Delft University of Technology

Understanding Viewpoint Biases in Web Search Results

Draws, T.A.

DOI
10.4233/uuid:1b177026-6af7-48f3-ba04-ab7109db3c36

Publication date
2023

Document Version
Final published version

Citation (APA)
Draws, T. A. (2023). *Understanding Viewpoint Biases in Web Search Results*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:1b177026-6af7-48f3-ba04-ab7109db3c36

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

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

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

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Understanding Viewpoint Biases in Web Search Results

Understanding Viewpoint Biases in Web Search Results

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen
op woensdag 13 december 2023 om 12:30 uur

door

Tim Alexander DRAWS

Master of Science in Psychologie, Universiteit van Amsterdam, Nederland,
geboren te Mannheim, Duitsland.

Dit proefschrift is goedgekeurd door de

Promotor: prof. dr. ir. G.J.P.M. Houben
Promotor: prof. dr. N. Tintarev

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. G.J.P.M. Houben,	Technische Universiteit Delft
Prof. dr. N. Tintarev,	Universiteit Maastricht
Onafhankelijke leden:	
Prof. dr. G. Kortuem,	Technische Universiteit Delft
Prof. dr. U. Kruschwitz,	Universität Regensburg, Duitsland
Prof. dr. C. Eickhoff,	Universität Tübingen, Duitsland
Prof. dr. M. Sanderson,	RMIT University, Australië
Prof. dr. M.M. Specht,	Technische Universiteit Delft, reservelid

Keywords: Web Search, Viewpoint Bias, Diversity, Debated Topics, Crowdsourcing, Cognitive Biases, User Behavior, Opinion Formation

Printed by: Print Service Ede

Cover by: Carol Wu. Mouse icon created by Ern from the Noun Project.

Style: TU Delft House Style with modifications by Tim Draws

Copyright © 2023 by T.A. Draws

SIKS Dissertation Series No. 2023-29
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

This activity has been financed by IBM and the Allowance for Top Consortia for Knowledge and Innovation (TKI's) of the Dutch ministry of economic affairs.

ISBN: 978-94-6366-778-4

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

For all those whose path is not as straight as others’.

Und für meine Eltern, Elvira und Joachim Draws,
meine größten Vorbilder,
die immer alles gegeben haben,
um uns ein gutes Leben zu ermöglichen.

Contents

1 Introduction **1**
1.1 Research Questions and Contributions . . . 3
1.2 Chapter Origins . . . 9

2 Background **13**
2.1 Preliminaries and Definitions . . . 13
2.1.1 Debated Topics . . . 13
2.1.2 Opinions and Viewpoints . . . 13
2.1.3 Viewpoint Diversity and Bias in Search Results. . . . 15
2.1.4 User Opinion Formation . . . 15
2.2 Related Work . . . 16
2.2.1 Viewpoint Representation . . . 16
2.2.2 Automatic Viewpoint Labeling . . . 17
2.2.3 Crowdsourcing Viewpoint Annotations . . . 19
2.2.4 Diversity, Fairness, and Viewpoint Biases in Search Results . . . 21
2.2.5 Search Result Viewpoint Biases and User Behavior. . . . 22
2.2.6 Summary . . . 23

I Representing Viewpoints 25

3 Automatic and Explainable Labeling of Search Results With Ternary Stance Labels **29**
3.1 Hypotheses . . . 31
3.2 Data. . . . 33
3.3 Search Result Stance Detection . . . 34
3.3.1 Stance Detection Models. . . . 34
3.3.2 Evaluation . . . 35
3.4 User Study Setup . . . 37
3.4.1 Materials. . . . 38
3.4.2 Variables . . . 39
3.4.3 Procedure . . . 40
3.4.4 Participants . . . 41
3.4.5 Statistical Analyses . . . 41
3.5 Results . . . 41
3.5.1 Descriptive Statistics . . . 42
3.5.2 Hypothesis Tests . . . 42
3.5.3 Exploratory Analyses. . . . 43
3.5.4 Qualitative Analyses . . . 45

3.6 Discussion . . . 46
3.6.1 Implications and Recommendations. . . . 47
3.6.2 Limitations. . . . 48
3.7 Conclusion . . . 49

4 Helping Users Discover Perspectives: Enhancing Stance Detection With Joint Topic Models 51
4.1 Data. . . . 53
4.1.1 Creating an Annotated Data Set . . . 53
4.1.2 Curating a Balanced Data Set . . . 54
4.2 Method . . . 54
4.2.1 Models . . . 55
4.2.2 Operationalization . . . 56
4.2.3 Procedure . . . 58
4.2.4 Hypotheses . . . 58
4.2.5 Statistical Analyses . . . 58
4.2.6 Participants . . . 59
4.3 Results . . . 60
4.3.1 Hypothesis Tests . . . 60
4.3.2 Exploratory Analyses . . . 61
4.4 Discussion . . . 63
4.5 Conclusion . . . 64

5 Comprehensive Viewpoint Representations 65
5.1 Novel Viewpoint Representation . . . 68
5.2 Obtaining Viewpoint Labels. . . . 68
5.2.1 Data . . . 69
5.2.2 Prior Considerations . . . 69
5.2.3 Task Setup . . . 69
5.2.4 Human Annotators . . . 70
5.2.5 Crowd Annotation Aggregation and Quality . . . 71
5.2.6 Gauging the Annotation Difficulty . . . 73
5.3 Analyzing Viewpoint Diversity . . . 73
5.3.1 Method . . . 74
5.3.2 Results . . . 74
5.4 User Evaluation of Viewpoint Label . . . 76
5.4.1 Method . . . 76
5.4.2 Results . . . 78
5.5 Discussion . . . 78
5.5.1 Guidelines for Obtaining Viewpoint Labels . . . 79
5.5.2 Implications . . . 80
5.5.3 Limitations. . . . 80
5.6 Conclusion . . . 80

II Crowdsourcing Viewpoint Annotations 83

6 A Checklist to Combat Cognitive Biases in Crowdsourcing 87
6.1 Introducing a Checklist . . . 88
6.1.1 Cognitive-Biases-in-Crowdsourcing Checklist . . . 89
6.1.2 How to Use the Proposed Checklist . . . 91
6.2 Case Study: Viewpoint Annotations for Search Results on Debated Topics . 94
6.3 Retrospective Analysis . . . 97
6.3.1 Paper Selection Criteria . . . 98
6.3.2 Method . . . 98
6.3.3 Results . . . 99
6.4 Discussion . . . 100
6.4.1 Limitations. . . . 100
6.4.2 Implications . . . 101
6.5 Conclusion . . . 102

7 Identifying Crowd Worker Biases in the Context of Debated Content 103
7.1 Crowdsourced Fact-Checking in Earlier Work . . . 105
7.2 Exploratory Study . . . 106
7.2.1 Data . . . 106
7.2.2 Data preprocessing . . . 107
7.2.3 Exploratory Analyses . . . 108
7.2.4 Hypotheses for the Novel Data Collection . . . 110
7.3 Methods . . . 112
7.3.1 Procedure . . . 112
7.3.2 Variables . . . 112
7.3.3 Crowd Workers . . . 113
7.3.4 Statistical Analyses . . . 114
7.4 Results . . . 114
7.4.1 Descriptive Statistics . . . 114
7.4.2 Hypothesis Tests . . . 116
7.4.3 Exploratory Analyses . . . 116
7.5 Discussion . . . 118
7.5.1 Key Findings . . . 118
7.5.2 Practical Implications . . . 119
7.5.3 Limitations. . . . 120
7.6 Conclusion . . . 120

III Viewpoint Bias Metrics for Search Results 121

8 Assessing Viewpoint Bias in Search Results Using Ranking Fairness Metrics 125
8.1 Measuring Fairness in Rankings . . . 127
8.1.1 Defining Fairness and Viewpoint Bias . . . 128
8.1.2 Desiderata and Practical Considerations for Metrics . . . 128
8.1.3 Ranking Fairness Metrics . . . 129

8.2 Simulation Study 131
  8.2.1 Generating Synthetic Rankings. 131
  8.2.2 Metric Behavior 133
8.3 Discussion 137
  8.3.1 Binomial Viewpoint Fairness. 137
  8.3.2 Multinomial Viewpoint Fairness 137
  8.3.3 Caveats and Limitations 138
8.4 Conclusion 138

9 Comprehensive Viewpoint Bias Evaluation and Viewpoint Diversification for Search Results 141
9.1 Evaluating Viewpoint Bias in Search Results 142
  9.1.1 Measuring Polarity, Stance, and Logic Bias 143
  9.1.2 Normalized Discounted Viewpoint Bias 145
9.2 Case Study 145
  9.2.1 Materials. 146
  9.2.2 Viewpoint Bias Evaluation Results 147
  9.2.3 Viewpoint Diversification 149
9.3 Discussion 151
9.4 Conclusion 152

IV How Search Result Viewpoint Biases Affect User Behavior 155

10 Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics 159
10.1 Hypotheses 161
  10.1.1 Vulnerability of Users to Opinion Change 162
  10.1.2 User Perception of Viewpoint Diversity 163
10.2 Data. 164
10.3 Method and Experimental Setup 165
  10.3.1 Materials. 165
  10.3.2 Variables 166
  10.3.3 Procedure 167
  10.3.4 Statistical Analyses. 168
  10.3.5 Participants 169
10.4 Results 169
  10.4.1 Descriptive Statistics. 169
  10.4.2 Hypothesis Tests 171
  10.4.3 Exploratory Findings. 172
10.5 Discussion 173
  10.5.1 Explaining SEME? 174
  10.5.2 Implications 175
  10.5.3 Caveats and Limitations 175
10.6 Conclusions. 176

11 Conclusion 177
11.1 Summary of Findings 177
11.2 Implications 180
11.3 Limitations and Future Work 182
11.4 Ethical Considerations 183
11.5 Concluding Remarks 184

Bibliography 185

List of Figures 224

List of Tables 226

Summaries 228
English Summary 228
Nederlandse Samenvatting. 229

Acknowledgements 230

Curriculum Vitæ 232

List of Publications 233

1

Introduction

Web search engines are embedded in our daily lives, providing fast and convenient access to the often overwhelming amount of resources that may satisfy users' information needs [135, 203]. In practice, however, search engines do not simply retrieve relevant resources; they act as information gatekeepers and, as a result, play an important role in shaping individual and collective knowledge [151]. This becomes more noticeable as people increasingly employ search engines to form opinions and seek advice on debated topics [54, 119, 120, 323, 369]. For example, users may search the web to help them decide whether to embrace veganism [120], what financial strategy to employ [369], or what to think of a particular political candidate [99]. A key difference between search that involves debated topics and other search contexts such as navigation [202] or learning [65] lies in its subjectivity: just like each user has a unique interpretation of the world accompanied by corresponding (pre-search) opinions [25], each search result conveys a unique perspective on any given debated topic. This means that it not only matters whether search results are *relevant* to the topic at hand but also what *viewpoints* they express. Recent research has demonstrated that the viewpoints users encounter on search engine results pages (SERPs) can play a crucial role in their opinion formation [10, 18, 46, 54, 99, 170, 223, 226, 283]. Hence, it is imperative to examine search results on debated topics and their effects on user behavior – and ultimately user opinions – more closely.

Aiming to build well-informed opinions by considering arguments from all sides, users without strong pre-search opinions often seek to encounter diverse viewpoints when searching the web for debated topics [7, 226]. Previous research has already considered search result *diversity* concerning different user intents [4, 64, 313] or *fairness* with respect to particular document attributes [35, 381, 391, 393]. These existing approaches, however, do not encompass *viewpoint diversity*, which is given when all provided search results are relevant to the same user intent (e.g., the debated topic *school uniforms*) but diverse concerning the viewpoints they express (i.e., giving equal attention to all existing viewpoints). Related work shows that web search results on debated topics are not always viewpoint-diverse and can include viewpoint biases (i.e., over-representing particular viewpoints, e.g., related to politics [284] or health information [370, 371]). Specifically, different data-related [13, 17, 254, 352, 372], algorithmic [114, 257, 284, 370, 371], presentational [26, 28, 386], and contextual biases [162, 181, 251, 389] can creep in at several points in the information retrieval (IR) process and skew users' viewpoint exposure. This stands in contrast to the perception of most users, who generally view their web search interactions as unbiased [54, 123, 283] and regard higher-ranked search results as more credible [143].

Viewpoint biases in search results can translate into biased user behavior, opinion formation, and decision-making. Smith and Rieh [328] argue that the strong trust users have gained in search engines to provide accurate and reliable resources can lead users to believe that they do not need to exert cognitive effort in the search process. Such low cognitive effort often goes hand-in-hand with mental heuristics and cognitive biases [349, 354, 377], e.g., web search users often pay much more attention to search results at high ranks and rarely examine results beyond the first SERP [168, 258]. Previous research has shown that combinations and interactions of search result and user biases can lead to noteworthy user behavior. For example, information seekers may fall prey to the *search engine manipulation effect* (SEME), where search result viewpoint biases interact with cognitive user biases so that users adopt whatever viewpoint is expressed by most highly-ranked search results [10, 37, 99, 274]. Phenomena such as SEME, which can occur without users' awareness [123], are unlikely what users aim for when searching the web for debated topics [123, 226, 283] and do not reflect responsible opinion formation [189, 265]. Thus, depending on the search topic and context, viewpoint-biased search results may have serious implications for individuals, businesses, and society.

Recent work has begun to assess search result viewpoint biases [121, 284, 370, 371] and their effects on user behavior and opinions [10, 99, 100, 274]. However, existing research in this area has faced four crucial limitations, which we aim to address in this dissertation. First, *viewpoint labels* for search results have typically followed simple viewpoint taxonomies (e.g., *against*/*neutral*/*in favor*) [121, 385] that are comparatively easy to obtain and handle but ignore important nuances between viewpoints. Part I of this dissertation focuses on developing novel viewpoint representations that are more comprehensive (i.e., incorporating nuanced notions of *stance* and *logic of evaluation*) yet topic-independent and computationally tractable. Second, viewpoint labels are often gathered using crowdsourced annotations without considering the cognitive biases of crowd workers that can reduce data quality [96, 156]. We address this issue in Part II by introducing tools and best practices for crowdsourcing (viewpoint) annotations. Third, to the best of our knowledge, no previous research has examined how to best *measure* viewpoint bias in search results, and earlier work has largely used simplified or manual annotation methods that are impractical for large-scale applications [99, 370, 385]. Part III explores using existing *fair ranking* methods and proposes novel metrics to automatically evaluate viewpoint bias in ranked search result lists. Fourth, while several studies have shown that observable search result viewpoint bias (see Part III) can lead to biased user behavior and opinion formation [10, 99, 274], it is currently unclear in what situations (e.g., different search contexts or bias degrees) these effects occur and what their underlying mechanisms are. Part IV presents a user study investigating the effects of search result viewpoint biases on user behavior across different search scenarios.

1.1. Research Questions and Contributions

We here describe the motivation, main research questions, and contributions of each part of the dissertation. Participants in our user studies always agreed to informed consent prior to commencing any task. All user studies in this dissertation were furthermore approved by the human research ethics committee at TU Delft and (aside from the studies described in Chapters 4 and 6) preregistered before data collection.[1]

Part I: Representing Viewpoints

A crucial decision when studying viewpoint biases in the web search context is how to *represent* viewpoints on debated topics and label search results accordingly. For example, when aiming to measure viewpoint bias in a search result list on the topic *school uniforms*, the search result assessor needs to know the viewpoints expressed by each individual search result. Previous human information interaction research has predominantly represented viewpoints by assigning binary (e.g., *democrat*/*republican*) or ternary stance labels (e.g., *against*/*neutral*/*in favor*) [121, 274, 385]. Such labels are feasible to obtain at scale (e.g., using crowdsourcing [233, 234] or automatic stance detection techniques [9, 294, 366]) and lend themselves to existing search result diversity [4, 64, 310] and ranking fairness methods [381, 391, 393]. However, binary and ternary stance labels are extremely generic categorizations of truly nuanced viewpoints as they do not incorporate any *degree* within or *reason* behind stances, e.g., they do not distinguish between *somewhat* or *strongly supporting* school uniforms and do not capture whether someone supports school uniforms for economic or functional reasons. These drawbacks greatly limit the insight that can be gained from such labels and motivate the development of more comprehensive viewpoint representations that capture nuanced differences between viewpoints while remaining computationally tractable. Part I of this dissertation, therefore, addresses the following research question:

$\mathbf{RQ_I}$ What label taxonomy can accurately represent viewpoints on debated topics?

Our first step in addressing **$\mathbf{RQ_I}$** is a user study investigating the feasibility of explainable, cross-topic stance detection for search results using current ternary stance labels (i.e., *against*/*neutral*/*in favor*; Chapter 3). We here aim to gauge the potential and limitations of currently available automated methods. Specifically, we apply stance detection techniques to search results on three debated topics (i.e., *atheism*, *intellectual property rights*, and *school uniforms*), generate explanations for stance predictions, and evaluate these explanations quantitatively and qualitatively with users. Our results show that, although some explanations help users interpret the behavior of stance detection models, users are often unsatisfied with the quality and amount of viewpoint-related information they receive. These findings indicate that considering more comprehensive viewpoint representations for search results could enable more meaningful viewpoint diversity analyses and better assist users in their opinion formation.

Chapter 4 introduces *perspectives* (i.e., reasons for opposing or supporting a topic) as an alternative viewpoint representation format to the typical binary or ternary stance labels. We assemble a corpus of debate forum entries on *abortion legalization* and apply unsupervised topic models to discover perspectives on why debaters oppose or support this issue. In a user study, we then evaluate whether users can identify the models' automatically generated *topics* as particular perspectives. We find that some topic models can indeed discover such user-identifiable perspectives and that users' pre-existing stances on abortion do not affect their ability to interpret the topic model output. Including perspectives as part of viewpoint representations could thus be feasible and useful.

---

[1] We preregistered user studies by publicly announcing the motivation, hypotheses, study setup, and statistical analysis plan for each user study before collecting data. Links to the individual preregistrations and online repositories can be found in the respective chapters.

Finally, Chapter 5 builds on the first two chapters and previous work in the communication sciences to propose a comprehensive viewpoint representation for human information interaction. This novel label consists of two dimensions: *stance* (i.e., a viewpoint's position regarding a debated topic, measured on a seven-point ordinal scale ranging from *strongly opposing* to *strongly supporting*) and *logic of evaluation* (i.e., stances' underlying reasons or perspectives categorized into seven topic-independent categories). Although current automatic methods cannot classify documents into such nuanced viewpoint labels, our seven-category stance and logic of evaluation taxonomies could feasibly be learned and predicted automatically.

Contributions of Part I: Representing Viewpoints. The first part of this dissertation contributes to the field of human information interaction. The findings we present here can thus be applied not only to web search results but also to similar use cases such as social media posts or podcasts. Specifically, Part I contributes the following:

- We conduct two user studies; one in which we quantitatively and qualitatively evaluate cross-topic stance detection explanations using state-of-the-art methods (Chapter 3) and one in which we investigate whether *topic models* can produce user-identifiable perspectives (i.e., underlying reasons for opposing or supporting a debated topic; Chapter 4). Based on these findings, we identify that more comprehensive viewpoint labels (compared to the classical ternary stance labels) could be useful and feasible.

- We propose a novel, two-dimensional viewpoint representation for human information interaction alongside guidelines on crowdsourcing corresponding viewpoint labels (Chapter 5). Such labels may be used to represent viewpoints expressed in documents such as web search results.

- We publish two data sets: one containing debate forum entries on *abortion legalization*, expert-annotated with *perspectives* (Chapter 4); and one containing tweets on three different debated topics (i.e., *atheism*, *Donald Trump*, and the *feminist movement*), annotated by crowd workers with our proposed two-dimensional viewpoint label (Chapter 5). Researchers and practitioners may use these data sets for purposes such as training machine learning models or conducting user studies.

Part II: Crowdsourcing Viewpoint Annotations

Assigning our proposed viewpoint label (see Part I) to search results at scale (e.g., to evaluate viewpoint bias or develop practical applications) is currently not possible using automatic methods and thus requires the input of crowd workers [112, 233]. Such crowdsourcing efforts typically involve collecting at least three judgments per search result from different crowd workers (e.g., asking workers what viewpoint a given search result expresses regarding *school uniforms*) and subsequently aggregating those into single labels. However, recent research has found that *cognitive worker biases* can strongly reduce data quality from subjective tasks such as annotating viewpoints [96, 156]. One example of this is the *confirmation bias*: workers may be more likely to annotate their personal viewpoint rather than other viewpoints because they pay more attention to document parts that (seem to) confirm their personal opinion [156, 245]. It is vital to reduce such cognitive worker biases when collecting viewpoint annotations for search results to prevent data biases and ensure high-quality research and practical applications. That is why Part II of this dissertation addresses the following research question:

RQ$_{\text{II}}$ What cognitive biases reduce crowd workers' abilities to correctly annotate web search results with viewpoint labels?

We begin addressing **RQ$_{\text{II}}$** in Chapter 6 by proposing a checklist to combat cognitive biases in crowdsourcing. For example, the *anchoring effect*, a commonly occurring cognitive bias, could lead crowd workers to judge a truly neutral document as *in favor* of school uniforms simply because it seems to be more in favor than previously seen documents [110, 248]. Our checklist, adapted from earlier work concerning business decision-making [169], comprises 12 items referring to particularly common or problematic (groups of) cognitive biases that may reduce the quality of crowdsourced data labels. We present a retrospective analysis of past crowdsourcing papers, showing that cognitive biases are rarely considered but may affect data quality for most tasks. Requesters can use our proposed checklist to inform their task design (e.g., to mitigate cognitive biases) and document potential influences of cognitive biases on the data they collect.

At the hand of a related but slightly different use case, i.e., crowdsourced fact-checking of politician statements, Chapter 7 presents a full application of the checklist we propose in Chapter 6. We apply our checklist to an existing data set of crowdsourced truthfulness annotations and crowd worker characteristics (e.g., political affiliation, level of education, and annotation confidence) to identify potential influences of cognitive biases in this context. Subsequently, we test our hypotheses by conducting a similar crowdsourcing study while measuring the cognitive biases we had identified.

Contributions of Part II: Crowdsourcing Viewpoint Annotations. Although we focus on generating viewpoint labels for web search results, the second part of this dissertation contributes to the general efforts toward higher-quality crowd worker annotations and more reliable crowdsourced data. Specifically, Part II makes the following contributions:

- We conduct a retrospective analysis of past crowdsourcing papers to investigate how often requesters consider the influence of cognitive biases and how often cognitive biases may affect data quality (Chapter 6). This analysis shows that cognitive crowd worker biases may be common but rarely considered by requesters.

- We propose a 12-item checklist to combat cognitive biases in crowdsourcing and demonstrate its use for collecting viewpoint labels for search results (Chapter 6). Requesters can now use this tool to document, assess, and mitigate cognitive crowd worker biases for the crowdsourcing tasks they design and the data they collect.
- We present a full application of our checklist for the use case of crowdsourced fact-checking, showing how to identify, assess, and mitigate potential influences of cognitive biases in crowdsourcing (Chapter 7).
- We propose guidelines to improve data quality when crowdsourcing truthfulness judgments and similar subjective annotations (e.g., viewpoint labels; Chapter 7).
- We publish a data set of crowdsourced truthfulness judgments alongside various worker characteristics and behavior (e.g., workers' political affiliations and annotation confidence; Chapter 7). This data set can be used to further investigate cognitive worker biases in this context and inform similar data collection efforts.

Part III: Viewpoint Bias Metrics for Search Results

Collecting search result data with accurate and comprehensive viewpoint labels (using the methods we propose in Parts I and II) enables researchers and practitioners to measure viewpoint bias in search results. For example, an assessor may wish to evaluate the viewpoint bias in a search result list on *school uniforms* to gauge its potential impact on users. Such viewpoint bias assessments are essential in scoping and understanding the general problem of viewpoint biases in current search engines, linking specific degrees of viewpoint bias to user behavior, and exploring how search result viewpoint diversity could potentially be improved. Much work has been devoted to measuring *diversity* concerning query subtopics (i.e., evaluating how well a ranked search result list covers all potential user intents given a query) [4, 64, 308] and *fairness* toward particular groups of search results (i.e., evaluating whether documents that express a particular viewpoint are represented equally across a ranking) [35, 114, 381, 391, 392, 393]. These methods consider the rank and particular characteristics (e.g., the subtopic relevance or protected attribute) of each search result. However, there is currently no notion of ideal viewpoint diversity or protected viewpoints for search results, and measuring viewpoint bias specifically [195] has received comparatively little attention. Recent research has looked at defining and measuring viewpoint diversity for the domain of news recommender systems [147, 148, 361], but more work is needed to translate these concepts into the web search paradigm and develop viewpoint bias metrics specifically for search results. Part III thus addresses the following research question:

RQ$_{\text{III}}$ What methods can evaluate viewpoint bias in search results?

In Chapter 8, we explore using existing ranking fairness metrics and propose a novel metric to assess viewpoint biases in search results. We make several practical considerations and design choices to adapt existing ranking fairness metrics to the search result viewpoint bias use case and show in simulation studies how different metrics behave for different degrees of viewpoint (ranking) bias. From these simulations, we derive guidelines for measuring viewpoint bias in search results using ranking fairness metrics. The novel ranking fairness metric we propose can accommodate multicategorical viewpoint labels (i.e., instead of the protected/unprotected dichotomy).

Previously developed methods such as ranking fairness metrics allow for measuring viewpoint biases in search results but cannot incorporate multidimensional viewpoint representations such as the one we propose in Part I. That is why Chapter 9 proposes a rank-aware viewpoint bias metric for search results that considers this more comprehensive viewpoint label. The novel metric, which measures bias as a deviation from viewpoint plurality, is founded upon a clear notion of viewpoint diversity and can be adapted to fit different topics or viewpoint structures.

Contributions of Part III: Viewpoint Bias Metrics for Search Results. The third part of this dissertation primarily contributes to the field of information retrieval by exploring how to measure viewpoint bias in web search results. Specifically, Part III makes the following contributions:

- We present a simulation study showing how different ranking fairness metrics behave under different degrees of viewpoint (ranking) bias in search results and derive guidelines for using ranking fairness metrics to measure search result viewpoint bias (Chapter 8). This simulation study demonstrates that measuring viewpoint bias using ranking fairness metrics is possible but has crucial limitations (e.g., such metrics are not applicable to multi-categorical or multi-dimensional viewpoint representations).

- We propose two metrics: one ranking fairness metric that can measure search result viewpoint bias when considering multi-categorical stance labels (Chapter 8) and one comprehensive viewpoint bias metric that accommodates multi-dimensional viewpoint labels (Chapter 9). Researchers and practitioners can use these novel metrics to comprehensively measure viewpoint bias in search results.

- We publish a viewpoint-annotated data set of search results on three different debated topics from two popular search engines and report on viewpoint biases in these search results (Chapter 9). This data set can be used for purposes such as the measurement or mitigation of search results viewpoint bias.

- We demonstrate how to increase viewpoint diversity in search results using existing diversification methods (Chapter 9). Our demonstration highlights that reducing viewpoint bias is not difficult and may not starkly reduce search result ranking utility.

Part IV: How Search Result Viewpoint Biases Affect User Behavior

Our contributions in Parts I, II, and III build the foundation for comprehensively measuring viewpoint biases in search results and connecting metric outcomes to user behavior. For example, viewpoint-biased search results concerning *school uniforms* could lead users to vote in favor of a school uniform mandate without properly considering the opposing side. Recent research has already demonstrated that severe viewpoint biases in search results can lead to phenomena such as SEME for users without strong pre-search opinions [10, 36, 37, 99, 274]. However, it is still unclear at what *degree* viewpoint biases begin to cause systematic user tendencies, what mechanisms cause these effects, and whether they occur across search topics. Understanding the relationship between search result viewpoint biases and user behavior in more detail contributes to a better understanding of when viewpoint biases may become problematic and how web search engines could support users in their opinion formation. Part IV thus investigates the following research question:

RQ$_{\mathbf{IV}}$ What cognitive processes underlie the effect of search result viewpoint bias on users' opinion formation?

