Explainability in AI Policies

A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK

Nannini, Luca; Balayn, Agathe; Smith, Adam Leon

DOI
10.1145/3593013.3594074

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the 6th ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023

Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK

Luca Nannini
lnannini@minsait.com
Minsait by Indra Sistemas SA, Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) Universidade de Santiago de Compostela
Madrid, Spain

Agathe Balayn
a.m.a.balayn@tudelft.nl
Delft University of Technology
Delft, Netherlands

Adam Leon Smith
adamleonsmith@protonmail.com
Dragonfly
Barcelona, Spain

ABSTRACT

Public attention towards explainability of artificial intelligence (AI) systems has been rising in recent years to offer methodologies for human oversight. This has translated into the proliferation of research outputs, such as from Explainable AI, to enhance transparency and control for system debugging and monitoring, and intelligibility of system process and output for user services. Yet, such outputs are difficult to adopt on a practical level due to a lack of a common regulatory baseline, and the contextual nature of explanations. Governmental policies are now attempting to tackle such exigence, however it remains unclear to what extent published communications, regulations, and standards adopt an informed perspective to support research, industry, and civil interests. In this study, we perform the first thematic and gap analysis of this plethora of policies and standards on explainability in the EU, US, and UK. Through a rigorous survey of policy documents, we first contribute an overview of governmental regulatory trajectories within AI explainability and its sociotechnical impacts. We find that policies are often informed by coarse notions and requirements for explanations. This might be due to the willingness to conciliate explanations foremost as a risk management tool for AI oversight, but also due to the lack of a consensus on what constitutes a valid algorithmic explanation, and how feasible the implementation and deployment of such explanations are across stakeholders of an organization. Informed by AI explainability research, we then conduct a gap analysis of existing policies, which leads us to formulate a set of recommendations on how to address explainability in regulations for AI systems, especially discussing the definition, feasibility, and usability of explanations, as well as allocating accountability to explanation providers.

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3594074

CCS CONCEPTS

• **Applied computing** → *Law*; • **Social and professional topics** → *Governmental regulations.*

KEYWORDS

Explainable AI, AI policy, social epistemology

ACM Reference Format:
Luca Nannini, Agathe Balayn, and Adam Leon Smith. 2023. Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3593013.3594074

1 INTRODUCTION

The ability of artificial intelligence (AI) systems to explain their decision-making processes, also known as "*explainability*", has become an increasingly important AI governance topic to ensure that systems are transparent and accountable [61]. Governments are now pacing up strategies to foster AI innovation while mitigating potential risks through explanations. On the other side, AI explainability has also become a prominent research focus in the machine learning and human-computer interaction communities [4, 71]. The research outputs around AI explainability oftentimes reveal the importance of the context of use, for which legislation might miss a clear and informed baseline to implement explainable AI methods appropriately. Hence, facing the plethora of recent policy documents, and the known complexity of the concept of AI explainability, one needs to understand to what extent the documents are informed by the research and are cognizant and account for current explainable AI challenges, so as to design policy documents that further foster a responsible use of AI in the future.

To shed light over the current state of policies around AI explainability and inform future regulatory trajectories, we conduct a rigorous, comprehensive analysis of existing governmental policies in the European Union (EU), the United States of America (US), and the United Kingdom (UK). In terms of official governmental communications, we consider commissioned white papers and guidelines, as well as bills and standardization documents impacting the fruition of AI explanations. Next, we identify key requirements for explainability in policy documents, considering aspects such

as transparency, interpretability, and explainability in decision-making. We then evaluate the suitability of each governmental policy through a gap analysis informed by research publications on AI explainability, and develop recommendations for addressing these issues.

Despite the diversity of policy documents, we observe primarily an urge to enact strategic plans for AI research and development (R&D) and a multifaceted consideration of explainability as a tool to downside innovation risks and harms within civil rights. We also find that these documents do not necessarily account for the complexity and recency of the concept of explainability in AI, especially in terms of definition, feasibility, and usability of AI explanations. This leads us to reflect upon the discretion of explanation providers, also connected to how ambiguities in policy terminology and procedures might further lead to implementation challenges and failures, and ethics-washing opportunities. Ultimately, our study provides a valuable starting point for understanding the current state of AI explainability in policies and standards, and finally brings together an unified view over regulatory approaches to AI explainability. It bears implications for policy-makers but also for AI researchers, calling for documents and research that would account for identified limitations in this ever-evolving field, while being specific enough to avoid ethics-washing.

The rest of the manuscript is organized as follows. Section 2 provides background knowledge on the rising attention towards AI explainability, from the lens of both technical Explainable AI approaches, and more comprehensive sociotechnical implementations. Section 3 details the mixed-method approach we followed for this study. Section 4 maps explainability regulations and standards across countries, performing a gap analysis as informed by potential intergovernmental discrepancies and constraints. Section 5 analyses discrepancies around the mapped dimensions of explainability documents, reasoning over their actuality in lights of related research publications. Finally, section 6 and section 7 reflects upon current research constraints, findings, and future research and policy directions.

2 BACKGROUND: TRANSLATING AI EXPLAINABILITY INTO REGULATIONS

2.1 Defining explainability and interpretability in AI research

Despite the relatively common use of the terms, there is no consensus on the definitions of *explainability* or *interpretability* in the AI standards community. The ambiguities with the definitions encompass the type of explanation, the reason for the explanation, the purpose of the explanation, and to whom the explanation is provided [14]. For the sake of clarity, we intend by *explainability* the communicative ability to deliver insightful information regarding the inner functioning of complex algorithmic architectures. The definitions of interpretability [30] revolve around either systems that are designed to be inherently interpretable [128], or model-agnostic approaches that are based on observing systems' inputs and outputs [126]. By *interpretability*, we refer to the human cognitive ability to inherently understand the functioning of an AI system, intended as the inner relations between its data inputs and its learned output computation functions. Once algorithmic architecture scales up, e.g., in terms of hyperparameters and features, interrelating complexity arises, jeopardizing direct interpretation. We detail below further distinctions as from multiple research domains for explainable AI systems.

Machine learning researchers have primarily focused on developing methods that allow to make the decision process of a machine learning (ML) model, going from an input data sample to the output label produced by the model, less of a black-box. These methods are often categorized along the following dimensions [39, 42]: the type of data (e.g., tabular [25], textual [41], visual [144]), task (e.g., classification [131], regression [90], recommendation [146]), and algorithm (e.g., different types of ML algorithms or deep learning (DL) architectures) the explanation applies to [16]; the nature of the explanation, especially the literature talks either about developing models that are inherently explainable [145] or methods that explain black-box models in a post-hoc manner [80]; family, as the modes of explanations might vary from reflecting associations between the input and output, to contrasting different input samples and their explanations, or to displaying causal explanations, e.g., through counterfactuals [138]; the scope of the explanations, especially literature often discusses local explanations about a single input sample or global explanations that refer to the overall model behavior [138]. Challenges in this research area especially revolve around developing faithful explanations for a broad set of ML models [81], ensuring these explanations are of high-fidelity (i.e., accurately reflect the model behavior) [6, 11, 134], and designing usable benchmarks to properly assess the proposed explanations [143] (difficulties arise in formulating the main properties a *good* explanation should fulfill [88]).

Human Computer Interaction (HCI) researchers have also investigated what makes a *good* explanation [98, 138], and how are explanations used in practice [13]. Besides the dimensions of explanations highlighted above, they have found additional dimensions that should be accounted for when developing or selecting explainability methods. Especially, the purposes of the explanations for each stakeholder within the AI lifecycle vary (e.g., a decision-subject might need explanations for contestability and recourse, a developer for model debugging, an auditor for assessing the model readiness for deployment), and would value best different types of explanation typologies [138]. The usability of the explanation [138] also varies across stakeholders, based e.g., on the explanations' complexity, completeness, interactiveness, or medium (e.g., visual or textual mode of explanation), etc. Studies with various stakeholders have been performed to understand their use of explanations in given scenarios or in their daily practices. Next to the potential usefulness of the explanations, these studies identified disparate use of a narrow set of explanations in practice, and various types of biases hindering a trustworthy use of the explanations (e.g., misinterpretations, confirmation bias, lack of critical attitude towards presented numbers, etc.) [8, 13, 47]. The challenge hence remains to make existing explanations appropriately usable by stakeholders with different backgrounds [48, 67, 140], as well as to better understand their needs and how new explanations should be designed to account for these needs [91].

2.2 Policy versus research

Explainability in AI can be considered a multidisciplinary field that involves various aspects such as technical, ethical, legal, and social [66]. The variety of XAI methods and purposes proposed in research reflects the diversity of categorical definitions for AI systems [129]. As a result, ambiguities pose risks to the public perception of these systems' capabilities and limitations [82, 84]. This might contribute to confusion and decreased awareness over how such systems are human-produced artifacts, reflecting prior sociotechnical instances and choices informing their design [49, 82]. This discrepancy is mirrored in how AI system developers, such as ML engineers, tend to differently define these systems and their capabilities in contrast to policymakers [85]. Such a discrepancy hence poses risks within the debate on regulating AI systems. This is indeed now signaled by the increasing attention within the field of AI ethics to operationalize human principles [72, 99], ascribable to the definition of *Second-wave of AI Ethics* [70]. This research attention aims to operationalize AI principles through a sociotechnical top-down approach to AI governance in organizations [10, 95]. In fact, only claiming to adhere to AI ethics principles might translate into ethics-washing practices if policies, such as accountability measures, are not implemented within a clear regulatory baseline and legal recourse mechanisms [18, 55]. For such reasons, establishing a proactive regulatory approach to define and regulate explainability of AI systems is crucial to ensure their ethical alignment with human values and principles [54]. Parallel to this second-wave, scholars are now inquiring over the suitability of explainability in governmental regulatory initiatives such as bills and enacted laws. Among the most relevant debates, the EU GDPR [51] proved to be the most debated ground for legal scholars to argue about the enactment of a possible "*right to an explanation*" within users affected by outputs of automated-decision making - and thus AI - systems [45, 141]. In a similar vein, attention is now geared towards other EU regulations, e.g., the AI Act draft and how AI interpretability is thereby defined [45, 66, 139]. Yet, up to now, to the best of our knowledge this attention has not addressed the AI regulatory trajectory within explainability informing policy documents. Similarly, no previous work has focused on comparing international approaches to AI explainability policies, reasoning over gaps and operationalization requirements. This perspective would offer an unprecedented background to visualize and evaluate regulatory approaches to AI explainability. This is the perspective we develop in the rest of this paper.

