions; (I7) regulators/DSCs and civil society seeking accountability (indirectly affected via outcomes and transparency, see Art. 86 DSA); (I8) content reporting users.

S4: Policy/Term of Service enforcement subsystem (rule-definition & operationalization)

Scope: policy drafting and updates, mapping and operationalizing inter- and supranational legal concepts to internal rules and building up legal moderation capacity (e.g. jurisdictional application of the definition of ‘illegal content’ in relation to EU Member States), enforcement thresholds (examples when a piece of content falls under an illegal content category), consistency mechanisms in enforcement, guidance/playbooks to educate moderators, testing with teams and open QA for policy implementation of LLMs.

Direct stakeholders: (D12) policy authors (e.g. Trust and safety teams); (D13) legal/compliance teams translating legal obligations into enforceable rules within the platform ecosystem; (D14) enforcement and policy-operations staff maintaining consistency across content categories/risks/regions/languages in not harmonized legal land-

scapes.

Indirect stakeholders: (I1–I3), insofar as rule design shapes protection; (I6) users whose content is evaluated under these rules (see enforcement and internal complaint handling mechanisms); (I7) platform external regulators/DSCs/Commission/auditors assessing compliance with risk assessment, consistency, transparency, and accountability.

S5: Data layer (storage, logs, similarity/near-duplicate handling)

Scope: media storage, metadata, event/log pipelines (impressions/shares/takedowns), similarity indexing and “equivalent/near-duplicate” tooling where present, reporting datasets used for transparency/audit.

Direct stakeholders: (D15) data engineering/analytics; (D16) security/abuse response; (D17) privacy/data protection and compliance officers; (D18) auditors/(vetted) researchers where access is granted.

Indirect stakeholders: (I1–I3) insofar as data capabilities enable detection, measurement, and remedy; (I7) regulators/DSCs relying on evidence quality and retention.

S6: Monetization/access-control subsystem (subscription gating, entitlements)

Scope: paid feature access, entitlement checks, pricing/payment flows, enforcement of moderation decisions as a mitigation measure (gate and access controls).

Direct stakeholders: (D19) paying subscribers; (D20) non-paying users; (D21) product teams implementing gating/feature access; (D22) billing and customer support for payments/access issues.

Indirect stakeholders: (I1–I3); (I5) lawful creators affected by restricted tooling; (I7) regulators/DSCs/Commission evaluating whether gating/paywall/access control actually mitigates risk or merely displaces it (since paying users still have access to the functionality).

S7: Creator tooling subsystem (standalone creator app + integration points)

Scope: standalone creation/editing app, export pipelines, interoperability with posting/reply surfaces, watermarking/metadata defaults, re-upload pathways.

Direct stakeholders: (D23) users creating content in the companion app; (D24) app product/engineering teams; (D25) integration teams managing cross-surface sharing and content provenance features.

Indirect stakeholders: (I1–I3), insofar as alternative creation pathways sustain abuse despite in-product restrictions; (I7) regulators/DSCs/Commission assessing whether mitigations cover all relevant risk areas.

Vulnerability lens via resilience (per stakeholder role). We operationalize vulnerability as *heightened risk to fundamental rights* arising from (i) reduced **personal resilience** (ability to anticipate, contest, and recover from platform-mediated harms) and/or (ii) reduced **structural resilience** (legal/technical/social scaffolding that enables protection, remedy, and accountability).

- **Targets of Non-Consensual Images (NCIS) (I1) — personal resilience:** limited practical capacity to monitor re-uploads; constrained ability to gather evidence; psychological distress and reduced agency; reliance on platform responsiveness and clear reporting channels, long-term consequences.

- **Targets of NCSI (I1) — structural resilience:** need for rapid containment, dependence on effective notice-and-action mechanisms, similarity handling, and escalation; cross-border enforcement frictions (same understanding of illegality in member states?); asymmetry of resources versus platform; limited transparency into moderation/ranking decisions, long-term consequences.
- **Minors / child targets (subset of I1/I2) — personal resilience:** reduced capacity to understand remedies and pursue complaints (we also acknowledge that some minors might be digitally more literate than some adults, however, we keep them under this categorization as indicating the potentially most vulnerable situation); limited practical capacity to monitor re-uploads; constrained ability to gather evidence; psychological distress and reduced agency; reliance on platform responsiveness and clear reporting channels heightened harms and long-term consequences.
- **Minors / child targets — structural resilience:** heightened need for rapid containment, specialized handling, and protective-by-default design; dependency on safeguards and reporting pathways tailored to minors and parents to act on their behalf (we acknowledge that in some situations children might not be able to reach out or do not want to reach out to an adult which can classify another vulnerable situation). CSAM regulations and prosecution of perpetrators. Liability of platform developers for allowing the system to produce sexualized content of minors.
- **Public officials (I3) — personal resilience:** exposure to coordinated intimidation; reputational harm with public salience; constraints on disengagement due to public role.
- **Public officials (I3) — structural resilience:** need for robust equivalence/near-duplicate handling and escalations; heightened public-interest scrutiny; potential links to electoral integrity and participation.

I.2.4. RISK IDENTIFICATION AND ASSESSMENT

We identify risks as subsystem-mediated pathways to *interferences* with Charter rights, and assess severity across: (i) **objective dimension** (normative seriousness of the infringement/interference), (ii) **subjective dimension** (experienced severity for affected (direct/indirect) stakeholders), (iii) **real-life consequences** (e.g. impacts, well-being, financial loss, and reversibility). Where legality under Member State law is not provided, we avoid definitive illegality determinations and focus on rights-relevant interferences and auditability.

FUNDAMENTAL-RIGHTS ASSESSMENT:

In this section we will only focus on a part of the assessment, limiting this assessment on the fundamental right of freedom of speech – Art. 11 of the Charter – for the context of the Austrian politician.

Short context description. A non-consensual sexualized edit of an Austrian parliamentarian's photo is generated and reposted; the depicted person requests takedown of the image and suppression of “similar/equivalent” re-uploads.

Right under test

Charter Art. 11 (Freedom of expression and information).

Subsystem(s) in scope

S3 Content moderation (takedown, account measures, escalation, appeals); S5 Data layer (similarity / near-duplicate handling for “equivalent” re-uploads); optionally S1 Interaction layer (reporting UI; statement-of-reasons surface).

Stakeholders (restricted party)

D2 Users posting/reposting images (removal/blocking restricts their ability to publish); I5 Lawful creators/users (risk of erroneous removal or chilling effects if equivalence matching is over-inclusive).

Stakeholders (protected party)

I1 Depicted/targeted person (the politician); optionally I2 disproportionately targeted groups (e.g., women/female politicians) as a vulnerability lens.

Measure / action assessed

Removal of the repost(s) and suppression of “similar/equivalent” variants (near-duplicates) following notice by the depicted person and persistent re-uploads.

Interference check (Art. 11)

Yes. The measure restricts posting and redistribution of the image and near-duplicates (a modality of expression). Spillover risk exists if similarity controls capture lawful content (e.g., commentary, satire, reporting).

Severity (scope / duration / reversibility)•Scope: case-linked (targets the NCSI content and near-duplicates), but can broaden via similarity matching (false positives).

•**Duration:** short-to-long lasting; depends on persistence controls and any time limits; *insufficient evidence (IE)* on the vignette.

•**Reversibility:** partly reversible via appeal/complaint mechanisms (Art. 20-21); *IE* on specific workflows and outcomes.

Probability

High probability of interference (enforcement is triggered by design); *IE* on the probability of over-removal due to missing error-rate evidence.

Art. 52(1) step 1: Provided for by law

IE. The vignette does not provide the applicable ToS/policy clauses invoked, however the use case might target Austrian criminal law under § 107c StGB (cybermobbing), or § 120a which criminalizes making an intimate image accessible to third parties or publishing it without the depicted person’s consent[615].

Art. 52(1) step 2: Legitimate aim

Yes (high confidence). Protecting the rights and freedoms of others: preventing harms to the depicted person (e.g. dignity, private life, data protection) in a scenario involving non-consensual sexualized imagery.

Art. 52(1) step 3: Necessity (least restrictive)

Plausible but IE. Given persistence through re-uploads and near-duplicates, limiting to exact-copy takedowns is likely insufficient. However, the vignette does not evidence whether less restrictive alternatives (narrower matching, visibility limits, staged enforcement) would be comparably effective.