We address **RQ$_{\mathbf{IV}}$** by conducting a user study investigating whether lower-degree viewpoint biases in search results can cause SEME and what cognitive biases may influence user behavior here (Chapter 10). Whereas previous research in this area has largely presented users with strongly biased search results, we expose users to single, top 10 SERPs that are overall viewpoint-balanced (i.e., five opposing and five supporting results) but are ranked with different degrees of bias (e.g., ranking all opposing results above the supporting results or ranking them in alternating fashion).

Contributions of Part IV: How Search Result Viewpoint Biases Affect User Behavior. The fourth and final part of this dissertation contributes to a better understanding of user behavior and opinion formation following interactions with viewpoint-biased web search results. This can inform future research and development of web search engines but may also prove useful for related domains such as news or online video recommendation. Specifically, Part IV contributes the following:

- We conduct a user study investigating the effect of viewpoint (ranking) bias for overall viewpoint-balanced top 10 SERPs on users' opinion formation (Chapter 10). This study provides exploratory evidence that viewpoint biases may not affect user behavior or opinion formation across search scenarios (e.g., when viewpoint bias is limited).

- We publish a viewpoint-annotated data set of search results on five different debated topics (Chapter 10). This data set can be used for purposes such as further exploring user behavior in this context and informing bias mitigation strategies.

Summary

In four parts that each address one core research question, this dissertation supports ongoing efforts toward a better understanding of search result viewpoint biases and their effects on user behavior. We present work concerning the representation of viewpoints in human information interaction, the collection of high-quality crowdsourced viewpoint annotations, viewpoint bias evaluation for search results, and how such viewpoint biases may affect users. Our contributions include novel theoretical concepts (e.g., a viewpoint representation; see Chapter 5), tools (e.g., a checklist to combat cognitive biases in crowdsourcing; see Chapter 6), metrics (e.g., a novel viewpoint bias metric for search results; see Chapter 9), user studies (e.g., investigating underlying mechanisms of the *search engine manipulation effect*; see Chapter 10), and resources (e.g., data sets of search results with viewpoint annotations; see Chapters 4, 5, or 9). We discuss the broader implications and limitations of our work in Chapter 11.

1.2. Chapter Origins

This dissertation comprises 11 chapters. Whereas the introduction (Chapter 1), background (Chapter 2), and conclusion (Chapter 11) contextualize and discuss our work, Chapters 3-10 are based on research papers, organized into four main parts analogous to our four research questions.

Part I: Representing Viewpoints

- Chapter 3 is based on a published, full conference paper: Tim Draws, Karthikeyan Natesan Ramamurthy, Ioana Baldini, Amit Dhurandhar, Inkit Padhi, Benjamin Timmermans, and Nava Tintarev. "Explainable Cross-Topic Stance Detection for Search Results". In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval.* CHIIR '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 221–235. DOI: `10.1145/3576840.3578296`

- Chapter 4 is based on a published workshop paper: Tim Draws, Jody Liu, and Nava Tintarev. "Helping Users Discover Perspectives: Enhancing Opinion Mining with Joint Topic Models". In: *2020 International Conference on Data Mining Workshops (ICDMW).* Sorrento, Italy: IEEE, Nov. 2020, pp. 23–30. DOI: `10.1109/ICDMW51313.2020.00013`

- Chapter 5 is based on a published, full conference paper: Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. "Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions with Debated Topics". In: *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval.* CHIIR '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 135–145. DOI: `10.1145/3498366.3505812`
  - 🏆 This paper won the Best Paper Award at the 2022 *ACM SIGIR Conference on Human Information Interaction (CHIIR).*

Part II: Crowdsourcing Viewpoint Annotations

- Chapter 6 is based on a published, full conference paper: Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. "A Checklist to Combat Cognitive Biases in Crowdsourcing". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing.* Vol. 9. HCOMP '21. 2021, pp. 48–59. DOI: `10.1609/hcomp.v9i1.18939`

  - 🏆 This paper won the Best Paper Award at the 2021 *AAAI Conference on Human Computation and Crowdsourcing (HCOMP).*

- Chapter 7 is based on a published, full conference paper: Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. "The Effects of Crowd Worker Biases in Fact-Checking Tasks". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 2114–2124. DOI: `10.1145/3531146.3534629`

Part III: Viewpoint Bias Metrics for Search Results

- Chapter 8 is based on a published workshop paper: Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. "Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics". In: *ACM SIGKDD Explorations Newsletter* 23.1 (May 2021), pp. 50–58. DOI: `10.1145/3468507.3468515`

- Chapter 9 is based on a published, full conference paper: Tim Draws, Nirmal Roy, Oana Inel, Alisa Rieger, Rishav Hada, Mehmet Orcun Yalcin, Benjamin Timmermans, and Nava Tintarev. "Viewpoint Diversity in Search Results". In: *Advances in Information Retrieval.* Ed. by Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo. Vol. 13980. Cham: Springer Nature Switzerland, 2023, pp. 279–297. DOI: `10.1007/978-3-031-28244-7_18`

Part IV: How Search Result Viewpoint Biases Affect User Behavior

- Chapter 10 is based on a published, full conference paper: Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. "This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* SIGIR '21. New York, NY, USA: Association for Computing Machinery, July 2021, pp. 295–305. DOI: `10.1145/3404835.3462851`

Additionally, this dissertation has benefited from the following published conference and workshop papers:

- Oana Inel, Tim Draws, and Lora Aroyo. "Collect, Measure, Repeat: Reliability Factors for Responsible AI Data Collection". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 11.1 (Nov. 2023), pp. 51–64. DOI: `10.1609/hcomp.v11i1.27547`
  - This paper won a Best Paper Honorable Mention at the 2023 *AAAI Conference on Human Computation and Crowdsourcing (HCOMP).*

- Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. “Disentangling Fairness Perceptions in Algorithmic Decision-Making: The Effects of Explanations, Human Oversight, and Contestability”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. New York, NY, USA: Association for Computing Machinery, 2023. DOI: `10.1145/3544548.3581161`
    - 🏆 This paper won a Best Paper Award at the 2023 *ACM CHI Conference on Human Factors in Computing Systems*.
- Zhangyi Wu, Tim Draws, Federico Cau, Francesco Barile, Alisa Rieger, and Nava Tintarev. “Explaining Search Result Stances to Opinionated People”. In: *Explainable Artificial Intelligence*. Ed. by Luca Longo. Communications in Computer and Information Science. Cham: Springer Nature Switzerland, 2023, pp. 573–596. DOI: `10.1007/978-3-031-44067-0_29`
- Markus Bink, Sebastian Schwarz, Tim Draws, and David Elsweiler. “Investigating the Influence of Featured Snippets on User Attitudes”. In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. CHIIR ’23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 211–220. DOI: `10.1145/3576840.3578323`
- Alisa Rieger, Tim Draws, Nava Tintarev, and Mariet Theune. “This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias”. In: *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*. HT ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 189–199. DOI: `10.1145/3465336.3475101`
    - 🏆 This paper won the Best Paper Award at the 2021 *ACM Conference on Hypertext and Social Media*.
- Fausto Giunchiglia, Styliani Kleanthous, Jahna Otterbacher, and Tim Draws. “Transparency Paths - Documenting the Diversity of User Perceptions”. In: *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 415–420. DOI: `10.1145/3450614.3463292`

2

Background

2.1. Preliminaries and Definitions

In our work, we assume that search results retrieved by typical, general-purpose web search engines may express viewpoint(s) concerning one or more debated topics. Users may consider these viewpoints to supplement their opinion formation. Below, we define and discuss the terminology we use to describe such interactions (e.g., debated topics, viewpoint bias, opinion formation) and outline the scope of this dissertation.

2.1.1. Debated Topics

Users increasingly employ web search to form opinions and seek practical advice [54, 119, 120]. Although such information seeking may sometimes involve shallow concerns (e.g., a celebrity's latest outfit), users often search the web for more impactful, *debated topics*: issues of ongoing discussion that are – at least according to some participants of the debate – unresolved. For example, users may search the web to help them decide whether to embrace veganism [120], what financial strategy to employ [369], or what to think of a particular political candidate [99]. Debated topics can include extremely one-sided matters that are scientifically answerable (e.g., whether the Earth is a sphere) or have large majority stances (e.g., whether societies should segregate people by race). In this dissertation, however, we focus on more balanced issues with legitimate arguments on both sides of the debate (e.g., whether students should wear school uniforms).

2.1.2. Opinions and Viewpoints

An individual's *opinion* is the collection of their thoughts, beliefs, or judgments about something or someone [252]. For example, someone may believe that students should wear school uniforms because, according to them, even though uniforms make it difficult for children to express themselves, uniforms prevent bullying and lead to better academic performance. We consider *viewpoints* instantiations of opinions, i.e., specific arguments that express one's opinion on a topic. For example, "I am in favor of school uniforms because they lead to better academic performance among students" could be a viewpoint

Table 2.1: The seven ordinal stance categories we consider.

Stance	**Description**	**Example (wrt. *school uniforms*)**
-3	strongly opposing	“A horrible idea! Students have to be able to wear what they want.”
-2	opposing	“We should not force students to wear uniforms.”
-1	somewhat opposing	“Despite the benefits of school uniforms, overall I’m against them.”
0	neutral	“There are good arguments for and against school uniforms.”
1	somewhat supporting	“Although school uniforms have their disadvantages, they have an overall positive impact.”
2	supporting	“I’m in favor of school uniforms; every school should have them.”
3	strongly supporting	“There is nothing wrong with school uniforms – there should be no other choice!”

Table 2.2: The seven logics of evaluation we consider, adapted from Baden and Springer [24]. Each logic represents a particular orientation of what is desired and can be used to either support or oppose a given claim.

Logic of Evaluation	**Good is…**	**Examples**
Inspired	…what is true, divine, and amazing	Righteous, pre-ordained, beautiful; false, uncreative, dull
Popular	…what is popular or what the people want	Preferred, popular, favourite; resented, feared, isolated
Moral	…what is social, fair, and moral	Solidary, responsible, just; inhumane, asocial, egoistic
Civic	…what is legal, accepted, and conventional	Legal, agreed, common; scandalous, deviant, inappropriate
Economic	…what is profitable and creates value	Beneficial, economic, affordable; wasted, costly, unproductive
Functional	…what works	Effective, necessary, quick; dysfunctional, inefficient, useless
Ecological	…what is sustainable and natural	Sustainable, organic; unnatural, irreversible

deducted from the opinion example above. Viewpoints, finally, can be dissected into *stances* and *logics of evaluation*. A *stance*, also sometimes referred to as *attitude*, describes the position a viewpoint takes on a topic on a scale ranging from *strongly opposing* to *strongly supporting*) (e.g., see Table 2.1) [337]. For example, the viewpoint example above expresses a *supporting* stance. A *logic of evaluation*, also sometimes referred to as *logic*, *perspective*, or *angle*, is a broad categorization of the reasoning behind a stance [25, 39]. In particular, Boltanski and Thévenot [39] introduced this term to represent argumentative reasons using seven topic-independent categories: *inspired*, *popular*, *moral*, *civic*, *economic*, *functional*, and *ecological* (see Table 2.2). Each logic stands for one fundamental way to reason within an argument, e.g., the viewpoint example above contains a *functional* logic (i.e., expressing the idea that school uniforms are effective in improving academic performance).[1]

Online content creators may have opinions on the topics they write about and consequently express viewpoints on these topics in their content. Search results retrieved in response to inquiries about debated topics (e.g., *school uniforms*) may thus contain a range of different viewpoints on these topics. Depending on what viewpoints (high-ranking) search results express, there may be a viewpoint bias on the SERPs that users receive. Users who perform this information-seeking may similarly have particular pre-search opinions and subscribe to corresponding viewpoints. The way users interact with and are influenced by such content, however, can heavily depend on their individual characteristics (e.g., users who already have a strong pre-search opinion on *school uniforms* may search in a more targeted fashion than uninformed, less opinionated users) [21, 245]. Our work follows previous research in the domain of search result viewpoint biases by focusing on users with *mild* pre-search opinions (i.e., users who are undecided or just somewhat oppose or support a topic) [10, 99, 123, 274, 372].

---

[1] Note that the terms described in this paragraph have been defined in many different ways and are often used interchangeably [24, 25, 41, 48, 56, 252, 337, 388]. Due to the ambiguity surrounding some of these terms, we here had to decide on a set of clear definitions that may not fully align with some related research.

2.1.3. Viewpoint Diversity and Bias in Search Results

Viewpoints can be expressed in text and may thus also appear in search results, i.e., on web pages. Specifically, any search result relevant to a debated topic may express anywhere from no viewpoints (e.g., purely descriptive documents) to large numbers of viewpoints (e.g., a debate forum where users argue in different directions). The aggregated viewpoints expressed by a document form the overall opinion it conveys. For example, an online article may argue that school uniforms impede children's natural desire to express themselves but that they make an overall positive contribution because they make children more equal and lead to better academic performance. This blog post may be labeled as overall *somewhat supporting* (see Table 2.1) the idea that students should wear school uniforms, arguing with *ecological*, *moral*, and *functional* logics (see Table 2.2).

One can gauge the overall viewpoint distribution across a ranked search result list by examining the viewpoints expressed in all documents across the ranks. Although *viewpoint diversity* has been conceptualized in different ways [147, 240, 345, 361], our work considers the *deliberative* notion of diversity, which posits that all users (e.g., no matter their pre-existing opinions) should be exposed to the greatest possible plurality of viewpoints (e.g., irrespective of different viewpoints' prevalence in society or the search engine's document index) [147]. This means that a maximally viewpoint-diverse search result list represents all topic-relevant viewpoints across its ranks. Note that this definition may not be applicable to all kinds of topics or queries – but we believe it most universally suits the kind of topic we consider in this dissertation (i.e., debated topics with legitimate arguments in both directions; e.g., *school uniforms*; see Section 2.1.1).

2.1.4. User Opinion Formation

Opinion formation describes the progressive process of developing one's opinion on a topic, e.g., to satisfy a personal interest or seek advice on an issue of individual, business-related, or societal concern [56]. It can occur in various contexts (e.g., reading the news, talking to peers, or searching the web) and may continue indefinitely as one's personal opinions change over time [56]. *Stance change* (or *attitude change*) is a particular form of opinion formation that primarily concerns one's stance, irrespective of whether a change in reasoning also occurred [271, 272]. For example, following a web search, a user could have changed their stance from somewhat opposing to strongly opposing school uniforms without necessarily adapting their logics of evaluation. Similarly, opinion formation can also occur without stance change, e.g., when someone adopts or removes a logic of evaluation in their argumentation but does not change their overall stance. We refer to web search with the intent of forming opinions as *search as opinion formation*.

2.2. Related Work

In this section, we describe and discuss related work concerning viewpoint representation, viewpoint labeling procedures, viewpoint diversity and biases in search results, and the effects of search result viewpoint biases on users.

2.2.1. Viewpoint Representation

Recent years have seen a stark increase in research concerning viewpoints and debated topics online. Inspired by calls to combat bias on the web [26, 259], this line of research has explored user interactions with debated topics in domains such as health [10, 123, 274, 370] or politics [99, 100, 248], social media [228], and the web in general [226]. An essential part of research concerning debated topics in human information interaction is how to represent viewpoints. That is, to measure viewpoint biases or examine what viewpoints users primarily interact with, each document (e.g., a search result or social media post) first needs to receive a *label* that reflects what viewpoint(s) it expresses. Previous human information interaction research has typically represented viewpoints as binary (e.g., *democrat*/*republican*) or ternary (e.g., *against*/*neutral*/*in favor*) stance categories [121, 274, 385]. For instance, Gezici et al. [121] used *against*/*in support* as well as *liberal*/*conservative* viewpoint categories, and Yom-Tov, Dumais, and Guo [385] classified users and documents into the political leanings *democrat*, *centrist*, or *republican*.

Although simple binary or ternary stance taxonomies enable cheap computation of metrics and are feasible to obtain at scale via crowdsourcing annotations (see Section 2.2.2), they are extremely generic categories that offer little detail. That is why recent work has explored alternative viewpoint label taxonomies, e.g., by representing viewpoints on continuous scales [195, 240]. However, despite adding more nuance to the *against*/*in favor* dichotomy, such labels are still lacking crucial information about viewpoints' underlying reasons (e.g., whether an argument reflects a moral or economic perspective on school uniforms). This notion of *perspective* as a dimension next to a viewpoint's stance has already been explored [5, 61] but often faced the limitation of these perspectives being highly topic-dependent (e.g., the debated topics *school uniforms* and *abortion legalization* have vastly different perspective spaces).

Drawing Inspiration from the Communication Sciences

Viewpoint diversity in public discourse is a long-standing subject of study in the communication sciences [23, 24, 25, 211, 220, 221, 276, 284, 358] that has already been applied to information access systems [147, 148, 236, 361]. Compared to human information interaction, the communication sciences have brought forward more advanced viewpoint representations. There, for instance, a common way to explore viewpoints is *framing*, whereby a viewpoint is typically analyzed on four different dimensions: *problem definition* (i.e., what is happening), *causal attribution* (i.e., who is responsible for the problem), *moral evaluation* (i.e., whether the problem is good or bad), and *treatment recommendation* (i.e., suggestions in response to the problem) [97]. More recent work has combined framing with the notion of *interpretative repertoires* to propose a topic-independent way of representing viewpoints [24, 25]. In this method, each *frame* (i.e., a viewpoint based on the four dimensions mentioned above) is seen as an instance of a more general way of interpreting the world (i.e., the interpretative repertoire). Building on the idea of “common worlds” proposed by Boltanski and Thévenot [39], Baden and Springer [24, 25] view frames as commensurable if they refer to the same repertoire commonly used in argumentation. For example, consider the phrases “*feminism is on the rise because women should be treated equally*” and “*stop attacking feminists, they are the ones who fight for fair treatment*”. These two phrases express different frames but have the same *logic of evaluation* (i.e., good is what is social, fair, and moral). This logic of evaluation is a key aspect of interpretative repertoires and offers a topic-independent way to represent perspectives behind the stances of viewpoints (see Table 2.2).

A drawback of analyzing viewpoints using framing or interpretative repertoires is that it usually requires a trained expert who performs manual annotation. This is impractical for human information interaction and related fields that need to obtain viewpoint representations at scale to enable cheap computation of metrics and algorithms. Although first attempts have been made to analyze the viewpoint diversity of content in hybrid [23] or automatic ways [240], to the best of our knowledge, no currently existing framework can reliably and cheaply obtain viewpoint labels that at least approximate the comprehensiveness of those typically handled in the communication sciences. Part I of this dissertation presents our work concerning this issue and proposes a novel viewpoint representation for human information interaction.

2.2.2. Automatic Viewpoint Labeling

Measuring search result viewpoint bias requires labeling each individual search result according to the viewpoint(s) it expresses. Depending on the selected viewpoint representation (see Section 2.2.1), researchers and practitioners can apply *stance detection* methods to automatically label search results for viewpoints. We discuss these and alternative methods below.

Stance Detection

Stance detection is the task of deriving a stance (e.g., *against*, *neutral*, or *in favor*) toward a claim (e.g., “students should have to wear school uniforms”) from text [366]. This implies that not just sentiment but also the *direction* of sentiment needs to be extracted. What makes stance detection challenging is that users may describe their stance in negative and positive ways. For example, both the following statements imply the same stance, but with different sentimental phrasing: *“I hate the terrible idea of school uniforms”* and *“It is fantastic that students do not wear school uniforms”*

Automatic stance detection is predominantly applied in a supervised, *target-specific* fashion; i.e., a text classifier is trained and evaluated on documents that all refer to a single topic or claim (often referred to as the *target*) [8]. For instance, previous work has built models to detect the stance on *atheism* or the *feminist movement* in tweets [75, 193, 199, 233]. Popular stance detection tasks, data sets, and models concern document types such as tweets [3, 70, 214, 233, 330, 342], microblogs [378], online debates [1, 242, 277, 332, 343], and news content [27, 106, 149, 214, 275]; featuring a wide range of topics and several different languages [8, 193, 317]. Due to the multiclass nature of stance detection (i.e., typically classifying documents into *against*, *neutral*, and *in favor*; although sometimes additional classes such as *other*/*unrelated* are added [141]), predictive performances are most commonly reported in terms of macro-f1 scores [193]. State-of-the-art target-specific stance detection models (e.g., applied to tweets and online forum posts) now regularly achieve macro-f1 scores ranging from .73 to .97 depending on document type and topic [127, 174, 287, 311]. Practical target-specific stance detection applications include handling rumors [60] and *fake news* [66, 140] related to specific topics on social media. However, web search interventions targeting the mitigation of undesired effects such as SEME require target-agnostic stance detection models to quickly respond to the large variety of debated topics users may search for.

Web search applications need to apply *cross-target* stance detection. In this variant, stance detection models are applied to data sets in which each document may refer to one of a variety of topics [8, 193]. Building models that can detect stances related to *any* topic in such a way usually leads to somewhat weaker predictive performances compared to target-specific models but makes stance detection applicable at scale. Macro-f1 scores for cross-target ternary stance detection (e.g., working with tweets or news articles) have ranged – again depending on document type – roughly from .450 to .750 [9, 11, 20, 142, 293, 376]. Although stance detection has thus far not been applied to openly available search result data, some data sets feature content similar to search results. The *Emergent* data set lends itself well to cross-target stance detection and is comparable to a search result data set: it contains a large number of news articles that have each been expert-annotated as *against*, *observing*, or *in favor* concerning one of 300 rumored claims [106]. Cross-topic stance detection models evaluated at the *Emergent* data set (and its follow-up version, the *2017 Fake News Challenge* data set [275]) have achieved macro-f1 scores of up to .756 [138, 306, 320]. We perform cross-target stance detection for search results in Chapter 3 (Part I) of this dissertation.

Alternative Methods to Extract More Than Mere Stance

Stance detection allows for considerable text comprehension concerning debated topics. To truly understand viewpoints on these topics, however, distilling the underlying reasons (i.e., perspectives) behind the different stances is essential. A technique that allows for more fine-grained opinion analysis is known as *argument mining*: here, arguments are automatically extracted from texts and subsequently divided into their different elements, e.g., claims, premises, and conclusions [200, 336, 362]. More recent work in the *argument retrieval* domain has built on this work by automatically extracting diverse arguments from document corpora [41, 42, 92, 260]. However, although argument mining and related approaches have started to get a finer grasp of distilling arguments from text, these methods are not yet able to classify arguments into different perspective categories or provide comprehensive debate summaries (e.g., extracting the most important *reasons* why people may oppose or support school uniforms).

One family of methods that could allow for better descriptions of perspectives compared to other approaches are *topic models* (e.g., *Latent Dirichlet Allocation*; LDA [38]). These (usually unsupervised) models aim to discover hidden structures in text corpora. By analyzing word co-occurrences across all documents in a corpus, they create a pre-specified number of *topics*. Each topic is a probability distribution over all words in the corpus. The probability density indicates how “typical” a given word is for the topic at hand. This way, topics can be described by their top-n highest-density words. Similarly, topic models also output per-document probability distributions over topics to indicate how “present” each topic is in a given document.

Joint topic models extend topic modeling (e.g., LDA) by adding components for more informative content extraction from text. For example, several joint topic models within opinion mining have proposed additional distributions or sentiment analysis features on top of LDA to extract more specific aspects. They include the *Topic-Aspect Model* (TAM) [262], the *Joint-Sentiment Topic model* (JST) [205], the *Viewpoint-Opinion Discovery Unified Model* (VODUM) [344], and the *Latent Argument Model* (LAM) [356]. Most joint topic models have not specifically been developed for the task of perspective discovery. However, their unsupervised nature and interpretable model output make all joint topic models mentioned above potential candidates in this respect. We perform and evaluate perspective discovery with joining topic models in Chapter 4 (Part I).

2.2.3. Crowdsourcing Viewpoint Annotations

To efficiently collect high-quality training data for stance detection models or annotate search results with viewpoint labels that cannot be automatically extracted using current methods (e.g., logics of evaluation; see Section 2.2.1), researchers may employ *crowd workers* [341]. Extensive experiments have been performed to crowdsource binary stance annotations (e.g., *against*/*in favor*) on news articles and tweets [66, 106, 204, 234, 330]. In addition, labels such as *neutral*, *neither in favor nor against*, or *I don't know* have been used to identify texts that do not take a stance, are unclear, unrelated, or ambiguous. Generally, the agreement percentages and the inter-rater reliability (IRR) values are substantial. For instance, Mohammad, Sobhani, and Kiritchenko [234] report an agreement percentage of 73% regarding the stance of the tweets, while Li et al. [204] report Krippendorff's α values of 0.60 and 0.81 when considering ternary and binary representations, respectively. Burscher et al. [51] used two trained annotators to identify pre-determined *frame types* (i.e., *conflict*, *morality*, *economic consequences*, and *human-interest*) in 156 political news articles. IRR c.f. Krippendorff's α ranged from 0.21 (*morality*) to 0.58 (*economic consequence*). Thus, human annotation of viewpoint labels is feasible, but its difficulty increases with label complexity. Furthermore, to the best of our knowledge, annotations for *logics of evaluation* have so far only been performed by experts in communication science and not yet by crowd annotators. We make a first attempt at crowdsourcing such annotations for logics of evaluation in Chapter 5 (Part I).

Early research by Snow et al. [329] showed that crowd workers can perform as well as domain experts in several natural language processing (NLP) tasks, such as event temporal ordering, word similarity, and affect recognition. However, collecting high-quality annotations from crowd workers is still challenging due to concerns posed by identifying, classifying, and counteracting crowd workers' biases and spamming behavior and patterns [68, 77]. Shah, Schwartz, and Hovy [324] name *label bias* as one of four core sources of bias in NLP models. Several criteria can influence the quality of crowdsourced annotations, including task and instructions clarity [113, 183, 374], task design [161], task difficulty [219], incentives [152], and quality control mechanisms [80, 94, 163, 225]. Moreover, there are qualitative differences between the ways in which previous work has aggregated crowdsourced data labels [263]. Recently developed annotation aggregation methods aim to improve overall label quality by weighting annotations according to different criteria such as individual annotator performance and biases [93, 94].

Following calls for making human-labeled data more reliable [117], several novel approaches have turned their attention to data documentation; i.e., by tackling issues such as reliability, transparency, and accountability in data collection practices. In the NLP field, Bender and Friedman [33] proposed *data statements*, a characterization for data sets that provides relevant details regarding the population involved in creating a given data set, how the data set is used in experimental work, and how potential biases in the data set might affect outcomes of the systems that are deployed with it. Gebru et al. [116] proposed *data sheets* for data sets, a companion document for data sets to exemplify the purpose and composition of the data set, who collected the data and how it was collected, as well as the intended use of the data set. Specifically for crowdsourcing annotations, Ramírez et al. [288] proposed a set of guidelines for reporting crowdsourcing experiments to better account for reproducibility purposes. Ramírez et al. [289] then followed up on this work by proposing a checklist that requesters can use to comprehensively report on their crowdsourced data sets. This body of research aligns with and facilitates current efforts towards more trustworthy artificial intelligence through better documentation [15, 382].

Data documentation approaches such as *data statements* [33] or *data sheets* [116] allow for a thorough assessment for many different types of potential biases (e.g., related to the distribution of crowd workers or the preprocessing of data). However, these methods usually do not consider the influence of *cognitive biases* on data collection.

Cognitive Biases in Crowdsourcing

Cognitive biases are general human tendencies toward irrational decision-making or deviation from norms under uncertainty [349]. For example, the *confirmation bias* is a tendency to specifically look for information that confirms one's preexisting beliefs [245]. Humans are especially vulnerable to cognitive biases when the cognitive demand of a situation exceeds their currently available cognitive resources [349]; e.g. when being confronted with too much or too little information to support a decision or when there is a need to act fast. This can also be the case in crowdsourcing tasks, where objectively "true" answers often do not exist [16].

Recent research has shown that different types of cognitive biases can negatively impact crowd workers' decision-making and thereby decrease the quality of crowdsourced data labels [96, 144, 156, 307]. For instance, this body of research shows that relevance judgments can be affected by displaying other crowd workers' judgments (i.e., *groupthink* or the *bandwagon effect*) or by revealing information on a single item in subsequent steps (i.e., the *anchoring effect* [96]). Other work demonstrated that crowd workers may be affected by their personal preexisting attitudes and stereotypes; e.g., when labeling images of faces [256] or when judging statements on debated topics (i.e., the *availability bias* and the *confirmation bias* [156]). Furthermore, Gadiraju et al. [111] found that crowd workers are often unaware of their actual level of competence, which may lead to *overconfidence*. Several strategies have been proposed to assess and mitigate cognitive biases in this context; e.g., by adapting the task design [29, 72, 96, 156].