3 METHODOLOGY: A THEMATIC AND GAP ANALYSIS OF AI EXPLAINABILITY POLICIES

To conduct our research, we first source policy documents from official governmental and affiliated agencies' websites of the European Union (EU), United States (US), and United Kingdom (UK)[1]. We collect four types of documents: *Communications*: related to AI governance strategies through public statements and releases; *Reports*: comprehensive studies, surveys, or official research papers that provide in-depth research information; *Regulations*: legally binding rules and guidelines that dictate how organizations must behave; *Standards*: technical specifications detailing implementation for AI explainability policies. We select the documents to review based on their level of relevance and availability of the data (e.g., document content under drafting might not be disclosed to the public, but titles and expected releases might be). We consider documentation produced from 2018 onwards to ensure that policies are tackling current AI developments. For the same reason, we consider the most up-to-date versions of the documents. We exclude contexts where explainability or interpretability are presented outside of direct AI system involvement, e.g., when a regulation uses the terms in auditing procedures, thereby intending explanations as being required as justifications between humans over business conduct, etc. Based on the documents (refer to Appendix Table 1, Table 2 for the complete list of documents reviewed), we examine the regulatory landscape in the EU, US, and UK, exploring the distinct ways in which each jurisdiction has been addressing the regulation of AI over time, with a specific focus on explainability (section 4). Our thematic analysis identifies common themes across the collected documents encompassing not only "explainability" but also associated concepts such as "transparency" and "trustworthiness", while reporting similarities and differences in how they are accounted for. The last stage of our research is to conduct a gap analysis. While comparing themes across documents already reveals a number of gaps, we complement this understanding of the policies with research publications. Based on the themes above, we search for relevant literature stemming from various research communities (algorithmic, human-computer interaction, ethics), to identify misalignment with policies, calling for future work.

[1]For the European Union, we refer to official websites such as e.g., the digital strategy of the European Commission https://digital-strategy.ec.europa.eu/, the official online law database Eur Lex https://eur-lex.europa.eu/, and standard bodies such as CEN http://www.cen.eu/, CENELEC http://www.cenelec.eu/, and ETSI http://www.etsi.org/. For the UK, we refer to https://www.gov.uk/, and related executive public bodies such as ICO https://ico.org.uk/, or research national institute as The Alan Turing Institute https://www.turing.ac.uk/. For the US, we refer to https://www.whitehouse.gov/, and governamental commissions such NSCAI https://www.nscai.gov/, or non-regulatory agencies such as NIST https://www.nist.gov/.

4 MAPPING THE AI REGULATORY LANDSCAPE FOR EXPLAINABILITY

4.1 European Union: a risk-based approach to explainability for AI oversight

The European Union (EU) is driven by a commitment to strengthened legal frameworks for the development, deployment, and utilization of AI in line with its values, including fundamental rights, encoded in the trustworthiness and safety of AI systems [35]. In this AI policy trajectory, the EU adopts a principle of regulatory proportionality, intended to be directly proportional to the severity of the adverse consequences of an AI system.

4.1.1 Setting up AI strategies in the Data Economy (2018-2020). In 2018, the European Commission (EC) acknowledged the potential of AI to drive economic growth and competitiveness through the *Communication on Artificial Intelligence for Europe* [33], outlining the first European strategy for its development and deployment. The document details a comprehensive approach to AI governance that balances promoting innovation with protecting citizens' rights and safety. The Communication states that the EU's approach to AI regulation is based on the principles of transparency, accountability, and human oversight. Regarding transparency, the document

approaches explainability as a high-level AI desiderata. The theme is mentioned under the principle of trust, increasing transparency, and mitigating bias and other risks, in a wider ethical and legal framework for AI development.

The first major discussion on the legitimacy of explanations over algorithmic decision-making systems, including AI ones, was found in the presumption of a "*right to an explanation*" within the EU *General Data Protection Regulation* (GDPR) [51] that came in force in 2018. The GDPR includes provisions for data subjects' rights such as accessing personal data and having them rectified in relation to automated decision-making. [2]. This is in conjunction to Article 22 and Recital 71, which state that individuals have the right not to be subject to a decision based solely on automated processing, including profiling, if it produces legal or similarly significantly effects on them. Art. 15(1)(h) emphasizes the importance of ensuring that individuals have *"meaningful"* information about the logic involved in the decision-making process alongside *"envisaged consequences"* of such processing for the individual. The provisions have been subject of an intense debate among experts, with some arguing that it provides a framework for ensuring transparency and explainability of automated decision-making [22, 94], while others have criticized it for being too vague and difficult to implement in practice [141]. To notice, the phrasing of articles reduces severely the casuistry of its enforcement, not setting a baseline over typology and sufficiency of explanations, thus leaning towards an illusion of remedy rather than a burden of proof for legal recourse [45, 46].

As a final note, in 2020 the Commission published the *White Paper on Artificial Intelligence* [34]. The document yet represents the first structural communication for an European strategy in AI to foster R&D competitiveness. Even if explainability was barely mentioned once for deep learning (DL) outcomes, the Paper outlines objectives and policy options to build a digital *ecosystem of excellence* through EU member coordination for research and innovation in AI across public sector, academia, industry, and small and medium enterprises (SMEs).

4.1.2 Principles for AI trustworthiness: the HLEG guidelines (2019). The High-Level Expert Group on Artificial Intelligence (HLEG) established in 2018, published the following year their *Ethics Guidelines for Trustworthy AI* [116], the first organic attempt by the EC to define AI explainability. Intended in a wider sociotechnical context, the term adopted was *"explicability"*. It was defined as a principle connected to transparency, crucial to creating and maintaining users' trust in the AI system. The need for transparency corresponds to the epistemic condition to comprehend and challenge decisions of AI systems along three coordinates: *Capabilities* - how system architecture is composed and what its functions are; *Purpose* - to what purposes these capabilities correspond, established a priori; *Decisions* - how and why are outputs processed. Worth noticing, a distinction is made between explainability per se and technical one. While the latter relates to system decisions to be understandable and traceable, the former includes also comprehending the purpose of the artifact in relation to the design choices followed. The context of an explanation is emphasized based on the expertise of the stakeholders involved directly (e.g., layperson, regulator, researcher). If capabilities and decisions are not deterministic, then indirect measures (e.g. stochastic model with surrogates) might assess impact and compliance with fundamental rights, measuring the need for explicability through a risk scale as informed by EU regulatory proportionality.

4.1.3 Moving into AI regulations: the AI Act Draft and the AI Liability Directive (2021-2024). The path inaugurated by the HLEG and the White Paper informed EC's committees and their intense work on reinforcing the EU AI Act draft [120] proposed in April 2021. The Act adopts a proportional risk-based approach to regulating AI systems, dividing them into three categories of risk with corresponding limitations, transparency requirements, and oversight mechanisms. Alongside definitions of risks assessments, transparency, and human oversight, the approach develops a taxonomy of AI actors within data quality requirements and technical documentation. The original draft included provisions for interpretability, specifically Article 13, requiring a sufficient degree of transparency mechanisms for *high-risk* systems and the attachment of instructions for use containing relevant, accessible, and understandable information to users regarding the characteristics, capabilities, and limitations of performance of these systems. The Act has undergone several revisions as a result of the EC work, prior to inter-institutional negotiations being held in 2023. In April 2022, the IMCO-LIBE committee proposed initial changes to the regulations[3] surrounding AI explainability during inspections [102]. In September 2022, the JURI committee stressed significant explainability mentions with an Opinion [103] addressing risk mitigation, user interaction and rights[4]. Yet, those revisions were not substantiated in the final draft as presented by the Permanent Representatives Committee on November 25, 2022 [122], including several revisions aimed at ensuring the traceability and interpretability of high-risk AI systems. The changes include the introduction of indirect measures such as strengthened technical documentation requirements and instructions for use with illustrative examples, as well as guidelines for collecting and interpreting system logs. Article 13, originally

[2]These rights can be exercised through the process of "*data subject access request*" (DSAR), allowing individuals to request information about how their personal data is being processed. In a similar vein, it requests organizations to carry out data protection impact assessments (DPIAs) for high-risk data processing activities, including those involving AI systems. These assessments are intended to identify and mitigate potential privacy risks associated with the processing of personal data, including the risk of discrimination, bias, or other forms of human rights violations.

[3]Specifically within IMCO-LIBE, Amendment 43 to Recital 80(d) strengthened the EC's ability to access and understand databases, algorithms, and source codes. Similarly, Amendment 26 added Article 68(f), outlining the EC's ability to access a provider's premises and request explanations for the use of AI systems in analyzing documents and records. Amendment 265 to the creation of Article 68(g) established explainability as a mechanism for illustration and correction during the preliminary finding phase in cases of non-compliance.

[4]For JURI, Recital 47(a) introduced transparency and explanation as a countermeasure for deterrent effects, Recital 80(a) expanded the possibility for an individual to invoke a right to explanation inspired by the EU GDPR, and Article 4(a) emphasized the value of transparency in AI system development in relation to traceability and explainability. Article 4(b) proposed to strengthen AI literacy strategies in organizations, while Recital 80(a) and Amendment 88 to Article 52(1) called for explicit communication of an AI system's capabilities and limitations. Additionally, Article 52 introduced the option of judicial redress for the harms and outputs of AI systems, including the *right to an explanation*. Article 69 also provided for the explicit inclusion of a right to an explanation, specifying the decision-making procedure, main parameters of the decision, and related input data. Furthermore, Article 13 was strengthened with amendments 47-60 to increase understanding of how the system works for providers and end-users, as well as the data processed, in order to have a comprehensive view of how decisions affect them.

stipulating a sufficient and appropriate degree of transparency for high-risk AI systems, was altered to include specific instructions for use and metrics related to the behavior of the system on certain groups of people and in relation to the sociotechnical context of adoption. Article 14(4)(c) was removed from the provision to have end users interpreting the characteristics of a system, introducing it not as a requirement but as a possibility of consulting them through interpretation methods. Overall, final amendments weakened concepts of transparency and explainability, shifting the focus on ensuring the traceability and oversight of high-risk AI systems rather than empowering end-users' explanation demands.