Art. 52(1) step 4: Proportionality (balancing)

IE overall. The protected-side interest is strong due to the seriousness and persistence of NCSI. The restricted-side interest is weaker for reposting the invasive image itself, but overbroad equivalence blocking can affect lawful expression (I5). Proportionality cannot be affirmed without evidence of safeguards (SoR quality, remedy/appeals, measurable false positives/negatives, transparency on thresholds).

Assurance / confidence

Legitimate aim: High. **Provided-by-law / necessity-as-least-restrictive / proportionality:** Low to IE on the vignette.

Minimum evidence required (IE → substantiated) •ToS/policy clauses invoked and user-facing definition of “equivalent/similar”; legal criminal law definition [615].

- Enforcement logs: time-to-action after notice/ identification; noticing person (e.g. Trusted Flagger); recurrence rate; suppression rate for re-uploads/near-duplicates; account's posting.
- Similarity thresholds and audit samples; false positive/false negative rates for equivalence matching and hashing.
- Remedy/appeals outcomes: overturn rate; turnaround times; common error categories.

I.2.5. DOCUMENTATION

- the scoped *subsystem(s)* and *stakeholder roles* used for the test (“per subsystem / per stakeholder”),
- the *Charter right(s)* under assessment and a concise description of the *interference mechanism* (i.e., the platform action or design choice that constrains the right),
- node-level outputs for the core test spine (interference → severity/probability → Art. 52(1) steps: legal basis, legitimate aim, necessity, proportionality),
- a set of *assurance/confidence tags* indicating whether each conclusion is supported by the vignette or remains *insufficiently evidenced (IE)*,
- an *evidence map* specifying the minimum provider-side and user-facing artefacts required wherever conclusions are underdetermined (e.g., ToS clauses invoked, statement-of-reasons text, appeal outcomes, similarity thresholds and error rates),
- mitigation notes, including proportionality-relevant *safeguards* (e.g., transparency and remedy pathways) and any identified spillover risks (e.g., false positives affecting lawful expression),

Table I.1: Walkthrough summary for the worked case: decision-tree outputs, including the Art. 11 (freedom of expression) test record.

ID	Output item	Subsys.	Rights implicated	Sev./Assur.
R1	NCSI at scale (gen→dist→persist)	S1,S2,S3,S5,(S7)	1/7/8 (+21/+24; +11 chill)	H / M–L
R2	Near-duplicate persistence after notice	S3,S5,(S2)	1/7/8 (+11 spillover)	H / M–L
R3	Paywall mitigation (displacement + collateral)	S6,S1,(S7) [†]	1/7/8; +11 (+16)	M–H / L
R4	Remedy/process bottlenecks (N&A, escalation, SoR, appeals)	S1,S3,S4,S5	1/7/8 (ind.); procedural fairness	M–H / IE
M1	Art. 52(1) proportionality test for paywall mitigation	S6,S1 [‡]	Protects 1/7/8; may limit +11	– / IE
FR1	Human-rights test (Art. 11): politician NCSI enforcement (takedown + “equivalent” near-duplicate suppression), incl. Art. 52(1) steps	S3,S5,(S1) [§]	+11 (balancing vs 1/7/8)	M / IE

Legend. Subsys.: S1=Interaction, S2=Curation, S3=Moderation, S4=Policy/ToS, S5=Data, S6=Access-control, S7=Creator app. Rights: numbers refer to Charter Articles; “+” indicates additional implicated rights. Sev.: H=High, M=Medium. Assur.: M–L=Medium/Low, L=Low, IE=Insufficient evidence.

[†] Depends on whether the creator app remains an alternative creation pathway (degree of displacement).

[‡] Proportionality requires evidence from S2/S3/S5/S7 (suitability, least-restrictive means, balancing, safeguards).

[§] Art. 11 enforcement depends on evidence for equivalence thresholds, false positives/negatives, SoR quality, and appeal outcomes.

- a consolidated list of decision-tree nodes flagged as *evidence-dependent* to support audit scoping and future data collection planning.

J

APPENDIX FOR CHAPTER 16

J.0.1. AI-ASSISTED LABELLING PROMPT (PRIMARY TARGET LABEL)

System / role: You are assisting with *descriptive* labelling of European Commission DSA enforcement communications. You must follow the instructions exactly.

Task: Assign *exactly one* PrimaryTarget label to the Commission item below, using only the Commission's own framing in the headline and lead summary. If the item plausibly fits multiple labels, choose the *single most central* theme as signalled by the headline/lead paragraph. If the theme cannot be inferred from the provided text, output UNCLEAR.

Allowed labels (choose one):

- MinorsProtection_AgeAssurance
- IllegalGoods_MarketplaceCompliance
- IllegalContent
- AdsTransparency_AdRepository_Monetisation
- ResearcherAccess_DataTransparency
- RecommenderSystems_EngagementRisks
- ElectionIntegrity_Disinformation_CrisisResponse
- GenerativeAI_Risks
- SecurityPrivacy
- UserReporting_Complaints_DarkPatterns
- ContentModerationResources
- Other (use only if clearly outside all categories)

- UNCLEAR

Input (record):

- **Date:** {YYYY-MM}
- **Platform(s):** {...}
- **Action type:** {RFI / preliminary findings / fine / proceedings / other}
- **Headline:** {...}
- **Lead summary:** {paste the first 2–5 sentences of the Commission text you stored for this record}

Output format (strict):

PrimaryTarget: <one label from the list above>

Rationale: <1-2 sentences, referencing only the given
Headline/Lead summary; no external knowledge>

Confidence: <High/Medium/Low>

Constraints:

- Use only the provided Commission text; do not infer beyond it.
- Do not cite articles or facts not explicitly present in the input.
- Do not output more than one label.

J.0.2. TABLE 1. DECISIONS BY CATEGORY (FINES, PRELIMINARY FINDINGS,
REQUESTS FOR INFORMATION)

Category	Date	Platform(s)	Commission action (title)	Useful details (abstract)	Source
Request for information	2023-10	Meta	Commission sends request for information to Meta under the Digital Services Act	The European Commission has formally sent Meta a request for information under the Digital Services Act (DSA). The Commission is requesting Meta to provide more information on the measures it has taken to comply with ob...	https://digital-strategy.ec.europa.eu/en/news/commission-sends-request-information-meta-under-digital-services-act-0
Request for information	2023-10	TikTok	Commission sends request for information to TikTok under the Digital Services Act	Today, the European Commission formally sent TikTok a request for information under the Digital Services Act (DSA). The Commission is requesting TikTok to provide more information on the measures it has taken to comply ...	https://digital-strategy.ec.europa.eu/en/news/commission-sends-request-information-tiktok-under-digital-services-act
Request for information	2023-10	X	The Commission sends request for information to X under DSA	Today the European Commission services formally sent X a request for information under the Digital Services Act (DSA).	https://ec.europa.eu/commission/presscorner/detail/en/ip_23_4953
Request for information	2023-11	AliExpress	Commission sends request for information to AliExpress under the Digital Services Act	The European Commission has formally sent AliExpress a request for information under the Digital Services Act (DSA). The Commission is requesting AliExpress to provide more information on the measures it has taken to co...	https://digital-strategy.ec.europa.eu/en/news/commission-sends-request-information-aliexpress-under-digital-services-act
Request for information	2023-11	Amazon	Commission sends request for information to Amazon under the Digital Services Act	The European Commission has formally sent Amazon a request for information under the Digital Services Act (DSA). The Commission is requesting Amazon to provide more information on the measures it has taken to comply wit...	https://digital-strategy.ec.europa.eu/en/news/commission-sends-request-information-amazon-under-digital-services-act