Despite this empirical knowledge of cognitive biases and how to mitigate some of them in the crowdsourcing context, few crowdsourcing studies consider the influence of cognitive biases on data quality — why? Cognitive biases are a vast and complex space that may be hard to navigate for requesters. What is lacking is a practical tool that helps requesters to identify which specific cognitive biases may be problematic in a given task. In practice, such a tool could aid requesters in describing, assessing, and mitigating the influence of the identified potentially problematic cognitive biases. It would contribute to the body of existing efforts towards more reliable and reusable human-labeled data [49, 116]. That is why we propose a checklist to combat cognitive biases in crowdsourcing (see Chapter 6 in Part II).

2.2.4. Diversity, Fairness, and Viewpoint Biases in Search Results

Obtaining viewpoint labels (see Sections 2.2.1, 2.2.2, and 2.2.3) enables the examination of viewpoint distributions across ranked search result lists from popular web search engines. Recent research has found that such search results may not always be viewpoint-diverse, i.e., they can reflect *viewpoint biases* [114, 284, 370, 371]. Such viewpoint biases can root in the overall search result index but become amplified by biased queries and rankings [114, 292, 371]. They typically involve an over-representation of particular viewpoints among (high-ranking) search results. For example, a search result list on the topic *school uniforms* could be considered viewpoint-biased if most high-ranking (e.g., 40 out of the top 50) search results support the idea of school uniforms while most search results expressing different viewpoints can only be found at lower ranks. Previous work has commonly measured viewpoint biases using simplified and topic-specific approaches. For instance, Puschmann [284] examined biases in politics-related search results by categorizing them into *source types* (e.g., media, government, party-affiliated); finding that some parties have more power over their representation than others (e.g., by having more party-affiliated web pages among high ranking search results). White [370] labeled search results on the efficacy of medical treatment as containing either *false* or *correct* information. They found that users may be misled by false information among high-ranking search results, e.g., because search engines – irrespective of the truth – tend to rank positive content higher than negative content.

The aforementioned findings demonstrate that search results on debated topics may often be viewpoint-biased. However, measuring search result viewpoint bias at scale requires topic-independent guidelines and metrics. We describe related work on such more generalizable metrics below; in particular, existing approaches to measure search result diversity, fairness, and viewpoint bias specifically.

Measuring Search Result Diversity

Much work in the IR domain has been devoted to measuring (and improving) the diversity concerning query *subtopics* in search results [2, 4, 64, 309, 310]. Such search result diversity metrics assess the degree to which a ranked search result list satisfies different user intents. For example, this applies to search terms such as `jaguar` or `apple`, which should return results related to the animal or fruit as well as the respective company. These diversity methods usually reward both *diversification* (i.e., absence of bias) and *relevance* of search results across a ranking. Doing so, they aim to maximize the *utility* of search results. However, a trade-off with document relevance is not necessarily desired when measuring search result *viewpoint bias*. The ultimate aim here is not to maximize utility for the user but to provide search results that reflect a broad range of opinions on a given debated topic. Although existing search result diversity metrics inspire our work (see Part II), they are thus impractical for measuring viewpoint bias.

Search Result Fairness

Recent work has developed *fair ranking* methods to deal with *bias* on the web. These approaches aim to evaluate [19, 35, 71, 195, 315, 381] or increase [35, 55, 326, 391] the *fairness* toward protected attributes in ranked lists of documents. For instance, a ranking assessor may apply a ranking fairness metric to measure the degree to which *female* workers are systematically ranked lower than others in a list of job candidate profiles. Previously proposed ranking fairness metrics commonly presuppose that a *fair* or *unbiased* ranking is one with *statistical parity* [381]. In the ranking context, statistical parity holds when membership in a protected group (e.g., *female*) has no influence on a document's position in the ranking [381]. Ranking fairness metrics typically assess statistical parity by comparing the group membership distribution (i.e., the share of protected and non-protected documents) for different top portions of a ranking (e.g., 10, 20, ...) with the ranking's overall group membership distribution. Like most search result diversity metrics, most ranking fairness metrics include a discount function for rank-awareness [315]. We formalize several ranking fairness metrics and explore how to use them for measuring search result viewpoint bias in Chapter 8 (Part III).

Apart from evaluating the divergence from statistical parity, earlier research has approached ranking fairness in at least two other notable ways. One line of research has defined *criteria* that rankings have to fulfill to be considered *fair* [315, 391]. Whether a ranking fulfills these criteria is assessed using null hypothesis significance testing. Our aim, however, is to quantify the *degree* of viewpoint bias in search result rankings. Kulshrestha et al. [195] introduce a metric that quantifies *ranking bias* (RB) related to continuous attributes instead of group memberships. Their metric considers the mean of a continuous variable at each step of its computation. We use the metric proposed by Kulshrestha et al. [195] for measuring search result viewpoint bias in Chapter 9 (Part III).

Measuring Search Result Viewpoint Bias

Building on work that measured diversity or fairness in search results concerning more general subtopics (see previous subsections), recent research has begun to evaluate *viewpoint bias* in ranked outputs. Various metrics have been adapted from existing IR practices to quantitatively evaluate ranked lists against democratic notions of diversity [147, 360, 361], though only few [360] crucially incorporate users' attention drop over the ranks [21, 99, 168, 258]. Existing diversity, fairness, or viewpoint bias metrics, however, have a key limitation when measuring viewpoint diversity: they cannot accommodate comprehensive, multi-dimensional viewpoint representations. Incorporating such more comprehensive viewpoint labels is crucial because stances and the reasons behind them can otherwise not be considered simultaneously (see Section 2.2.1). That is why we introduce a novel metric to evaluate search result viewpoint bias in Chapter 9 (Part III).

2.2.5. Search Result Viewpoint Biases and User Behavior

Search result rankings strongly affect user preferences and behavior [133, 168, 171, 177, 248, 250, 258]. For example, users exhibit a *position bias* when exploring search results: they tend to pay more attention to results at higher ranks [258] and are more likely to click and examine them [168, 177, 250]. As a result, users usually do not even examine search results beyond the first result page [43].

More recent research has shown that this preference for consuming higher-ranked search results can affect the behavior of users with mild pre-search opinions [10, 99, 100, 274]. Specifically, these studies suggest that users tend to adopt whatever viewpoints are most prominent among the search results they consume. Viewpoint biases in search results (e.g., related to school uniforms) could thereby have systematic effects on user opinions and behavior (e.g., voting in favor of a school uniform mandate). This phenomenon – often referred to as the *search engine manipulation effect* (SEME) – can occur even after single search sessions, across topics (e.g., political elections, medical treatment, and vaccinations) [10, 99, 274, 371], and without users' awareness [123]. Given that real search result rankings are often biased concerning viewpoints (see Section 2.2.4), SEME is a pressing concern [26, 54]. We seek to better understand the underlying mechanisms of SEME and identify under what circumstances it occurs in Chapter 10 (Part IV).

2.2.6. Summary

We have discussed related research concerning viewpoint representation, viewpoint labeling procedures, viewpoint biases in search results, and the effects of search result viewpoint biases on users. Although considerable research efforts have been made in each of these directions, we have highlighted crucial limitations and research gaps that we aim to address in this dissertation. Specifically, we seek to develop comprehensive viewpoint representations that incorporate stances as well as logics of evaluation (Part I), combat cognitive biases in crowdsourcing viewpoint annotations (Part II), develop metrics to measure viewpoint bias in search results (Part III), and better understand how search result viewpoint biases affect user behavior (Part IV).

I

Representing Viewpoints

Measuring viewpoint biases in search results, studying their effects on user behavior, and supporting users in their search for debated topics requires a meaningful and scalable representation of viewpoints. For example, when examining whether users interact with viewpoint-biased search results related to *school uniforms*, each search result first needs a label that reflects what viewpoint it expresses regarding this topic. A fundamental part of research and practical applications in this space, therefore, is deciding how to represent viewpoints and then label search results accordingly. Earlier work in human information interaction has predominantly opted for binary or ternary stance labels (e.g., *against*/*neutral*/*in favor*) [121, 274, 385] as viewpoint representations. Such simple labels are feasible to obtain at scale (e.g., using crowdsourcing [233, 234] or automatic stance detection techniques [9, 294, 366]) and lend themselves to existing search result diversity [4, 64, 310] and ranking fairness methods [381, 391, 393]. However, binary and ternary stance labels are fairly generic categorizations of truly nuanced viewpoints as they do not incorporate any *degree* within or *reason* behind stances, e.g., they do not distinguish between *somewhat* or *strongly supporting* school uniforms and do not capture whether someone supports school uniforms for economic or functional reasons. These drawbacks greatly limit the insight practitioners and users can gain from such labels. Aside from being feasible to obtain and computationally tractable, viewpoint labels should capture nuanced differences between viewpoints in an explainable, human-understandable fashion. This first part of the dissertation addresses the issue of representing viewpoints by asking the following research question:

$\mathbf{RQ_I}$ What label taxonomy can accurately represent viewpoints on debated topics?

Part I consists of three chapters that each examine or propose different viewpoint representations. We begin in Chapter 3 with a user study investigating the feasibility of explainable, cross-topic stance detection for search results using current ternary stance labels (i.e., *against*/*neutral*/*in favor*). Our aim here is to gauge the potential and limitations of currently available automated methods. Specifically, we predict stance labels for search results, generate explanations for those predictions, and evaluate the explanations quantitatively and qualitatively with users. We find that, although some explanations help users interpret the behavior of (ternary) stance detection models, many users are unsatisfied with the quality and amount of viewpoint-related information they receive. Chapter 4 then introduces *perspectives* (i.e., reasons for opposing or supporting a debated topic) as an alternative or supplementary viewpoint representation format to the typical binary or ternary stance labels. We apply unsupervised *topic models* to debate forum entries and subsequently evaluate whether users can identify the topic model output as such perspectives. Our results show that some topic models can indeed discover user-identifiable perspectives. We thus conclude that considering perspectives, if generalized to be applicable across topics, as part of viewpoint representations could be feasible and useful. Building on this idea, we propose a comprehensive, topic-independent viewpoint representation for human information interaction in Chapter 5. This novel label consists of two dimensions: *stance* (i.e., a viewpoint's position regarding a debated topic, measured on a seven-point ordinal scale ranging from *strongly opposing* to *strongly supporting*) and *logic of evaluation* (i.e., the stance's underlying reasons or perspectives categorized into seven topic-independent categories).

3

Automatic and Explainable Labeling of Search Results With Ternary Stance Labels

---

This chapter is based on a published, full conference paper: Tim Draws, Karthikeyan Natesan Ramamurthy, Ioana Baldini, Amit Dhurandhar, Inkit Padhi, Benjamin Timmermans, and Nava Tintarev. "Explainable Cross-Topic Stance Detection for Search Results". In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. CHIIR '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 221–235. DOI: `10.1145/3576840.3578296`.

Tim Draws primarily planned and carried out the conceptualization, investigation, methodology, project administration, visualization, and write-up of the work described in the paper referenced above. His co-authors supervised him during the project and made edits to the writing.

Measuring viewpoint bias or studying user interactions in web search on debated topics requires labeling search results according to the viewpoints they express. Previous research has predominantly approached this by representing viewpoints using a ternary stance taxonomy, e.g., labeling each search result as either *against*, *neutral*, or *in favor* with respect to a debated topic [121, 274, 385](see Section 2.2.1). Despite not capturing nuanced differences between viewpoints (e.g., *against* could mean somewhat or strongly opposing a topic), ternary stance labels are a straightforward and arguably easy-to-understand way of categorizing content into viewpoint categories. They can also feasibly be obtained using automatic *stance detection* techniques [20, 193, 293, 306, 366]. Next to large-scale viewpoint bias evaluations and user behavior studies, automatically identifying the stances of search results in this way would allow for interventions that support users in navigating online debates (e.g., by displaying warning labels for viewpoint biases or labeling individual search results) [57, 100, 109, 298, 304, 379]. Moreover, stance predictions could be supplemented by explanations [166], which may help users decide whether they can trust the stance labels and identify stance-specific patterns.

A problem with automated methods for ternary stance labels is that their predictive accuracy in the context of search results is currently still unclear. Search results and the web pages they refer to are much more diverse (e.g., concerning text length and language) and less straightforward compared to the document types typically handled by automatic stance detection models (e.g., tweets or microblogs; see Section 2.2.2). Although stance detection has been applied to search results before [304], earlier work examining user interactions with debated topics has predominantly assigned stance labels via human annotations or proxy measures rather than stance detection [99, 385]. Applying explainable artificial intelligence (XAI) methods to search result stance predictions could help users and practitioners understand when and how stance detection methods fail in this context and support improvement efforts. However, despite large efforts toward developing such explainability methods [218], only a few works have focused on explaining stance detection specifically [166]. Text classification explanations are moreover not always easily interpretable by users [318], and it is currently not known what types of stance label explanations users would require. Thus, it is unclear how accurately stance detection models can predict stances expressed in search results and whether current XAI methods can produce helpful explanations for those predictions.

This chapter reports on a preregistered user study investigating whether and how explanations for automatic stance detection models can help users in their online information interactions. In particular, we aim to explore the possibilities and limitations of automated methods for ternary stance labels in the web search context. Two research questions guide our work:

$\text{RQ}_{\text{I.1}}$ Are current stance detection methods sufficiently explainable for users when applied to web search results?

$\text{RQ}_{\text{I.2}}$ What explanation visualization techniques can best explain stance detection for search results?

We address these research questions by first training and evaluating 10 different stance detection models (i.e., using classical machine learning and transformer-based language models) on a data set containing 1204 search results on 11 different debated topics (e.g., *school uniforms*; see Sections 3.2 and 3.3). Our evaluations show predictive performances comparable to the state of the art from several approaches, with *RoBERTa-base*, *BERT-base*, linear SVM, and logistic regression delivering some of the highest macro-f1 scores. We then investigate the explainability of these four models in a preregistered user study. Specifically, we ask users to *forward-simulate* model predictions, i.e., identify what the models have predicted based on explanations but without knowing the true or predicted labels (Section 3.4). We find that some explainability methods (e.g., LIME) can produce human-interpretable explanations for some stance detection methods (e.g., RoBERTa-base) most of the time and significantly more interpretable than randomly generated explanations. A qualitative analysis further reveals potential application areas, challenges, and improvements for such explanations. We discuss the implications and limitations of our findings in Section 3.6.

Supplementary material such as preregistrations, data sets, task screenshots, and code related to this chapter is available at `https://osf.io/fyvqu`.

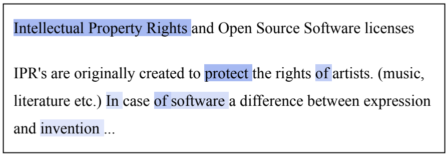
3.1. Hypotheses

Although many methods have been proposed to explain the behavior of *natural language processing* (NLP) models generally (i.e., from abstract global explanations such as *Submodular Pick LIME* [295] and *behavioral probes* [207] to local explanations such as *SHAP* [339], *SEA* [296], or *input reduction* [105]), user-focused solutions often involve explaining specific model predictions. How a particular model prediction came about can be explained in multiple ways, e.g., by adding influential examples [187, 282] or counterfactuals [305]. Jayaram and Allaway [166] recently proposed supplementing attention weights with crowdsourced human rationales to explain predictions of stance detection models. Arguably the most common and straightforward way to explain specific text classification predictions, however, is to produce *input feature explanations*. These explanations consist of token-wise importance attributes [218] that can be derived from XAI methods such as *LIME* [295], *integrated gradients* [340], or Grad-CAM [319].

Explanation quality can be measured in numerous ways, from *application-oriented* evaluations that focus on specific use cases (e.g., using human-annotated ground truth data sets) to *functionality-oriented* evaluations that inspect how well explanations reflect a model's technical process (i.e., often referred to as *faithfulness* or *fidelity*) [78, 79, 217, 218, 281]. A commonly chosen path when aiming to evaluate explanations directly with users whilst avoiding the cost of creating a ground truth data set is to conduct *human-oriented* evaluations. These evaluation tasks typically ask users to either choose the best of several models or perform *forward simulation*, i.e., to recreate model predictions based on explanations [79, 164, 218].[1] Despite some earlier work pointing to a general lack of interpretability among deep learning models [58, 105], it has been demonstrated

---

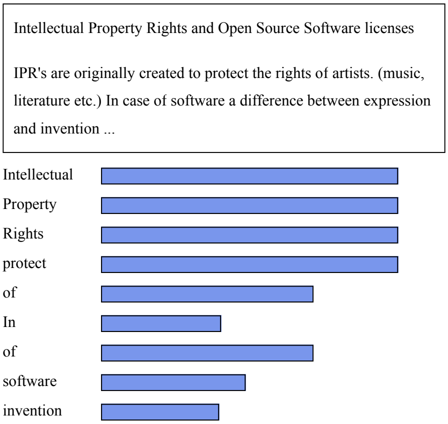
[1] A successful forward simulation occurs when, given an explanation but without knowing a prediction outcome, a user correctly identifies what the model has predicted (regardless of whether that model prediction was correct). For instance, suppose a stance detection model incorrectly classifies a search result whose title says "People who argue that school uniforms support academic performance are all wrong" as *in favor* of school uniforms. A user may successfully simulate this prediction by identifying that the model incorrectly predicted *in favor* when given an explanation highlighting the "support academic performance" passage.



Intellectual Property Rights and Open Source Software licenses

IPR's are originally created to protect the rights of artists. (music, literature etc.) In case of software a difference between expression and invention ...

Figure 3.1: Example of a salience-based explanation (using BERT-base and LIME) from our user study.



Intellectual Property Rights and Open Source Software licenses

IPR's are originally created to protect the rights of artists. (music, literature etc.) In case of software a difference between expression and invention ...

Figure 3.2: Example of a bar plot explanation (using BERT-base and LIME) from our user study.

that explanations can help users simulate the predictions of artificial intelligence (AI) systems [180, 278, 384]. In the NLP domain specifically, earlier work suggests that explanations help users to better understand models [145, 244]. Jayaram and Allaway [166] created explanations for stance detection models based on human-annotated rationales and found users deemed such explanations congruent with model predictions and sufficient. We expected that users would also be able to *simulate*[2] search result stance predictions when provided with automatically generated model explanations with greater accuracy than when provided with pseudo-explanations (i.e., a baseline that looks like a proper explanation but really only highlights words at random); hence we hypothesized:

H$_{\mathbf{I.1}}$ Users can simulate the predictions of stance detection models for search results with greater accuracy when provided with a model-specific explanation than a pseudo-explanation that highlights random words.

Input feature explanations are typically visualized using one of two techniques: as *salience-based explanations* that highlight words or tokens directly in the relevant document depending on their importance [69, 218, 318] (see Figure 3.1) or bar plots that indicate the token- or word-wise importance individually [318] (see Figure 3.2). Although salience-based explanations are often seen as an intuitive way to explain text classification models' predictions [69, 218], Schuff et al. [318] recently demonstrated that end users may find those explanations difficult to understand and less intuitive than bar plots. We thus expected that there would be a difference in simulatability for search results stance predictions depending on whether users see salience-based or bar plot explanations.

H$_{\mathbf{I.2}}$ Users' ability to simulate stance detection model's decisions differs depending on the way in which the explanation is visualized.

---

[2] By *simulate*, we here mean successfully performing forward simulations (see Footnote 1).

Table 3.1: The topic and stance distributions in our data set.

Topic	N	Stance Distribution Against – Neutral – In Favor
Zoos	48	50% – 6% – 44%
Bottled water	48	46% – 15% – 40%
Vegetarianism	45	38% – 31% – 31%
Homework benefits	45	47% – 18% – 36%
Obesity as a disease	48	33% – 25% – 42%
Milk health benefits	49	29% – 37% – 35%
Social networking sites	50	42% – 26% – 32%
Cell phone radiation safety	50	56% – 20% – 24%
Intellectual property rights	299	13% – 19% – 69%
School uniforms	276	28% – 29% – 43%
Atheism	246	22% – 46% – 32%
Total	1204	27% – 28% – 45%

3.2. Data

To train, test, and explain stance detection models, we assembled a data set containing search results on 11 debated topics (see Table 3.1). We obtained these data by combining three different data sets we had created as part of earlier work (see Rieger et al. [298] and Chapters 9 and 10). These previously created data sets included URLs, titles, snippets, and stance labels for a total of 1453 search results, which we had retrieved via API or web crawling from two popular web search engines. Stance labels had been assigned on seven-point Likert scales (i.e., ranging from -3 to 3 and thus including three degrees of opposing or supporting a topic) via crowdsourcing in two cases (i.e., taking the median annotation of at least three crowd workers with satisfactory inter-rater reliability; Krippendorff's $\alpha = \{.78, .79\}$; see Rieger et al. [298] and Chapter 10) or expert annotation in one case (i.e., mostly single annotations; Krippendorff's $\alpha = .90$; see Chapter 9). We mapped these seven-point stance labels into the three categories *against* $(-3, -2, -1)$, *neutral* (0), and *in favor* $(1, 2, 3)$ because automatic stance detection methods typically consider this ternary label taxonomy [193]. Using the provided URLs, we crawled the full web page text bodies (stripped of any HTML tags) for all search results. We here dropped 249 search results from the data as their text bodies could not be retrieved, leaving 1204 search results. Finally, we concatenated each search result's title, snippet, and text body (in this order) into single documents and removed all other information from the data aside from the documents' stance labels.

Table 3.1 shows the stance distribution per topic in our final data set. These 1204 annotated search results provide a ground truth for stance detection – both for evaluating classification performance (Section 3.3) and to inform a user study where participants forward simulate stance detection models' predictions based on provided explanations (Section 3.4).

3.3. Search Result Stance Detection

Explanations for stance detection models' predictions inevitably depend on the models' predictive performance. To ensure a realistic explanation pipeline in the context of search results, we first investigate the performance of current stance detection approaches and determine which methods may work particularly well here. This section thus describes the implementation and evaluation of 10 different stance detection models that we applied to our data (see Section 3.2). We measured the models' test set macro-accuracy, -precision, -recall, and -f1 scores across different model initializations and data splits, and compared their performance to the state of the art on other data sets (e.g., containing news articles or tweets). Finally, we selected four particularly well-performing models to generate explanations for.

Note that our core focus here was not to maximize predictive performance but instead to try a broad range of methods on this novel type of data (i.e., web search results). That is why we conducted only limited experimentation and hyperparameter tuning.

3.3.1. Stance Detection Models

We implemented two different types of models to perform stance detection on our search result data (see Section 3.2): *transformer-based language models* and *classical machine learning models*. Although transformer-based language models have recently dominated text classification and other NLP tasks [125], classical machine learning models such as logistic regression continue to demonstrate competitive predictive performances while remaining highly interpretable [229, 299]. It is thus relevant to investigate the performance-explainability trade-offs between these two model types.

Transformer-based Language Models

We implemented five pretrained language models, fine-tuning each of them on our search result data in 10 epochs and using a learning rate of 0.00003.[3] Each model considered the first 512 tokens per document (or 1024 tokens in the case of *Longformer*).

- **BERT**-base [74]: one of the most commonly used pretrained language models [146, 210, 312] and often used for stance detection [8, 141, 142, 173, 293, 311, 317].
- **DistilBERT**-base [312]: a light version of BERT that allows for much faster fine-tuning and inference, yet often with comparable predictive performance [312]. DistilBERT has been used for stance detection before [222] and also performed well on the related task of news classification [52].
- **RoBERTa**-base [210]: an improved version of BERT that has been trained for a longer time and on more data. RoBERTa has also often been used for stance detection [141, 320, 394].

---

[3] We tried different model types (e.g., base and large) and hyperparameter values but observed only marginal improvements beyond these settings.

- **DeBERTa**-base [146]: another improved version of BERT that focuses on disentangling attention mechanisms. Although DeBERTa has so far not been used for stance detection, it has been implemented for the related tasks of agreement detection in online debates [277] and fake news detection [325].
- **Longformer**-base [32]: an adaptation of RoBERTa to handle long texts and thus potentially better suited for search results and the (often long-form) web pages they refer to. Whereas all above models only considered their maximum of 512 tokens, our Longformer implementation considered the first 1024 tokens per document. Longformer has already been used for rumor stance detection on different kinds of social media posts [179].

Classical Machine Learning Models

We applied five classical machine learning models to a *tfidf* feature matrix we had created from a preprocessed version of our data set.[4] This matrix considered all unigrams with a document frequency between 0.005 and 0.8.[5]

- **Logistic regression**: an inherently interpretable model (i.e., coefficients reflect feature importance) that has often been used for stance detection in previous research [63, 142, 158, 193, 198, 348].
- **Linear support vector machine** (linear SVM): arguably the most common stance detection approach before the advent of transformer-based language models [76, 193, 198, 208, 209, 232, 261, 348, 373]. We used linear rather than kernel SVM because it performed slightly better during testing and is inherently interpretable.
- **Random forest**: a tree-based ensemble model that is often used for stance detection [193, 198, 208, 209, 348].
- **Gradient boosting**: another tree-based model commonly used for stance detection [193, 208, 348].
- **Naive Bayes**: a fully interpretable and highly simple machine learning model that has been used for stance detection in earlier work [193, 198, 209, 232] and lends itself to forming a baseline.

3.3.2. Evaluation

To enable a thorough and fair comparison between stance detection models, we used different random seeds to create 10 different 80-10-10 (i.e., train, validation, test) splits of our data. We then fine-tuned/trained each of the 10 models we consider (as described in Section 3.3.1)[6] a total of 100 times (i.e., 10 times using different random seeds that

---

[4] Aside from removing long (>127 characters) and stop words, this preprocessing involved lemmatization and stemming (all using the `nltk` library [212]).

[5] We decided to include only unigrams here as experiments wherein we included bi- and trigrams did not show improved model performances.

[6] For models that do not need validation data for training, we added the 10% validation data to the 80% training data, thus using 90% of the data for training in these cases.

Table 3.2: Mean test set performances (± standard error) of stance detection models over 100 trials (i.e., using 10 different seeds controlling any model randomness for each of 10 different data splits; best scores in each column are bold).

	Mean Macro-			
Model	**Accuracy**	**Precision**	**Recall**	**F1**
RoBERTa	**.770** (±.004)	**.652** (±.005)	**.641** (±.005)	**.640** (±.005)
DeBERTa	.757 (±.007)	.609 (±.016)	.617 (±.011)	.603 (±.014)
BERT	.741 (±.004)	.614 (±.005)	.598 (±.006)	.596 (±.006)
Linear SVM	.741 (±.002)	.604 (±.004)	.589 (±.003)	.590 (±.004)
DistilBERT	.737 (±.003)	.602 (±.005)	.591 (±.005)	.589 (±.005)
Longformer	.747 (±.005)	.598 (±.014)	.598 (±.009)	.587 (±.012)
Logistic Regr.	.719 (±.002)	.584 (±.004)	.542 (±.003)	.542 (±.003)
Random Forest	.687 (±.003)	.551 (±.005)	.477 (±.004)	.469 (±.005)
Grad. Boosting	.668 (±.002)	.569 (±.007)	.434 (±.003)	.405 (±.004)
Naive Bayes	.651 (±.003)	.520 (±.008)	.404 (±.003)	.360 (±.004)

control model randomness on each of the 10 different data splits).[7] Each time we had fine-tuned/trained a model, we produced predictions for the unseen test set and subsequently computed the macro-accuracy, -precision, -recall, and -f1 score for those test set predictions. Table 3.2 shows each model's performance averaged over the 100 trials.[8] To compare our results with previous research on stance detection (see Section 2.2.2), we focus on mean macro-f1 scores for the evaluation.