As a final mention, in October 2022, the EC advanced a *Proposal for an AI liability directive* [121] to define accountability allocation within non-contractual civil liability for damage involving AI systems. Whenever an allegedly damaged claimant suspects non-compliance of an AI system output, then it is required to establish a causal link as a burden of proof. Art. 4(2)(b) details that if an AI system is considered high-risk, opaque, and complex (i.e., not allowing transparency requirements of the AI Act's Art.13), therefore explainability is mandated from an EU court not within the system (e.g., through XAI methods) but to the AI deployer through an order to disclose proportional evidence necessary (e.g., logs, documentation, and datasets). This is done to preserve the confidentiality of trade secrets, as in Recital 16, it is said that the AI Act does not specify any right for an injured person to access that information.

4.2 United States: explainability research game changer, loose policy measures

In the United States of America (US), the approach to AI regulation has relied mostly on self-regulation, where industry stakeholders develop best practices and guidelines for the use of AI [69] even if critics pointed a loose legal oversight [56]. The US approach to AI explainability had a pioneering role through the announcement in 2016 of the Defense Advanced Research Projects Agency (DARPA) of the federal funding BAA 16-53, inaugurating the first research program in *Explainable AI* (XAI) [114].

4.2.1 The road to an AI leadership. Strong in its position as a global technological powerhouse, the US AI regulatory trajectory started to gain momentum in February of 2019, when the Executive Order (EO 13859) on *Maintaining American Leadership in AI* was announced by the White House [108]. Despite the nature of the order addressing specifically a national R&D strategy to keep up its global competitiveness, the policy tone thereby advocated for ethical, responsible, and transparent development of AI. This balance informed the *Guidance for Regulation of AI Applications* issued by the Office of Management and Budget (OMB) later in 2020 [109]. The Guidance stated that AI systems should be designed accordingly to be trustworthy, auditable, and thus explainable - especially for the mentions of transparency and disclosure reported in section (8.). The interpretability of AI systems is also briefly announced in the Appendix to enhance transparency for oversight. But rather than leveraging XAI methodologies, indirect regulatory processes are proposed for that aim, i.e., impact analysis, public consultation, and risk assessments. A more structured answer to EO 13859 was elaborated by NIST. Indeed, the Order mandated the agency to develop a plan [106] to tackle Federal priorities for a robust and safe AI R&D while promoting international standardization activities. In the plan, explainability is ascertained below the concept of trustworthiness, one of nine key areas of focus identified for AI standards. These governmental policy communications found a wider output in the *Final Report* [118] issued by the National Security Commission on AI (NSCAI) on March 1st, 2021. The federal commission, created in August 2018 and dismantled in October 2021, had the mandate to advise the US President and Congress on AI R&D for national security and defense needs, reflecting the willingness to maintain its technological competitiveness through workforce development, international cooperation, and AI ethics. There, explainability is only mentioned among areas of R&D, alongside remarks for the transparency and accountability of AI deployment in national security applications. These strategic documents reflect a lack of informed perspective over AI explainability for civil rights, filled only partially by the release in October 2022 of the *Blueprint for the AI Bill of Rights* by the Office of Science and Technology Policy (OSTP) [104]. Rather than industry oriented, the Blueprint can be seen as a civil rights framework for ensuring that automated decision-making systems (ADMs) are used in ways that respect American values such as privacy, autonomy, and other civil liberties. Under the principle of *Notice and Explanation*, automated decisions shall be justified through "*clear, timely, understandable, and accessible*" use in connection to valid explanations tailored to purpose, audience target, and level of risks.

In terms of bills, the US R&D AI trajectory found legislative output in the enactment of the *National AI Initiative Act* on January 1st, 2021 [2]. The Act remarked the duty (Sec. 22A(c)(2)) to establish, within NIST a voluntary risk management framework by 2023 [107], detailing also common definitions and characterizations of aspects of AI trustworthiness such as explainability. Yet for civil rights and litigation in March of 2022 a relevant bill was introduced for possible enforcement of explanations within ADM/AI systems. In Section 4 of the *Algorithmic Accountability Act* [3], companies deploying such systems will be required to perform impact assessments. Among the most impacting provisions figure requirements for business explanations over data collection and maintenance (Sec. 4(7)(A)(ii)) and evaluation standards. Also, there is a need to deliver end-users explanations of system features contributing to the decision output, as well as overall information about the system and process (Sec. 4(8)(B)(i)). Interestingly, provision Sec. 4(11)(C) further addresses the need to engage stakeholders to provide feedback on improvements for the ADM/AI system and also for their explainability. To conclude, in the context of Commission oversight, Sec.5(1)(H)(i) mandates the submission of documentation from the impact assessment over system transparency and explainability measures.

4.2.2 Explainability within NIST. Despite the loose mentions of explainability in AI R&D as shown in policy communications, the theme is considered part of NIST's "*Fundamental AI Research*" [105]. The working group for explainability has so far released two publications in 2021. In April, a first paper (NISTIR 8312)[123] reviewed human psychological traits that characterize users within algorithmic explanations. The document, unique in its genre, exposes psychological properties of interpretation and explanation related to human mental representations. The former term is said to be

useful for policymakers and general users for AI system oversight, while the latter is valuable to developers for debugging and design improvement. Yet, the distinction is loose, and the value of individual differences is underlined, as well as the recommendation to design interpretable algorithms contextualizing data in relation to human cognition representations. This perspective in cognitive psychology and user interaction informed the second paper (NISTIR 8367) [23] in September 2021 on four principles for explainable AI systems. Through the acknowledgment that explanation types should differ based on users, the principles indicate AI explanations to be designed as meaningful to humans, accurate over system inherent process, and expressing system capabilities and knowledge limits.

As set in the *National AI Initiative Act*, these reports informed the release in January 2023 of the first version of the *Risk Management* for AI systems (AI RMF 1.0) [107]. The framework incorporates trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and system. Explainability and interpretability are named as characteristics of *AI Trustworthiness* in Section 3. These characteristics are evaluated within a few AI risks and trade-offs, e.g., enhancing system interpretability against predictive accuracy or achieving privacy. Yet, the framework just provides a coarse overview of the terminology, referencing the previous NIST's papers, without detailed descriptions of explainability methods nor specific requirements for users and contexts of AI systems, thus lacking a properly informed baseline. Yet, in the second part of the framework, four functions are advanced to evaluate and mitigate AI risks (i.e., *Govern*, *Map*, *Measure*, *Manage*). Explainability is found below *Measure 2.9* function for quantifying AI risks. In other words, the concept diverges from previous NIST approaches being narrowed down in its functionalities, as related only to the evaluation phase for trustworthy characteristics. Similarly, interpretation is briefly considered for AI system output within its context, without specifying further baseline or approaches also directed to different targets, e.g., system capabilities and complexity.

4.3 United Kingdom: explainability guidance for government and industry

The approach of the United Kingdom (UK) to AI regulation is still evolving, and ongoing discussions are taking place on the need for specific legislation to address the risks and challenges posed by AI, with a focus on explainability addressed foremost to industry adoption. Yet, the UK government is well grounded ensuring that AI systems are developed and deployed in a way that is safe, trustworthy, and respects the rights and interests of individuals [60].

Guiding organizations for explanability. During the implementation period of Brexit, in 2018 the UK government established the Centre for Data Ethics and Innovation (CDEI), the first governmental body worldwide to advise on the ethical and societal implications of AI and data-driven technologies. In 2019, the UK released a *National Data Strategy* [57] advocating for the adoption and use of safe and explainable data sources, in parallel to an AI Sector Deal [58]. The Deal outlined a public AI R&D strategy to enhance its market competitiveness while reinforcing digital infrastructures. In the plan, it was announced the intention to set up a research collaboration between the Alan Turing Institute (ATI) and the Information Commissioners Office (ICO) for guidance in explaining AI decisions. In 2020, that guidance was released in a well-structured report named *Explaining Decisions with AI* [75]. In that particularly prolific year for the ICO, it also released reports on AI auditing frameworks [111] and AI data protection [110]. The guidance on explainability is the first comprehensive governmental technical report ever produced detailing AI explainability definition, methods, process, and impacts. The document is structured in three parts, respectively addressed to compliance teams, technical teams, and management. This reflects the priority given by ICO & ATI to provide industry guidance on explainable AI. Yet, their major contribution is to be found in the establishment in 2019 of *Project ExplAIn* [74], the first engagement strategy on AI explainability for organizations, implementing the guidance through workbooks and workshops.

Leading AI industry standards. The UK is also taking a novel approach to widening participation in AI standards, led by the Alan Turing Institute, the National Physics Laboratory (NPL), and the British Standards Institute (BSI). To notice, in May 2021 a guidance on ethics, transparency, and accountability of ADM systems was released by the Central Digital & Data Office (CDDO), the Cabinet Office, and the Office of Artificial Intelligence [115]. Their major output is the development of a cross-government standard hub for algorithmic transparency for government departments and public sector bodies (Pillar 3 of the National Strategy Plan) [60] announced in November 2021 [29]. Informed by its close collaborations within AI standardization bodies in the EU and worldwide, the UK plans to integrate future standards in its AI innovation strategies [59, 60]. This direction found output in the first algorithmic transparency standard for government departments and public sectors in November 2021 [29] alongside a report from The Alan Turing Institute [7] in July 2022, calling for a coordinated approach to increase AI readiness and proposing the creation of an *AI and Regulation Common Capacity Hub* to facilitate regulatory collaborations for policymakers.