Continued on next page

Category	Date	Platform(s)	Commission action (title)	Useful details (abstract)	Source
Request for information	2023-11	Meta; Snap	Commission sends requests for information to Meta and Snap under the Digital Services Act	The European Commission has formally sent Meta and Snap requests for information under the Digital Services Act (DSA). The Commission is requesting the companies to provide more information on the measures they have tak...	https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-meta-and-snap-under-digital-services-act
Request for information	2023-11	TikTok; YouTube	Commission sends requests for information to TikTok and YouTube under the Digital Services Act	Today, the European Commission has formally sent TikTok and YouTube requests for information under the Digital Services Act (DSA). The Commission is requesting the companies to provide more information on the measures t...	https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-tiktok-and-youtube-under-digital-services-act
Request for information	2023-12	Apple; Google	Commission sends request for information to Apple and Google under the Digital Services Act	The European Commission has formally sent requests for information under the Digital Services Act (DSA) to Apple and Google. The Commission is requesting the providers of these services to provide more information on ho...	https://digital-strategy.ec.europa.eu/en/news/commission-sends-request-information-apple-and-google-under-digital-services-act
Request for information	2023-12	Apple; Google	Commission sends request for information to Apple and Google under the Digital Services Act	The European Commission has formally sent requests for information under the Digital Services Act (DSA) to Apple and Google. The Commission is requesting the providers of these services to provide more information on ho...	https://digital-strategy.ec.europa.eu/en/news/commission-sends-request-information-apple-and-google-under-digital-services-act
Request for information	2023-12	Meta	Commission sends request for information to Meta under the Digital Services Act	Today, the European Commission has formally sent Meta a request for information under the Digital Services Act (DSA).	https://digital-strategy.ec.europa.eu/en/news/commission-sends-request-information-meta-under-digital-services-act

Continued on next page

Category	Date	Platform(s)	Commission action (title)	Useful details (abstract)	Source
Request for information	2024-01	Very; Large; Online; Platforms; Search; Engines	Commission sends requests for information to 17 Very Large Online Platforms and Search Engines under the Digital Services Act	The Commission has sent formal requests for information under the Digital Services Act (DSA) to 17 of the Very Large Online Platforms and Search Engines (VLOPs and VLOSEs) designated on 25 April 2023, namely AliExpress,...	https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-17-very-large-online-platforms-and-search-engines-under
Request for information	2024-03	LinkedIn	Commission sends request for information to LinkedIn on potentially targeted advertising based on sensitive data under Digital Services Act	The European Commission has formally sent LinkedIn a request for information under the Digital Services Act (DSA), asking for more details on how their service complies with the prohibition of presenting advertisements ...	https://digital-strategy.ec.europa.eu/en/news/commission-sends-request-information-linkedin-potentially-targeted-advertising-based-sensitive-data
Request for information	2024-03	Meta	Commission sends request for information to Meta under the Digital Services Act	Today, the Commission has formally sent Meta a request for information (RFI) under the Digital Services Act (DSA). The Commission is requesting Meta to provide more information related to the Subscription for no Ads opt...	https://digital-strategy.ec.europa.eu/en/news/commission-sends-request-information-meta-under-digital-services-act-1
Request for information	2024-03	Very; Large; Online; Platforms; Search; Engines	Commission sends requests for information on generative AI risks to 6 Very Large Online Platforms and 2 Very Large Online Search Engines un...	The Commission has formally sent requests for information under the Digital Services Act (DSA) to Bing and Google Search (Very Large Online Search Engines, or VLOSEs), as well as to Facebook, Instagram, Snapchat, TikTok...	https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-generative-ai-risks-6-very-large-online-platforms-and-2-very

Continued on next page

Category	Date	Platform(s)	Commission action (title)	Useful details (abstract)	Source
Request for information	2024-04	TikTok; Lite; France; Spain	Commission sends request for information to TikTok regarding the launch of TikTok Lite in France and Spain	The Commission has sent TikTok a request for information under the Digital Services Act (DSA), asking for more details on the risk assessment the provider of TikTok should have carried out before deploying the new app T...	https://digital-strategy.ec.europa.eu/en/news/commission-sends-request-information-tiktok-regarding-launch-tiktok-lite-france-and-spain
Request for information	2024-05	Microsoft; Bing	Commission compels Microsoft to provide information under the Digital Services Act on generative AI risks on Bing	The Commission steps up its enforcement actions against Microsoft: following an initial request for information sent on 14 March regarding specific risks stemming from Bing's generative AI features, notably "Copilot in ..."	https://digital-strategy.ec.europa.eu/en/news/commission-compels-microsoft-provide-information-under-digital-services-act-generative-ai-risks
Request for information	2024-05	X	Commission requests information from X on decreasing content moderation resources under the Digital Services Act	The Commission has sent X a request for information (RFI) under the Digital Services Act (DSA).	https://digital-strategy.ec.europa.eu/en/news/commission-requests-information-x-decreasing-content-moderation-resources-under-digital-services
Request for information	2024-06	Pornhub; XVideos; Stripchat	Commission sends request for information on illegal content and protection of minors to Pornhub, XVideos and Stripchat under the Digital Se...	The Commission has sent Pornhub, Stripchat and XVideos requests for information under the Digital Services Act (DSA).	https://digital-strategy.ec.europa.eu/en/news/commission-sends-request-information-illegal-content-and-protection-minors-pornhub-xvideos-and

Continued on next page

Category	Date	Platform(s)	Commission action (title)	Useful details (abstract)	Source
Request for information	2024-06	Temu; Shein	Commission requests information from online marketplaces Temu and Shein on compliance with the Digital Services Act	The Commission has sent the online marketplaces Temu and Shein formal requests for information (RFIs) under the Digital Services Act (DSA).	https://digital-strategy.ec.europa.eu/en/news/commission-requests-information-online-marketplaces-temu-and-shein-compliance-digital-services-act
Preliminary findings	2024-07	X	Commission sends preliminary findings to X for breach of DSA	Today, the Commission has informed X of its preliminary view that it is in breach of the Digital Services Act (DSA) in areas linked to dark patterns, advertising transparency and data access for researchers.	https://ec.europa.eu/commission/presscorner/detail/en/ip_24_3761
Request for information	2024-07	Amazon	Commission requests information to Amazon under the Digital Services Act	The Commission has sent Amazon a request for information (RFI) under the Digital Services Act (DSA).	https://digital-strategy.ec.europa.eu/en/news/commission-requests-information-amazon-under-digital-services-act
Request for information	2024-08	Meta	Commission sends request for information to Meta under the Digital Services Act	The Commission has sent Meta a request for information (RFI) under the Digital Services Act (DSA).	https://digital-strategy.ec.europa.eu/en/news/commission-sends-request-information-meta-under-digital-services-act-2
Request for information	2024-10	Pornhub; Stripchat; XVideos	Commission requests information under the Digital Services Act to Pornhub, Stripchat and XVideos on their transparency reports and adverts	The Commission has sent requests for information (RFI) to the providers of Pornhub, Stripchat and XVideos under the Digital Services Act (DSA).	https://digital-strategy.ec.europa.eu/en/news/commission-requests-information-under-digital-services-act-pornhub-stripchat-and-xvideos-their

Continued on next page

Category	Date	Platform(s)	Commission action (title)	Useful details (abstract)	Source
Request for information	2024-10	Temu	Commission requests information under the Digital Services Act from Temu on traders selling illegal products on its marketplace	The Commission has sent a request for information (RFI) to Temu under the Digital Services Act (DSA), requesting the platform to provide detailed information and internal documents on the mitigation measures taken again...	https://digital-strategy.ec.europa.eu/en/news/commission-requests-information-under-digital-services-act-temu-traders-selling-illegal-products
Request for information	2024-10	YouTube; Snapchat; TikTok	Commission sends requests for information to YouTube, Snapchat, and TikTok on recommender systems under the Digital Services Act	The Commission sent a request for information to YouTube, Snapchat, and TikTok under the Digital Services Act (DSA), asking the platforms to provide more information on the design and functioning of their recommender sy...	https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-youtube-snapchat-and-tiktok-recommender-systems-under-digital
Request for information	2024-11	TikTok	Commission sends additional request for information to TikTok under the Digital Services Act	The European Commission sent TikTok a request for information (RFI) under the Digital Services Act (DSA).	https://digital-strategy.ec.europa.eu/en/news/commission-sends-additional-request-information-tiktok-under-digital-services-act
Request for information	2025-02	Shein	Commission requests information from Shein on illegal products and its recommender system	The Commission has sent a request for information (RFI) to Shein under the powers granted by the Digital Services Act (DSA).	https://digital-strategy.ec.europa.eu/en/news/commission-requests-information-shein-illegal-products-and-its-recommender-system
Fines	2025-12	X	Commission fines X EUR 120 million under the Digital Services Act	Today, the Commission has issued a fine of EUR 120 million to X for breaching its transparency obligations under the Digital Services Act (DSA).	https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2934

ACKNOWLEDGEMENTS

A dissertation carries a single name on its cover, but it is never the work of one person alone. This thesis is the result of several years of research, collaboration, discussion, disagreement, experimentation, rewriting, and learning. It was shaped by the people who supervised me, wrote papers with me, shared their expertise and data, participated in studies and workshops, read unfinished drafts, questioned my assumptions, and supported me through the less visible parts of doing a PhD. I have been extraordinarily fortunate in the number and diversity of people from whom I have been able to learn.