As expected, transformer-based language models (mean macro-f1 = [.586, .640]) performed considerably better than classical machine learning models (mean macro-f1 = [.360, .590]). Pairwise one-sided Wilcoxon signed-rank tests between models show that RoBERTa significantly outperformed all other models aside from DeBERTa (mean macro-f1 = .640; all $p_{\text{adj}} < 0.005$).[9] DeBERTa and Longformer both delivered strong predictive performances in most of the 100 trials but had their average scores greatly reduced by occasional bad runs (see Figure 3.3). This was especially surprising in the case of Longformer, as Longformer had twice as much training data available as the other transformer-based language models (i.e., the first 1024 instead of 512 tokens per document). Linear SVM delivered the best predictions among classical machine learning models, outperforming all other models of this type (mean macro-f1 = .590; all $p_{\text{adj}} < 0.005$).

Our macro-f1 scores ranging up to .640 are comparable to cross-target stance detec-

---

[7] For deterministic models such as naive Bayes or logistic regression, the 10 model initializations for any particular data split were identical.

[8] Due to an error in our evaluation code, we had initially computed f1 scores incorrectly. The (corrected) f1 values reported in this chapter thus slightly differ from Draws et al. [85], i.e., the published paper that this chapter is based on. Although the corrected f1 values are overall somewhat lower, the corrections did not affect any other results or change our conclusions.

[9] We Bonferroni-adjusted all p-values reported here to correct for multiple testing.

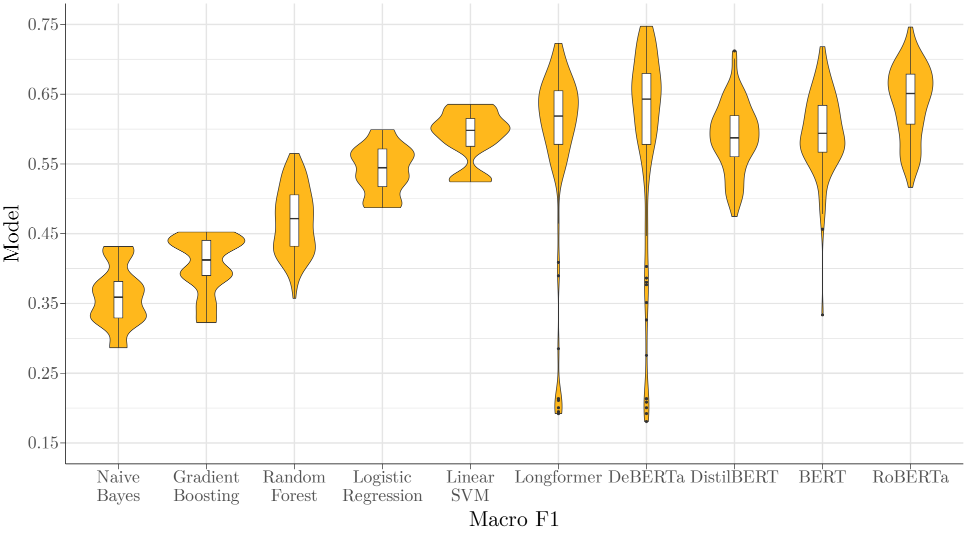


Figure 3.3: Distributions of macro-f1 scores across stance detection models (see also Table 3.2). Box plots (white) show medians and interquartile ranges, while violin plots (green) show how macro-f1 scores were distributed over the 100 runs.

tion conducted on similar (but much larger) data sets, where recent work has achieved macro-f1 scores ranging from around .450 to .750 (see Section 2.2.2). Moreover, the 8% performance increase from linear SVM to RoBERTa in our experiment aligns with earlier work that has found similar differences between classical machine learning models and transformer-based language models for cross-target stance detection [293].

3.4. User Study Setup

To investigate the explainability of stance detection models in the web search context (**RQ$_{\mathbf{I.1}}$** and **RQ$_{\mathbf{I.2}}$**), we applied several different XAI methods to four of the best-performing models we had implemented (see Section 3.3). Specifically, we here considered the two best-performing methods (i.e., in terms of mean macro-f1 score) from each of the two model types; that is, the top two transformer-based language models (i.e., **RoBERTa-base** and **BERT-base**; see Table 3.2)[10] and the top two classical machine learning models (i.e., **linear SVM** and **logistic regression**). The motivation here was to assemble a group of models that has strong overall predictive performance and represents a broad range of existing methods, yet is small enough to efficiently conduct a meaningful user study without too many different conditions. Furthermore, although we had trained and evaluated the models using 10 different data splits (see Section 3.3.2), we generated explanations for only one specific scenario, i.e., using the data split where the four selected models performed best overall (see Table 3.3). The aim here was to reduce the complexity of

---

[10] In our initial evaluation (which was also reported in the published paper, see Draws et al. [85] and Footnote 8), BERT-base was the second-best-performing model. Although DeBERTa turned out to perform slightly better than BERT after we corrected the macro-f1 values, BERT still showed a strong performance.

Table 3.3: Mean test set performances (± standard error; except for deterministic models) over 10 random seeds on the data split where the four selected models performed best.

	Mean Macro-			
Model	**Accuracy**	**Precision**	**Recall**	**F1**
RoBERTa-base	.793(±.004)	.697(±.003)	.676(±.009)	.677(±.009)
BERT-base	.771(±.011)	.654(±.015)	.643(±.015)	.645(±.016)
Linear SVM	.774	.648	.629	.635
Logistic Regr.	.730	.597	.550	.557

explanation evaluations while maintaining comparability between stance detection models. For the non-deterministic models RoBERTa and BERT, we chose their respective best-performing initializations on the selected data split. The remainder of this section describes how we created and visualized input feature explanations and evaluated their quality in a preregistered, online user study.

3.4.1. Materials

Input Feature Explanations

To enable an explainability comparison between the four stance detection models we selected (i.e., RoBERTa-base, BERT-base, logistic regression, and linear SVM), we created explanations for 20 test set documents (i.e., using the data split where these four models performed best overall) for which all four models made *the same stance prediction* (i.e., 10 correct and 10 incorrect predictions). This allowed us to directly compare the different explanations by looking at how many predictions users could successfully simulate. We obtained feature attributions for specific predictions from transformer-based language models by applying three different XAI methods (i.e., **integrated gradients** [340] and **Grad-CAM** [319]; both using *Captum* [188]; and **LIME** [295]). For the two classical machine learning models we considered (i.e., logistic regression and linear SVM), we obtained feature attributions from the **model coefficients** as these models are inherently interpretable. Moreover, to create a baseline, we also generated one set of **random feature attributions** for each document. Each of the 20 selected test set documents thus received a total of 3 (XAI methods) × 2 (transformer-based language models) + 2 (inherent coefficients of classical machine learning models) + 1 (random feature attributions) = 9 sets of feature attributions.

We mapped feature (token) attributions onto the original text by assigning each word the relevant token attribution (or 0 if there was none). To words that consisted of several tokens, we assigned the maximum attribution among the tokens it consisted of. We finally performed a min-max normalization on the word-wise attributions for each document to bring attributions from all methods to the same scale. This process resulted in nine sets of explanations indicating per-word importance for each of the 20 documents.

Explanation Visualization Techniques

Our aim was to visualize the nine different sets of input feature explanations per document in ways that are (1) intuitively understandable for users and (2) integratable into a search engine user interface. That is why we decided to consider not the full documents but only the title and snippet (thus only the top portion; see Section 3.2) of each document for the explanation visualizations, as this is what could be shown on a regular SERP. To further limit cognitive load for users and make methods better comparable, we set all negative feature attributions to 0. We created two different visualizations:

1. **Salience-based explanations over search results** (see Figure 3.1) highlighted words depending on their attributions. The darker the shade of a word highlight, the greater the word's importance in the model prediction. Words whose (normalized) attributions were below a threshold of 0.25 were not highlighted.
2. **Bar plot explanations below search results** (see Figure 3.2) visualized each word's attribution with a bar. The longer the bar next to a word, the greater the word's importance in the model prediction. Words whose (normalized) attributions were below a threshold of 0.25 were not listed in the bar plot.

3.4.2. Variables

Our study showed each participant the same set of 20 search results for which we had created explanations (see Section 3.4.1). However, participants saw different explanations for those search results depending on the conditions (i.e., explanation content and explanation visualization) they had been randomly assigned to. We evaluated participants' proportion of successful simulations and additionally measured several descriptive and exploratory variables.

Independent Variables

These variables were used to test our hypotheses $\mathbf{H_{I.1}}$ and $\mathbf{H_{I.2}}$ (see Section 3.1).

- **Explanation content** (between-subjects, categorical). Each participant saw explanations stemming from only one of the nine different stance detection model/XAI method combinations we considered (i.e., integrated gradients, GradCam, or LIME explanations from either of the two transformer-based language models, coefficients from either of the two classical machine learning models, or random explanations).
- **Explanation visualization** (between-subjects, categorical). Each participant saw explanation content visualized in one of two ways: either salience-based or as bar plots.

Dependent Variable

Both of our hypotheses $\mathbf{H_{I.1}}$ and $\mathbf{H_{I.2}}$ had the same dependent variable (see Section 3.1).

- **Simulation proportion** (continuous). We recorded the number of times each participant had correctly identified the stance detection models' predictions and divided that by the total number of documents (20).

Descriptive and Exploratory Variables

We used these measurements to describe our sample and for exploratory analyses, but we did not conduct any conclusive hypothesis tests on them.

- **Demographics** (categorical). We asked participants to state their gender, age group, and level of education from multiple choices. Each of these items included a "prefer not to say" option.
- **Attitudes** (ordinal). We recorded participants' attitudes on each of the debated topics mentioned in the 20 search results they saw (i.e., nine of the eleven topics in Table 3.1) by asking participants to indicate these attitudes on seven-point Likert scales ranging from "strongly disagree" to "strongly agree".
- **Simulation rationale** (open text). We asked participants to shortly describe their rationale behind each of the 20 simulations.
- **Simulation confidence** (continuous). Participants reported their confidence in each of their simulations on a seven-point Likert scale from "extremely unconfident" to "extremely confident".
- **Explanation quality perceptions** (ordinal). We asked participants to state on seven-point Likert scales the degrees to which they (1) understood what was expected of them in this task, (2) felt that the explanations helped them understand the AI system's decisions, and (3) believed that such explanations (if they have good quality) could make a useful feature in search engines.
- **Textual feedback** (open text). We asked participants to provide feedback on the explanations in three items:
  - "Who would benefit most from stance label explanations for search results? If you don't think such explanations are helpful to anyone, why not?"
  - "In what situations do you think users would benefit from such explanations?" (optional)
  - "What would need to change for such explanations to be (more) useful in web search?" (optional)

3.4.3. Procedure

Participants of our study went through three subsequent steps. First, participants stated their gender, age group, and level of education. We here also asked participants for their attitudes concerning each debated topic (see Section 3.4.1; including one attention check where we specifically instructed participants on what option to select from a Likert scale). Second, we randomly assigned participants to one of the nine **explanation content** conditions and one of the two **explanation visualization** conditions, gave them a task introduction, and then presented them – one by one – with the 20 search results. Each search result was accompanied by one of the nine different explanations displayed using one of the two visualization techniques depending on the conditions participants had been assigned to. Below each search result, we asked participants to (1) simulate the stance detection model's prediction, (2) describe their rationale behind the simulation, and (3) state their confidence in the simulation. Third, next to another attention check, we measured participants' perceived explanation quality in three different Likert scale items and asked them to provide textual feedback (see Section 3.4.2).

3.4.4. Participants

Before conducting the study, we had computed a required sample size of 290 using the software *G*Power* [104] for an ANOVA; specifying the default effect size of 0.25, a significance threshold of $\alpha = \frac{0.05}{2} = 0.025$ (i.e., due to testing multiple hypotheses), a desired power of 0.8, $(9 \times 2) = 18$ groups, and the respective degrees of freedom for the two hypothesis tests (regarding $\mathbf{H_{I.1}}$ and $\mathbf{H_{I.2}}$) we aimed to conduct. We eventually recruited 302 participants from *Prolific* (`https://prolific.co`), who were all above 18 years of age and had high proficiency in English (i.e., as reported by *Prolific*). The task was hosted on *Qualtrics* (`https://www.qualtrics.com`). Each participant was allowed to participate only once and rewarded \$5 for completing the study (i.e., equivalent to an hourly wage of \$11.26 considering the median completion time of 26:39 minutes). We excluded observations from 11 participants from data analysis because they had failed at least one of the attention checks in the task, thus leaving 291 observations to be statistically analyzed.

3.4.5. Statistical Analyses

To test our two hypotheses (see Section 3.1), we conducted an ANOVA with the two between-subjects-factors *explanation content* (to test $\mathbf{H_{I.1}}$) and *explanation visualization* (to test $\mathbf{H_{I.2}}$) as independent variables and *simulation proportion* as the dependent variable. Because we were testing two hypotheses as part of this study, we applied a Bonferroni correction to our significance threshold, reducing it to $\frac{0.05}{2} = 0.025$. We additionally conducted Tukey posthoc tests to analyze pairwise differences in case there was a main effect in the ANOVA (i.e., here thus adjusting our p-values automatically so that the significance threshold could remain at 0.05). Bayesian hypothesis tests[11] (e.g., to quantify evidence in favor of null hypotheses) and exploratory analyses (e.g., to note any unforeseen trends in the data) further helped us to better understand our results. Using *Atlas.ti* (`https://atlasti.com`), we finally conducted a *reflexive thematic* (qualitative) analysis [47] of the participants' textual answers to systematically dissect their feedback.

3.5. Results

This section describes the results of the user study we conducted to evaluate explanations for stance detection models in the web search context (see Section 3.4; $\mathbf{RQ_{I.1}}$ and $\mathbf{RQ_{I.2}}$). We report the results of our preregistered hypothesis tests as well as exploratory and qualitative analyses that may help interpret our findings.

---

[11] We denote Bayes Factors as BF_{10} or BF_{01} depending on whether they quantify evidence in favor of the alternative or the null hypothesis, respectively, and interpret them according to the guide proposed by Lee and Wagenmakers [201], who adapted it from Jeffreys [167].

3.5.1. Descriptive Statistics

Among the 291 recruited participants who passed both attention checks and were thus eligible for statistical analysis (see Section 3.4.4), 140 (48%) identified as female, 141 (49%) as male, and 9 (3%) as non-binary/third gender, while one participant (< 1%) preferred not to state their gender. Participants were rather young, with most (237; 81%) being under 35 years of age, although there were at least some participants from all age groups until 84 years. There was a diversity of education levels among participants, as only about half of them (146; 50%) had completed a university degree. While seven participants held a doctorate degree, six participants did not hold a high school diploma. Participants' attitudes on the nine debated topics present in the 20 search results they saw were reasonably balanced: across topics, there were always at least 5% who opposed and at least 20% who supported the topic. The average number of highlighted or listed words across *explanation content* conditions was 11.41 (SD = 3.62) and ranged from 8.10 (SD = 6.83, integrated gradients for RoBERTa) to 17.00 (SD = 5.01, random explanations).

Nearly all participants (270; 93%) stated that they understood what was expected from them in this task (i.e., by selecting "somewhat agree", "agree", or "strongly agree" for the relevant item). A majority of participants (216; 74%) at least somewhat agreed that the explanations helped them understand the stance detection model's predictions, with 57(20%) participants strongly agreeing and only 10(3%) participants strongly disagreeing here. Similarly, 217(75%) participants at least somewhat agreed that the explanations they saw (if they have good quality) could make a useful feature in search engines. Participants' overall mean simulation proportion across conditions was .54; slightly above a proportion of .50 that participants would have achieved had they always selected the true instead of (as instructed) the predicted stance label, as half of the shown explanations were for incorrect predictions (see Section 3.4.1). They reported a mean simulation confidence of 1.11 (i.e., on a scale ranging from -3/extremely unconfident to 3/extremely confident). Examining participants' simulation rationales indicated that participants indeed understood the task and were interpreting the explanations according to the highlighted or listed words (e.g., "*The word help could be a positive meaning for the AI*").

3.5.2. Hypothesis Tests

Figure 3.4 shows the mean simulation proportion per explanation content, split by explanation visualization technique. Whereas the difference between explanation types was significant ($\mathbf{H_{I.1}}$; $F = 25.615, p < .001, \eta_p^2 = .42$; see Section 3.4.5 for our analysis plan), the difference between explanation visualization techniques was not ($\mathbf{H_{I.2}}$; $F = .105, p = .746, \eta_p^2 < .01$). A Bayesian ANOVA further strengthened these findings, revealing extremely strong evidence for a difference between explanation types ($\mathbf{H_{I.1}}$; $\mathrm{BF}_{10} = 4.28 \times 10^{26}$) and moderate evidence for the null hypothesis that there is no difference between visualization techniques here ($\mathbf{H_{I.2}}$; $\mathrm{BF}_{01} = 6.36$).

Pairwise Tukey posthoc tests between explanation content conditions showed that five explanation types (i.e., coefficients for logistic regression and linear SVM, LIME for RoBERTa and BERT, and integrated gradients for BERT) led to significantly greater simulation proportions (M = [.576, .682], SE = [.019, .028]) than the random explanations (M = .452, SE = .019; $p_{\mathrm{adj}} = [< .001, .015]$). However, there were no significant differences among these five best-performing explanation types. We also found no significant differences

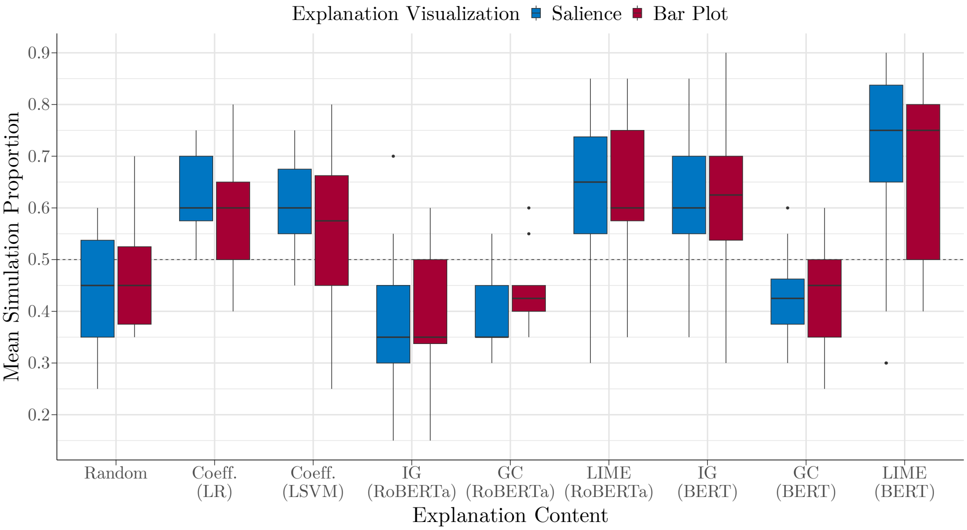


Figure 3.4: Mean simulation proportion per explanation content, split by explanation visualization (Coeff. = coefficients, LR = logistic regression, LSVM = linear SVM, IG = integrated gradients, GC = Grad-CAM). The dotted line reflects always selecting the true instead of (as instructed) the predicted stance label (i.e., 10 out of 20 explanations were for incorrect predictions).

between the remaining three explanation types (i.e., integrated gradients for RoBERTa and Grad-CAM for both RoBERTa and BERT; M = [.373, .424], SE = [.016, .022]) and the random explanations or each other. Our results thus suggest that explanations generated from logistic regression and Linear SVM coefficients, LIME for RoBERTa and BERT, and integrated gradients for BERT lead to greater simulation proportions among users than other methods or random explanations. Moreover, these five methods all led to median simulation proportions above 0.5 (see Figure 3.4), indicating that most participants who saw these explanations did better than if they had tried to predict the true stance labels themselves.

3.5.3. Exploratory Analyses

We conducted exploratory analyses in addition to the hypothesis tests described above to better understand our results. The aim of these additional analyses is to shed light on whether the differences in simulation proportion (see Section 3.5.2) are reflected in participants' subjective experiences (i.e., whether the explanations were indeed helpful for participants). Note that the analyses below were not preregistered as we conducted them after inspecting the data.

Simulation Proportion Regarding Correct Versus Incorrect Predictions

Although the set of 20 explanations we showed to participants included equal amounts of correct and incorrect predictions and many participants' simulation proportions were greater than if they had tried to predict stance labels themselves (see Sections 3.5.1 and 3.5.2), we conducted a separate analysis to test whether participants tended to assign the true instead of (as instructed) the predicted stance label. Had this been the case, participants' simulation proportions would be higher for correct than for incorrect model predictions. We thus performed a paired-samples t-test between participants' simulation proportions for the 10 correct versus the 10 incorrect predictions. Participants' mean simulation proportions were .535 and .542 for correct and incorrect predictions, respectively. This difference was not significant (delta = .007, $t = -0.525$, $p = .600$, $d = -0.03$), with a Bayesian t-test suggesting that participants simulation proportions for explanations of correct and incorrect predictions may be the same ($BF_{01} = 13.28$).

Relationship Between Simulation Confidence and Simulation Proportion

Our main analyses (see Section 3.5.2) measured explanation quality by participants' simulation proportions (i.e., reflecting the degree to which users can understand model predictions based on explanations), but that does not necessarily mean that participants *realized* when they correctly identified model predictions. To investigate whether participants grasped their ability to simulate model predictions, we looked at the relationship between participants' simulation proportions and their mean confidence (i.e., Likert scale items ranging from -3/extremely unconfident to 3/extremely confident; averaged over 20 items per participant). A Pearson correlation analysis revealed a significant association between these two variables ($r = .17$, $p = 0.003$), suggesting that participants were more confident in their simulations when they had stronger simulation proportions. Users thus may have a sense of their ability to make correct simulations; however, we note that this positive correlation was also rather weak. An ANOVA did not reveal any exploratory evidence for differences in participants' mean confidence across explanations ($F = 0.782$, $p = .619$, $\eta_p^2 = 0.02$) or explanation visualization techniques ($F = 1.462$, $p = .228$, $\eta_p^2 = 0.01$).

Differences in Explanation Quality Perceptions

Simulation proportion and confidence measure participants' ability to correctly simulate stance detection model predictions but do not necessarily speak to participants' *perceived* or *subjective* explanation quality. As with simulation confidence, we found exploratory evidence for a positive relationship between simulation proportion and the degree to which participants felt that the explanations *helped them to understand the model's predictions* ($r = .20$, $p < 0.001$). We did not find any evidence for differences between explanations or explanation visualization techniques regarding participants explanation quality perceptions, though. Given that participants' overall simulation confidence and perceived usefulness was rather high (see Section 3.5.1), participants across conditions may have felt that the explanations shown to them are useful even when they did not help them to successfully simulate model predictions. There was no sign of a relationship between simulation proportion and participants' perception that *explanations for search results could make a useful feature in search engines if they have a good quality.* Participants may have thus judged the general usefulness of such explanations independently from their experience in the task.

3.5.4. Qualitative Analyses

We conducted a qualitative, *reflected thematic analysis* [47] on participants' textual feedback to gain insights regarding where participants could see such explanations applied and what improvement suggestions they may have. To perform this analysis, one author generated response codes for participants' textual feedback in an inductive fashion and grouped them into code clusters. This resulted in the identification of **four web search scenarios** where stance label explanations could be especially helpful, **three user groups** who may particularly benefit from stance label explanations in search results, **two concerns** about such explanations, and **two ways** in which stance label explanations for search results could be **improved** according to our participants. We report on these themes below, indicating in brackets how many of our 291 participants mentioned a given theme.

Web Search Scenarios

A common theme among our participants was that explanations for search result stance labels could be used by those who intend to **research** debated topics, i.e., for school or university assignments (13), to prepare for a debate (9), to write an essay (3), or for academic work (29; e.g., "*to facilitate literature reviews*"). Participants also emphasized that stance label explanations for search results could help ordinary users in **forming opinions** by organizing the landscape of arguments on topics (26), enabling users to identify biased search results (3), and offering a diversity of viewpoints (18; e.g., "*I think that this would be a great tool for people to have the option to take a look contrasting perspectives about a subject.*") Related to this, participants believed that such explanations can lead users to **better understand** the topics or viewpoints they are searching about (8) and how search engines work (4; e.g., "[...] *why a result was given to them*"). Participants finally remarked that stance label explanations for search results deliver great **utility** by helping users to save time (46; e.g., "*it helps users to think quickly*") and teaching them how to search in a more targeted fashion (18; e.g., "[...] *a summary in that sense would make it easier to choose what you want to actually read and spend your time on*").

User Groups

Many participants thought that search result stance label explanations could help web search users in general (54; e.g., "*I think everyone that uses search engines would benefit from these explanations* [...]"). Additionally, participants identified three main user groups for whom stance label explanations may be particularly helpful: **neurodivergent users** who have trouble comprehending complex topics (14; e.g., "*those with learning difficulties*"), **researching users** such as students (33), teachers (5), academics (56), content creators (3), debaters (1), or journalists (6; e.g., "*Journalist or researchers who need to filter a lot of material*"), and **industry users and practitioners** who work directly with stance detection models (14; e.g., "*AI/ML model auditors*") or seek to inform business decisions (7; e.g., "*people who search for quick answers and information, advertising companies and generally the marketing section* [...]").

Concerns

Despite the largely positive feedback (see also Section 3.5.1), participants' answers contained two themes involving concerns surrounding stance label explanations for search results. The first aspect some participants found problematic was **bad explanation quality**; specifically, participants stated that explanations missed context (1), contained overwhelming amounts of information (2), sometimes highlighted wrong or misleading words (8; e.g., "*i cant see that we can be sure they are accurate based on AI decisions*"), or were just not useful in general (7; e.g., "[...] *they are difficult to understand*"). Although we gathered such feedback from all participants, i.e., including those who saw randomly generated explanations, these comments indicate that explanation quality may be a key concern for web search users. The second problematic aspect participants saw involved the explanations' **influence on users**: they believed that explanations could induce biased behavior in users by providing too much information and thereby discouraging critical thinking (22; e.g., [...] "*it should be up to the individual to make their own mind up rather than be pushed into believing what the author writes*"). Participants were particularly concerned about users' *confirmation bias*, i.e., that stance label explanations would lead more users to just consume content they already agree with (13; e.g., "[...] *If someone is trying to prove their point (whether it is in an everyday discussion, or in science), they could be biased in finding arguments for their point of view because they could easily filter for search results that suit their opinion*"). Concerned participants were distributed across conditions; that is, we did not observe any qualitative differences regarding participants' concerns between explanation content or visualization conditions.

Improvement Suggestions

Partly in line with their concerns surrounding stance label explanations for search results, participants described two main improvement suggestion themes. One of these was rather straightforward: explanations should have **better quality**, i.e., predictions should be highly accurate and explanations should be more consistent in highlighting key terms (20; e.g., "*accuracy must be top notch*" or "*improve the keywords chosen by the AI*"), explanations should highlight words in a smart fashion (4; e.g., "*omit repeating words*" or "*Maybe linking words together* [...]"), stop words and other neutral terms should be ignored (9; e.g., "*Cut out generic words like, the and it etc.*"), and explanations should be simpler and clearer in general (7; e.g., "*just a quick guide, don't get too bogged down in details*"). Some participants, on the other hand, wished for **more extensive explanations**, i.e., supplementing search result stance label explanations with a clear labeling system or description for what makes a stance on the topic at hand (2), confidence scores for stance label predictions (2), more context (4; e.g., "samples could have been a little longer"), or just more information in general (11; e.g., "*Examples of how it works, decisions that were made based on the algorithm*"). We again observed no differences regarding improvement suggestions between conditions. As previous research has pointed out [166], a key issue for the future development of stance label explanations for search results thus seems to be trading off simplicity and clarity with providing information that is extensive enough for users to fully comprehend the stance label predictions.

3.6. Discussion

This chapter has presented a preregistered user study investigating the quality of stance label explanations for web search results. We first applied 10 different stance detection models to search result data and found that several transformer-based language models (e.g., RoBERTa and BERT) significantly outperformed classical machine learning models (e.g., linear SVM and logistic regression) in terms of predictive quality (Section 3.3.2). Subsequently, we asked 291 participants in a user study to *forward-simulate* 20 different stance detection model predictions given different kinds of explanations, i.e., to identify – based on the explanation but without knowing the true or predicted labels – what the model had predicted in these 20 cases (Section 3.4). Our results showed differences between explanation types regarding participants' proportions of correctly simulated predictions ($\mathbf{RQ_{I.1}}$; Section 3.5): several XAI methods (i.e., coefficients from inherently interpretable models, LIME for transformer-based language models, and integrated gradients for BERT) led to significantly higher simulation proportions than other methods or randomly generated explanations. However, we found no evidence for any differences among these best-performing explanations or between explanation visualization techniques ($\mathbf{RQ_{I.2}}$). The remainder of this section pairs these findings with results from our exploratory and qualitative analyses to paint a comprehensive picture of how web search engines could implement stance label explanations to assist their users in navigating debated topics in search results.