4.4 A perspective on standards for explainability

The EU has a complex system of actors involved in the development and implementation of standards for AI regulation. The EC serves as the executive branch and plays a crucial role in shaping the regulatory framework. The European Standards Organizations (ESOs) such as CEN, CENELEC, and ETSI are responsible for creating standards in support of the EU AI Act (AIA)[5] alongside a variety of

[5] In the context of AI regulation, the CEN (*European Committee for Standardization*) and the CENELEC (*European Committee for Electrotechnical Standardization*) play a crucial role in defining the technical standards and safeguards needed for the effective implementation of regulations. CEN is responsible for creating European standards in the areas of consumer goods, engineering, healthcare, environment, and other related fields, while CENELEC is responsible for setting standards in the areas of electrotechnical engineering and applications. Under the New Legislative Framework in the EU, standards can be referenced in the Official Journal of the European Union, in order to provide presumption of conformity with particular legislation. These standards provide the technical details to support conformity assessment of products and services. In terms of digital society, CEN-CENELEC often expand their work in the international landscape, collaborating with ISO and IEC.

stakeholders including national standards bodies, European stakeholder organizations, and harmonized standards consultants as outlined in a report from 2021 from the EC on the EU AI standardization landscape [38] and in the 2022 *Rolling Plan for ICT Standardization* [37]. The EC, on December 2022, provided a draft standardization request to CEN-CENELEC outlining requirements for standards to support the presumption of conformity [1]. CEN-CENELEC may choose to develop their own standards, but can also adopt the ISO/IEC standards as sufficient to meet the needs of the standardization request. For further development within AI explainability, attention should be given to CEN-CENELEC's Joint Technical Committee 21 *'Artificial Intelligence'* (JTC 21) and the AI work program communicated in 2020, where explainability was mentioned among key research themes for standards development [27]. With CEN-CENELEC, the European Telecommunications Standards Institute (ETSI) is the third body that will work on advising for the AI Act. Since September 2019, the ETSI has a Strategic Advisory Board on Artificial Intelligence (SAI) [50]. The ETSI ISG SAI has identified several areas for standardization, including transparency and explainability, and aims to produce several deliverables related to the AI Act draft [100], including an existing deliverable on the *Experiential Networked Intelligence* (ENI) system architecture and two forthcoming deliverables on explainability and transparency of AI processing [ETSI ISG SAI GR 007] [53] and traceability of AI models [ETSI ISG SAI GR 010][6].

On the US side, NIST has produced two papers to reflect on principles and methods for explainability informed by research perspectives [23, 123], yet this seems not to have been transposed consistently in the *Risk Management Framework v1.0* [107]. Despite the non-binding nature of NIST, their guidance could be in the near future reinforced by standardization initiatives from the US Federal Trade Commission (FTC) given their activities addressing irregular ADM practices, and by future work for enacting the Algorithmic Accountability Act [83, 113].

Similarly to the US, the UK has not substantiated yet binding standards or regulations affecting AI explainability. Also informed by the recent settlement with the Standard Hub [7, 29], the UK is able to take an equivalent approach to the EU in selecting standards that provide a presumption of conformity. Despite leaving the European Union, the UK remains a member of ESOs such as CEN/CENELEC[7] while actively engaging in the draft of international AI standards [59].

Yet, as mentioned earlier, governments tend to refer to international standardization activities in AI provided by the International Standard Association (ISO) jointly with International Electrotechnical Commission (IEC), alongside the Institute of Electrical and Electronics Engineers (IEEE). The technical committee ISO/IEC JTC 1/SC 42 '*Artificial Intelligence*' [76] focuses on the standardization AI program within ISO/IEC Joint Technical Committee JTC 1. As the central advocate for AI standardization, this committee provides direction to JTC 1, IEC, and ISO committees working on the development of AI applications. Connected to the principles of trust and transparency, their document ISO/IEC TR 24028:2020 [78] published in May 2020 considers explainability as a mitigation measure to AI vulnerabilities and threats, surveying existing approaches in explainability methods and evaluations. Currently under development and expected to be released during 2024, document ISO/IEC AWI TS6254 [77] will describe explainability methods for AI systems accordingly to different stakeholders (i.e., academia, industry, policymakers, end-users among others), while ISO/IEC AWI 12792 [79] will define a taxonomy of AI system transparency information elements. For IEEE, the standardization activities are more scattered and not piloted by a single committee. For such a reason, different documents affecting AI explainability are proposed, such as a standard for algorithmic bias consideration over output interpretability (CC/S2ESC/ALGB-WG P7003) [137], a guide for XAI techniques, application, and evaluation (C/AISC/XAI P2894) [9], a standard for transparency of autonomous systems (VT/ITS 7001-2021) [73], and a standard for mandatory requirements to recognize an AI system as explainable (CIS/SC/XAI WG) [62]. Interestingly, the XAI guide was supposed to be released in 2022 but no further updates are presented, while the standard is expected for July 2024 at the latest.

5 DISCUSSION: REGULATING EXPLAINABILITY

The European Union has made significant efforts in policy communication and legislation related to data and AI, but the GDPR and the proposed AI Act do not contain clear requirements for interpretability and explainability for end users. There is some emphasis on explainability for oversight purposes, but the lack of technical reports and standards in this area highlights a need for further development. In the United States, there is a significant investment in AI R&D and some mention of explainability in relation to civil rights, but the current regulatory landscape is largely focused on driving innovation rather than protecting human rights. The Algorithmic Accountability Act draft includes some provisions for end-user rights to explanation, but the policy focus remains on innovation. In the United Kingdom, the regulatory landscape for AI is limited to data protection, and there is a lack of clear enforcement or standardization in the area of explainability. Currently, the focus seems to be on providing guidance to the industry and promoting innovation, with limited attention paid to end-user rights and protections. These focuses call for a more balanced approach that balances commercial interests with human rights and the need for trustworthy AI.

5.1 Main themes stemming from our analysis

Governmental approaches to AI explainability have been marked by a growing recognition of the importance of interpretability and transparency in AI systems, and a commitment to the development of explainable AI systems that can be audited and held accountable in the wider governance framework of AI risks management. NIST or ICO approaches to explainability research and organizational implementation yet did not substantiate actionable regulatory standards. Yet, provisions such as the EU GDPR, *AI Act Act* draft, *AI*

[6]The work on these deliverables is available within the ETSI Technical Committee on Intelligent Transport Systems.

[7]By promoting AI standardization, the UK is also involving organizations that do not normally have the time or resources to contribute to the development process. In January 2023, the Standards Hub led a full-day workshop attended by 20 selected experts from the government and industry. Experts from ISO/IEC SC 42 and CEN/CENELEC attended to lead a discussion on the definitions of explainability, interpretability, and transparency. This resulted in a written contribution that is being submitted to SC 42 for future consideration.

Liability Directive proposal, and also the US *Accountability Act* draft might be considered a first legal baseline for explainability operationalization, but technical requirements might benefit from less coarse specifications spanning also towards end-user services, and not just oversight. Indeed, as witnessed by casuistry and feasibility within the EU GDPR (Sec. 4.1.1), challenges might arise in establishing a legal baseline for explanations, balancing between explainability method criteria, model inherent complexity, stakeholders' interests and expertise, as well as contextual organizational factors that might introduce further tensions connected to enhanced transparency. As an example, interpretability as a design criterion might face objections under the EU directive for trade secret integrity and intellectual property rights. While transparency is desirable from a deontological perspective to respect EU user rights, it may be problematic from a consequentialist perspective to unilaterally maximize AI system transparency when the same users are not subjected to oversight and legal remedy. In this regard, end-user liability can be ascertained from the need, as expressed in AI Act's Article 29(1), for instructions from the provider to the end-user on the intended use of high-risk AI systems. Additionally, Article 13(1) does not take into account the possibility of using interpretability as a burden of proof by third parties affected by the AI system. These organizational factors, such as trade secret integrity, intellectual property rights, and privacy of third parties could be intended as safeguards and thus incentivize organizations' R&D, without feeling pressured to be continuously subject to explanation requirements. In this vein, an organization might feel less threatened by not being mandated for explanations over their AI system process and output to users, since the latter might deploy them as a burden-of-proof under litigation [20]. Rather, they might prefer being subject to clear regulatory practices to indirectly assess that AI models, despite their inherent complexity, can be considered safe, fair, and thus compliant with the rule of law.

Even before that, bills seem to leave open the assessment of transparency and context of use for explanations to end users, upon provider discretion e.g., in Art.13 of the EU AI Act, at the discretion of the provider of high-risk systems only (Sec. 4.1.3). The criterion of "*appropriate*" with regard to the type of transparency seems to suggest an understanding directed towards legal adequacy for oversight rather than a generic end-user. The major challenge is balancing model complexity, end-user expertise, as well as legal and commercial constraints. Such ambiguity might also stem from a wider sense of terminology ambiguity that characterized the production of communications and reports across governments and commissioned standardization bodies. As an example, we noted that NIST's papers on explainability provided an in-depth research analysis, but this seemed not to substantiate in their *Risk Management Framework* nor the White House's *Blueprint* (Sec. 4.2.2). On this line, we could further trace back this ambiguity to the lack of official coordination among agencies and their inner working groups, as for example within IEEE where two different working groups have been proposed to respectively draft a guide (C/AISC/XAI with P2894 [9]) and a standard (CIS/SC/XAI WG with P2976 [62]) to AI explainability (Sec. 4.4). In the rest of the discussion, we further detail tensions and limits we detected in policies and standards as informed by academic research.

5.2 Accounting for the recent nascence of the concept of explainability in research

One important limitation of current documents is that, while they recognize the importance of explainability for civil right enhancement, their high-level mentions in policy communications hint at the lack of an informed perspective of explainability as a research object still nascent, novel, and complex, and far from being a "solved problem". As is currently, an AI developer cannot easily implement "explainability" in their systems, as the meaning of it and the methods for it might not exist. If a body envisions developing standards for explainability, once again, the task could not be performed yet entirely for the reasons and tensions we envision below.

Theoretical understanding of explainability. Explainability being a nascent concept, it remains unclear what should be considered a *good* explanation [44, 68, 87]. Early HCI works have shown that explainability is contextual, different stakeholders might need different types of explanations depending on their purpose [47, 92], on the domain of application, as well as on a plethora of human factors (e.g., AI literacy, cultural background) that still need to be characterized [13]. On the other side, policy documents often mention explainability in AI strategies without providing any specification in terms of stakeholders, purpose, nature of explanations, etc. Leaving these concepts undefined in policies while they are also yet to be comprehensively understood from research, allows for ambiguity that might be beneficial to the development of new methods, but also detrimental to their responsible use.

Practical feasibility of implementing explainability. Practically, the designers and developers of an AI system might encounter obstacles when building a system with explainability in mind. While policy documents discuss three types of explainability for an AI system, the vast majority of research papers has only focused on one of them, technical explainability. Besides, explainability methods are data-, task-, and algorithm- specific. Not all currently used AI systems in production have now been associated with explainability methods, as a large majority of the research now focuses on DL technology (whereas organizations still rely greatly on traditional ML models). Hence, one would not necessarily be supported in developing the right types of explainability. This narrow focus probably stems from the way the research community is organized, prioritizing algorithmic research especially around DL, to other types of research in terms of rewards and venues for publication, methodologies employed and taught, recognition from peers, and organizational incentives [130]. Furthermore, for the explainability methods that do exist, it is now well-known that they suffer from several issues hindering their use in practice. They are often of low-fidelity [5], brittle to various types of perturbations such as adversarial attacks [20, 63, 135], and inconsistent [89], not always allowing to observe all issues (e.g., spurious correlations) an ML model might suffer from [6]. Yet, it remains challenging for researchers to develop better explanations, as developing appropriate objectives and benchmarks for explainability is not even a solved problem until now [97].