SUPERVISION AND INTELLECTUAL GUIDANCE

First and foremost, I would like to thank **Ben Wagner**. Ben has been central to this PhD from its beginning and is also a co-author of several of the works that became part of this dissertation. I am deeply grateful for the freedom he gave me to develop this research agenda and to follow questions that did not always fit neatly within a single discipline. Working on the Digital Services Act required moving continuously between law, platform governance, empirical research, computer science, HCI, political science, and questions of power and accountability. Ben encouraged precisely this kind of boundary-crossing research. I am particularly thankful for his willingness to entertain new ideas, even when they initially consisted of little more than an observation that something in the DSA's implementation did not quite add up. Many of the projects in this dissertation began in exactly this way. Our collaborations on transparency, consistency, content moderation, reporting mechanisms, electoral risks, and digital rights gave me the opportunity to develop these observations into systematic research questions. I also learned from Ben to look beyond individual provisions and technical systems and to consider the institutional and political structures in which they operate. Most importantly, I am grateful for the trust and independence I was given throughout the PhD. That freedom allowed this dissertation to become what it is: not an examination of one isolated provision of the Digital Services Act, but an attempt to understand how a complex regulatory system can actually be evaluated empirically.

I would also like to thank **Hans de Bruijn**, whose way of thinking has influenced this dissertation far beyond our collaboration on the Instagram transparency study. Hans has a remarkable ability to look at a highly complex problem, resist being distracted by its complexity, and reduce it to the few questions that actually matter. During my time at TU Delft, I repeatedly learned from this approach: first understand the complexity, but then impose structure on it. I am particularly grateful for the intellectual discipline Hans brought to my work. When a project threatened to become too complicated, his questions often brought it back to its essentials: What are the three main questions? What are the three findings somebody should remember after reading the paper? And, most

importantly, *so what?* What changes once we know this, why does the finding matter, and what larger conclusion can legitimately be drawn from it? I am also grateful to Hans for embodying something I came to value greatly about doing this PhD at TU Delft: the expectation that complex societal and technological problems should not merely be described, but analysed in a way that makes their underlying structure visible and their implications understandable. His conceptual perspective, his insistence on clarity, and his ability to ask the deceptively simple question that exposes what an argument is still missing have made me a more disciplined researcher. Thank you, Hans, for repeatedly reminding me that understanding complexity is only the first step; the harder task is to explain clearly what matters within it, what we have learned, and why anyone should care.

DOCTORAL COMMITTEE

I would also like to thank the members of my doctoral committee for taking the time to read this dissertation so carefully and for engaging with a thesis that deliberately moves across disciplinary boundaries.

I am especially grateful to **Dave Murray-Rust** for the perspective he has brought to this PhD over the years, both through the **AI Futures Lab** and as a member of my committee. Dave repeatedly encouraged me to look beyond the question of whether a technological or regulatory system formally works and to ask how it is actually encountered, understood, and experienced by people. His design and human-centred perspective helped broaden the way I approached many of the questions in this dissertation, particularly where legal rules eventually become interfaces, procedures, and concrete choices for users. I have also greatly valued Dave's openness to experimental and interdisciplinary ideas, his ability to see possibilities in projects that were still only half formed, and the many conversations through which those ideas became clearer. Thank you for helping create an academic environment in which critique and constructive thinking could comfortably sit alongside one another.

I would like to thank **Michel van Eeten** for the perspective he brought to this thesis as both a member of my committee and through my growing connection to the Cyber Security group at TPM. Michel's approach to research has been an important reminder that claims about complex digital systems ultimately have to survive contact with evidence. I also learned from Michel's willingness to ask direct methodological questions: What can we actually observe? What does the evidence allow us to conclude? And where are we still making assumptions? Beyond the thesis itself, I am grateful for the openness with which Michel welcomed me into the Cyber Security community and for the conversations about research, academic careers, and the different ways in which empirical work can have impact.

I am very grateful to **Tarlach McGonagle** for reading this work through the lens of media law, freedom of expression, and human rights, and for helping situate the DSA questions

addressed in this dissertation within the much longer history of governing media and communication.

I would like to thank **Linnet Taylor** for bringing questions of data governance, justice, representation, and power into the discussion and for encouraging a perspective that reaches beyond compliance alone to ask who is represented, protected, or potentially left out by digital governance.

Finally, I am grateful to **Stefania Milan** for bringing a critical perspective on data, infrastructures, participation, and digital power to the committee, and for engaging with the broader political implications of the systems and regulatory mechanisms examined throughout this dissertation.

I feel extremely fortunate to have had a committee that reflects the interdisciplinary character of the thesis itself.

THE PARANYMPHS

Sofie-Amalie Torp Dideriksen I am deeply grateful to have you as a wonderful collaborator and as a cherished colleague. I have greatly valued our conversations about research, university life, and everything beyond it, as well as the generous intellectual environment we shared at TU Delft. In moments when academic life felt like a race, you reminded me that, as Taylor Swift puts it, “only the young can run”—but also that meaningful things can be created patiently, one careful stitch at a time. Thank you for the ideas, laughter, encouragement, and knitting wisdom along the way.

I am especially grateful to **Yana Angelova**, both for her friendship and for agreeing to be my paranymp. Yana is an exceptionally bright, thoughtful, and inspiring researcher, but also simply a wonderfully cool person to have beside me. Her warmth, humour, encouragement, and calm presence have meant a great deal to me during the final stages of this PhD. I feel very fortunate that she will accompany me on the day of my defence.

THE PEOPLE BEHIND THIS DISSERTATION

I am deeply grateful to the many academic communities that gave this PhD an intellectual home. First and foremost, I want to thank my colleagues at **TU Delft**, particularly within the Faculty of **Technology, Policy and Management (TPM)**. I am especially grateful to the **Multi-Actor Systems** and **Organisation and Governance** communities for providing an environment in which complex technological questions could be understood through the institutions, relationships, and competing interests that shape them. One of the things I have valued most about Delft is the ease with which research can move between disciplines. Law, governance, computer science, cybersecurity, design, and social science regularly met around the same questions, and many of the ideas in this dissertation grew out of precisely these encounters.

Within this community, I am particularly grateful to **Martijn Warnier** for his leadership, openness, and support, and for helping to create an environment in which interdisciplinary ideas could flourish. I thank **Gerdien de Vries** for our many thoughtful and energising conversations and for always offering an open ear. I am grateful to **Wijnand Veeneman** for the conversations over tea that brought both insight and a welcome moment of calm; to **Haiko van der Voort** for the pleasure of teaching together, his thoughtful advice, and the pleasure of sharing patisserie that made the work considerably sweeter; and to **Tamara Metze** for her generous support throughout the years and across several projects.

Dearest **Cyber Security group**, I may have arrived as something of a party crasher, but by the end of this PhD, you had become one of my academic homes at Delft. I am especially grateful to **Michel van Eeten** for welcoming me so warmly and for making me feel that there was room for my somewhat different perspective, and to everyone else who made the group such a lively and generous community. Being part of this team broadened how I think about research and introduced me to questions, methods, and academic communities I might otherwise never have encountered. More importantly, it gave me wonderful colleagues with whom to share ideas, spirited discussions, academic fiestas, and successes both large and small. It has been a true pleasure to get to know you all over the years. Thank you for making me feel so thoroughly at home.

The **AI Futures Lab on Rights and Justice** has been at the centre of my time at Delft and became much more than simply the institutional home of my PhD. I am especially grateful to **Ben Wagner** and **Dave Murray-Rust** for creating and sustaining a space in which interdisciplinary, critical, and occasionally unconventional research ideas could develop. The Lab gave me the freedom to move between law, governance, design, computer science, rights, and technology, and many of the ideas and collaborations that shaped this dissertation grew out of that openness. I am particularly thankful to **Sofie-Amalie Torp Dideriksen** and **Mahan Mehrvarz** for sharing so much of the everyday experience of building and growing the Lab. I am equally grateful to the many people who joined and expanded the community over the years, as well as the wider group of associated researchers and faculty around the Lab. Thank you for unfinished ideas discussed over lunch, comments on drafts, spontaneous research conversations, workshops, shared projects, conference trips, Lab Talks, and the many moments in between. Watching the Lab grow, and seeing new people, disciplines, and perspectives become part of it, has been one of the most enjoyable parts of my time at TU Delft. It made the PhD feel considerably less like an individual project and much more like research developed within a genuine intellectual community.