3.6.1. Implications and Recommendations

Can stance labels for search results be *sufficiently* explained using current methods?
Most participants in our user study felt that the explanations helped them understand stance detection model predictions and that such explanations could make a useful feature in web search (Section 3.5.1). Our hypothesis tests confirm that explanations from at least some XAI methods can lead users to better understand model predictions than randomly generated explanations (Section 3.5.2). Moreover, participants' simulation proportions were positively related to their simulation confidence ratings and feelings that the explanation helps them understand model predictions (Section 3.5.3). This suggests that simulation proportion may be a good proxy for explanation quality in the user's eye. Our qualitative analyses underline the potential usefulness of stance label explanations for search results as participants could imagine a range of potential application areas and user groups who may particularly benefit from such explanations (Section 3.5.4). Given the stronger predictive performance of transformer-based language models and no apparent explainability differences between stance detection model types in this context, models such as RoBERTa and BERT, coupled with XAI methods such as LIME, may be prime candidates for this endeavor. However, participants also pointed to weaknesses and concerns surrounding search result stance label explanations that need to be dealt with for these explanations to be truly useful. Especially extending the amount of viewpoint-related information (e.g., going beyond the ternary *against*/*neutral*/*in favor* taxonomy) may be worth exploring in this context.

What would explainable stance labels for search results ideally look like?
None of our analyses (including a null hypothesis significance test; see Section 3.5.2) point to any difference in simulation proportion, explanation quality, or preference between the two explanation visualization techniques we had implemented (i.e., salience-based and bar plot explanations). Although our between-subjects user study design meant that we could not show both explanation visualizations to participants for direct comparisons

and related research suggested otherwise [318], our findings incline us to assume that there is indeed no difference between these two methods in the web search context. Salience-based explanation visualizations over the search results may, however, still be the better option in this case as they do not require any additional space on the SERP.

Our qualitative analyses send at least two clear messages regarding the future development and implementation of search results (Section 3.5.4). First, explanations have to be of high quality, i.e., highlight key terms and relate them to each other while ignoring irrelevant terms such as stop words. The number of words that were highlighted in an explanation did not seem to matter to participants as two of the worst-performing explanation types featured the least and most highlighted words on average, respectively (see Section 3.5.1); indicating that users care primarily about the *quality* of word highlights. This not only means that predictive model performance has to be high but also that the explanation content (i.e., the word attributions) has to clearly describe the model's reasoning in a human-like way [166, 383]. Second, there is a concern that stance label explanations negatively influence user behavior and thereby contribute toward the fragmentation of society. Such concerns could be alleviated by supplementing explanations with more nuanced viewpoint labels [25] or information about stance detection and XAI methods, (cognitive) biases in web search [21], and the topic at hand [215, 379]. Nevertheless, these solutions have to be rigorously evaluated before implementation to ensure that they do not do more harm than good [395, 396].

3.6.2. Limitations

We acknowledge that our work is limited in several important ways. First, in line with most previous work on stance detection (see Section 2.2.2), we have considered a simple, ternary taxonomy for stance classification (i.e., *against*, *neutral*, *in favor*). Recent work has represented stances in more comprehensive ways (e.g., on continuous [195] or ordinal scales; see Chapters 5, 6, 8, 9, and 10). Second, our participant sample (see Sections 3.4.4 and 3.5.1) may not be representative of internet users in general. Users may respond differently to search results depending on beliefs and common practices in their countries or cultures – in particular concerning debated topics such as school uniforms. Third, we exposed users to individual search results and explanations rather than a more realistic scenario (e.g., with top 10 search result pages). This may have led users to judge search results differently than they normally would. Fourth, for consistency, the topics in our data were all based on claims formulated in a positive direction (e.g., *in favor* on *vegetarianism* meant *supporting* the idea that one should be vegetarian; see Table 3.1 and Section 3.2). Users may get confused if they conceptualize topics in other ways (e.g., "vegetarianism is unhealthy") and find that stance labels do not match their preconceived notions (e.g., *in favor* suggesting that vegetarianism is healthy). Fifth, we only looked at two different explanation visualization techniques (i.e., salience-based and bar plot explanations), while other existing or novel explanation styles, such as more complex textual explanations, may also be suitable. Sixth, although our search results came from different search engines and featured 11 topics, we did not have much data at hand, and search results had been annotated in part by experts and in part by crowd workers, which may reduce the overall data quality and coherence (see Section 3.2).

Aside from the above points, it is worth noting that modern generative artificial intelligence (GenAI) systems such as ChatGPT [290] were not yet available when we conducted the study presented in this chapter. These systems (and the large language models that underlie them) have recently caused a massive shift in natural language processing and generation [45]. In the context of web search on debated topics, GenAI systems may be superior to the models we tested in terms of both detecting and explaining stances of search results. We believe that using GenAI to help users navigate debated topics in web search is an interesting and promising direction for future work.

3.7. Conclusion

Efforts toward more reliable, bias-free, and trustworthy interactions with debated topics for web search users would greatly benefit from automatic and explainable cross-topic stance detection methods. Such methods could support users in navigating debated topics online by organizing search results into viewpoint categories in a human-understandable fashion. In this chapter, we have presented a preregistered user study investigating the feasibility and ideal implementation of search result stance label explanations based on a ternary stance label taxonomy. Our findings suggest that automatic stance detection for search results is possible and show that at least some explainability methods can help users, e.g., in deciding when stance detection methods are wrong and identifying stance-specific patterns. Qualitative analyses revealed potential web search scenarios (e.g., search as opinion formation) and user groups (e.g., neurodivergent people and researchers) where such explanations could be particularly helpful. However, they also uncovered important user concerns and improvement suggestions, e.g., many users wanted more extensive explanations of the viewpoints they engaged with. Including more nuance and information regarding the viewpoints expressed in search results may facilitate stance detection and explanation efforts in the web search context. In the following Chapters 4 and 5, we explore more comprehensive viewpoint labels compared to the ternary taxonomy we considered here.

4

Helping Users Discover Perspectives: Enhancing Stance Detection With Joint Topic Models

---

This chapter is based on a published workshop paper: Tim Draws, Jody Liu, and Nava Tintarev. “Helping Users Discover Perspectives: Enhancing Opinion Mining with Joint Topic Models”. In: *2020 International Conference on Data Mining Workshops (ICDMW)*. Sorrento, Italy: IEEE, Nov. 2020, pp. 23–30. DOI: `10.1109/ICDMW51313.2020.00013`.

Tim Draws and Jody Liu collaborated in planning the conceptualization, investigation, and methodology of the work described in the paper referenced above. The implementation was primarily carried out by Jody Liu, under the supervision of Tim Draws and Nava Tintarev. Tim Draws administered the project, guided the research design, data collection, and data analysis, and contributed most of the writing.

Previous human information interaction research has commonly represented viewpoints expressed in search results using ternary stance labels (e.g., *against*/*neutral*/*in favor*; see Section 2.2.1). As we have shown in Chapter 3, such simple labels are feasible to obtain and explainable but offer limited information according to users. Adding *perspectives* (i.e., underlying reasons or arguments) could be a way to enrich existing viewpoint representations. For example, a commonly debated claim is “abortion should be legal”. To express a viewpoint concerning this claim primarily means to take a *stance* (i.e., supporting or opposing the legalization of abortion). However, the same stance can be supported by different perspectives [356], e.g., someone supporting the legalization of abortion could take the perspective that “...reproductive choice empowers women by giving them control over their own bodies” or instead that “... personhood begins after a fetus is able to survive outside the womb or after birth, not at conception.”

In contrast to stances, which can be automatically identified in text using stance detection techniques (see Section 2.2.2 and Chapter 3), perspectives are much harder to distill. Specifically, such *perspective discovery* is challenging due to the unstructured nature of perspectives and the debated topics they concern, e.g., topics such as *abortion* and *school uniforms* have vastly different argumentative spaces that may change over time depending on ongoing socio-political developments. High-quality, topic-specific perspective labels are thus hard and expensive to obtain and maintain at scale (e.g., using crowdsourcing), which makes supervised learning (as used in many stance detection techniques) infeasible for perspective discovery.

A family of unsupervised methods that could potentially perform perspective discovery is *topic models*. Topic models aim to find hidden patterns in unstructured corpora of textual documents. Part of the output of a topic model is a pre-defined number of probability distributions (i.e., topics) over all words in the corpus it has been applied to. In practice, topics can then be described by selecting a number of words (e.g., 10) based on the highest probability density in each topic distribution. When applying a topic model to a corpus of opinionated documents that all relate to the same controversial issue, these topics could be seen as perspectives (i.e., rather than content subjects). An example of an output perspective related to abortion could be *{woman, choice, body, fetus, control, pregnant, birth, baby, fetus, sex}*. Especially promising in this respect are so-called *joint topic models*, which add additional components (e.g., some form of sentiment analysis) to the classical topic modeling approach.

Several joint topic models have been used for tasks such as sentiment analysis and stance detection (see Section 2.2.2), and their ability to compute topics informed by constructs such as sentiment also makes them promising candidates for automatic perspective discovery. However, to the best of our knowledge, whether joint topic models can indeed perform this task, i.e., output topic keywords that human users can identify as perspectives concerning a debated topic, has not been evaluated yet. It is moreover unclear whether the confirmation bias, where pre-existing opinions skew how humans interpret information [245], could impede users’ understanding of topic model output as perspectives. Knowing whether (and which) topic models can perform perspective discovery and what other factors may get in the way of users successfully recognizing topics as perspectives would contribute toward more comprehensive automatic viewpoint detection methods. We study whether joint topic models can distill such human-identifiable

Table 4.1: Abortion perspectives in the final data set. Whereas the perspectives p_1, p_2, and p_3 support the legalization of abortion, p_4, p_5, and p_6 oppose it.

Persp.	**Description**
p_1	Reproductive choice empowers women by giving them control over their own bodies.
p_2	Personhood begins after a fetus becomes 'viable' (able to survive outside the womb) or after birth, not at conception.
p_3	A baby should not come into the world unwanted.
p_4	Abortion is murder, because unborn babies are human beings with a right to life.
p_5	Abortion is the killing of a human being, which defies the word of God.
p_6	If women become pregnant, they should accept the responsibility that comes with producing a child.

perspectives and whether the confirmation bias can impede user interpretation of topic model output in this context. Two research questions guide our work:

$\mathrm{RQ}_{\mathrm{I.3}}$ Can joint topic models support users in discovering perspectives in a corpus of opinionated documents?

$\mathrm{RQ}_{\mathrm{I.4}}$ Do users interpret the output of joint topic models in line with their personal pre-existing stance?

To address these research questions, we created a data set from debate forum entries on the topic *abortion legalization* and assigned perspective labels to the forum posts via an expert annotation procedure. We applied several different topic models to this data set. In a user study, we then evaluated whether these topic models are effective in helping people to identify perspectives that exist in the data. We find that at least one joint topic model (i.e., the *Topic-Aspect Model*; TAM [262]) can help users distill perspectives from text. Our results furthermore contain no evidence for a tendency of users to interpret topic model output in line with their personal pre-existing attitude.

All material related to this research (e.g., annotated data set, code, and results) is openly available at `https://osf.io/uns63`.

4.1. Data

For this study, we created a perspective-annotated data set consisting of debate forum entries on the topic *abortion*. The data set is openly available on our repository.

4.1.1. Creating an Annotated Data Set

We retrieved a total of 2934 opinionated documents on the topic *abortion* from an online debate platform.[1] On this platform, users can participate in openly held debates by posting their opinions in either the supporting or opposing category.

---

[1] `https://debate.org`, retrieved May 2020

Each document in our data set was assessed by a human annotator to (1) ensure that all documents are written in English, (2) remove ambiguous documents (such as spam and unclear stance position), and (3) assign a *perspective label* to each document. These perspective labels were taken from the website *ProCon*.[2] *ProCon* provides a list of 31 perspectives that exist in the abortion debate (i.e., categorized into *Pro* and *Con*). In the annotation process, it became clear that two perspectives listed at *ProCon* were difficult to distinguish. We therefore merged these two perspectives into one.[3]

We controlled the annotation quality by having a randomly selected 10% documents annotated by another, independent annotator. The results of this quality control suggested that the main annotator was reliable (Krippendorff's $\alpha = 0.81$).[4]

4.1.2. Curating a Balanced Data Set

For our user study, we aimed to curate a data set balanced in terms of stances and perspectives. To create this final data set, we picked documents from the raw annotated data to include (1) an equal amount of supporting as well as opposing documents, and (2) an equal amount of documents across six selected perspectives. We selected these six perspectives (i.e., three supporting and three opposing the legalization of abortion; see Table 4.3) because they were the most commonly occurring perspectives in the data.

We created the final data set by randomly picking 100 documents from each of the six perspectives listed above. Here we only considered documents that had *uniquely* been annotated with the perspective at hand; thus excluding documents that expressed several different perspectives at once. This resulted in a corpus of 600 documents that was balanced in terms of stances and perspectives. To prepare the final data set for topic modeling, we applied several pre-processing steps. First, we removed any contractions, punctuation, and digits. Second, we lowercased the text and removed stop words. Third, we applied a spelling checker and performed lemmatization. Fourth, we replaced words preceded by "not", "no", "never", or "none" with their antonyms, removed non-sentiment words that do not appear in the subjectivity lexicon *SentiWordNet* [22], and added bigrams and trigrams.

4.2. Method

We applied six different models (i.e., four joint topic models and two baseline models; see Table 4.2) to the data set containing 600 perspective-annotated documents (see Section 4.1) and showed parts of the output to participants in a user study. Using sets of keywords, participants had to identify the six correct perspectives that are present in the data. Specifically, participants saw the top ten keywords for each of the six topics that the model at hand had computed.[5]

---

[2] `https://abortion.procon.org`, retrieved May 2020

[3] We formulated this merged perspective as *Abortion is murder, because unborn babies are human beings with a right to life.* (see Table 4.3).

[4] For the annotation reliability metric Krippendorff's α, a score of 0.8 or higher is desired [390].

[5] It is common practice to represent the output of topic models by the top ten keywords. Accordingly, for our study, we decided that ten words should be enough for participants to understand what the topic is about, but at the same time not too much so that participants are not overwhelmed.

Table 4.2: Models used in the user study.

Model	Description	Implementation
TF-IDF	A baseline model created by randomly distributing generally important words from the corpus over six groups.	`Sklearn` [264]
LDA	A baseline topic model that computes bag-of-words topics to describe themes in text.	Blei, Ng, and Jordan [38]; `Gensim` [291]
TAM	Joint topic model that performs LDA and adds additional distributions and processes to group tokens into *background*, *topic-specific*, and *perspective-specific* tokens.	Paul and Girju [262]
JST	Joint topic model that performs LDA and groups tokens according to a subjectivity lexicon.	Lin and He [205]
VODUM	Joint topic model that performs LDA and groups tokens according to POS-tags.	Thonet et al. [344]
LAM	Joint topic model that performs LDA and groups tokens according to a subjectivity lexicon and POS-tags.	Vilares and He [356]

4.2.1. Models

We evaluated four different joint topic models in terms of their ability to help users discover perspectives in corpora of opinionated documents. These joint topic models were the *Topic-Aspect Model* (TAM) [262], the *Joint-Sentiment Topic model* (JST) [205], the *Viewpoint-Opinion Discovery Unified Model* (VODUM) [344], and the *Latent Argument Model* (LAM) [356] (see Section 2.2.2). Each performs *Latent Dirichlet Allocation* (LDA) [38] and adds an additional component where tokens are grouped in a particular way (see Table 4.2).

To compare the joint topic models to a baseline, we evaluated two additional models (see Table 4.2). First, we added a regular topic model (i.e., LDA) to test the impact of the components that the joint topic models add on top of LDA. Second, we created a model whose output merely *resembled* that of a topic model by randomly distributing the top 60 words in the corpus (according to *term frequency-inverse document frequency*; TF-IDF) over 6 sets. The purpose of this TF-IDF model was to create a “control condition” in which the presented output consists of incoherent groups of words that can still vaguely be associated with the topic *abortion*.

Aside from the TF-IDF model, all models were computed using the original approach and code proposed by their respective authors. In terms of their core topic modeling functionality, each model used similar hyperparameter values to those with which topic models are typically configured [131, 285]. The hyperparameter values were: 1000 iterations, $\beta = 0.01$, number of topics $k = 6$ (i.e., to reflect six different perspectives), and $\alpha = 50/k$.

4.2.2. Operationalization

To compare the models introduced above and investigate the research questions $\mathbf{RQ_{I.3}}$ and $\mathbf{RQ_{I.4}}$, we conducted an online between-subjects user study. We measured the following variables:

Independent Variable

- *Model.* Each participant saw the output of one of six different models they had randomly been assigned to (see Table 4.2 for a model overview).

Dependent Variables

- *Number of correct perspectives found* (*nCor*). This variable measured how many of the six perspectives that truly exist in the corpus were found by participants based on the model output they saw. It could take on seven different values (i.e., integers ranging from 0 to 6).
- *Number of opposing perspectives selected* (*nOpp*). This variable measured the selected number of perspectives that oppose abortion. Similar to *nCor*, it could take on seven different values (i.e., integers ranging from 0 to 6).[6]

Individual Differences

We measured several variables that reflected individual differences among participants. These variables were later used to get a better idea of the sample and (in part) to answer $\mathbf{RQ_{I.4}}$.

- *Gender.* Selectable from multiple choices.
- *Age.* Selectable by using a slider.
- *Pre-existing stance.* Participants responded to the item "*In my opinion, abortion should be legal*" by selecting the appropriate option from a five-point Likert scale ranging from "strongly disagree" to "strongly agree."[7]
- *Pre-existing knowledge.* Participants responded to the item "*I have good knowledge about the abortion debate*" by selecting the appropriate option from a five-point Likert scale ranging from "strongly disagree" to "strongly agree."

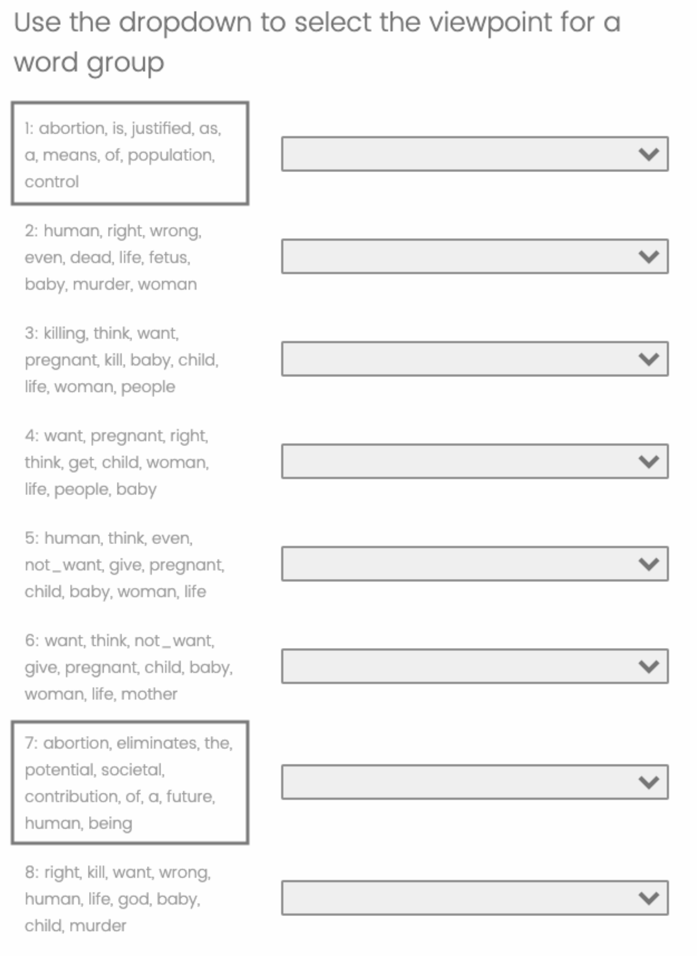
Exploratory Measurements

We used three additional items to measure the overall user experience with the task and to understand the possible potential a topic model has for a user. Participants could respond to each item by selecting the appropriate option from a five-point Likert scale ranging from "strongly disagree" to "strongly agree." The results from these items were used for exploratory analyses.

---

[6] Here, we excluded topics that were used as attention checks. We do not compute the number of supporting perspectives selected due to symmetry.

[7] Additionally, participants had the option to select an "I don't know" option. This option was also available for pre-existing knowledge.



Use the dropdown to select the viewpoint for a word group

1: abortion, is, justified, as, a, means, of, population, control

2: human, right, wrong, even, dead, life, fetus, baby, murder, woman

3: killing, think, want, pregnant, kill, baby, child, life, woman, people

4: want, pregnant, right, think, get, child, woman, life, people, baby

5: human, think, even, not_want, give, pregnant, child, baby, woman, life

6: want, think, not_want, give, pregnant, child, baby, woman, life, mother

7: abortion, eliminates, the, potential, societal, contribution, of, a, future, human, being

8: right, kill, want, wrong, human, life, god, baby, child, murder

Figure 4.1: Screenshot of the main task. Word groups 1 and 7 are the two honeypot topics.

- *Perceived usefulness.* To measure the general perceived usefulness of a model that can perform perspective discovery, participants responded to the item "*A model that can automatically show all viewpoints is useful to quickly understand a debate.*"

- *Perceived awareness increase.* We measured whether participants experienced an increased awareness of the different perspectives related to *abortion* by asking them to respond to the item "*I'm now better aware of the possible viewpoints than before.*"

- *Confidence in task performance.* To measure participants' confidence in whether the model helped them make the right choices, participants responded to the item "*I'm confident that I've correctly assigned the viewpoints to the word groups.*"

4.2.3. Procedure

Our study consisted of an online task that we set up using the platform *Qualtrics* (`https://qualtrics.com`). Participants then went through three subsequent steps:

Step 1. Participants stated their age, gender, as well as pre-existing stance and knowledge related to *abortion*.

Step 2. Participants did the main task. We randomly assigned each participant to one of the six models we aimed to test. After reading an introduction, participants were shown a list of 16 different perspectives. This list of 16 perspectives contained the six perspectives that were part of the corpus and ten other abortion perspectives taken from *ProCon* (see Section 4.1). Below the list of perspectives was the output of the model participants had been assigned. This output consisted of six "topics" represented by ten keywords (see Section 4.2.1). Additionally, we mixed two *honeypot topics* into the output. Each of these honeypot topics was a set of keywords that matched one of the 16 perspectives word for word. Participants were instructed to match each set of keywords with one of the 16 abortion perspectives by selecting it from a drop-down menu (see Figure 4.1).

Step 3. We assessed participants' experience with the task. Specifically, we measured *perceived model usefulness*, *perceived awareness increase*, and *confidence in task performance*. Additionally, participants were given the option to provide feedback using an open-text field.

4.2.4. Hypotheses

Given our two research questions $\mathbf{RQ_{I.3}}$ and $\mathbf{RQ_{I.4}}$ as well as the operationalization and study procedure described above, we defined two hypotheses:

- $\mathbf{H_{I.3}}$ Users find more correct perspectives when being exposed to the output of a joint topic model compared to the output of a regular topic model or baseline.
- $\mathbf{H_{I.4}}$ Users are more likely to identify sets of keywords as perspectives that are in line with their personal stance compared to perspectives that they do not agree with.

4.2.5. Statistical Analyses

Here, we describe the statistical analyses that we used to investigate $\mathbf{H_{I.3}}$ and $\mathbf{H_{I.4}}$. All analyses were performed using either the open-source statistical software *JASP* [165] or *R* [159]. The *JASP* file and *R* code are openly available on our repository.

Investigating $\mathbf{H_{I.3}}$

We performed a one-way analysis of variance (ANOVA) with *Model* as the independent and *nCor* as the dependent variable. This was to test the null hypothesis that there is no difference between models in terms of how many correct perspectives users could identify based on their output (i.e., the alternative hypothesis here was $\mathbf{H_{I.3}}$). Additionally, we checked the assumptions of normality and heterogeneity of variances using the Shapiro-Wilk and Levene's tests, respectively. In case the data did not meet the assumptions for the classical ANOVA, we would conduct a Kruskal-Wallis test as a non-parametric alternative.

Table 4.3: Participant's pre-existing abortion stance.

"In my opinion, abortion should be legal."	n	Percent
Strongly disagree	16	10.1
Somewhat disagree	19	12.0
Neutral	16	10.1
Somewhat agree	26	16.5
Strongly agree	81	51.3
Total	158	100.0

In case we found a significant main effect of *Model* on *nCor*, we would perform posthoc tests to study which models specifically differ from each other. Because this series of posthoc tests would involve testing multiple (i.e., $\binom{6}{2} = 15$) hypotheses, we would apply a Bonferroni correction to the traditional significance threshold of 0.05 and therefore only regard p-values below $\frac{0.05}{15} = 0.003$ as significant.

Investigating $\mathbf{H_{I.4}}$

We computed the Spearman rank correlation – a non-parametric test for the correlation between two variables [335] – between *Pre-existing stance* and *nOpp*. The null hypothesis in this test was that there is no correlation between these variables (i.e., the alternative hypothesis here was $\mathbf{H_{I.4}}$). Similar to other correlation coefficients, the Spearman rank correlation coefficient ranges from -1 to 1.

4.2.6. Participants

To determine the required sample size for our study, we conducted a power analysis using the open-source software *G*Power* [104]. Here, we specified an effect size $f = 0.3$, a significance threshold $\alpha = 0.05$, a statistical power of 0.8, and a group size of 6 (i.e., due to testing six different models). This resulted in a required sample size of at least 150 participants. Based on a short pilot study, we estimated that we would exclude about 10% of the participants due to failed honeypot checks. We thus recruited 170 native English speakers from the online participant pool *Prolific*.[8] Here, we also applied an abortion-stance pre-screening offered by *Prolific* to make the sample more balanced in terms of participant's personal attitude towards abortion (i.e., recruiting 135 "pro-life" and 135 "pro-choice" participants). After excluding some participants due to failing both honeypot checks, 158 participants remained in the study.[9]

Participants had a mean age of 33.34 (ages ranged from 18 to 64). 49.4% were male and 50.6% female. Surprisingly, despite applying the abortion-specific pre-screening offered by *Prolific* to approximate a 50/50 ratio in terms of participants who support/oppose abortion, participants in our sample turned out to largely support the legalization of

---

[8] `https://prolific.co`

[9] To pass a honeypot check, participants had to allocate the right perspective to the honeypot topic that matched this perspective word for word (see Section 4.2.3.).

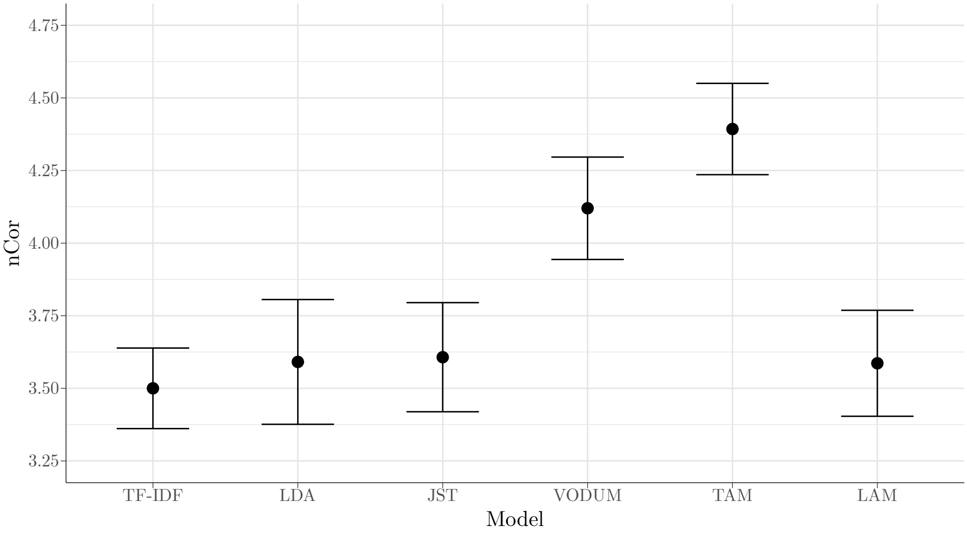


Figure 4.2: Mean *nCor* (i.e., the mean number of correctly identified perspectives) per model. The error bars represent the standard error.

abortion (see Table 4.3). Most participants believed that they were familiar with the topic, with 57.8% responding with either "strongly agree" or "somewhat agree."