Usability of explainability for stakeholders. Assuming that there would exist methods for making an AI system explainable, additional challenges arise. On one side, AI designers and developers

might not be aware of these methods (the research/practice disconnect is a well-known one, more broadly than simply for AI) and might not be able to correctly use them for their system (methods might require more or less algorithmic knowledge, coding abilities, etc.) [13, 16, 17]. On the other side, it is also well-studied that those who exploit the explanations resulting from the explainability methods implemented within the AI system might not be well-supported to do so [17, 86]. Again, they might lack the knowledge to understand them [80], and might fall into traps from various cognitive biases [68, 86, 97] leading them to blindly trust the AI systems. For instance, it has been found that explanation receivers might be deceived with tailored explanations sounding factual from an epistemic perspective and satisfying prior beliefs, leveraging conditions of confirmation or automation bias [11] or illusions of explanatory depth [32]. This might turn out to be a costly organization choice if a biased, unfactual explanation could be deployed as a burden of proof during litigation by an end-user, e.g., as reported in the EU AI Liability Directive.

5.3 Accountability allocation of explanations

Besides the theoretical and practical challenges discussed above to develop and use explainable AI systems, another set of concerns revolve around organizational obstacles towards explainable AI. Various tensions exist with making a system explainable. As also reported by NIST, at an algorithm-level it is now well understood that a complex, intricate, contextual, trade-off exists between explainability and other desirable properties of an AI system such as accuracy [15], or privacy-preservation [24]. As it is often said that organizations might prefer accuracy as it allows for more effective and efficient business processes, this might come as an obstacle to making models explainable.

Discretion of the service providers. More broadly, several publications have posed the existence of organizational factors [93, 125] that might become obstacles to developers willing to make models more explainable, i.e., absence of incentives, scarcity of time, burdensome computational costs of explanations might be inefficient [31, 43]. The publications report these factors for fairness objectives, but these could be easily transposed. On the other hand, policy documents (e.g., Art. 13(1), EU AI Act) appear to leave at the discretion of the service providers to design AI systems for explainability, and especially for high-fidelity, high-explainability-level, and usable explanations. Similarly to what is currently discussed around the regulations of AI towards fairness [12], uninterested providers might implement the easiest solutions, that might not actually be enough for supporting explainability in effect.

Gaming opportunities and ethics washing. Connected to implementing the easiest solution to ensure legal compliance, this practice can be ascertained to the concept of *ethics bluewashing* [55]. Given the discretion upon organizations reported in the analyzed policies, AI explanation typologies and targets could spotlight part of data and system design that can be considered as desirable and compliant. Explanations could be deployed to vaguely describe AI system properties rather than to inform on relevant actions the decision-subject shall take to remedy to her condition, or on what design rationale informed the development and maintenance of the AI system [26, 127]. Similarly, *ethics dumping* or *shopping* of explanations could justify organization behavior while appealing to good practices or laws present in a single region, being endorsed not by internationally recognized standardization bodies but by other stakeholders with more "appealing" code of ethics and standards [55, 142]. Explanations should be then operationalized accordingly to the rule of law affecting recipients of explanations, establishing clear baselines to reprove accountability measures, e.g. in the US *Accountability Act* and EU *AI liability* drafts. Yet, obliging to legal baselines should be not considered equivalent to truly endorsing "ethics best practices" [18, 64, 70]. Future enactments of AI regulations paired with data and service ones (e.g., EU will enforce its *Digital Services Act* [119] in 2023) will constitute an interesting benchmark to evaluate ground truths of algorithmic explanations, hopefully contributing to allocate responsibilities over non-factual information.

6 AVENUES FOR FUTURE WORK

6.1 Methodological limitations of our research

Given our analysis scope, we centered our focus only on the EU, US, and UK regulatory landscapes, thus not fully capturing the global landscape of AI explainability policies and recommendations outside of those regions, e.g., G20 members or countries currently less engaged in drafting AI policies yet still affected [96, 124]. While the analyzed governmental and official standardization bodies play a crucial role in shaping the legal and technical landscape of AI explainability, future research should account for the influence of further actors, such as industry associations, civil society organizations, and academic experts.

In terms of influences, our analysis tackles explainability in the context of AI policies. Future work should account for other important factors interrelating to influence AI explainability in practice, e.g., data and digital platform laws, trade secrets, and privacy concerns among others.

We also note how, at this point in time, our analysis is based on a relevant quantity of documents under development. Therefore, final versions of the discussed drafts here might differ in the next future, presenting new amendments and unforeseen tensions within AI explainability policies. As the AI field is evolving, our analysis and recommendations should be re-evaluated as new regulations and policies will be mirrored accordingly. Future research should tackle these issues by examining how they interact with explainability tensions and recommendations hereby outlined.

6.2 Recommendations for future work

- *Disambiguating terminologies.* Across policy documents, we note a great diversity of terminological ambiguities. Most documents do not clearly define the scope of technological systems they tackle (e.g., mentioning either “AI”, “ML”, “ADM”, etc.), the relevant concepts surrounding explainability such as its definition and typology, nor its purpose in relation to its various stakeholders (e.g., “end-users” is often used but it is unclear whether it refers to end-users of the AI system, or of the explanation which would encompass decision subjects, developers, assessors, etc.). Prior works have already illustrated terminological confusions in AI ethics and policy [85, 101]. We second their suggestions

and emphasize the need for precise concept definitions in policy documents to foster actionable recommendations.

- *Accounting for the complexity of explainability.* As discussed in section 5, explainability is a recent concept, that is complex and prone to be gamed, and still iterated over in research. Hence, we recommend future policy documents to remain cognizant of these particular characteristics also discussed around AI fairness [12] and privacy [40, 65], and to strive for flexibility to adapt to changes, while also remaining specific enough to enforce relevant constraints.
- *Fostering further research.* We emphasize the importance of a reciprocal influence between policy and research. While we identified limitations of current policy documents based on research outputs (and recognize the complexity of the problem since it requires collaborations between various roles, technical and legislative, researchers and policy-makers, etc.), we also propose that research should cater to prioritizing needs stemming from these documents. This might entail not developing more efficient AI systems but re-focus on different types of explainability.
- *Looking ahead towards explanations for AI innovations.* To conclude, a brief mention needs to be addressed for regulations to be "future-proof" within new AI advanced systems. These systems, such as e.g., generative and foundation models, might pose exceptional challenges within their interpretability, being composed by an aggregation of multiple AI systems stacking up the complexity of inspection [19]. Similarly, explanations provided by large language models might not reflect factual information, delivering explanations for user recommendations without previous fact-checking [21, 133]. This might lead to cascade effects of societal harms not addressed by policymakers, connected to misinformation and public opinion deception, potential eroding adoption of AI explanations due to jeopardized public perception of their trustworthiness and stricter deployment [54, 136]. Similarly, AI systems built via novel AutoML tools [132] might require different types of explanations that need to be accounted for in future policies.

7 CONCLUSION

In this study, we rigorously surveyed policy documents and standards stemming from the EU, US, and UK, and contrasted them with research publications around AI explainability. We contribute the first, valuable, starting point for understanding the current state of AI explainability in policies and standards. We particularly found that policy trajectories prioritize AI innovation under a risk management lens rather than truly empowering users with explanations. Despite the rising attention for civil rights, we find that documents might fall short in addressing the complexity and sociotechnical impact of explainable AI, leading to missed-opportunities over its development and implementation. With this awareness, we advocate a series of recommendations for future policies. We hope that, in the long term, this informed perspective will foster new standards and guidelines specifically tailored to the different contexts in which explainable AI becomes relevant, and will inspire further necessary research directions.

ACKNOWLEDGMENTS

Funding contribution from the ITN project NL4XAI (*Natural Language for Explainable AI*). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621. This document reflects the views of the author(s) and does not necessarily reflect the views or policy of the European Commission. The REA cannot be held responsible for any use that may be made of the information this document contains.

REFERENCES

[1] European Commission Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs. 2022. Draft standardisation request to the European Standardisation Organisations in support of safe and trustworthy artificial intelligence. https://ec.europa.eu/docsroom/documents/52376

[2] 116th Congress of the United States of America. 2020. National Artificial Intelligence Initiative Act of 2020. https://www.congress.gov/116/crpt/hrpt617/CRPT-116hrpt617.pdf#page=1210

[3] 117th Congress of the United States of America. 2022. Algorithmic Accountability Act of 2022, H.R.6580. https://www.congress.gov/bill/117th-congress/house-bill/6580/text

[4] Ashraf M. Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan S. Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox (Eds.). ACM, 582. https://doi.org/10.1145/3173574.3174156

[5] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 9525–9536. https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html

[6] Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. 2022. Post hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=xNOVfCCvDpM

[7] Mhairi Aitken, David Leslie, Florian Ostmann, Jacob Pratt, Helen Margetts, and Cosmina Dorobantu. 2022. Common Regulatory Capacity for AI. (2022). https://doi.org/10.5281/zenodo.6838946

[8] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *IUI '20: 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, March 17-20, 2020*, Fabio Paternò, Nuria Oliver, Cristina Conati, Lucio Davide Spano, and Nava Tintarev (Eds.). ACM, 275–285. https://doi.org/10.1145/3377325.3377519

[9] XAI Explainable Artificial Intelligence IEEE Computer Society (IEEE C/AISC/XAI) Artificial Intelligence Standards Committee. 2020. IEEE P2894 - Guide for an Architectural Framework for Explainable Artificial Intelligence. https://standards.ieee.org/ieee/2894/10284/.