I owe an especially large debt of gratitude to **Rita Gsenger**. Few people appear across as many different parts of the intellectual journey behind this dissertation. Our collaboration repeatedly required us to translate between legal concepts and empirical categories. I am grateful for all the work that is less visible in a final publication: developing ideas, organising projects, discussing coding decisions, revising arguments, preparing workshops, dealing with deadlines, responding to reviews, and repeatedly returning to a manuscript

until it finally says what we intended it to say. Our collaborations have influenced not only individual chapters but also how I think about interdisciplinary research more generally. Thank you, Rita, for being such an important part of this journey and such a wonderful friend.

I am deeply grateful to **Simon Parkin**, whose perspective has been particularly important in moving my work from asking whether a platform technically provides a regulatory mechanism to asking whether a person can actually use it. Our collaboration on reporting mechanisms under Article 16 DSA was formative for the user-facing strand of this dissertation. Simon helped bring HCI, usability, user journeys, and interface evaluation (some might call it even legal heuristic walkthroughs) into a project that could otherwise easily have remained a predominantly legal analysis. I greatly value his ability to identify the practical implication hidden inside an abstract argument and to ask what a system actually requires a user, practitioner, or institution to do. His feedback, methodological expertise, and willingness to work through difficult research-design questions (without searching for agreement in all parts) have strengthened several parts of this dissertation. Thank you also for the many conversations, comments on drafts, and collaborative ideas.

I am especially grateful to **Paola Lopez**, both as a co-author and as a friend. Our early work on algorithmic decision-making and the AMS algorithm was an important part of my academic development and interdisciplinary work. Not many people can say they know each other through a presumably discriminatory algorithm. Beyond the research itself, I am grateful for the friendship that grew around our collaboration: for the many conversations about academia and everything outside it, for sharing ideas, frustrations, and happy places. It has been particularly meaningful to look back at those early projects and see how many of the questions we discussed then continued to accompany me throughout this dissertation. Thank you, Paola, for being part of both the research and the journey around it.

I would like to thank **Shreyan Biswas** for our collaboration and for the many stimulating conversations about artificial intelligence – how these systems work, how they reshape digital platforms, and how they should be studied and (in my opinion) governed. Our work on assessing generative AI systems under the Digital Services Act challenged us to reconsider categories, UI design of reporting mechanisms, or explanations for users, and to explore how regulatory and auditing approaches must evolve alongside rapidly changing technologies. I am especially grateful for Shreyan's technical insight, intellectual curiosity, and openness to thinking across disciplinary boundaries. Our conversations continually helped me to see familiar regulatory questions from new perspectives – thank you for that.

Merci **Ujwal Gadiraju** for enriching our work with his enthusiasm, openness, and expertise in human computation and human-computer interaction, which brought an essential perspective to our discussions about how emerging AI systems can be meaningfully evaluated, how their behaviour is shaped by human judgment (and meaningful forms of oversight), and how these dynamics intersect with regulatory expectations. I

greatly valued his intellectual generosity and the opportunity to think together across technical, human-centred, and regulatory perspectives.

I am grateful to **Scott** for our collaboration on sampling methodologies for systemic-risk assessments and independent DSA audits. I greatly valued Scott's contribution to translating broad regulatory requirements into concrete methodological choices. Our collaboration strengthened how I think about the relationship between legal obligations, sampling design, and the role of statistics in evidence in DSA auditing.

The **Weizenbaum Institute** has been an especially important part of my academic journey. I am also especially thankful to **Rita Hermann-Gsenger**, **Paola Lopez**, **Rainer Rehak**, and **Esther Görnemann**, and **Volker Stocker**, who made the Weizenbaum community such an important and enjoyable part of these years. My **Weizenbaum fellowship** gave me the rare opportunity to spend concentrated time within this community, exchange ideas across disciplines, and develop collaborations outside my immediate institutional environment. I am grateful not only for the research that resulted from this time, but for the people and friendships that made the Institute feel like another academic home.

I am equally thankful for the connections beyond Delft that became part of this PhD, particularly the colleagues at **IT:U**. I am especially grateful to **Till Winkler**, for the many conversations, ideas, and collaborations that developed over these years. I greatly valued Till's perspective, curiosity, flexibility, and willingness to explore new ideas together.

The **ReSocial project** provided another important community during these years. I am particularly grateful to **Gianclaudio Malgieri** for bringing this wonderfully interdisciplinary group together, for including me in it, and for the enthusiasm, openness, and energy with which he approaches both research and collaboration. Our conversations about vulnerability, fundamental rights, platform regulation, and the DSA have given me new ways of thinking about many of the questions addressed in this dissertation. I am equally thankful to **Nina Baranowska**, **Maria-Lucia Reborean**, **Manon Carrere**, and **Lorena Sánchez Chamorro**, and the wider ReSocial community. I greatly valued being part of a project.

I am thankful to **Hannah Ruschemeier** for her friendship, support, and guidance throughout this PhD. I have greatly valued having someone with whom I could discuss not only questions of law and research, but also the much less straightforward questions about academic careers, opportunities, and what comes next. Her encouragement and career advice came at important moments and helped me think more clearly about the choices ahead of me. Thank you, Hannah, for being someone I could turn to for both a thoughtful legal perspective and a very good conversation.

I would like to thank **Matthias C. Kettemann** for the collaborations that preceded and accompanied this PhD and for helping shape my thinking about platform governance as a legal and institutional field. We worked together on platform governance during crisis

situations, deplatforming, legal interpretations in content moderation, and electoral risks. Those projects were important precursors to the questions pursued in this dissertation. I am grateful for Matthias's legal expertise, intellectual generosity, and for the opportunities to work together across several different stages of my academic development.

I am also grateful to the community at **WU Vienna**, which was an important part of my academic development before this PhD. My time there gave me some of my earliest experiences of interdisciplinary research, collaborative projects, teaching, academic debate, and the everyday life of being part of a university community. I am thankful to the mentors, colleagues, collaborators, and friends I met at WU for the conversations, encouragement, criticism, shared projects, and many hours spent discussing ideas that were often still far from becoming papers. I remain very grateful for that time and for all the people at WU who were part of it.

Across all of these communities, I have been fortunate to work with people who were willing to cross disciplinary boundaries, challenge assumptions, share expertise generously, and entertain research questions that did not always have obvious answers. This dissertation is much richer because of them.

RESEARCH PARTICIPANTS AND PRACTITIONERS

The research in this dissertation would not have been possible without the many people who shared their time, expertise, and practical experience with me. I am deeply grateful to everyone who participated in interviews, workshops, discussions, and informal exchanges throughout the course of this PhD.

In particular, I would like to thank representatives of the **European Commission, Digital Services Coordinators, out-of-court dispute settlement bodies, Trusted Flagger organisations, civil-society organisations and NGOs**, as well as researchers and other practitioners working on the implementation of the Digital Services Act. Their willingness to speak openly about emerging procedures, institutional challenges, resource constraints, platform relationships, and the practical realities of implementing the DSA made it possible to study the Regulation as it operates in practice rather than only as it appears on paper.

Many of these exchanges took place in formal interviews and workshops, while others happened during conferences, research visits, meetings, or over coffee. Together, they provided perspectives that could never have been obtained from legal texts, transparency reports, or datasets alone. I am particularly grateful for the openness with which people discussed both what is working and what remains difficult in this still-evolving regulatory system.

THE WORK BEHIND THE WORK

A PhD depends on an enormous amount of labour that rarely appears in its bibliography. I want to thank the administrative and support staff who helped with research ethics,

contracts, finances, travel, teaching, events, data management, publication procedures, and ultimately the doctoral process itself. Research institutions run because people perform this work, and I am grateful to everyone who made it possible for me to concentrate on the parts that appear in this book.

I am especially grateful to the secretariat and management-support team of **Multi-Actor Systems**. In particular, I want to thank **Maaïke Vijverberg** for the great start with her, the tea time and the talks about riding and horses and life in general, **Jolanda Berbee** for all her cheers and great energy and help, **Joy van Eijk** for the laughter and appreciation of colors, and **Linda de Deugd** for her support, as well as **Joyce van Velzen** who always had an open ear.

I am particularly grateful to **Nicolas Dintzner** for his patient and pragmatic guidance on research data management, documentation, and the many practical details involved in conducting and preserving research responsibly and many nerdy chats about digital regulations. His support made institutional requirements fun, considerably easier to understand, and put into practice.