4.3. Results

This section presents the results of the hypothesis tests outlined in Section 4.2.5 and several exploratory findings.

4.3.1. Hypothesis Tests

Participants find more correct perspectives when using TAM. We find that models differed in terms of how many of the six correct perspectives participants were able to identify. The ANOVA showed a significant main effect of *Model* on *nCor* ($\mathbf{H_{I.3}}$; $F = 4.399, df = 5, p < 0.001, \eta^2 = 0.126$). Table 4.4 and Figure 4.2 show the descriptive differences between the models with the highest mean *nCor* for TAM (4.39). However, although the assumption of heterogeneity of variances held according to Levene's test ($F = 0.768, df = 5, p = 0.574$), the Shapiro-Wilk test suggested that the data were non-normal ($W = 0.905$, $p < 0.001$). We thus conducted a Kruskal-Wallis test as a non-parametric alternative to the classical ANOVA, which confirmed the results of the ANOVA ($X^2 = 20.611, df = 5, p < 0.001$). We therefore reject the null hypothesis that there is no difference between the models in terms of correctly identified perspectives.

Due to the non-normality in our data, we conducted a series of non-parametric posthoc analyses (i.e., Mann-Whitney U tests) to study the individual differences between the models. The results show that only TAM led to significantly more correctly identified perspectives compared to the TF-IDF baseline model. Aside from that, the only significant

Table 4.4: Descriptive statistics of the user study. Here, *n* refers to the number of participants, mean *nCor* to the mean number of correctly identified perspectives per model (ranging from 0 to 6), and SE to the standard error.

Model	***n***	**Mean *nCor***	**SE**
TF-IDF	26	3.50	0.18
LDA	22	3.59	0.19
JST	28	3.61	0.17
VODUM	25	4.12	0.18
TAM	28	4.39	0.17
LAM	29	3.59	0.17
Total	158		

difference we found was the one between TAM and LAM.

No evidence for user tendency to interpret model output in line with personal stance. We did not find a significant correlation between *pre-existing stance* and *nOpp* ($\mathbf{H_{I.4}}$; $\rho = 0.122$, $p = 0.163$). Based on these results, we cannot reject the null hypothesis that these two variables do not correlate. Our results thus do *not* suggest that users are more likely to interpret the output of topic models in line with their personal stance.

4.3.2. Exploratory Analyses

Figure 4.3 illustrates the normalized distribution of the chosen perspectives per topic model. It displays all perspectives that could be chosen for the task (excluding the two honeypot checks). Some perspectives in the data (e.g., p_5) were more often identified than other perspectives (e.g., p_2). Furthermore, we also see differences between the models that may help explain the results from the hypothesis tests. For instance, Figure 4.3 also shows that, compared to the other models, TAM was a lot more successful in describing perspectives p_1, p_2, and p_6. TAM also did not lead people to false perspectives as much as other models did, e.g., regarding p_{10} and p_{12}.

Table 4.5 shows descriptive statistics of the exploratory measurements as described in Section 4.2.2. Overall, participants reported a high perceived usefulness of a model that can perform perspective discovery (mean = 3.82, sd = 1.06), indicating that they generally understood and approved this method. Participants felt across models that their awareness of the different perspectives had increased (mean = 3.47, sd = 1.13 respectively), although this could be due to seeing the list of 16 possible perspectives as opposed to a result of model performance. Confidence in task performance was not as high, with participants reporting moderate task performance confidence across models (mean = 2.83, sd = 1.14). This indicates that none of the models performed so well as to clearly communicate the different perspectives to users.

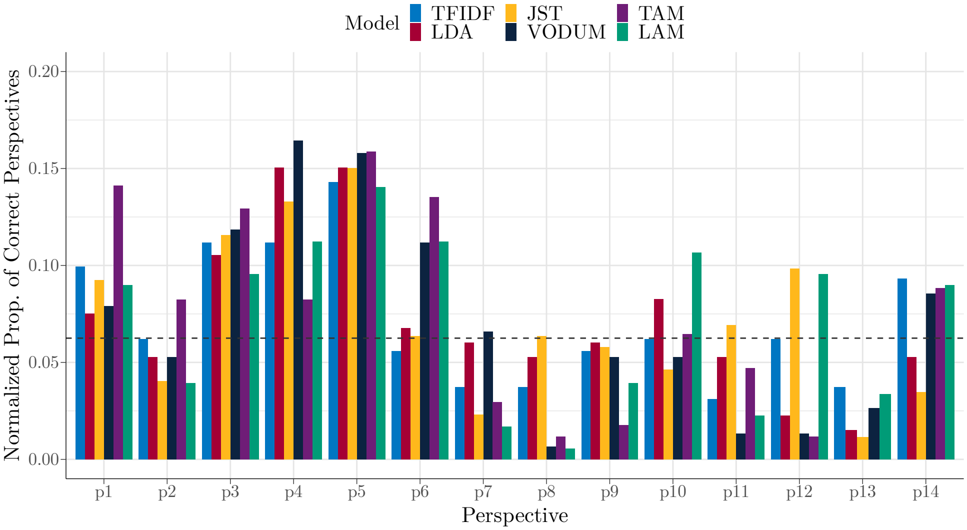


Figure 4.3: Normalized distribution of how often each available perspective was chosen (excluding the two honeypot checks). Whereas p_1 - p_6 were actually present in the corpus (see Table 4.3), the remaining perspectives were not. The horizontal line at $\frac{1}{16} = 0.0625$ shows the expected proportion for random selection.

Table 4.5: Descriptive statistics (mean ± standard deviation) on the exploratory measurements. Responses are from five-point Likert scales with 1 = "strongly disagree" and 5 = "strongly agree."

Model	**Perceived Usefulness**	**Perspective Awareness**	**Confidence**
TFIDF	3.63 (± 1.27)	3.38 (± 1.27)	2.46 (± 1.24)
LDA	3.60 (± 1.00)	3.32 (± 1.09)	2.68 (± 1.17)
TAM	3.64 (± 1.16)	3.50 (± 1.11)	3.18 (± 1.12)
VODUM	4.20 (± 0.76)	3.80 (± 1.00)	2.72 (± 0.94)
JST	4.04 (± 1.00)	3.68 (± 0.94)	3.25 (± 1.08)
LAM	3.79 (± 1.01)	3.17 (± 1.31)	2.62 (± 1.12)
Overall	3.82 (± 1.27)	3.47 (± 1.13)	2.83 (± 1.14)

Table 4.6: The six topics computed by TAM.

Topic	**Topic Words**
t_1	woman, choice, body, fetus, control, pregnant, birth, baby, foetus, sex
t_2	fetus, human, brain, person, fetus_not, cell, murder, alive, killing, egg
t_3	sex, woman, pregnant, parent, forced, child, want, child_not, option, unwanted
t_4	god, life, wrong, child, womb, baby, murder, killing, kill, creation
t_5	want, woman, sex, not, responsibility, child, get, not_want, pregnant, choice
t_6	life, god, begin, baby, life_begin, choice, choose, use, protection, responsibility

4.4. Discussion

We evaluated several joint topic models for the task of perspective discovery. Our results suggest that the *Topic-Aspect Model* (TAM) [262] can distill perspectives (i.e., in the form of topic keywords) better than a TF-IDF baseline model. Specifically, users in our study were able to correctly identify more perspectives in our data set when given TAM-generated keywords compared to baseline keywords. We find no evidence for a tendency of users towards interpreting model output in line with their personal stance.

Why did TAM perform better than other models? It seems that participants tried to find keywords in topics that explicitly appear in the perspective expression. For example, a topic containing the words *God* and *kill* is easily matched with perspective p_5 in our study (i.e., *Abortion is the killing of a human being, which defies the word of God*). Whereas all models were able to distill this particular perspective quite well (see Figure 4.3), TAM also excelled at this task for other perspectives. Table 4.6 shows the TAM model output, where the fourth topic (t_4) could be interpreted as a representation of p_5.

Outputting perspective-relevant keywords per topic seems to be a useful ingredient for a topic model that performs perspective discovery. Unlike the other joint topic models, TAM is *designed* to distinguish common words appearing in any document and words being more topic-/perspective-specific. Models that use sentiment lexica to group words, such as the *Joint-Sentiment Topic model* (JST) [205] and the *Latent Argument Model* (LAM) [356], contained more sentiment words in their topic and were therefore less effective in discovering perspectives.

Limitations

Our study is subject to several limitations. First, we created a data set containing debate forum entries with perspective annotations. This enabled us to curate a corpus of 600 documents that were balanced in terms of stance and perspectives. Such a scenario is unlikely to occur in real-world applications, where "mainstream" perspectives appear much more often than others. Second, despite our best efforts to control for it, our sample was not balanced in terms of pre-existing stance on the legalization of abortion: most participants turned out to support it. Third, we only evaluated one highly politicized, commonly debated topic (i.e., abortion). It could be questioned whether the models we tested behave similarly on other, less divisive claims (e.g., *zoos should exist* or *social* *media is good for our society*). Fourth, although our results only show a difference between TAM and two other models, descriptive statistics suggest that there could be more subtle differences (see Figure 4.2). If these differences truly exist, they could be discovered with a larger sample than the 158 participants we included in our study.

4.5. Conclusion

Simple viewpoint representations such as ternary stance labels are feasible to obtain and explainable but offer limited information (see Chapter 3). In this chapter, we introduced *perspectives* (i.e., underlying reasons for stances on debated topics) as an alternative or supplementary viewpoint representation. We then investigated whether joint topic models can help users distill perspectives from a corpus of opinionated documents and found that the *Topic-Aspect Model* (TAM) [262] can indeed produce such human-identifiable perspectives (i.e., in the form of topic keywords). Furthermore, we found no evidence of users interpreting model output in line with their personal stance.

Our findings suggest that stances on debated topics include different perspectives, that separating content into such perspectives makes intuitive sense to users, and that joint topic models have the potential to perform perspective discovery in a human-understandable fashion. If used in this way, assigning perspective labels (using joint topic models) could prove helpful in many different areas, including policy-making or helping people overcome biases when participating in (online) debates. A key limitation of the work we presented in this chapter is that perspectives are highly dependent on the topic (i.e., *abortion legalization*) and the data we considered (i.e., debate forum posts). It may be difficult to simultaneously differentiate viewpoints as well as topics in more diverse data sets. Chapter 5, the next and final chapter of Part I, explores how stances and perspectives can be combined into a comprehensive yet simple and topic-independent viewpoint representation.

5

Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions With Debated Topics

---

This chapter is based on a published, full conference paper: Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. “Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions with Debated Topics”. In: *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*. CHIIR ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 135–145. DOI: `10.1145/3498366.3505812`.

Tim Draws primarily planned and carried out the conceptualization, investigation, project administration, visualization, and write-up of the work described in the paper referenced above. Tim and Oana Inel collaborated in implementing and describing the methodology. The remaining co-authors supervised Tim during the project and made edits to the writing.

In Chapter 3 of this dissertation, we have examined the possibilities and limitations of ternary stance labels (i.e., *against*/*neutral*/*in favor*), which are the most common way to represent viewpoints in current human information interaction research. We have shown that existing stance detection methods can generate such labels in an automatic and explainable fashion but that they often lack nuanced information regarding the viewpoints they represent. In Chapter 4, we have shown that topic models could enhance stance detection by discovering perspectives (i.e., the reasons that underlie stances); however, such perspectives are highly topic-dependent as arguments can differ vastly across topics. Recent research in the communication sciences has argued that viewpoints are complex constructs with multiple dimensions and can vary in a plurality of ways [23, 24, 25, 39]. For instance, classifying documents using a ternary stance taxonomy removes any notion of a *degree* to which a viewpoint may oppose or support a topic, e.g., somewhat or strongly supporting the feminist movement. Moreover, two tweets may support feminism for different reasons (i.e., *perspectives*) that relate to broader ways of viewing the world (e.g., focusing on ethical or economic aspects of empowering women) [25, 39]. Viewpoint representations should capture this latent richness in a human-understandable, topic-independent fashion to allow for more nuanced viewpoint bias analyses, user studies, and practical applications.

Obtaining a deep understanding of user interactions with debated topics may require information such as whether a user interacted with search results that are overall “strongly opposing” or just “somewhat opposing” feminism. Similarly, interventions for diverse news reading could more effectively expose users to alternative perspectives when knowing which reasons for opposing or supporting a topic different search results convey. Enabling such advanced analyses could unlock greater potential for research and practical applications in human information interaction. To this end, we argue that more comprehensive viewpoint representations are needed.

Recent human information interaction research has already begun to represent viewpoints in alternative formats (e.g., as continuous stance scales [195]) and measure nuanced differences between documents [240, 345]. For instance, Mulder et al. [240] drew from the communication sciences to operationalize *framing*, a concept that represents viewpoints across four different dimensions (i.e., *problem definition*, *causal attribution*, *moral evaluation*, and *treatment recommendation*). They used different automatic methods to compute a distance function considering these four dimensions, gauging the viewpoint similarity between news articles. This earlier work shows – albeit via a distance function instead of a label – that the richer viewpoint notions handled in the communication sciences can be practically applied in human information interaction. However, to the best of our knowledge, no current method translates comprehensive viewpoint representations into practical viewpoint labels applicable to user interactions with debated topics. What is needed is a standard, go-to framework that yields computationally tractable and feasibly obtainable labels (e.g., using crowdsourcing) but is significantly more comprehensive than currently used methods. The present chapter aims to fill this research gap by answering four research questions:

RQ$_{\mathbf{I.5}}$ What label represents viewpoints in a comprehensive yet relatively simple and topic-independent fashion?

RQ$_{\textbf{I.6}}$ Can crowd workers reliably assign our proposed viewpoint label to textual documents?

RQ$_{\textbf{I.7}}$ Do cognitive biases affect crowd workers when assigning our proposed viewpoint label?

RQ$_{\textbf{I.8}}$ Is our proposed viewpoint representation more meaningful compared to binary stance labels?

To address **RQ$_{\textbf{I.5}}$**, inspired by work from the communication sciences, we developed a topic-independent, two-dimensional viewpoint representation that incorporates a viewpoints' *stance* (i.e., the degree to which it supports or opposes a claim) and *logic of evaluation* (i.e., a generalized taxonomy for perspectives or underlying reasons; see Section 5.1). We then tasked crowd workers to assign our novel viewpoint label to tweets on several debated topics. Analyses of this crowdsourcing task suggest that crowd workers can perform this task reliably (**RQ$_{\textbf{I.6}}$**), and there was no evidence that cognitive biases would have affected the results (**RQ$_{\textbf{I.7}}$**; see Section 5.2). We further demonstrate in a viewpoint diversity analysis of the tweets that our proposed viewpoint label enables more nuanced insights (i.e., by reflecting stance degrees and logics of evaluation) compared to binary or ternary stance labels (**RQ$_{\textbf{I.8}}$**; see Section 5.3). Finally, we report on a user study where participants saw sets of tweets, diversified either based on our proposed viewpoint representation or a binary stance label. Exploratory results of this user study suggest that users judge sets of tweets as more viewpoint-diverse when the sets are diversified based on our proposed label compared to the baseline – as long as these sets do not contain too many extreme viewpoints (**RQ$_{\textbf{I.8}}$**; see Section 5.4).

Supplementary material such as data sets, task screenshots, and analysis code related to this chapter is available at `https://osf.io/pjws9/`.

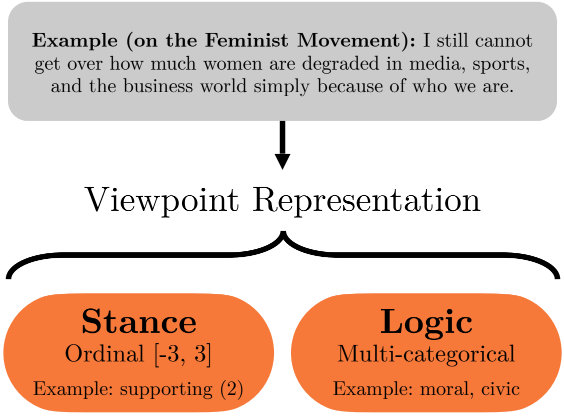

Example (on the Feminist Movement): I still cannot get over how much women are degraded in media, sports, and the business world simply because of who we are.
Viewpoint Representation
Stance
Ordinal [-3, 3]
Example: supporting (2)
Logic
Multi-categorical
Example: moral, civic

Figure 5.1: Proposed viewpoint representation at the example of a tweet from the *SemEval 2016 Stance Detection* data set [233]. A viewpoint is evaluated on two dimensions: *stance* (i.e., on a seven-point ordinal scale ranging from “strongly opposing” to “strongly supporting” a topic) and *logic of evaluation* (i.e., in a multi-categorical format to include all logics present; see Table 2.2).

5.1. Novel Viewpoint Representation

We propose a novel viewpoint representation for human information interaction that improves upon binary and ternary stance labels by reflecting a viewpoint's *stance* on a more nuanced level and a viewpoint's *logic of evaluation* as a second dimension (see Figure 5.1). Thereby, our proposed viewpoint representation is more comprehensive than existing methods. We detail the two dimensions of our proposed representation below.

Stance

The first dimension in our proposed viewpoint representation is a viewpoint's *stance*; i.e., its moral evaluation of the topic at hand. For example, consider the tweet displayed in Figure 5.1. This tweet is clearly *in favor* of the feminist movement and was therefore classified accordingly in the *SemEval 2016 Stance Detection* data set [233] (see Section 5.2.1). In our proposed framework, however, stances are represented on a seven-point ordinal scale ranging from "strongly opposing" (-3) to "strongly supporting" (3; see Chapters 8 and 10). This representation reflects a viewpoint's general orientation similar to the standard binary approach but also the *degree* to which a viewpoint opposes or supports a topic. For instance, we may label the tweet in Figure 5.1 as "supporting" (2), meaning that it takes a clear stance in favor of feminism but does not do so to an extreme extent.

Logic of Evaluation

The second dimension of our proposed viewpoint representation is a viewpoint's *logic of evaluation* (or simply *logic*), a construct that we borrow from the communication sciences [24, 25, 39] (see Section 2.2.1). A viewpoint's logic of evaluation reflects the general perspective behind the stance: it describes *why* a stance is taken. For example, the statements "women should be treated fairly" and "empowering women would benefit the economy" both arguably support feminism but do so for different reasons. Whereas the first one refers to fairness (i.e., using a *moral* logic), the second one refers to value creation (i.e., using an *economic* logic). Baden and Springer [24] mention seven different logics that a viewpoint can include: *inspired*, *popular*, *moral*, *civic*, *economic*, *functional*, and *ecological* (see Table 2.2). Each of these seven logics represents a particular maxim according to which a problem may be evaluated. For instance, an *ecological* logic is employed when the viewpoint refers to something that is supposedly (not) sustainable or natural; e.g., opposing feminism by expressing that "equal treatment of men and women is unnatural". Classifying viewpoints into logics of evaluation thus allows for entirely topic-independent descriptions of the latent perspectives that viewpoints embody. Note that any document may refer to one or several of the seven logics. For example, the example tweet in Figure 5.1 refers to a *moral* logic (i.e., arguing that women are treated unfairly) and a *civic* logic (i.e., suggesting that this is not acceptable). This type of information is lacking when using a standard binary viewpoint label.

5.2. Obtaining Viewpoint Labels

In this section, we report on a crowdsourcing study in which we collected viewpoint labels according to our proposed framework (see Section 5.1) for 169 tweets from the *SemEval 2016 Stance Detection* data set [234, 331]. We describe the data, task setup, and process of collecting the annotations. Furthermore, we analyze whether workers were able to assign viewpoint labels reliably and whether they were influenced by cognitive biases when annotating.

5.2.1. Data

One of the most utilized data sets for stance classification is the *SemEval 2016 Stance Detection* data set, which consists of 4870 tweets on six different debated topics: *atheism*, *climate change*, *Donald Trump*, *feminist movement*, *Hillary Clinton*, and *legalization of abortion* [234, 331]. It was originally created for the *SemEval 2016 Stance Detection Challenge* [233], which invited contributors to create automatic methods for classifying tweets into four stance categories: *in favor*, *neutral*, *against*, and *none* (i.e., no stance). All tweets in the *SemEval 2016 Stance Detection* data set are annotated for their stance (i.e., using the same four categories) and relevance concerning the target topic.

We aimed to collect annotations according to the viewpoint representation we propose in Section 5.1 for a subset of the tweets contained in the *SemEval 2016 Stance Detection* data set. Specifically, we selected all 169 tweets that at least 90% of the original annotators judged as relevant to the topics *atheism* (16), *Donald Trump* (54), or *feminist movement* (99). We chose these three topics to limit expenses (i.e., allowing for more annotations per tweet) while maintaining topical diversity (i.e., they cover diverse topics such as religion, politics, and social and political movements) and relevance in online discussions and information sharing platforms.

5.2.2. Prior Considerations

Aside from collecting our proposed viewpoint label for the 169 tweets in our final data set, we also aimed to investigate whether cognitive biases can affect crowd workers when assigning these labels. We additionally applied the cognitive-biases-in-crowdsourcing checklist we propose in Chapter 6 and concluded that two different cognitive biases might affect crowd workers in our task. First, we were concerned about a *halo effect*, in which irrelevant pieces of information affect crowd workers' annotations. We were particularly concerned that crowd workers with pre-existing solid knowledge on the topic might rate viewpoints as more extreme (i.e., more readily placing tweets into the "opposing camp" or "supporting camp"). Second, we suspected that the *confirmation bias* could affect crowd workers if they had a tendency to label tweets in line with their personal stance (i.e., looking for attitude-confirming evidence). We thus decided to incorporate measurements of personal knowledge and stance concerning the given topic in our task design.

5.2.3. Task Setup

We designed a human intelligence task (HIT) to obtain viewpoint annotations in our proposed format. First, crowd workers were presented with one of the three topics (i.e., *atheism*, *Donald Trump*, or the *feminist movement*) and asked for their personal knowledge and stance on it (see Section 5.2.2). We measured these constructs on seven-point Likert scales ranging from "non-existent" to "expert" (knowledge) and from "strongly opposing" to "strongly supporting" (stance). Crowd workers then saw one of the 169 tweets in our data set relevant to the same topic.

The main task for crowd workers was to evaluate the viewpoint expressed in the tweet in the three subsequent steps: they (1) described the expressed viewpoint in their own words, (2) judged its stance regarding the topic on a seven-point Likert scale ranging from “strongly opposing” to “strongly supporting”, and (3) selected which logic(s) applied (see Section 5.1). In step (3), the seven logics were displayed as completions of a sentence; e.g., *“Fundamentally, the viewpoint contained in the tweet is that Feminist Movement is (not) in line with… what is social, fair, or moral.”* (i.e., indicating a moral logic, c.f., Table 2.2). Crowd workers could obtain more information (including examples) about any given logic by hovering over the respective option. In this last step, participants first selected the viewpoint’s main logic by choosing one of the seven categories and then had the option to select any other logic that may also apply. We also added a mandatory attention check (i.e., an item where we explicitly told crowd workers which option to select) and an option to give feedback in open-text form. We published the task on *Amazon Mechanical Turk* (MTurk).[1]

5.2.4. Human Annotators

Crowd Annotators

A total of 66 crowd annotators annotated our HITs (i.e., consisting of one tweet). They had a *Master* status on MTurk, a HIT approval rate of at least 95%, and at least 500 accepted HITs. Furthermore, we only allowed crowd workers from a selection of 30 countries that either have English as their main language (e.g., the United States) or that have high English proficiency according to the EF English Proficiency Index[2] (e.g., The Netherlands and Denmark). These constraints ensured a high-quality pool of annotators with good English understanding (i.e., our tweets are in English). Furthermore, we excluded nine annotations for which the crowd worker failed the mandatory attention check. The final sample consisted of 1197 annotations from 66 different crowd annotators. Crowd workers were allowed to submit as many HITs as they wished and were rewarded with \$0.50 for each completed HIT. Each tweet received between six and eight annotations (mean = 7.08, sd = 0.30). On average, crowd workers reported a good knowledge across the three topics *atheism* (mean = 1.70, sd = 1.24), *Donald Trump* (mean = 1.91, sd = 1.05), and the *feminist movement* (mean = 1.54, sd = 1.19).[3] Regarding personal stance, they slightly supported *atheism* (mean = 0.62, sd = 1.98), opposed *Donald Trump* (mean = -1.49, sd = 1.96), and were approximately neutral towards *feminism* (mean = -0.14, sd = 2.07).

Expert Annotators

To evaluate the quality of the crowd annotations, we created a ground truth data set consisting of 34 tweets (i.e., 20% of the tweets used in our study). We aimed to avoid bias by randomly selecting the tweets for each of the three topics of interest and proportional to the total number of tweets for a given topic (i.e., 4 on *atheism*, 11 on *Donald Trump*, and 19 on the *feminist movement*). First, two expert annotators (i.e., authors of the paper that this chapter is based on; with a background in computer science and familiar with the logics depicted in Table 2.2) independently annotated the 34 tweets. The two experts annotated the 34 tweets using the same task that was provided to the crowd annotators, i.e., on

---

[1] `https://www.mturk.com`

[2] `https://www.ef.com/wwen/epi/`

[3] We here represent the seven Likert points as integers on an ordinal scale $[-3, 3]$.

MTurk. We computed the inter-rater reliability of the two experts concerning tweets' stances and logics using Krippendorff's α [191]. The reasons for choosing this metric were three-fold: it is (1) applicable on both ordinal and nominal values (i.e., our data is ordinal - stance and nominal - logics), (2) deals with missing data (not all annotators annotate all examples), and (3) generalizes to any number of annotators. Regarding stance, the two experts had a high IRR score of 0.84, while in terms of logics their agreement varied from almost no agreement (e.g., *popular*, *functional*, *inspired*, *civic*, and *financial* logics have α values below 0.07) to high agreement (e.g., 0.58 for *moral*) and perfect agreement (e.g., 1.0 for *ecological*). The two experts then discussed the annotations with a third expert who has a background in communication science (also a co-author of the paper that this chapter is based on). In the discussion session, all 34 tweets in the ground truth were individually discussed until an agreement was reached regarding the applicable stances and logics.

5.2.5. Crowd Annotation Aggregation and Quality

As described in Section 5.2.3, we asked crowd annotators to judge the *stance* and the *logic(s)* of each tweet. In this section, we report on the aggregation of the crowd annotations to identify the collective stance and logics for each tweet, as well as on the quality of the annotations gathered in our crowdsourcing study.

Tweet Viewpoint Stance

To aggregate stance annotations, we represented the seven options from the Likert scale as integers $[-3, 3]$ and assigned each tweet the median annotation value (i.e., rounded to integer). Crowd annotators largely agreed on the extent to which a tweet opposes or supports a particular stance. Their IRR score on the tweet stance (c.f. Krippendorff's α) is 0.69 on the entire data set and 0.72 on the expert-annotated data set of 34 tweets. We also compared the aggregated crowd and expert stance on the tweets. In this case, the IRR score c.f. Krippendorff's α is 0.84, further emphasizing the crowd's reliability in annotating stances of tweets using our more complex representation, i.e., on an ordinal scale ranging from -3 to 3. The crowd's micro F1-score in terms of stance was 0.53 when using the ordinal scale ranging from -3 to 3 and 0.97 when using a ternary scale (against, neutral, in favor). The aggregated stance labels from crowd workers matched the stance indication contained in the original *SemEval 2016 Stance Detection* data set in 97% of cases. Five tweets that had all originally been classified as *in favor* of feminism or atheism were annotated as *neutral* (0) in our data (e.g., "*Just been putting the finishing touches to a feminist-themed cryptic crossword... Standard. #crosswords*").