[10] Jacqui Ayling and Adriane Chapman. 2022. Putting AI ethics to work: are the tools fit for purpose? *AI Ethics* 2, 3 (2022), 405–429. https://doi.org/10.1007/s43681-021-00084-x

[11] Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. 2022. The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, 1194–1206. https://doi.org/10.1145/3531146.3533179

[12] Agathe Balayn and Seda Gürses. 2021. *Beyond Debiasing: Regulating AI and its inequalities.* Technical Report. https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf

[13] Agathe Balayn, Natasa Rikalo, Christoph Lofi, Jie Yang, and Alessandro Bozzon. 2022. How can Explainability Methods be Used to Support Bug Identification in Computer Vision Models?. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, Simone D. J. Barbosa, Cliff Lampe, Caroline Appert, David A. Shamma, Steven Mark Drucker, Julie R. Williamson, and Koji Yatani (Eds.). ACM, 184:1–184:16. https://doi.org/10.1145/3491102.3517474

[14] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel

Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[15] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. 2022. It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, 248–266. https://doi.org/10.1145/3531146.3533090

[16] Vaishak Belle and Ioannis Papantonis. 2021. Principles and Practice of Explainable Machine Learning. *Frontiers Big Data* 4 (2021), 688969. https://doi.org/10.3389/fdata.2021.688969

[17] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna (Eds.). ACM, 648–657. https://doi.org/10.1145/3351095.3375624

[18] Elettra Bietti. 2020. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna (Eds.). ACM, 210–219. https://doi.org/10.1145/3351095.3372860

[19] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. https://doi.org/10.48550/ARXIV.2108.07258

[20] Sebastian Bordt, Michèle Finck, Eric Raidl, and Ulrike von Luxburg. 2022. Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, 891–905. https://doi.org/10.1145/3531146.3533153

[21] Samuel R. Bowman. 2023. Eight Things to Know about Large Language Models. *CoRR* abs/2304.00612 (2023). https://doi.org/10.48550/arXiv.2304.00612 arXiv:2304.00612

[22] Maja Brkan and Grégory Bonnet. 2020. Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: Of Black Boxes, White Boxes and Fata Morganas. *European Journal of Risk Regulation* 11, 1 (March 2020), 18–50. https://doi.org/10.1017/err.2020.10

[23] David Broniatowski. 2021. *Psychological Foundations of Explainability and Interpretability in Artificial Intelligence*. Technical Report. https://doi.org/10.6028/NIST.IR.8367

[24] Tobias Budig, Selina Herrmann, and Alexander Dietz. 2020. Trade-offs between privacy-preserving and explainable machine learning in healthcare. In *Seminar Paper, Inst. Appl. Informat. Formal Description Methods (AIFB), KIT Dept. Econom. Manage., Karlsruhe, Germany*. https://doi.org/10.5445/IR/1000138902

[25] Nadia Burkart and Marco F. Huber. 2021. A Survey on the Explainability of Supervised Machine Learning. *J. Artif. Intell. Res.* 70 (2021), 245–317. https://doi.org/10.1613/jair.1.12228

[26] Federico Cabitza, Andrea Campagner, Gianclaudio Malgieri, Chiara Natali, David Schneeberger, Karl Stoeger, and Andreas Holzinger. 2023. Quod erat demonstrandum? - Towards a typology of the concept of explanation for the design of explainable AI. *Expert Systems with Applications* 213 (2023), 118888. https://doi.org/10.1016/j.eswa.2022.118888

[27] CEN-CENELEC. 2020. Focus Group Report - Road Map on Artificial Intelligence (AI). https://www.standict.eu/node/4854 Accessed on: 2023-01-30.

[28] CEN-CENELEC. 2021. Joint Technical Committee 21 (JTC 21) 'Artificial Intelligence'. https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/ Accessed on: 2023-01-30.

[29] Central Digital & Data Office & Centre for Data Ethics & Innovation. 2021. Algorithmic Transparency Recording Standard - Guidance. https://www.gov.uk/government/publications/algorithmic-transparency-template

[30] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M. Rao, Troy D. Kelley, Dave Braines, Murat Sensoy, Christopher J. Willis, and Prudhvi Gurram. 2017. Interpretability of deep learning models: A survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, San Francisco, CA, 1–6. https://doi.org/10.1109/UIC-ATC.2017.8397411

[31] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. 2019. L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=S1E3Ko09F7

[32] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 307–317. https://doi.org/10.1145/3397481.3450644

[33] European Commission. 2018. Communication Artificial Intelligence for Europe: Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/library/communication-artificial-intelligence-europe

[34] European Commission. 2020. White paper on artificial intelligence: A European approach to excellence and trust. *Brussels, 1st edn. European Commission, Brussels* (2020). https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

[35] European Commission. 2021. Communication on Fostering a European approach to Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence

[36] European Commission. 2021. Coordinated Plan on Artificial Intelligence 2021 Review. https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review

[37] European Commission. 2022. Rolling Plan for ICT Standardisation 2022. https://joinup.ec.europa.eu/collection/rolling-plan-ict-standardisation/rolling-plan-2022

[38] European Commission, Joint Research Centre, S Nativi, and S De Nigris. 2021. *AI Watch, AI standardisation landscape state of play and link to the EC proposal for an AI regulatory framework*. Technical Report. https://doi.org/10.2760/376602

[39] Roberto Confalonieri, Ludovik Coba, Benedikt Wagner, and Tarek R. Besold. 2021. A historical perspective of explainable Artificial Intelligence. *WIREs Data Mining Knowl. Discov.* 11, 1 (2021). https://doi.org/10.1002/widm.1391

[40] George Danezis, Josep Domingo-Ferrer, Marit Hansen, Jaap-Henk Hoepman, Daniel Le Métayer, Rodica Tirtea, and Stefan Schiffner. 2015. Privacy and Data Protection by Design - from policy to engineering. *CoRR* abs/1501.03726 (2015). arXiv:1501.03726 http://arxiv.org/abs/1501.03726

[41] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (Dec. 2020), 447–459. https://aclanthology.org/2020.aacl-main.46

[42] Arun Das and Paul Rad. 2020. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *CoRR* abs/2006.11371 (2020). arXiv:2006.11371 https://arxiv.org/abs/2006.11371

[43] Hans de Bruijn, Martijn Warnier, and Marijn Janssen. 2022. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly* 39, 2 (2022), 101666. https://doi.org/10.1016/j.giq.2021.101666

[44] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *DIS '21: Designing Interactive Systems Conference 2021, Virtual Event, USA, 28 June, July 2, 2021*, Wendy Ju, Lora Oehlberg, Sean Follmer, Sarah E. Fox, and Stacey Kuznetsov (Eds.). ACM, 1591–1602. https://doi.org/10.1145/3461778.3462131

[45] Martin Ebers. 2022. Explainable AI in the European Union: An Overview of the Current Legal Framework(s). *The Swedish Law and Informatics Research Institute* (March 2022), 103–132. https://doi.org/10.53292/208f5901.ff492fb3

[46] Lilian Edwards and Michael Veale. 2018. Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"? *IEEE Security & Privacy* 16, 3 (2018), 46–54. https://doi.org/10.1109/MSP.2018.2701152

[47] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael J. Muller, and Mark O. Riedl. 2021. The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. *CoRR* abs/2107.13509 (2021). arXiv:2107.13509 https://arxiv.org/abs/2107.13509

[48] Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence - 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19-24, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12424)*, Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones (Eds.). Springer, 449–466. https://doi.org/10.1007/978-3-030-60117-1_33

[49] M. C. Elish and danah boyd. 2018. Situating methods in the magic of Big Data and AI. *Communication Monographs* 85, 1 (2018), 57–80. https://doi.org/10.1080/03637751.2017.1375130

[50] European Telecommunications Standards Institute (ETSI). 2019. ETSI Securing Artificial Intelligence (SAI) Committee. https://www.etsi.org/technologies/securing-artificial-intelligence

[51] European Commission. 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

[52] Experiential Networked Intelligence (ENI) European Telecommunications Standards Institute (ETSI). 2021. ETSI GS ENI 005 V2.1.1 - Experiential Networked Intelligence (ENI), System Architecture. https://www.etsi.org/deliver/etsi_gs/ENI/001_099/005/02.01.01_60/gs_ENI005v020101p.pdf

[53] Securing Artificial Intelligence (SAI) European Telecommunications Standards Institute (ETSI). 2023. DGR/SAI-007' Work Item: Explicability and transparency of AI processing. https://portal.etsi.org/webapp/WorkProgram/Report_WorkItem.asp?WKI_ID=63078

[54] Mark Fenwick, Erik P. M. Vermeulen, and Marcelo Corrales. 2018. *Business and Regulatory Responses to Artificial Intelligence: Dynamic Regulation, Innovation Ecosystems and the Strategic Management of Disruptive Technology.* Springer Singapore, Singapore, 81–103. https://doi.org/10.1007/978-981-13-2874-9_4

[55] Luciano Floridi. 2019. Translating Principles Into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy and Technology* 32, 2 (2019), 185–193. https://doi.org/10.1007/s13347-019-00354-x

[56] Center for AI and Digital Policy (CAIDP). 2022. Artificial Intelligence and Democratic Values Index - February, 2022. https://caidp.org/reports/aidv-2021/

[57] Department for Digital, Culture, and Media & Sport of the United Kingdom. 2019. National Data Strategy. https://www.gov.uk/guidance/national-data-strategy

[58] Department for Digital, Culture, Media & Sport, Department for Business, and Energy & Industrial Strategy of the United Kingdom. 2019. AI Sector Deal - Policy paper. https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal

[59] Department for Digital, Culture, Media & Sport, Department for Business, Energy & Industrial Strategy, and Office for Artificial Intelligence of the United Kingdom. 2022. Establishing a pro-innovation approach to regulating AI - Policy paper presented to UK Parliament. https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement#fnref:35

[60] Department for Digital, Culture, Media & Sport, Department for Business, Energy & Industrial Strategy, and Office for Artificial Intelligence of the United Kingdom. 2022. National AI Strategy - AI Action Plan. https://www.gov.uk/government/publications/national-ai-strategy-ai-action-plan/national-ai-strategy-ai-action-plan

[61] Stanford Institute for Human-Centered Artificial Intelligence (HAI). 2022. The 2022 AI Index Report - Measuring trends in Artificial Intelligence. https://aiindex.stanford.edu/report/

[62] Standard for XAI eXplainable AI Working Group IEEE Computational Intelligence Society/ Standards Committee (IEEE CIS/SC/XAI WG). 2024. IEEE CIS/SC/XAI WG P2976 - Standard for XAI – eXplainable Artificial Intelligence - for Achieving Clarity and Interoperability of AI Systems Design. https://standards.ieee.org/ieee/2976/10522/.