I also want to thank **Charlotte Boelens** and **Taylor Stone** from the **TU Delft AI Initiative** for their warmth, encouragement, and the many ways in which they connected people across Delft's AI community. Their work helped create the wider intellectual and social environment in which the AI Futures Lab could flourish and made the Initiative feel like a genuine community rather than simply an institutional programme.

FRIENDS

A PhD can very easily expand until it occupies every available corner of life. More broadly, I am grateful to all my friends for tolerating an unreasonable amount of conversation about the Digital Services Act while continuing to remind me that there were far more interesting things to talk about.

I am therefore especially grateful to the friends who made sure that mine never did. I am deeply thankful to **Maria Scheibelhofer** and **Dieter Giesing** for their friendship, generosity, and for providing such a constant sense of living in the moment and greater perspective throughout these years. Thank you for the dinners, conversations, advice, laughter, invitations, late nights, and the many moments in which academic deadlines suddenly seemed considerably less important.

A very special thank you goes to **Friedrich Lachmayer**, for many years of mentorship and for teaching me to think about law not only through text, but also through structures, relations, and images. Long before legal visualisation became relevant to how I approached my own research, Friedrich showed me that complicated legal ideas can be taken apart, rearranged, and made visible in entirely different ways. I am equally grateful for the many conversations that accompanied this learning – especially the long talks at Café Hummel in Vienna, where discussions could move effortlessly from legal informatics

and visualisation to academia, Vienna, and life more generally. His intellectual curiosity and generosity have accompanied me for many years and left a lasting mark on how I think about law and research.

I am equally grateful to **Herbert Lachmayer**, who demonstrated an entirely different lesson about academic life: that intellectual seriousness does not require giving up *grandezza*, style, curiosity, or *joie de vivre*. Herbert showed me that an academic career can encompass ideas, art, culture, conversation, performance, and an enormous appetite for the world around it. His way of moving between disciplines and intellectual worlds – and of doing so with unmistakable style – has always been a reminder that there is no single correct way to build an academic life. I am grateful for that example and for the encouragement to make an academic career one's own rather than merely following the path already laid out by others.

Dear **Noha** – Thank you for the conversations about stories, people, work, and the strange situations journalism so often produces, but also for the time spent around horses, where the pace and priorities are entirely different. I have always valued being able to move so naturally with you between those worlds and cities. Thank you for the friendship, shared experiences.

A particularly important counterweight to academic life came from horse riding and **Manege De Wildhoef**. The stables gave me a place where papers, deadlines, and reviewer comments became temporarily irrelevant and where attention had to shift completely to something immediate and physical. I am especially grateful for the beach rides, long hours at the stables, weekend rides, and horseback-riding weekends that punctuated these years and gave them a rhythm entirely separate from academic calendars. There are few better ways to put a reviewer comment into perspective than spending hours on horseback, riding through the dunes or along the beach, with very different things demanding one's attention. I am grateful to the horses, the riders, and everyone at Manege De Wildhoef who made the stables such a welcome escape and such an important part of life alongside the PhD.

FAMILY

Finally, and most importantly, I want to thank my family.

Mom, this thesis quite literally used to contain only the words “Thanks mom!” on this page. Thank you for supporting me long before there was a PhD to support, for encouraging my curiosity, and for trusting that all the different things I wanted to do might somehow eventually fit together (or not). Some of the most important conversations behind this journey happened around the kitchen table: about music, work, politics, decisions, worries, new ideas, things that had gone well, and things that had not. You always asked the right questions, and helped me work out what actually mattered. Thank you for being there through the moves, deadlines, successes, non-weekends, and uncertain decisions, while always giving me the freedom to find my own way – or as Thomas D

would say: "Denn wenn ihr beteuert mich zu verstehen dann lasst mich gehen – Und wir werden uns wiedersehen!". The independence that finishing a doctorate represents rests on a foundation of your support, trust, patience, and love. So, after all the additional words, the first version was perhaps still the most accurate one: *Thanks, Mom.*

I am deeply grateful to **my dad** for the intellectual curiosity he encouraged in me from the very beginning. Some of my most vivid memories are of learning together: asking questions, building things, repeating and testing vocabulary in Russian or French, trying to understand how something worked, and being taken seriously even when I was still a child. That willingness to explain, discuss, and then continue discussing has stayed with me. I am grateful for the many intellectual conversations we have had since – about economics, gardens, technology, society, history, music, and whatever new question happened to occupy us – and for teaching me that curiosity is not something one eventually grows out of. A considerable part of the person who wanted to spend several years investigating why complicated systems do not quite work as they claim to probably began in those conversations.

I am equally grateful to my stepfather, **Gernot**, for bringing a set of beautiful sensibilities into my life: an appreciation for light in an even wider sense, and the details through which things acquire their particular design and main character. Thank you for the time in Vienna and Mariazell, for many wonderful moments there, and for our shared passion for piano music. Music has been one of the threads running quietly alongside this PhD, and I am grateful for having someone with whom the difference between one Bach interpretation of Gregory Sokolov and Tatiana Nikolayeva, or Martha Argerich and Daniel Barenboim can be a perfectly reasonable subject for long conversations.

My grandmother, **Maria**, gave me another inheritance that has accompanied me throughout my life: books, opera, and a love of culture that never treated any of these things as luxuries. I am particularly grateful for the love of literature we shared. Through her, opera, bookshelves, and classical style became parts of life one could inhabit, return to, argue about, and continue discovering. Thank you!

I also want to remember my grandmother, **Chris**, with enormous affection for something that sounds simple but is anything but: her extraordinary ability to be always in a good mood. Her warmth, optimism, and capacity to enjoy the people and things around her provided a very different kind of lesson about how to move through life, and one that is particularly useful during a PhD.

I am very grateful to my aunt, **Regina**, for all the things that belong specifically to having a wonderful aunt: for support, interest, family conversations, encouragement, and for being someone who has remained present throughout the different stages of my life. To **Michael**, thank you for being much more like a brother than a cousin.

I am also deeply grateful to **Lorenz** for our shared enthusiasm for multinational music lyrics and AI research. And to **Marie**, thank you for being such a wonderful stepsister. Families are not defined only by the relationships one starts with, but also by the ones one is fortunate enough to gain along the way.

AND FINALLY

This dissertation is about evaluating compliance. It is consequently filled with things that do not align: reports that contradict datasets, rules that are difficult to operationalise, interfaces that make rights unnecessarily hard to exercise, categories that cannot be reconciled, evidence that is missing, and institutions still figuring out what a new regulatory framework requires of them.

Spending several years looking closely at these gaps could easily produce cynicism. Instead, I repeatedly encountered people – in academia, civil society, public institutions, journalism, and elsewhere – who were willing to do the much less visible work of trying to make these systems better. They answered questions, challenged platforms, built methods, collected evidence, represented users, scrutinised regulators, designed alternatives, and kept asking whether the systems we have are actually the systems we want. This thesis owes a great deal to them.

And to everyone who contributed to these years through a paper, a dataset, an interview, a workshop, a review, an argument, an introduction, an email, a conversation, a friendship, or simply by being there when it mattered: thank you.

*Marie-Therese Sekwenz
The Hague, August 2026*

AUTHORSHIP CONTRIBUTIONS

This dissertation contains several research chapters based on published or submitted papers. Below, I outline my specific contributions to each chapter, aligned with the structure of Parts I–IV of the thesis.

PART I — CONCEPTUAL ELEMENTS OF DSA COMPLIANCE

CHAPTER 2 — THE DIGITAL SERVICES ACT: RULES, RIGHTS, AND RISKS

I conducted the doctrinal legal analysis of the DSA's regulatory architecture, including transparency obligations, user and procedural rights, institutional actors, and governance mechanisms. I authored the full sections (e.g. on reporting mechanisms, transparency or trusted flaggers), synthesised relevant legal and interdisciplinary literature, and situated the DSA within the broader EU digital regulatory landscape.

CHAPTER 3 — THE PROCESS OF CONTENT MODERATION AND USER ENGAGEMENT

I analysed how the DSA restructures content moderation as a socio-technical and legal process. I developed the moderation process model, mapped platform and user roles and obligations, integrated case law and regulatory guidance, and authored the chapter.