Tweet Viewpoint Logic

Annotating logics to each tweet was the more difficult task for crowd workers, as the interpretation of logics could be somewhat subjective and ambiguous. Moreover, a given tweet may contain multiple different logics with different degrees of relevance or intensity, so attaching a single logic to each tweet is not optimal. These observations led us to analyze the crowd annotations regarding the logic(s) of the tweets with the disagreement-aware metrics called *CrowdTruth* [93, 94], which compute quality scores for input units (i.e., tweets), crowd annotators, and target annotations (i.e., the seven logics).[4] When applying the metrics, we considered the main logic as well as all additional logics that a crowd annotator selected.

The CrowdTruth metrics assume that the three main components of the crowdsourcing task (i.e., tweet, crowd annotators, and logics) are mutually dependent. For instance, a difficult tweet can make crowd annotators disagree, but this does not necessarily mean that their answers' quality is poor (i.e., annotators can fill in each others' gaps by adding logics that others have missed). Thus, c.f. the CrowdTruth metrics, the quality of a tweet is weighted by the quality of the crowd annotators that annotated the tweet and of the target annotations, i.e., the logics, and vice versa. The answers of a crowd annotator who constantly disagrees with the other crowd annotators will have a lower weight in the final aggregation of answers. These quality scores are computed in a loop, using a dynamic programming approach, until convergence. Each quality score ranges from 0 and 1, where higher values indicate higher quality or clarity.

Upon applying the CrowdTruth metrics, we thus had (1) crowd annotators quality scores, (2) tweet quality scores, and (3) tweet-logic scores. A tweet-logic score is computed for each tweet and each logic, expressing the likelihood of the logic to be expressed by the tweet. We evaluated the crowd's performance in terms of the micro-F1 score [270], using the 34 tweets for which we collected ground truth data from expert annotators (see Section 5.2.4).[5] For this, we use the tweet-logic score as a threshold to differentiate between positive and negative samples (i.e., logics expressed and not expressed in a tweet). We experimented with threshold values between 0 and 1, in increments of 0.01, and computed the crowd's micro-F1 score for each such threshold. We generally observe that the lower the threshold (i.e., considering more logics to be expressed in a tweet), the higher the crowd micro-F1 score. For example, the micro-F1 score is equal to 0.67 at a threshold of 0.01 and equal to 0.02 at a threshold of 1. Based on this analysis, we considered a threshold of 0.25 as optimal (micro-F1 = 0.61) to have a more balanced performance concerning recall and precision and still eliminate logics that are considered applicable by only a few crowd annotators or crowd annotators with low-quality scores. The final viewpoint label per tweet thus comprised of two dimensions: the median stance annotation and a vector of all logics that passed the aforementioned threshold (see Figure 5.1 for an example).

Compared to stance, logic annotations generated substantially more disagreement, resulting in much lower Krippendorff's α values (0.23 or lower on both the main and the expert-annotated data set). The crowd agreed most on the *moral* and *functional* logics. When compared to the expert logics on the 34 tweets in our ground truth, we observe similar agreements as for the experts. Specifically, we found perfect agreement for the *ecological* logic, moderate to high agreement for the *moral* (Krippenforff's $\alpha = 0.58$) and *popular* ($\alpha = 0.36$) logics, and low agreement for the other logics.

---

[4] `https://github.com/CrowdTruth/CrowdTruth-core`

[5] We compute micro-F1 scores because we deal with a multi-label classification problem, where logics are not equally represented across the data set. We also consider all logics equally important, and we are interested to see how the crowd performs across logics.

5.2.6. Gauging the Annotation Difficulty

To better understand the difficulty of our task, we had eight different crowd workers annotate between one and 103 tweets *twice*. We ensured here that there was always a considerable amount of time and other HITs between the first and second annotation of a tweet. We found that workers were largely consistent in their two annotations of the same tweet. Overall, annotators did not diverge more than one point on the stance scale in 89% of cases and assigned precisely the same set of logics in 38% of cases. The average Jaccard distance of logics annotation pairs was 0.44, indicating that workers may often have missed or added a logic in their second annotation compared to their first, but usually annotated with some degree of overlap.[6] For example, if a worker first only assigned [*inspired*] to a tweet but annotated [*inspired*, *popular*] at the second time, the Jaccard distance between the two annotations was 0.5. This also shows that the low inter-rater reliability scores reported in the previous paragraph may give a somewhat misleading image regarding the task difficulty. In sum, workers were fairly consistent when annotating a tweet for the second time but may have missed certain logics that other crowd workers detected.

Checking for Cognitive Biases

As explained in Section 5.2.2, we tested whether specific cognitive biases (i.e., the *halo effect* and the *confirmation bias*) influence crowd workers when assigning our proposed viewpoint label. The halo effect we were concerned about would have taken place if workers' knowledge of the topic at hand had influenced the variance of their stance annotations (i.e., placing tweets in either *extremely opposing* or *extremely supporting* "camps"). However, we found no evidence of this effect from a Spearman correlation analysis between workers' self-reported knowledge on their assigned topic and the standard deviation of their stance annotations ($\rho = 0.14$, $p = 0.5$).[7] A confirmation bias in our task could have meant that crowd workers look for information that confirms their pre-existing beliefs and thus annotate stances in line with their personal stance. However, we also found no evidence for a confirmation bias from a Spearman correlation analysis between workers' self-reported stance on their assigned topic and their mean stance annotation ($\rho = 0.04$, $p = 0.8$).

5.3. Analyzing Viewpoint Diversity

This section presents a viewpoint diversity analysis of the data described in Section 5.2.1 using the two-dimensional viewpoint labels we collected (see Section 5.2). The aim of this analysis is to obtain insights into the discussions surrounding the three debated topics (i.e., *atheism*, *Donald Trump*, *feminist movement*) and to showcase the depth of understanding that our proposed viewpoint representation provides. For each topic, we analyzed (1) the stance distribution, (2) the logics distribution, and (3) how the different logics relate to each other within the online discussions.

---

[6] There was no overlap concerning logics annotations in 24% of cases.

[7] To ensure independence of observations for this analysis, we here only considered one stance (on one topic) per crowd worker.

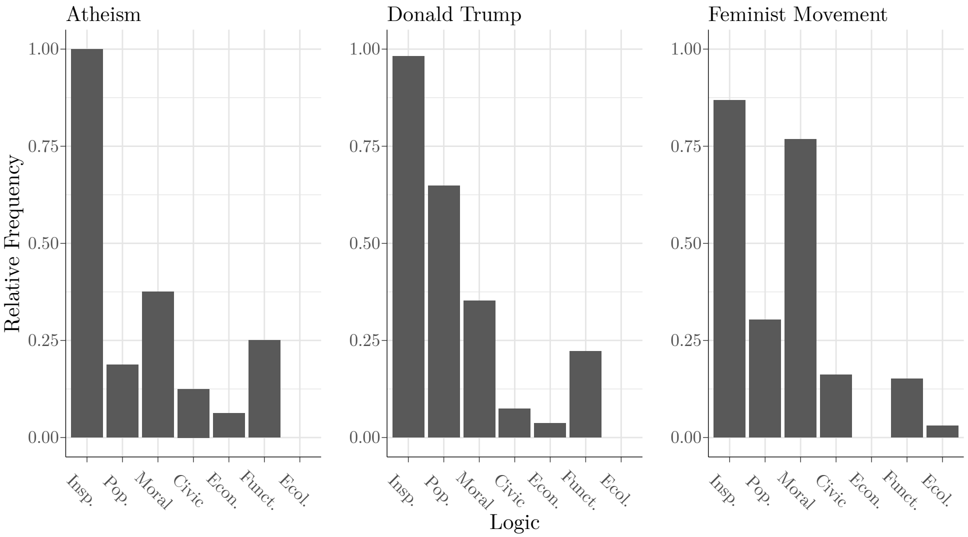


Figure 5.2: Relative frequency of the seven different logics across the topics *atheism*, *Donald Trump*, and *feminist movement*.

5.3.1. Method

We analyzed the viewpoint diversity in our tweets data set in a qualitative fashion. Aside from their raw format, we examined the data using two different visualizations. The first visualization (Figure 5.2) shows – per topic – the relative frequency of the seven different logics across all tweets. We computed the relative frequency by dividing the number of tweets in which a logic appears by the total number of tweets within that topic. This provides a visual overview of the relative importance of the different logics.

Our second visualization (Figure 5.3) shows network plots of tweets and their logic similarity. Each node in the networks represents a tweet and is colored according to the tweet's stance. To create the networks, we first computed Jaccard similarity matrices of all tweets within each topic based on the logics they refer to. This meant that two tweets would receive a high similarity index if they used similar or the same logics in their argumentation but a low similarity index if they used distinct logics. We then used the similarity matrices as weight matrices for the networks such that stronger edges indicate stronger logic similarities between tweets. For better visibility of meaningful similarities, we omitted all edges with Jaccard indexes of 0.4 or lower. The networks visualize how tweets cluster together in terms of the logics they refer to. This helped us investigate structural similarities in the argumentation used on either side of the debated topics.

5.3.2. Results

This section presents the results of our viewpoint diversity analysis for each topic.

Atheism. Of the 16 relevant tweets in our data set, only two were labeled as either "somewhat opposing" or "strongly opposing" atheism. The remaining 14 tweets received

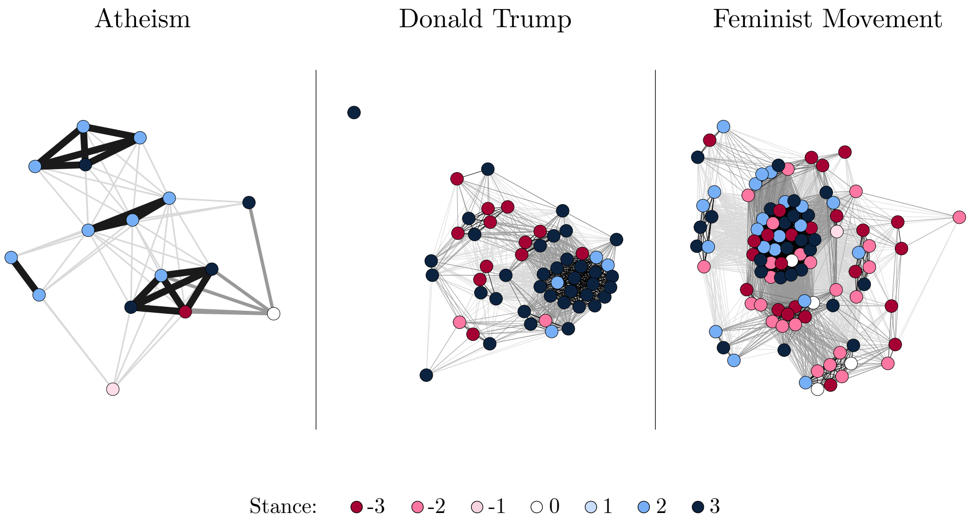


Figure 5.3: Network plots of all tweets divided into the three topics *atheism* (left-hand panel), *Donald Trump* (central panel), and the *feminist movement* (right-hand panel). Each node is a single tweet, whereby its color indicates the stance. Edges indicate the Jaccard similarity between tweets based on the assigned logics (i.e., the stronger the edge, the greater the similarity).

"neutral" (1), "supporting" (9), or "strongly supporting" (4) labels. The left-hand panel of Figure 5.2 shows the relative importance of the different logics that were used when discussing *atheism*. Whereas the *inspired* logic was found in every tweet, all other logics appeared in 0% to 31% of tweets. The network plot in the left-hand panel of Figure 5.3 shows three main clusters of at least three tweets that evaluate atheism. These tweets would refer solely to an *inspired* logic (e.g., "*[...] which god? Yours? not mine. oh wait i don't have one. #LoveWins*") or combine an *inspired* logic with either a *functional* logic (e.g., "*If God = Miraculous And Miracles = Impossible Then God = Impossible #logic #reason #science #RT*") or a *moral* logic (e.g., "*Serious question for my atheist libertarians: How can rights exist without God? #ChristianLibertarian*").

Donald Trump. Our data set contains 54 tweets from 2016 that evaluated *Donald Trump*. Compared to the other topics, the discussion around Donald Trump is much more polarized, as 89% of tweets are either strongly supporting or strongly opposing Donald Trump. The bar plot in the central panel of Figure 5.2 shows that the *inspired* and *popular* logics were used most often. Conversely, only a few tweets in our data set express viewpoints that refer to an *economic* or *ecological* logic. The network plot in the central panel of Figure 5.3 shows that most tweets are highly similar to each other (i.e., they cluster closely together). The largest cluster consists almost entirely of tweets that strongly support Donald Trump and represents a combination of the *inspired* and *popular* logics (e.g., "*[...] We have got to take our country back. It's time! Win it Mr. Trump*"). Tweets in a similar cluster that is almost entirely in favor of Donald Trump combine the *inspired* and *popular* logics with a *functional* logic (e.g., "*[...] Hell I'm from the UK and I believe realDonaldTrump would make an amazing WORLD Leader*"). Arguments on the opposing side, in contrast, were usually made by taking a *moral* aspect into account (e.g., "*Donald Trump needs to stop embarrassing himself. Racist assholes...*").

Feminist Movement. The majority of tweets in our data set (99) evaluate the *feminist movement*. Here, the stance distribution is comparatively balanced with 46% supporting, 4% neutral, and 50% opposing tweets, only half of which are at the extreme ends of the stance spectrum. The bar plot in the right-hand panel of Figure 5.2 shows that feminism was discussed using similar logics compared to the other topics, but that the *moral* logic is noticeably more important here. This is also reflected in the network plot displayed in the right-hand panel of Figure 5.3. The largest cluster contains tweets that combine *inspired* and *moral* logics to argue in both directions (e.g., "*I think it's okay for a woman to take a mans name if she wants to. #genderequality*"). Many tweets that support feminism argue exclusively using a moral logic ("*I shouldn't have to be holding a man's hand to be left alone on the street. #catcalling #streetharassment #equality*"). On the other hand, tweets opposing *feminism* tend to use the *inspired* logic more often (e.g., "*All the feminist block me because I speak true.*") and sometimes combine that with other logics such as the *popular* one (e.g., "*[...] Most feminists don't know what they are fighting for?! Most ego maniac's who want they're 15 minutes of fame. #c4news*").

5.4. User Evaluation of Viewpoint Label

We have shown that our proposed viewpoint label is obtainable via crowdsourcing with acceptable reliability (as measured by Krippendorff's α; see Section 5.2) and that it enables in-depth viewpoint analyses (see Section 5.3). However, it is still unclear whether this approach can also help to create noticeably superior outcomes from the user's perspective. To test whether using our proposed viewpoint representation can more meaningfully organize online discussions (i.e., addressing $\mathbf{RQ_{I.8}}$), we conducted a user study. We presented users with sets of tweets that were diversified based on either our proposed viewpoint label or a binary stance label and asked them which set was more viewpoint-diverse. The user study had been preregistered before any data collection.[8]

5.4.1. Method

Data

For this user study, we considered tweets that were part of the data set described in Section 5.2.1 and that related to the topic *feminist movement*. We only focused on the *feminist movement* here because the other two topics had comparably few relevant tweets and skewed stance distributions, which hindered diversification efforts. We further excluded five feminism-related tweets that had received a neutral stance label (4) or that were the only ones in their stance category (i.e., one *somewhat opposing* tweet).

---

[8] The preregistration is available at `https://osf.io/cn8qa`.

Sets of Tweets

We assembled a total of 10 different sets of tweets from the data set described above. Each set contained six tweets on the *feminist movement* and was created using one of two different sampling algorithms. The first algorithm diversified tweets using our proposed viewpoint label: after sampling one random tweet as the first element in the set, this algorithm added the five remaining tweets by always picking the tweet with the maximum average Jaccard distance to the tweets that were already in the set. It did this in such a way that the stance distribution was as balanced as possible, i.e., including at least one of the available stance categories. The second algorithm diversified tweets based on the original binary label contained in the *SemEval 2016 Stance Detection* data set and therefore randomly sampled three *against* and three *in favor* tweets to create a set. We created five such sets per algorithm.

Procedure

The user study consisted of two steps. First, participants read an informed consent and stated their gender and age group from multiple choices. Second, we presented participants with a scenario: they were co-organizing a debating event aiming to bring people of diverse viewpoints together. It was explained that two methods are being tested to diversify the table seat allocations based on attendees' recent tweets on the feminist movement. Participants then saw two random sets of tweets (i.e., one per sampling algorithm; in random order) graphically arranged in a circle to imitate a table seat allocation (see our repository for screenshots). A border surrounding each tweet was colored red (*against*) or blue (*in favor*) depending on the tweet's stance label in the original *SemEval 2016 Stance Detection* data set. We asked participants to judge which table had a greater viewpoint diversity and shortly explain their choice in an open text field.

Analysis

Our hypothesis for this study was that users would judge tweet sets created with the sampling algorithm based on our proposed viewpoint label as more diverse. To test this hypothesis, we conducted a binomial test with a test value of 0.5 (i.e., testing the null hypothesis that users choose tables at random).

Participants

We conducted a power analysis before data collection to gauge the required number of participants for this user study. Using the software *G*Power* [104], we specified that we expect a medium effect size (i.e., Cohen's $g = 0.15$), handle a significance threshold of $\alpha = 0.05$, and aim for a statistical power of $\beta = 0.8$ in a two-tailed binomial test. This resulted in a required sample size of 90 participants, which we thus recruited from *Prolific*.[9] All participants were native English speakers above 18 years of age. We paid $0.70 per participation (an average of $10.33 per hour) while allowing each participant to only judge one pair of tweet sets.

---

[9] `https://prolific.co`

5.4.2. Results

Among the 90 participants we had recruited, 58 (64%) were female, 31 (34%) were male, and one (1%) was non-binary. Participants' age distribution was somewhat skewed towards younger ages, with only 7 participants being older than 44 years of age. Most participants (56%) judged the sets of tweets that had been diversified based on our proposed viewpoint label as more diverse than the sets sampled based on a binary label. However, the binomial test was not significant ($p = 0.34$). We thus did not find any evidence for a difference between the two types of tweet sets.

Exploratory analysis

To help explain why we did not find a significant difference between the two types of tweet sets, we collected additional data and conducted a second exploratory analysis. One potential reason we suspected could have led to the insignificant results was an overestimation of the effect size in our initial required sample size computation (see Section 5.4.1). To address this potential issue of insufficient power, we adjusted the sample size calculation to detect a smaller effect (i.e., Cohen's $g = 0.1$) rather than a medium effect. We thus recruited an additional 110 participants (i.e., raising the sample size to 200), who went through the same procedure as the first 90.

Another suspected reason for the insignificant result concerned spurious variation in the tweets. Upon closer examination of the results, we noticed that most participants judged four out of the five tweet sets diversified based on our proposed viewpoint representation as more diverse. However, for one particular tweet set pair, our diversification was judged as more diverse only five out of eighteen times. Participants stated that this set contained many extreme opinions and that therefore it did not seem like a good discussion would result from this set. Indeed, our method had assembled a set containing four extreme viewpoints (i.e., *strongly opposing* and *strongly supporting*) and only two mild viewpoints. This was different in all other sets, which had no more than 50% extreme viewpoints. We therefore excluded data from participants who had annotated this set from this exploratory analysis.

Ninety-seven (61%) out of the remaining 159 participants judged the sets diversified using our proposed viewpoint label as more diverse, a proportion significantly higher than random ($p = 0.007$).[10] Note that these analyses are exploratory as we conducted them outwith the preregistration and after examining the main results.

5.5. Discussion

We have proposed a novel viewpoint representation for human information interaction that overcomes the limitations of currently used binary stance labels in two crucial ways. First, instead of classifying viewpoints into generic stance categories, it represents a viewpoint's stance on a more nuanced, seven-point ordinal scale ranging from "strongly opposing" to "strongly supporting". Second, it includes a viewpoint's logic(s) of evaluation (i.e., a notion that we borrow from the communication sciences), representing underlying reasons or perspectives using seven general categories. Our proposed viewpoint representation thus incorporates important aspects of viewpoints identified by the communication sciences in two dimensions while remaining topic-independent (**RQ**$_{\mathbf{I.5}}$; see Section 5.1). We have shown that workers can assign this novel viewpoint label with satisfactory reliability (**RQ**$_{\mathbf{I.6}}$) and found no evidence for an influence of cognitive biases (i.e., the *halo effect* and the *confirmation bias*) in this context (**RQ**$_{\mathbf{I.7}}$; see Section 5.2). Furthermore, in a viewpoint diversity analysis of tweets and a user study, we have demonstrated that our proposed viewpoint representation, while subtle, is more comprehensive and meaningful compared to binary stance labels (**RQ**$_{\mathbf{I.8}}$; see Sections 5.3 and 5.4). Our exploratory analyses further suggest that the diversification algorithm must be tuned correctly concerning stance; i.e., including too many extreme opinions from either side of the spectrum may lead users to find the diversification less meaningful.

---

[10] Without removing the problematic set of tweets, the binomial test was not significant even in the larger sample of 200 participants ($p = 0.1$).

5.5.1. Guidelines for Obtaining Viewpoint Labels

Our crowdsourcing study has shown that workers are sufficiently reliable when annotating our proposed viewpoint label. However, especially with respect to assigning logics of evaluation or when dealing with ambiguous tweets, this task can be difficult. Worker feedback on our task included comments such as "*The tweet doesn't really mention the logic behind the support*," "*This one doesn't seem to make any sort of argument*," and "*Really have to read between the lines with this one honestly*" Based on our experience, we therefore propose a set of guidelines that requesters should follow when aiming to obtain annotations of our proposed viewpoint label:

1. Given the difficulty of the task and in line with earlier work on this topic [157] (see also Section 2.2.3 and Part II of this dissertation), we recommend setting the worker requirements rather high; e.g., *Master* workers from MTurk.

2. While crowd workers seem to have no trouble annotating stance even on a seven-point ordinal scale, the logic(s) of evaluation can be hard to interpret. Requesters should ensure that all logics are well-explained and include several examples as well as relevant words to look for (see Table 2.2).

3. In line with recent work on similar subjective crowdsourcing tasks [176], we recommend collecting more than the standard three annotations per document (i.e., at least six). Disagreement might still be high in this case, but we found that crowd workers fill in each other's gaps by identifying logics that others may have missed. When aggregating six or more annotations in a weighted fashion, the final labels are comparable with expert evaluations (see Section 5.2).

4. When collecting difficult viewpoint representations such as logics of evaluation, requesters should consider training campaigns for crowd workers to build a pool of knowledgeable and reliable annotators over time (i.e., as proposed by earlier work such as Wais et al. [364]).

5. Asking the crowd workers to justify their answer or describe the viewpoint in their own words has been shown to increase the quality of their annotations [196]. In a workflow setting [59], a crowd worker could use such rationales to approve or reject a certain logic of evaluation provided by a different crowd worker.

5.5.2. Implications

The two-dimensional viewpoint representation we propose has implications for human information interaction research and practical applications that concern user interactions with debated topics. For instance, it may lead to a better understanding of attitude change in web search by providing insight into nuanced shifts of stance. The dynamics of discussions on social media may similarly be studied in more depth when considering which logics of evaluation drive conversations (e.g., to automatically determine where exactly people of different stances disagree). From a practical point of view, ranking bias metrics and re-ranking algorithms may take both dimensions of our proposed viewpoint representation into account, e.g., for a richer notion of viewpoint diversity in a list of recommended news items. In the same way, user interface interventions that aim to mitigate user biases in content consumption could benefit from comprehensive viewpoint representations by taking nuanced stances and logics into account when highlighting, hiding, or explaining documents.

5.5.3. Limitations

Crowd annotators in our study were reliable in annotating tweets' stances but often disagreed regarding logics of evaluation. However, workers were still able to perform as well as expert annotators, whose annotations were also often in disagreement. The discussion session conducted by the experts, however, proved beneficial to reach consensus, and we consider the lack of discussion among crowd annotators as a limitation. Another limitation of our approach is that crowdsourcing studies can become expensive when large amounts of data need to be annotated and even more so when the task is difficult. Finally, we acknowledge that our study considered a limited set of debated topics, and that our results do not necessarily generalize to other controversial issues (i.e., especially those that are much more or less relevant in particular countries or cultural contexts).

5.6. Conclusion

In this chapter, we have proposed a novel, two-dimensional viewpoint representation for human information interaction, inspired by our findings from Chapters 3 and 4 as well as research from the communication sciences. The proposed two-dimensional viewpoint representation consists of a viewpoint's *stance* on a nuanced level which reflects the degree to which a viewpoint opposes or supports a topic, and a viewpoint's *logic of evaluation*, which reflects the perspective behind the stance. We efficiently collected such viewpoint's stances and logics in a crowdsourcing study with acceptable reliability. In a viewpoint diversity analysis and user study, we further showed that our proposed viewpoint representation could be more meaningful in representing diverse opinions on a topic compared to binary stance labels (i.e., *against*/*in favor*). Although we used tweets as a simple use case in this work, our proposed viewpoint representation could also be applied to other document types, such as news articles, podcasts, or web search results. We hope that our work enables researchers and practitioners to represent viewpoints in a more detailed fashion, eventually leading to a better understanding and more effective interventions related to user interactions with debated topics on the web.

Later parts of this dissertation will make use of our novel, two-dimensional viewpoint representation to measure viewpoint bias in web search results (Chapter 9). The viewpoint bias evaluations we conduct in Part III (i.e., Chapters 8 and 9) are another demonstration of how the viewpoint representation we proposed here allows for more comprehensive insights. Before that, however, Part II will cover how to reliably obtain viewpoint labels for search results and similar document types from crowd workers.

II

Crowdsourcing Viewpoint Annotations

Creating training data sets for automatic text classification methods (see Section 2.2.2 and Chapter 3) or obtaining viewpoint labels for which no automatic methods are currently available (e.g., the two-dimensional viewpoint representation we propose in Chapter 5) requires the input of crowd workers [112, 233] (see Section 2.2.3). Such crowdsourcing annotation efforts typically involve collecting at least three judgments per search result from different crowd workers (e.g., asking workers what viewpoint a given search result expresses regarding *school uniforms*) and subsequently aggregating those into single labels. However, crowd workers' *cognitive biases* can strongly reduce data quality from subjective tasks such as annotating viewpoints [96, 156]. One example of this is the *confirmation bias*: workers may be more likely to annotate their personal viewpoint rather than other viewpoints because they pay more attention to document parts that (seem to) confirm their personal opinion [156, 245]. It is vital to identify such cognitive worker biases when collecting viewpoint annotations for search results to prevent data biases and ensure high-quality research and practical applications. That is why Part II of this dissertation addresses the following research question:

$\mathbf{RQ_{II}}$ What cognitive biases reduce crowd workers' abilities to correctly annotate web search results with viewpoint labels?

We begin addressing **$\mathbf{RQ_{II}}$** in Chapter 6 by proposing a checklist to combat cognitive biases in crowdsourcing. Our checklist, adapted from earlier work concerning business decision-making, comprises 12 items referring to particularly common or problematic (groups of) cognitive biases that may reduce the quality of crowdsourced data labels. We present a retrospective analysis of past crowdsourcing papers, showing that cognitive biases are rarely considered but may affect data quality for most tasks. Requesters can use our proposed checklist to inform their task design (e.g., to mitigate cognitive biases) and document potential influences of cognitive biases on the data they collect. Chapter 7 then presents a full application of the checklist to a related, but slightly different, use case: crowdsourced fact-checking of politician statements. We apply our checklist to an existing data set of crowdsourced truthfulness annotations and crowd worker characteristics (e.g., political affiliation, level of education, and annotation confidence) to identify potential influences of cognitive biases in this context. Subsequently, we test our hypotheses by conducting a similar crowdsourcing study while measuring the cognitive biases we had identified. Our findings demonstrate the presence of several cognitive biases (e.g., the *affect heuristic* and *overconfidence*) that may reduce the quality of subjective crowd worker annotations such as truthfulness judgments or viewpoint labels.

6

A Checklist to Combat Cognitive Biases in Crowdsourcing

This chapter is based on a published, full conference paper: Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. “A Checklist to Combat Cognitive Biases in Crowdsourcing”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 9. HCOMP ’21. 2021, pp. 48–59. DOI: `10.1609/hcomp.v9i1.18939`.