[63] Amirata Ghorbani, Abubakar Abid, and James Y. Zou. 2019. Interpretation of Neural Networks Is Fragile. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 3681–3688. https://doi.org/10.1609/aaai.v33i01.33013681

[64] Ben Green. 2021. The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics in Practice. *J. Soc. Comput.* 2, 3 (2021), 209–225. https://doi.org/10.23919/JSC.2021.0018

[65] Seda Gürses, Carmela Troncoso, and Claudia Diaz. 2011. Engineering privacy by design. *Computers, Privacy & Data Protection* 14, 3 (2011), 25. https://software.imdea.org/~carmela.troncoso/papers/Gurses-CPDP11.pdf

[66] Philipp Hacker and Jan-Hendrik Passoth. 2022. Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond. In *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek (Eds.). Springer International Publishing, Cham, 343–373. https://doi.org/10.1007/978-3-031-04083-2_17

[67] Sophia Hadash, Martijn C. Willemsen, Chris Snijders, and Wijnand A. IJsselsteijn. 2022. Improving understandability of feature contributions in model-agnostic explainable AI tools. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, Simone D. J. Barbosa, Cliff Lampe, Caroline Appert, David A. Shamma, Steven Mark Drucker, Julie R. Williamson, and Koji Yatani (Eds.). ACM, 487:1–487:9. https://doi.org/10.1145/3491102.3517650

[68] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. 5540–5552. https://doi.org/10.18653/v1/2020.acl-main.491

[69] Jeffrey Heer. 2018. The Partnership on AI. *AI Matters* 4, 3 (oct 2018), 25–26. https://doi.org/10.1145/3284751.3284760

[70] Merve Hickok. 2021. Lessons learned from AI ethics principles for future actions. *AI and Ethics* 1, 1 (2021), 41–47. https://doi.org/10.1007/s43681-020-00008-1

[71] Andreas Holzinger, Peter Kieseberg, Edgar R. Weippl, and A Min Tjoa. 2018. Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. In *Machine Learning and Knowledge Extraction - Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27-30, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 11015)*, Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar R. Weippl (Eds.). Springer, 1–8. https://doi.org/10.1007/978-3-319-99740-7_1

[72] Javier Camacho Ibáñez and Mónica Villas Olmeda. 2022. Operationalising AI Ethics: How Are Companies Bridging the Gap between Practice and Principles? An Exploratory Study. *AI Soc.* 37, 4 (dec 2022), 1663–1687. https://doi.org/10.1007/s00146-021-01267-0

[73] Intelligent Transporation Systems IEEE Vehicular Technology Society (IEEE VT/ITS). 2022. IEEE 7001-2021 - Standard for Transparency of Autonomous Systems. https://standards.ieee.org/ieee/7001/6929/.

[74] Information Commissioner's Office (ICO) of the United Kingdom and The Alan Turing Institute. 2019. Project ExplAIn - Interim Report. https://ico.org.uk/media/about-the-ico/documents/2615039/project-explain-20190603.pdf

[75] Information Commissioner's Office (ICO) of the United Kingdom and The Alan Turing Institute. 2020. Explaining decisions made with AI. https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-ai/

[76] JTC 1 /SC 42 Artificial Intelligence International Standards Association (ISO). 2017. ISO/IEC JTC 1/SC 42 - Artificial intelligence Committee. https://www.iso.org/committee/6794475.html.

[77] Securing Artificial Intelligence (SAI) International Standards Association (ISO). 2023. ISO/IEC AWI TS 6254 -Information technology — Artificial intelligence — Objectives and approaches for explainability of ML models and AI systems. https://www.iso.org/standard/82148.html.

[78] International Standards Association (ISO). 2020. ISO/IEC TR 24028:2020 - Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence. https://www.iso.org/standard/77608.html.

[79] International Standards Association (ISO). 2025. ISO/IEC AWI 12792 - Information technology — Artificial intelligence — Transparency taxonomy of AI systems. https://www.iso.org/standard/84111.html.

[80] Sérgio M. Jesus, Catarina Belém, Vladimir Balayan, João Bento, Pedro Saleiro, Pedro Bizarro, and João Gama. 2021. How can I choose an explainer?: An Application-grounded Evaluation of Post-hoc Explanations. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, Madeleine Clare Elish, William Isaac, and Richard S. Zemel (Eds.). ACM, 805–815. https://doi.org/10.1145/3442188.3445941

[81] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. 2022. Evaluating Explainable AI on a Multi-Modal Medical Imaging Task: Can Existing Algorithms Fulfill Clinical Requirements?. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 11945–11953. https://ojs.aaai.org/index.php/AAAI/article/view/21452

[82] Deborah G. Johnson and Mario Verdicchio. 2017. Reframing AI Discourse. *Minds Mach.* 27, 4 (dec 2017), 575–590. https://doi.org/10.1007/s11023-017-9417-6

[83] Kate Kaye. 2022. This Senate bill would force companies to audit AI used for housing and loans. https://www.protocol.com/enterprise/revised-algorithmic-accountability-bill-ai

[84] Patrick Gage Kelley, Yongwei Yang, Courtney Heldreth, Christopher Moessner, Aaron Sedley, Andreas Kramm, David T. Newman, and Allison Woodruff. 2021. Exciting, Useful, Worrying, Futuristic: Public Perception of Artificial Intelligence

in 8 Countries. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 627–637. https://doi.org/10.1145/3461702.3462605

[85] P. M. Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. 2020. Defining AI in Policy versus Practice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/3375627.3375835

[86] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. 2022. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *CoRR* abs/2202.01602 (2022). arXiv:2202.01602 https://arxiv.org/abs/2202.01602

[87] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An Evaluation of the Human-Interpretability of Explanation. *CoRR* abs/1902.00006 (2019). arXiv:1902.00006 http://arxiv.org/abs/1902.00006

[88] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. 2019. Human Evaluation of Models Built for Interpretability. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2019, Stevenson, WA, USA, October 28-30, 2019*, Edith Law and Jennifer Wortman Vaughan (Eds.). AAAI Press, 59–67. https://ojs.aaai.org/index.php/HCOMP/article/view/5280

[89] Eunjin Lee, David Braines, Mitchell Stiffler, Adam Hudler, and Daniel Harborne. 2019. Developing the sensitivity of LIME for better machine learning explanation. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, Tien Pham (Ed.), Vol. 11006. International Society for Optics and Photonics, SPIE, 1100610. https://doi.org/10.1117/12.2520149

[90] Simon Letzgus, Patrick Wagner, Jonas Lederer, Wojciech Samek, Klaus-Robert Müller, and Grégoire Montavon. 2022. Toward Explainable Artificial Intelligence for Regression Models: A methodological perspective. *IEEE Signal Process. Mag.* 39, 4 (2022), 40–58. https://doi.org/10.1109/MSP.2022.3153277

[91] Q. Vera Liao, Daniel M. Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik (Eds.). ACM, 1–15. https://doi.org/10.1145/3313831.3376590

[92] Q. Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. In *Proceedings of the Tenth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2022, virtual, November 6-10, 2022*, Jane Hsu and Ming Yin (Eds.). AAAI Press, 147–159. https://ojs.aaai.org/index.php/HCOMP/article/view/21995

[93] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik (Eds.). ACM, 1–14. https://doi.org/10.1145/3313831.3376445

[94] Gianclaudio Malgieri. 2019. Automated Decision-Making in the EU Member States: The Right to Explanation and Other "Suitable Safeguards" in the National Legislations. *Computer Law & Security Review* 35, 5 (Oct. 2019), 105327. https://doi.org/10.1016/j.clsr.2019.05.002

[95] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature machine intelligence* 1, 11 (2019), 501–507. https://doi.org/10.1038/s42256-019-0114-4

[96] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology* 33 (2020), 659–684. https://doi.org/10.1007/s13347-020-00405-8

[97] Sina Mohseni, Jeremy E. Block, and Eric D. Ragan. 2021. Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark. In *IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021*, Tracy Hammond, Katrien Verbert, Dennis Parra, Bart P. Knijnenburg, John O'Donovan, and Paul Teale (Eds.). ACM, 22–31. https://doi.org/10.1145/3397481.3450689

[98] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst.* 11, 3-4 (2021), 24:1–24:45. https://doi.org/10.1145/3387166

[99] Jessica Morley, Libby Kinsey, Anat Elhalal, Francesca Garcia, Marta Ziosi, and Luciano Floridi. 2023. Operationalising AI ethics: barriers, enablers and next steps. *AI Soc.* 38, 1 (2023), 411–423. https://doi.org/10.1007/s00146-021-01308-8

[100] Markus Mueck, Raymond Forbes, Scott Cadzow, Suno Wood, and European Telecommunications Standards Institute (ETSI) Gazis, Evangelos. 2022. ETSI Activities in the field of Artificial Intelligence - Preparing the implementation of the European AI Act. https://www.etsi.org/images/files/ETSIWhitePapers/ETSI-WP52-ETSI-activities-in-the-field-of-AI.pdf. (2022).

[101] Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. 2019. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proc. ACM Hum. Comput. Interact.* 3, CSCW (2019), 119:1–119:36. https://doi.org/10.1145/3359221

[102] IMCO-LIBE Committees of European Parliament. 2022-04-20. Draft Report on the proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts.

[103] JURI Committee of European Parliament. 2022-09-12. Opinion of the Committee on Legal Affairs for the Committee on the Internal Market and Consumer Protection and the Committee on Civil Liberties, Justice and Home Affairs on the 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.