CHAPTER 4 — WICKED BY DESIGN: REGULATING AND AUDITING PLATFORMS UNDER THE DSA

I developed the chapter's theoretical argument on socio-technical and socio-legal complexity in DSA compliance. I analysed fragmented reporting mechanisms, multi-actor dependencies, and algorithmic systems used in moderation, and produced the conceptual diagrams. I authored the chapter.

PART II — METHODOLOGICAL FRAMEWORKS TO ASSESS DSA COMPLIANCE

CHAPTER 5 — A HOLISTIC CONSISTENCY FRAMEWORK FOR DSA COMPLIANCE EVALUATION

I developed the consistency-analysis approach (internal, external, and historical consistency) and designed its application to DSA transparency artefacts. I conducted the Instagram case study and authored the chapters under the supervision of my co-authors.

CHAPTER 6 — SYSTEMIC RISK ASSESSMENT, INDEPENDENT AUDITS, AND SAMPLING METHODOLOGIES

I designed the mixed-methods audit model by integrating sampling strategies with legal content analysis. I reviewed platform audit reports, defined methodological requirements for systemic risk assessments (Art. 34) and independent audits (Art. 37), and led the conceptual development and project coordination. I assessed and curated the underlying materials and drafted core parts of the paper where I had interdisciplinary support by my co-authors.

CHAPTER 7 — DEVELOPING A LEGAL CODEBOOK FOR ELECTORAL AND POLITICAL CONTENT

I constructed the legal codebook for assessing political and electoral content by synthesising the DSA, the TTPA, and the Commission's electoral guidelines. I operationalised Art. 34(1)(c) into empirical annotation categories, validated the codebook, and led the doctrinal analysis and publication coordination. During the process my co-author was able to add valuable insights and refined the work.

CHAPTER 8 — SCALING UP DSA AUDITING: THE ROLE OF LARGE LANGUAGE MODELS

I designed the interdisciplinary workshop as well as contributed with the funding for it, and evaluated with the participants LLM-supported auditing workflows through different exercises and debates. I developed the workshop exercises for AI-assisted auditing, led the data collection, and co-conducted the analysis. I conceptualised and co-authored the paper as lead author together with the interdisciplinary group of workshop experts.

PART III — CONDUCTING EMPIRICAL EVALUATION OF DSA COMPLIANCE

CHAPTER 9 — EVALUATING THE SYSTEMIC RISK OF DISSEMINATING ILLEGAL CONTENT

I co-developed the extensive codebook and co-conducted the empirical study by annotating the corpus of platform moderation decisions. I co-developed the legal taxonomy (Austrian and German law), collected and systematised relevant legal provisions and Terms of Service clauses as well as examples of jurisprudence, and performed the annotation as coder 1. I participated in all alignment meetings and contributed to the interpretation of moderation decisions. I authored the legal-doctrinal sections and substantial parts of the chapter.

CHAPTER 10 — EVALUATING DSA REPORTING MECHANISMS: USER PATHWAYS AND DESIGN COMPLIANCE

I mapped the reporting interfaces of Facebook and TikTok and produced the legal-heuristic walkthrough materials as a novel evaluation methodology. I conceptualised the research design, designed and recruited the expert interviews, analysed Art 16 DSA

pathways, and identified Art. 16 design tensions that informed the interview protocol and analysis together with my co-authors. I authored substantial parts of the chapter.

CHAPTER 11 — AUTOMATING DSA USER SUPPORT: LLMs FOR ILLEGAL CONTENT REPORTING

I co-designed the experiment evaluating AI-assisted reporting under German law and supported the project funding for the user study. I contributed to the qualitative annotation of the data and designed the evaluation criteria, and authored key parts of the chapter analysing implications for Art. 16 and Art. 35 compliance.

CHAPTER 12 — EVALUATING SYSTEMIC ELECTORAL RISKS UNDER THE DIGITAL SERVICES ACT

I operationalised electoral-risk categories and applied the legal codebook from Chapter 7 to the 2021 German federal election. I contributed to extending the codebook to capture negative effects on the electoral process, conducted annotation and agreement meetings, supported the alignment of electoral categories, and authored substantial parts of the chapter.

PART IV — CONCLUSION, IMPACT, AND FUTURE WORK

CHAPTER 13 — EU DIGITAL REGULATION AS INDUSTRY-SHAPING POLICY

I conducted the governance and political-economy analysis, contextualised the DSA within global regulatory diffusion and the Brussels Effect, and authored the chapter.

CHAPTER 14 — ENFORCEMENT: DIMENSIONS OF DSA COMPLIANCE

I synthesized the first two years of DSA enforcement by developing the chapter's conceptual framing, compiling and analyzing the underlying dataset, and integrating these materials into a coherent account of key compliance dimensions for an interdisciplinary work together with my co-authors.

LIST OF PUBLICATIONS

PUBLISHED AND ACCEPTED RESEARCH AND WORKSHOP PAPERS

1. Sekwenz, M.-T., Wagner, B., & de Bruijn, H. (2026). *From Reports to Reality: Testing Consistency in Instagram's Digital Services Act Compliance Data*. *ACM Journal on Responsible Computing*. doi:10.1145/3821526
2. Sekwenz, M.-T., Hermann-Gsenger, R., Biswas, S., & Gadiraju, U. (2026). *Assessing Compliance for Generative AI Systems under the Digital Services Act*. Position paper accepted at the AI CHAOS! workshop, ACM Intelligent User Interfaces (IUI 2026), 23–26 March 2026, Paphos, Cyprus. [Workshop programme and accepted papers](#)
3. Mehrvarz, M., Sekwenz, M.-T., Murray-Rust, D., & Wagner, B. (2026). *Socio-Technical Fiction and Legal Exegesis: Exploring the AI Act in the Near Future of Agentic AI*. In *Companion Publication of the 2026 ACM Designing Interactive Systems Conference (DIS Companion 2026)*. ACM. doi:10.1145/3802974.3809460
4. Sekwenz, M.-T. (accepted; forthcoming 2026). *When Is Content “AI-Generated Enough”? Labelling Synthetic Media under the Digital Services Act and the AI Act*. European Conference on the Impact of Artificial Intelligence and Robotics (ECAI 2026), 2–4 September 2026, Ghent.
5. Sekwenz, M.-T., & Parkin, S. (2026). *Making Reporting Approachable: User-Friendly Reporting Mechanisms for Platform Flagging*. Position paper accepted at the CHI 2026 workshop *Clicks, Complaints, and Consequences: Designing Effective Online Safety Mechanisms*. [Workshop paper](#)
6. Sekwenz, M.-T., Biswas, S., Hermann-Gsenger, R., & Gadiraju, U. (2026). *AI at the Front Lines of Platform Governance: Using LLMs to Support Illegal Content Reporting under the Digital Services Act*. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2848–2884). ACM. doi:10.1145/3805689.3812301
7. Sekwenz, M.-T. (2026). *Wicked by Design: Regulating and Auditing Platforms under the Digital Services Act*. In E. Schweighofer, S. Eder, F. Costantini, & F. Schmautzer (Eds.), *Mensch–Maschine–Kooperation im Cyberspace / Human–Machine Cooperation in Cyberspace: Tagungsband des 29. Internationalen Rechtsinformatik Symposions IRIS 2026* (pp. 441–451). Editions Weblaw. [Published proceedings](#)
8. Sekwenz, M.-T., Wagner, B., & Parkin, S. (2025). *“It is unfair, and it would be unwise to expect the user to know the law!”—Evaluating reporting mechanisms under the Digital Services Act*. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 532–546). ACM. doi:10.1145/3715275.3732036
9. Sekwenz, M.-T., Gsenger, R., Stocker, V., Görnemann, E., Talypova, D., Parkin, S., Greminger, L., & Smaragdakis, G. (2025). *Can't LLMs do that? Supporting Third-Party Audits under the DSA: Exploring Large Language Models for Systemic Risk Evaluation of the Digital Services Act in an Interdisciplinary Setting*. In *CHIWORK 2025 Adjunct Proceedings*. ACM. doi:10.1145/3707640.3731929
10. Sekwenz, M.-T., & Gsenger, R. (2025). *Mapping Compliance: A Taxonomy for Political Content Analysis under the EU's Digital Electoral Framework*. *Jusletter IT*, 245–256. doi:10.38023/2623a364-7b74-4230-a3cd-c331b9999719

11. Sanchez, A., Gsenger, R., Montgomery, S., Sekwenz, M.-T., Pridmore, J., & Wagner, B. (2025). *Playing with Politics: Preliminary Results from Interactive Interventions on AI and Democracy in Five Countries with 2024 Elections*. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA 2025)*. ACM. doi:10.1145/3706599.3719861
12. Hanif, H., Constantino Torres, J. E., Sekwenz, M.-T., van Eeten, M. J. G., Ubacht, J., Wagner, B., & Zhauniarovich, Y. (2024). *Navigating the EU AI Act Maze Using a Decision-Tree Approach*. *ACM Journal on Responsible Computing*. <https://doi.org/10.1145/3677174>
13. Torkamaan, H., Steinert, S., Pera, M. S., Kudina, O., Kernan Freire, S., Verma, H., Kelly, S., Sekwenz, M.-T., Yang, J., van Nunen, K. L. L., Warnier, M., Brazier, F. M., & Oviedo-Trespalacios, O. (2024). *Challenges and Future Directions for Integration of Large Language Models into Socio-Technical Systems*. *Behaviour & Information Technology*. <https://doi.org/10.1080/0144929X.2024.2431068>
14. Wagner, B., Kettemann, M. C., Tiedeke, A.-S., Rachinger, F., & Sekwenz, M.-T. (2024). *Mapping Interpretations of the Law in Online Content Moderation in Germany*. *Computer Law & Security Review*, 55, 106054. <https://doi.org/10.1016/j.clsr.2024.106054>
15. Hanif, H., Constantino, J., Sekwenz, M.-T., van Eeten, M., Ubacht, J., Wagner, B., & Zhauniarovich, Y. (2023). *Tough Decisions? Supporting System Classification According to the AI Act*. In *Legal Knowledge and Information Systems (JURIX 2023)* (Frontiers in Artificial Intelligence and Applications, Vol. 379, pp. 353–358). IOS Press. doi:10.3233/FAIA230987
16. Kübler, J., Sekwenz, M.-T., Rachinger, F., König, A., Gsenger, R., Pírková, E., Wagner, B., Kettemann, M. C., Krennerich, M., & Ferro, C. (2023). *The 2021 German Federal Election on Social Media: Analysing Electoral Risks Created by Twitter and Facebook*. In *Proceedings of HICSS 2023*. doi:10.24251/HICSS.2023.493
17. Eren, E., Hondrich, L., Huang, L., Imana, B., Kettemann, M. C., Kuai, J., Mattiuzzo, M., Pirang, A., Pop Stefaniya, A., Rzepka, S., Sekwenz, M.-T., Siebert, Z., Stapel, S., & Weckner, F. (2021). *Increasing Fairness in Targeted Advertising: The Risk of Gender Stereotyping by Job Ad Algorithms*. Zenodo. doi:10.5281/zenodo.4619894
18. Wagner, B., Lopez, P., Cech, F., Grill, G., & Sekwenz, M.-T. (2020). *Der AMS-Algorithmus*. *Juridikum*, 2020(2), 191–202. doi:10.33196/juridikum202002019101
19. Wagner, B., Rozgonyi, K., Sekwenz, M.-T., Cobbe, J., & Singh, J. (2020). *Regulating Transparency? Facebook, Twitter and the German Network Enforcement Act*. In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency* (pp. 261–271). ACM. doi:10.1145/3351095.3372856

FORTHCOMING, SUBMITTED, AND PREPRINT WORK

1. Sekwenz, M.-T., Gsenger, R., Dahlgren, S., & Wagner, B. (2025, preprint; forthcoming). *Doing Audits Right? The Role of Sampling and Legal Content Analysis in Systemic Risk Assessments and Independent Audits in the Digital Services Act*. *ACM Journal on Responsible Computing*. doi:10.48550/arXiv.2505.03601
2. Prinianaki, E., Sekwenz, M.-T., & Wagner, B. (forthcoming). *Assessing Risks on Online Platforms: Thinking Conceptually about Very Large Online Platform Data*. European University Institute book chapter.
3. Sekwenz, M.-T., Beadle, K., & Parkin, S. (2026, preprint). *“There is literally zero funding”: Understanding the Emerging Role of Trusted Flaggers under the EU Digital Services Act*. doi:10.48550/arXiv.2603.29874

BOOK CHAPTERS

1. Sekwenz, M.-T., & Gsenger, R. (2025). *The Digital Services Act: Online Risks, Transparency and Data Access*. <https://doi.org/10.5771/9783748943990-115>
2. Sekwenz, M.-T. (2025). *Avant-Garde—Wie Leo Reisinger die Regulierung von Inhaltsmoderation auf Online-Plattformen unter dem Digital Services Act vorausgriff*. In M. Holzweber, F. Lachmayer, & G. Scheffbeck (Eds.), *Anfänge, Entfaltung und Perspektiven der Rechtsinformatik in Österreich: Gedenkschrift für Leo Reisinger* (pp. 136–145). Danzig & Unfried. ISBN 978-3-9907200-6-6. [Publisher page](#)
3. Wagner, B., Sanchez, A., Sekwenz, M.-T., Dideriksen, S. A. T., & Murray-Rust, D. S. (2023). *The Politics of Digital (Human) Rights*. In *Oxford Research Encyclopedia of International Studies*. [doi:10.1093/acrefore/9780190846626.013.694](https://doi.org/10.1093/acrefore/9780190846626.013.694)
4. Kettemann, M. C., & Sekwenz, M.-T. (2022). *Pandemics and Platforms: Private Governance of (Dis)Information in Crisis Situations*. In M. C. Kettemann & K. Lachmayer (Eds.), *Pandemocracy in Europe: Power and People in Times of COVID-19* (pp. 263–282). Hart Publishing. [doi:10.5040/9781509946396.ch-013](https://doi.org/10.5040/9781509946396.ch-013)
5. Kettemann, M. C., Rachinger, F., & Sekwenz, M.-T. (2022). *Deplatforming*. In *Regulierung von Kommunikationsplattformen—Aktuelle Fragen der Umsetzung* (pp. 78–89). MANZ. [Publisher page](#)

EDITED VOLUME

1. Gsenger, R., & Sekwenz, M.-T. (Eds.). (2025). *Digital Decade: How the EU Shapes Digitalisation und Entscheidungsverfahren* (Schriftenreihe des Weizenbaum-Instituts, Vol. 3).

ABOUT THE AUTHOR

Marie-Therese Sekwenz is a PhD Candidate at Delft University of Technology (TU Delft), Faculty of Technology, Policy and Management, and Deputy Director of the AI Futures Lab on Rights and Justice. Her research lies at the intersection of law, governance, and human-computer interaction, with a focus on the implementation and enforcement of the EU's Digital Services Act (DSA). She investigates how regulatory requirements for online platforms, such as transparency obligations, risk assessments, and independent audits, are translated into practice, and what these developments mean for user rights and platform accountability. Marie-Therese's dissertation combines empirical legal research with socio-technical methods, including content analysis, experimental interface studies, and interdisciplinary workshops. She has developed frameworks for auditing systemic risks, measuring consistency across transparency reports, and evaluating the effectiveness of content moderation processes under the DSA. Her work aims to contribute to both academic debates and policy discussions on platform governance, providing tools and insights that support regulators, auditors, and civil society actors in holding powerful digital platforms accountable.

Before beginning her PhD, Marie-Therese earned a master's degree in Information Systems from the Vienna University of Economics and Business, where her undergraduate studies combined business administration and law. She has worked in cultural project management at the Vienna Konzerthaus and Vienna State Opera and as a journalist for the Austrian public broadcaster ORF, contributing to the Ö1 radio programmes *Diagonal* and *Radiokolleg*. Her earlier academic experience includes positions at the Vienna University of Economics and Business and the Leibniz Institute for Media Research | Hans-Bredow-Institut in Hamburg, as well as collaborations with the Alexander von Humboldt Institute for Internet and Society in Berlin. Alongside her doctoral research, she contributes to interdisciplinary collaborations, edited volumes, and teaching that bring together regulation, technology, and human-centred design.

Her dissertation examines how the DSA's reporting, transparency, and accountability mechanisms function in practice and how they can be evaluated systematically and holistically. Drawing on transparency reports, systemic-risk assessments, independent audit reports, and Statements of Reasons, it develops and applies a multilevel consistency framework for evaluating compliance across regulatory documents. It also explores how large language models can support content-reporting and compliance-auditing processes. Through empirical case studies of large online platforms such as Instagram, the dissertation offers theoretical insights into digital regulation and practical tools for strengthening user rights and democratic oversight online.



N

W

D

E

I

L

S

E

I

G

O

D

K

Y

VOTE