[104] Office of Science and Technology Policy of the White House of United States of America. 2022. Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[105] National Institute of Standards and Technology (NIST) of United States of America. 2021. Artificial Intelligence: AI Fundamental Research - Explainability. https://www.nist.gov/artificial-intelligence/ai-fundamental-research-explainability

[106] National Institute of Standards and Technology (NIST) of United States of America. 2021. U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools - Prepared in response to Executive Order 13859 Submitted on August 9, 2019. https://www.nist.gov/news-events/news/2019/08/plan-outlines-priorities-federal-agency-engagement-ai-standards-development

[107] National Institute of Standards and Technology (NIST) of United States of America. 2023. Risk Management Framework v1.0. https://doi.org/10.6028/NIST.AI.100-1

[108] Executive Office of the President of United States of America. 2019. Maintaining American Leadership in Artificial Intelligence - Executive Order 13859 of February 11, 2019. https://www.federalregister.gov/d/2019-02544

[109] Executive Office of the President of United States of America Office of Management and Budget. 2020. Guidance for Regulation of Artificial Intelligence Applications. https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf

[110] Information Commissioner's Office (ICO) of the United Kingdom. 2020. Guidance on AI and data protection. https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-artificial-intelligence-and-data-protection/

[111] Information Commissioner's Office (ICO) of the United Kingdom. 2020. Guidance on the AI auditing framework: Draft guidance for consultation. https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

[112] Parliament of the United Kingdom. 2018. Data Protection Act 2018. https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

[113] Federal Trade Commission of the United States of America. 2022. Trade Regulation Rule on Commercial Surveillance and Data Security (Billing Code: 6750-01-P). https://www.federalregister.gov/documents/2022/08/22/2022-17752/trade-regulation-rule-on-commercial-surveillance-and-data-security

[114] Defense Advanced Research Projects Agency of United States of America. 2016-08-050. Federal Contract Opportunity for Explainable Artificial Intelligence (XAI) DARPA-BAA-16-53. https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

[115] Cabinet Office & Central Digital & Data Office & Office for Artificial Intelligence. 2021. Ethics, Transparency and Accountability Framework for Automated Decision-Making - Guidance. https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making/ethics-transparency-and-accountability-framework-for-automated-decision-making

[116] High-Level Expert Group on Artificial Intelligence. 2019. Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[117] High-Level Expert Group on Artificial Intelligence. 2020. The Assessment List for Trustworthy Artificial Intelligence (ALTAI). https://ec.europa.eu/futurium/en/ai-alliance-consultation

[118] National Security Commission on Artificial Intelligence of United States of America. 2021. Final Report. https://www.nscai.gov/2021-final-report/

[119] European Parliament and Council. 2020-12-15. Proposal for a 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.

[120] European Parliament and Council. 2021-04-21. 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.

[121] European Parliament and Council. 2022-09-28. Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive). https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en
[122] European Parliament and Council. 2022-11-25. 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 - General Approach.
[123] P. Jonathon Phillips, Carina Hahn, Peter Fontana, Amy Yates, Kristen K. Greene, David Broniatowski, and Mark A. Przybocki. 2021. Four Principles of Explainable Artificial Intelligence. https://doi.org/10.6028/NIST.IR.8312
[124] Marie-Therese Png. 2022. At the Tensions of South and North: Critical Roles of Global South Stakeholders in AI Governance. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1434–1445. https://doi.org/10.1145/3531146.3533200
[125] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proc. ACM Hum. Comput. Interact.* 5, CSCW1 (2021), 7:1–7:23. https://doi.org/10.1145/3449081
[126] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-Agnostic Interpretability of Machine Learning. http://arxiv.org/abs/1606.05386 arXiv:1606.05386 [cs, stat].
[127] Scott Robbins. 2019. A misdirected principle with a catch: explicability for AI. *Minds and Machines* 29, 4 (2019), 495–514. https://doi.org/10.1007/s11023-019-09509-3
[128] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x
[129] Stuart Russell and Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson. http://aima.cs.berkeley.edu/
[130] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen K. Paritosh, and Lora Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker (Eds.). ACM, 39:1–39:15. https://doi.org/10.1145/3411764.3445518
[131] Wojciech Samek and Klaus-Robert Müller. 2019. Towards Explainable Artificial Intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (Eds.). Lecture Notes in Computer Science, Vol. 11700. Springer, 5–22. https://doi.org/10.1007/978-3-030-28954-6_1
[132] Shubhra Kanti Karmaker Santu, Md. Mahadi Hassan, Micah J. Smith, Lei Xu, Chengxiang Zhai, and Kalyan Veeramachaneni. 2022. AutoML to Date and Beyond: Challenges and Opportunities. *ACM Comput. Surv.* 54, 8 (2022), 175:1–175:36. https://doi.org/10.1145/3470918
[133] Chirag Shah and Emily M. Bender. 2022. Situating Search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval* (Regensburg, Germany) *(CHIIR '22)*. Association for Computing Machinery, New York, NY, USA, 221–232. https://doi.org/10.1145/3498366.3505816
[134] Leon Sixt, Maximilian Granz, and Tim Landgraf. 2020. When Explanations Lie: Why Many Modified BP Attributions Fail. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 9046–9057. http://proceedings.mlr.press/v119/sixt20a.html
[135] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, Annette N. Markham, Julia Powles, Toby Walsh, and Anne L. Washington (Eds.). ACM, 180–186. https://doi.org/10.1145/3375627.3375830
[136] Nathalie A. Smuha. 2021. Beyond the individual: governing AI's societal harm. *Internet Policy Rev.* 10, 3 (2021). https://doi.org/10.14763/2021.3.1574
[137] Algorithmic Bias Working Group IEEE Computer Society (IEEE C/S2ESC/ALGB-WG) Software & Systems Engineering Standards Committee. 2017. IEEE P7003 - Algorithmic Bias Considerations. https://standards.ieee.org/ieee/7003/6980/.
[138] Kacper Sokol and Peter A. Flach. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna (Eds.). ACM, 56–67. https://doi.org/10.1145/3351095.3372870
[139] Francesco Sovrano, Salvatore Sapienza, Monica Palmirani, and Fabio Vitali. 2022. Metrics, Explainability and the European AI Act Proposal. *J* 5 (Feb. 2022), 126–138. https://doi.org/10.3390/j5010010
[140] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021*, Tracy Hammond, Katrien Verbert, Dennis Parra, Bart P. Knijnenburg, John O'Donovan, and Paul Teale (Eds.). ACM, 109–119. https://doi.org/10.1145/3397481.3450662
[141] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law* 7, 2 (May 2017), 76–99. https://doi.org/10.1093/idpl/ipx005
[142] Ben Wagner. 2018. *Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping?* Amsterdam University Press, Amsterdam, 84–89. https://doi.org/10.1515/9789048550180-016
[143] Mengjiao Yang and Been Kim. 2019. BIM: Towards quantitative evaluation of interpretability methods with ground truth. *arXiv preprint arXiv:1907.09701* (2019).
[144] Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. 2022. Explainability of Deep Vision-Based Autonomous Driving Systems: Review and Challenges. *Int. J. Comput. Vis.* 130, 10 (2022), 2425–2452. https://doi.org/10.1007/s11263-022-01657-x
[145] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. 2018. Interpretable Convolutional Neural Networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 8827–8836. https://doi.org/10.1109/CVPR.2018.00920
[146] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. *Found. Trends Inf. Retr.* 14, 1 (2020), 1–101. https://doi.org/10.1561/1500000066

A APPENDIX - TABLES OF POLICY REFERENCES CONSULTED

In the following page, Table 1 reports a non-exhaustive list of policy documents such as communications, reports, and regulations involving AI Explainability. For each document is reported year of announcement, while within the Regulations columns, square brackets denote status.

Table 2 reports a comprehensive representation of major standardization activities affecting AI Explainability. The asterisk sign denotes either expected publication release or, if delayed, latest announcement for expected publication release. Note that *INT* in the Area column refers to international standardization organizations, while the column *Delegation* refers to specific committee and/or working groups responsible for standard documents provision.

Table 1: Policy communications, reports, and regulations affecting AI explainability.

Area	Communications	Reports	Regulations
EU	• 2018, EC AI for Europe [33] • 2020, White Paper on AI [34] • 2021, EC Fostering a European approach to AI [35] • 2021, EC Coordinated Plan on AI 2021 Review [36] • 2022, EC Rolling Plan for ICT Standardisation - AI [37]	• 2019, HLEG Guidelines for Trustworthy AI [116] • 2020, HLEG Assessment List for Trustworthy Artificial Intelligence (ALTAI) [117] • 2021, JRC AI Standardization Landscape [38]	• 2018, EU GDPR [Enactment] [51] • 2021, AI Act [First Draft] [120] • 2022, AI Liability Directive [Proposal] [121] • 2022, AI Act [Final Draft - General Approach] [122]
US	• 2019, WH - Executive Order on Maintaining American Leadership in AI [108] • 2020, WH - OMB - Guidance for Regulation of AI Applications [109] • 2022, Blueprint for AI Bill of Rights [104]	• 2021, NSCAI Final Report [118] • 2021, NIST Federal Engagement Plan [106] • 2021, NISTIR 8367 [23] • 2021, NISTIR 8312 [123] • 2023, NIST AI RMF v1.0 [107]	• 2021, National AI Initiative Act [Enactment] [2] • 2022, Algorithmic Accountability Act [Draft] [3]
UK	• 2019, National Data Strategy [57] • 2019, AI Sector Deal [58] • 2019, ICO & Alan Turing Institute Project ExplAIn [74] • 2022, National AI Strategy - AI Action Plan [60] • 2022, Establishing a pro-innovation approach to regulating AI [59]	• 2020, ICO, Guidance on the AI Auditing Framework [111] • 2020, ICO, Guidance on AI and Data Protection [110] • 2020, ICO & Alan Turing Institute, Explaining Decisions with AI [75] • 2022, Alan Turing Institute, Common Regulatory Capacity for AI [7]	• 2018, Data Protection Act [Enactment] [112]

Table 2: Standards affecting AI explainability.

Area	Standardization Body	Delegation	Document(s)	Publication	Release Date
EU	**CEN - CENELEC**	JTC 21 AI [28]	Explainability, Verifiability [27]	Proposal	TBD
	ETSI	ISG SAI [50]	DGR/SAI-007 [53, 100]	Under Appr.	2023-03*
			DGR/SAI-010 [100]	Proposal	TBD
		GS ENI	GS ENI 005 v2.1.1 [52, 100]	Published	2021
US	**NIST**	NIST Interagency	NISTIR 8312 (Paper) [123]	Published	2021-09
			NISTIR 8367 (Paper) [23]	Published	2021-04
			RMF V1.0 [107]	Published	2023-01
UK	**BSI - NPL**	AI Standard Hub (CDDO, CDEI)	Algorithmic Transparency Recording Standard (Guide) [29]	Published	2021-12
		CDDO, Cabinet Office, Office for AI	Ethics, Transparency and Acc. Framework for ADM (Guide) [115]	Published	2021-05
INT	**ISO/IEC**	JTC 1/SC 42 AI [76]	ISO/IEC AWI 12792 [79]	Under Dev.	2025-02
			ISO/IEC AWI TS 6254 [77]	Under Dev.	2024-02
			ISO/IEC TR 24028:2020 [78]	Published	2020-05
	IEEE	CIS/SC/XAI WG	P2976 [62]	Inititation	2024-07*
		VT/ITS	7001-2021 [73]	Published	2022-03
		C/AISC/XAI	P2894 (Guide) [9]	Under Dev.	2022-03*
		C/S2ESC/ALGB-WG	P7003 [137]	Published	2017-02