Tim Draws primarily planned and carried out the conceptualization, investigation, methodology, project administration, visualization, and write-up of the work described in the paper referenced above. Tim, Alisa Rieger, and Oana Inel collaborated in implementing and describing the retrospective analysis in Section 6.3. The remaining co-authors supervised Tim during the project. All co-authors made edits to the writing.

Conducting high-quality viewpoint bias evaluations for search results or comprehensively examining user interactions with viewpoints in search results requires human-labeled search result data (see Section 2.2.3 and Part I). However, the quality of this (typically crowdsourced) data can be compromised by different types of systemic biases [102, 117]. Prominent biases when crowdsourcing data labels include unequal representations of demographic attributes among annotators [29] or linguistic biases that lead to stereotypical annotations [255]. Another crucial yet relatively less-considered source of poor data quality is *cognitive biases* of crowd workers. Cognitive biases are general human tendencies towards irrationality when making decisions under uncertainty [349], and crowdsourcing is unlikely to be an exception from these tendencies (i.e., as crowd workers typically deal with at least some degree of uncertainty concerning the correctness of the labels they assign). Recent research has indeed shown that cognitive biases such as the *confirmation bias* or *anchoring effect* can negatively affect the quality of crowdsourced annotations [96, 156].

Despite the empirical knowledge of the negative influence of crowd workers' cognitive biases on their annotation quality, requesters typically design crowdsourcing tasks without explicitly considering this influence. Existing data documentation approaches (e.g., Gebru et al. [116]) aim to make (human-labeled) data sets more reliable by clearly describing the process and purpose of data collection but have so far not included cognitive bias assessments. Moreover, although several methods have been proposed to mitigate cognitive biases in crowdsourcing [96, 156], it is currently unclear when different mitigation strategies may be applicable; i.e., there is no protocol by which requesters can identify the specific cognitive biases that may be problematic given a particular task at hand. The large variety and complexity of cognitive biases that have been identified to date make this a difficult space to navigate [150]. Requesters need a practical tool that can help them assess which specific cognitive biases may affect crowd workers in a given task at hand so that targeted assessment and mitigation strategies for these biases can be implemented.

In this chapter, we propose a 12-item checklist, adapted from the domain of business psychology [169], for combating commonly occurring cognitive biases in crowdsourcing. Each item in this checklist targets a different cognitive bias that may affect crowd workers when labeling data. We explain each bias using a running example of a relevance judgment task and demonstrate the practical application of the checklist through a case study on viewpoint annotations for search results. Finally, by carrying out a large-scale retrospective analysis of relevant studies published at the *AAAI Conference on Human Computation and Crowdsourcing* (HCOMP) over the last three years, we found that cognitive biases apply to a vast majority of crowdsourcing studies but are rarely assessed, accounted for, or reported.

Supplementary material such as data sets, task screenshots, and analysis code related to this chapter is available at `https://osf.io/rbucj`.

6.1. Introducing a Checklist

Assessing or controlling for cognitive biases in crowdsourcing is currently not straightforward. Identifying which (and how) specific cognitive biases may harm data quality is important but requires a thorough consideration of the task in combination with potentially problematic cognitive biases. However, a plethora of different cognitive biases have been identified to date [150], and for many cognitive biases, it is still unclear whether or how they may affect crowd workers. This makes navigating the space of cognitive biases extremely complex for requesters. One way to reduce such complexity is to compile a checklist [115].

Kahneman, Lovallo, and Sibony [169] developed a 12-item checklist for combating cognitive biases in business decisions. Given a recommended or planned decision, this checklist aims to assist decision-makers in ensuring that their conclusions are as unbiased as possible. Such business decisions may involve, for instance, overhauling a company's pricing structure or acquiring a competitor. Each question in the checklist targets a different cognitive bias (e.g., the *confirmation bias* or *loss aversion*) that may lead to bad decisions in such situations. The 12 items are meant to cover the majority of potential judgment errors that could occur while ensuring that the checklist is concise and easy to use. Although in this case applied to a business context, cognitive biases are general patterns of behavior that humans exhibit when making decisions under uncertainty [349]. The basic decision heuristics mentioned in the checklist developed by Kahneman, Lovallo, and Sibony [169] therefore apply to crowd workers just as well as to business decision-makers.

We adapted the checklist developed by Kahneman, Lovallo, and Sibony [169] to the context of crowdsourcing human-labeled data by reformulating each of the 12 items to suit the crowdsourcing context. Thus, whereas this adapted checklist practically concerns the same cognitive biases that are mentioned in the original version, it mentions how each of these biases could manifest when conducting a crowdsourcing task. The 12-item checklist we propose is a practical tool that requesters can use to identify potential cognitive biases in the crowdsourcing tasks they design. Each bias in the checklist is accompanied by a guiding question that gives a specific pointer to its applicability. For further illustration, we consider the running example of a simple task in which crowd workers are asked to provide binary relevance judgments on products related to the query "paella pan." We describe the checklist's intended use and future development in the subsections below.

6.1.1. Cognitive-Biases-in-Crowdsourcing Checklist

1. **Self-interest Bias.** *Does my task offer any room for motivated errors?* That is, could crowd workers have some financial, social, or other self-interest-related incentive to judge particular items differently than others? Crowd workers may (subconsciously) fall prey to self-interest bias due to inadvertent incentives and pricing schemes. For example, if workers receive a financial bonus for each "paella pan"-relevant product they find. Other examples include *social desirability* (i.e., when crowd workers are more likely to make incorrect decisions because other people may examine them [14]) and *satisficing* (i.e., exerting only the minimum required amount of effort into conducting a task to save time or resources [172]).

2. **Affect Heuristic.** *Could crowd workers be swayed by the degree to which they 'like' the items they annotate?* For example, crowd workers may be more likely to judge

products of a particular brand they like as relevant, independent from the products' true relevance to “paella pan”. Phenomena such as *priming effects* (i.e., responding differently depending on a previously presented stimulus) and the *familiarity bias* (i.e., greater favorability towards familiar things or concepts) can play a role here [239].

3. **Groupthink or Bandwagon Effect.** *Does my task design give crowd workers some notion of other people's evaluation of the items they annotate?* For example, crowd workers may judge products as more likely to be relevant to “paella pan” when they see that a majority of other crowd workers have judged this product as being relevant or if it has received high ratings from consumers [96].

4. **Salience Bias.** *Could crowd workers' judgments be affected by the salience of particular information?* For example, crowd workers may be more likely to judge products as relevant to “paella pan” if they stand out in an unrelated way (e.g., caps lock titles or high-quality images).

5. **Confirmation Bias.** *Could crowd workers be overly influenced by preconceived notions of the items they annotate?* For example, crowd workers who have a false preexisting idea of what a paella pan is may exhibit confirmation bias if they conduct the task by looking specifically for information that confirms this belief.

6. **Availability Bias.** *Does my task involve judgments related to concepts or people that are likely to elicit stereotypical associations?* For example, crowd workers may be more likely to judge Spanish products as relevant to “paella pan” because they can easily recall numerous examples of the paella dish in Spanish contexts.

7. **Anchoring Effect.** *Is there a possibility that crowd workers overly focus on a specific reference point (i.e., an anchor) when making judgments?* For example, if the first of several products that crowd workers are exposed to are clearly not paella pans (e.g., products unrelated to kitchenware), the first item that somewhat resembles a paella pan (e.g., a regular saucepan) may be more likely to be judged as relevant compared to when the same item was shown in a sequence of actual paella pans. Note that the anchoring effect can also occur within a single human intelligence task (HIT); e.g., when workers are overly influenced by the first information they see (i.e., *primacy effect*), such as the product title, or the last information they see before making their judgment (i.e., *recency effect*).

8. **Halo Effect.** *Does my task involve judgments that could be influenced by irrelevant pieces of information?* For example, crowd workers may be more likely to judge products as relevant to “paella pan” if these products seem suitable for similar dishes (e.g., risotto). This encompasses related biases such as the *decoy effect*, where the choice between two options is affected by the introduction of a (potentially irrelevant) third choice, or the *ambiguity effect*, where (potentially irrelevant) missing information affects crowd workers' decision-making [96].

9. **Sunk Cost Fallacy.** *Is the time required to complete my task and what it requires from crowd workers clear at the onset?* The more time and effort crowd workers

invest in a task, the more they may want to complete it, despite potentially already having lost interest in the task. This is undesirable as uninterested crowd workers may abandon a task after investing efforts or complete the task with sub-optimal performance [137]. For example, assuming that crowd workers have to annotate the relevance of 50 different products before completing the task but are not aware of the task length beforehand, their performance may deteriorate in the later stages.

10. **Overconfidence or Optimism Bias.** *Is there a possibility that crowd workers overestimate their ability to perform my task?* For example, it arguably takes a particular level of cooking knowledge to distinguish a paella pan from a regular frying pan or wok. Crowd workers who have never learned about these distinctions may not perceive the task of assigning "paella pan"-relevance judgments to products as hard but may actually not be skilled enough to give high-quality annotations here. This is related to the *Dunning-Kruger effect*, which posits that people with low ability concerning a task tend to be overconfident about their projected performance in it [111, 192].

11. **Disaster Neglect.** *Have crowd workers who commit to my task, been properly informed about the consequences of their participation?* The task selection process is often fairly arbitrary, which means that workers may not realize potential negative effects of committing to a task that they don't have expertise on [95]. For example, crowd workers may commit to doing "paella pan"-relevance judgments for products on a whim without considering the potential reputation loss and bad annotation quality that could follow if they do not perform well.

12. **Loss Aversion.** *Does my task design give crowd workers a reason to suspect that they may not get paid (fairly) after executing my task?* Due to loss aversion, crowd workers may not select such tasks or abandon them early, leading to a skewed distribution of participants or task starvation [103]. For example, if a crowd worker suspects that annotating products in a task will only earn them money if they perform at a particular level, they may abandon the task early to avoid wasting their time and effort [137].

6.1.2. How to Use the Proposed Checklist

Here, we give a few pointers regarding the checklist's usage.

When should I apply this checklist? The optimal point to use the checklist is *before data collection*. This allows requesters to not only alert themselves to potential limitations of the data to be collected but also allows for appropriate changes to the task design. If the data have already been collected, requesters may, however, still use the checklist to determine whether cognitive biases may have affected the data in some way (i.e., led to poor data quality or whether the data potentially encodes said biases). The checklist we propose can thus also augment data documentation approaches such as *data sheets* [116].

I applied the checklist to my task design and found at least one potential cognitive bias — now what? The identification of at least one potential cognitive bias in a task design at hand may call for three different actions. First, requesters may want to use this information to *assess* the influence of the identified cognitive biases. The aim behind this would be to check whether these biases truly affect crowd workers during the task. Second, requesters may adapt their task design to *mitigate* the identified cognitive biases. Such adaptations could –at least in some cases– be an easy way to increase data quality without compromising the task design in meaningful ways or vastly elongating the task. Third, especially if data have already been collected from the task at hand, requesters may use the checklist to better *document* their data sets by providing detailed limitations. Pointing out specific cognitive biases that may have affected crowd workers can contribute towards a more accurate data description and thereby make data more reusable. We discuss each of these three actions in more detail below.

How can I *assess* the influence of cognitive biases in my task? Suppose we conclude that our task on relevance judgments for products with respect to the term “paella pan” potentially elicits the *affect heuristic*: we suspect that crowd workers may be more likely to judge products as relevant if they like those products. Previous research suggests that monitoring crowd workers’ biases is best done during data collection [118]. We may thus enhance the task design by collecting additional metadata to assess whether crowd workers make erroneous judgments due to the *affect heuristic*. Specifically, we could add an item that measures the degree of crowd workers’ personal favorability towards each product they annotate. This would then allow us to approximate the influence of the *affect heuristic* in multiple different ways. For example, we may use a quantitative measure that compares how crowd workers rate items of high and low favorability or conduct a statistical hypothesis test that assesses whether there is a relationship between product favorability and relevance judgments.

How to exactly measure or test for cognitive biases in this context has to be decided individually per suspected bias and the particular crowdsourcing task at hand. To the best of our knowledge, no standard assessments for particular cognitive biases exist in this space. It is nevertheless important to decide on a specific criterion that establishes whether (and perhaps to what degree) bias is present so that appropriate action can be taken. Below are a few pointers to potential ways of developing such a criterion:

- *Statistical hypothesis tests* are a straightforward way to analyze the presence of systematic patterns in data (e.g., differences between groups or correlations). A caveat of this approach is that failing to reject a null hypothesis may not necessarily mean that no bias is present. That is why we recommend considering not only classical null hypothesis significance testing (i.e., where null hypotheses may be rejected after examining the p-value) but also Bayesian hypothesis testing, which allows for quantification of evidence in favor of either null or alternative hypotheses [363].

- *Self-created or adapted metrics* can be used to quantitatively measure patterns or occurrences in data. Here, it is useful to set one or multiple specific thresholds that reflect bias severity before data collection. This can help to decide when the degree of bias is too extreme.

- Statistical techniques such as *structural equation modeling (SEM)* [350] or *network analysis* [98] may be used to analyze relationships between several factors simultaneously.

It should be pointed out that any such test or metric will only approximate the true, latent influence of the cognitive bias one may wish to assess for. Therefore, we recommend constructing a procedure that consists of several tests and measurements, which build the criterion together. Another useful approach may be to add sanity checks (e.g., by manually evaluating samples of individual cases that show high and low bias according to the criterion). Note also that many statistical procedures (i.e., especially in hypothesis testing) underlie assumptions [253]. For instance, to satisfy the assumption of independence of observations, data may have to be aggregated per crowd worker before conducting a hypothesis test. Requesters should further be aware of common pitfalls in hypothesis testing, such as misinterpretation of the p-value or statistical power [129].

How can I *mitigate* the influence of cognitive biases in my task? Earlier work has already explored the mitigation of cognitive biases in crowdsourcing tasks. For instance, Eickhoff [96] showed –through the lens of a standard relevance judgment task– how requesters may deal with biases related to *groupthink*, *anchoring*, and the *halo effect*. Hube, Fetahu, and Gadiraju [156] investigated how requesters could preempt *confirmation bias* when crowdsourcing subjective judgments related to opinions on debated topics. Next to adapting the task design, requesters may consider improving the data (i.e., the item selection) or changing the worker requirements. Especially difficult tasks may sometimes require non-ambiguous items or particularly qualified workers. Eventually, however, mitigating cognitive biases in crowdsourcing will often require a unique solution that fits the particular task design and suspected cognitive bias at hand. We recommend combining any mitigation efforts with assessments for the suspected cognitive biases to ensure that they have been mitigated successfully.

How can I *document* the influence of cognitive biases in my task? Especially if data have already been collected when applying the checklist, requesters may wish to at least document the potential influence of cognitive biases to make their data more reusable. We recommend augmenting the checklist with general data documentation approaches such as *data sheets* [116]. Requesters can add the checklist we propose under a separate section in the data documentation and discuss each bias's potential influence.

Further development and context of this checklist. A few more things should be pointed out to put the usage of this checklist into perspective. First, the checklist – as we propose it in this chapter – is unlikely to be exhaustive. We expect that novel research will demonstrate how cognitive biases that we do not yet mention can affect crowd workers. That is why, in contrast to the original checklist developed by Kahneman, Lovallo, and Sibony [169], we host the latest version of the checklist we propose on an online repository that is open to anyone's contributions.[1] Second, in contrast to our running example, it is unlikely that all of the mentioned biases occur in every crowdsourcing task. We merely posit that any of the 12 mentioned biases could (but do not necessarily do) take place in crowdsourcing. Third, we recommend using more general data documentation approaches such as *data sheets* [116] in tandem with this checklist. Answering questions about the population of crowd workers or the purpose of the (to-be-)collected data set can help distill potential issues. If the collected data is part of a larger study, we recommend preregistering the research project [247].

---

[1] `https://osf.io/rbucj`

6.2. Case Study: Viewpoint Annotations for Search Results on Debated Topics

This section demonstrates the practical application of the checklist we propose at the hand of a case study. Our aim in this case study was to collect viewpoint annotations from crowd workers for search results on debated topics.[2] Such data is useful, for example, when aiming to measure viewpoint bias in ranked search result lists (see Part III) or study the effects of viewpoint-biased search result rankings on user attitudes (see Part IV). We had retrieved search results on nine different debated topics from *Bing*:[3]

1. *Are social networking sites good for our society?*
2. *Should zoos exist?*
3. *Is cell phone radiation safe?*
4. *Should bottled water be banned?*
5. *Is obesity a disease?*
6. *Is Drinking Milk Healthy for Humans?*
7. *Is Homework Beneficial?*
8. *Should People Become Vegetarian?*
9. *Should Students Have to Wear School Uniforms?*

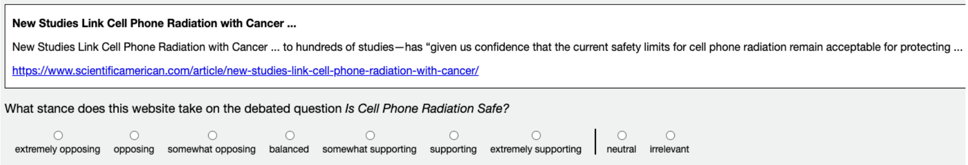
We designed a task wherein crowd workers would be randomly assigned to one of the nine debated topics and see a set of search results related to it. The search results were presented similarly compared to regular search engines (i.e., with a title, snippet, and clickable URL; see Figure 6.1). Crowd workers would be asked to label each search result for its viewpoint towards the debated topic. Table 6.1 shows the viewpoint representation we considered: a one-dimensional taxonomy of the overall stance that a document expresses, ranging from “strongly opposing” to “strongly supporting.”[4] Crowd workers would be tasked to annotate the stance of each search result on seven-point Likert scales (i.e., “What stance does this website take on the debated question [*topic*]?”). We also included attention checks between the search results, in which we specifically instructed participants on what option to select on the Likert scale. Full data sheets for the data we collected from this task are available on our repository.

---

[2] We had first collected these data to study user behavior in web search on debated topics; see Chapter 10 and Rieger et al. [298].

[3] `https://bing.com`

[4] We included two additional options, *neutral* and *irrelevant* (see Figure 6.1), for search results that did not express any stance or that were found to be irrelevant to the topic, respectively.



New Studies Link Cell Phone Radiation with Cancer ...

New Studies Link Cell Phone Radiation with Cancer ... to hundreds of studies—has “given us confidence that the current safety limits for cell phone radiation remain acceptable for protecting ...

https://www.scientificamerican.com/article/new-studies-link-cell-phone-radiation-with-cancer/

What stance does this website take on the debated question *Is Cell Phone Radiation Safe?*

○ extremely opposing ○ opposing ○ somewhat opposing ○ balanced ○ somewhat supporting ○ supporting ○ extremely supporting | ○ neutral ○ irrelevant

Figure 6.1: Example item to collect viewpoint (i.e., stance label) annotations for search results in our case study.

Table 6.1: The stance label taxonomy we considered as viewpoint representation in our case study. Crowd workers could assign a stance label to each search result by selecting one of the seven options ranging from “strongly opposing” ($l = -3$) to “strongly supporting” ($l = 3$).

l	**Label**	**Example (topic: “Zoos should exist”)**
-3	strongly opposing	“Horrible places! All zoos should be closed ASAP.”
-2	opposing	“We should strive towards closing all zoos.”
-1	somewhat opposing	“Despite the benefits of zoos, overall I’m against them.”
0	balanced	“These are the main arguments for and against Zoos.”
+1	somewhat supporting	“Although zoos are not great, they benefit society.”
+2	supporting	“I’m in favor of zoos, let’s keep them.”
+3	strongly supporting	“There is nothing wrong with zoos – open more!”

From walking through the checklist before data collection, we could derive that crowd workers’ judgments may be affected by three cognitive biases in our task:[5]

1. **Confirmation bias.** We suspected that crowd workers’ preexisting attitudes on their assigned topics may affect their annotations. Specifically, we were concerned that crowd workers might interpret their own attitudes into the content they see (especially for ambiguous search results).
2. **Anchoring bias.** Another concern was that crowd workers’ first judgment would act as a reference point for the search results to come and thus affect subsequent annotations. Practically, this would have meant that crowd workers’ judgments tend towards whatever annotation they gave to the first item they saw.
3. **Halo effect.** We also suspected that crowd workers’ preexisting knowledge of their assigned topics may affect their annotations. A halo effect could have occurred if crowd workers have strong preconceived notions about particular subtopics or search result sources, causing them to prematurely rate search results as more extreme (i.e., placing search results into the “opposing camp” or “supporting camp”).

We decided to conduct a pilot study to collect annotations for search results on two of the nine debated topics (i.e., *Should zoos exist?* and *Are social networking sites good for*

---

[5] Arguably, other biases that are mentioned in the checklist (e.g., the affect heuristic or the availability bias) could have affected crowd workers’ judgments as well. For conciseness, however, we keep it to the three biases we mention here.

our society?) while assessing the cognitive biases mentioned above. We enhanced our task design with two additional items that collect the necessary contextual metadata. First, to be able to assess the confirmation bias, we measured crowd workers' *personal stance* (i.e., on a seven-point Likert scale ranging from "strongly opposing" to "strongly supporting") on their assigned topic. Second, to enable an assessment of the halo effect, we measured crowd workers' *perceived knowledge* (i.e., on a seven-point Likert scale ranging from "non-existent" to "excellent") of their assigned topic. Assessing the anchoring effect did not require collecting additional metadata.

We published our task on *Amazon Mechanical Turk*[6] to collect stance label annotations for all 643 search results related to the two topics mentioned above. We recruited workers who were located in the United States and had a task approval rate of at least 95%. Crowd workers were paid \$2 for completing the task and could earn a \$0.50 bonus if they clicked on at least half of the links provided in the search results and if they passed both attention checks. We excluded annotations from crowd workers who did not pass at least one of the attention checks.

The data set collected in this pilot study (D_1) contains 1994 annotations from 109 different crowd workers for 643 different search results. Each search result in D_1 pertains to either of the two debated topics *Should Zoos Exist?* or *Are Social Networking Sites Good for Our Society?* and was annotated by two to eleven different crowd workers. Specifically, whereas 92% of search results received three stance annotations, 2% received only two, and 6% received four or more annotations. The low inter-rater reliability between crowd workers who annotated D_1 (Krippendorff's $\alpha = 0.21$) indicates that D_1 contains considerable amounts of noise.[7] Applying the cognitive-biases-in-crowdsourcing checklist and testing for cognitive biases is one way to investigate possible contributing factors to this low data quality.

We conducted several statistical hypothesis tests on D_1 to analyze whether (1) the *confirmation bias*, (2) the *anchoring effect*, or (3) the *halo effect* might have had an influence on crowd workers' annotations.

Confirmation bias. To check whether there was confirmation bias, we conducted classical and Bayesian correlation analyses between crowd workers' pre-existing stance on their assigned topic and their mean stance annotation. We found a significant Spearman correlation ($\rho = 0.27$, $p = 0.002$) and strong evidence in favor of a correlation as part of a Bayesian correlation analysis ($\text{BF}_{10} = 12.49$).[8]

Anchoring effect. We analyzed the influence of a potential anchoring effect by conducting classical and Bayesian hypothesis tests, this time between crowd workers' first annotation and the mean of their remaining annotations. Here, we found a significant Spearman correlation ($\rho = 0.31$, $p < 0.001$) and extreme evidence in favor of a correlation as part of the Bayesian correlation analysis ($\text{BF}_{10} = 195.33$).

---

[6] https://mturk.com

[7] Krippendorff's alpha accounts for missing annotations, so items can vary in terms of how many people annotated them.

[8] We used the *R* package *BayesFactor* [237] to perform Bayesian analyses. We interpret the strength of evidence from Bayes Factors in line with the guidelines proposed by Lee and Wagenmakers [201], who adapted them from Jeffreys [167].

Halo effect. To analyze whether there might have been a halo effect, we compared the range of annotations between crowd workers with lower versus higher knowledge on the topic. We defined "lower knowledge" as low or medium self-reported knowledge on their assigned topic (i.e., the bottom two and central three options on the Likert scale), and "higher knowledge" with those who indicated the top two options on the Likert scale. We did not find any evidence in favor of such a difference as the classical t-test was not significant ($t = 0.42$, $p = 0.68$). A Bayesian t-test revealed moderate evidence in favor of the null hypothesis that there was no difference between these groups ($\mathrm{BF}_{01} = 4.61$).

The analyses we conducted on D_1 rang the alarm bell for a potential influence of confirmation bias as well as the anchoring effect. We aimed to use this knowledge to inform the collection of labels for search results on all nine debated topics. Further exploratory analyses (e.g., looking at the agreement on different types of search results) led us to suspect that the main source of bias in our crowdsourcing task may have been ambiguous items. Whereas search results that took a clear stance (e.g., "Why Zoos Are An Important Part Of Responsible Wildlife") were often rated quite unanimously, workers diverged when confronted with less strongly opinionated search results (e.g., "Are Zoos Good Or Bad For Animals/Wildlife?"). We thus decided to manually select only those search results that we judged as non-ambiguous or opinionated for our final data collection. Moreover, we suspected that we may have had underestimated the difficulty of the task, so we increased the worker requirements to a HIT approval rate of greater than 98% as well as *Master* status at MTurk.[9] This considerably shrunk our pool of potential crowd workers, so we eased restrictions on workers' location by including other countries where English is spoken as the first or second language by most people (e.g., Australia and Germany). With these changes, we again published our task on Amazon Mechanical Turk.

Selecting only opinionated search results for the nine debated topics resulted in a data set of 480 different search results (approximately balanced over the nine topics). A total of 56 crowd workers provided 1499 stance annotations for this second, final data set of search results (D_2). Each search result pertained to just one of the nine different debated topics and was annotated by three to seven different crowd workers. Inter-rater reliability for this data set is satisfactory (Krippendorff's $\alpha = 0.79$). In contrast to D_1, we did not find evidence for the confirmation bias ($\rho = -0.07$, $p = 0.68$) or the anchoring effect ($\rho = 0.04$, $p = 0.76$) in D_2. Bayesian analyses revealed that the null hypotheses (i.e., that there were no confirmation bias and no anchoring effect) explained the data better than the respective alternative hypotheses ($\mathrm{BF}_{01} = 2.31$ and 3.31, respectively).

6.3. Retrospective Analysis

Although the examples and case study we have presented so far relate to specific use cases of crowdsourcing subjective judgments (e.g., relevance judgments), there is reason to expect that cognitive biases occur across different types of crowdsourcing tasks. Cognitive biases are general phenomena that occur when humans make decisions under uncertainty [349], and the checklist we propose covers several different ways in which biases could affect crowd workers (e.g., related to their personal gains, losses, or abilities as well as simple heuristics they may apply while conducting the task). Therefore, in this section, we apply our proposed checklist to a set of 27 recent research papers in the crowdsourcing domain. We assess which biases may have been present in the reported studies and whether their (potential) influence was reported upon. By means of this analysis, we aim to show that cognitive biases are often impactful while their influence is not considered in crowdsourcing task designs and publicly available data sets that contain human judgments.

---

[9] Amazon MTurk awards particularly well-performing crowd workers with a *Master* status. This acknowledges the high quality these workers deliver and allows them to earn higher rewards.

6.3.1. Paper Selection Criteria

We selected research papers for this analysis based on four criteria that all needed to be met for a paper to be included:

1. We selected papers from the 2018, 2019, and 2020 *AAAI Conference on Human Computation and Crowdsourcing* (HCOMP) proceedings, as HCOMP is among the most important venues for research in this area.
2. Papers had to include an online crowdsourcing study in which data was collected (i.e., not using external data).
3. The crowdsourcing task(s) described in the paper had to concern some form of labeling or evaluating data objects in a constrained, closed format (i.e., the crowd workers were given a well-defined answer space). For example, we would include a task that asked crowd workers to judge products as “relevant” or “non-relevant” to the term “paella pan” but we would exclude a task that asked crowd workers to describe products in open text fields.
4. We only included papers in which crowd workers were paid for completing the described task(s).

The selection criteria above were developed by three authors of the paper that this chapter is based on, who acted as independent experts in this study. Using a test sample of 16 papers, the experts ensured that they reached agreement on which papers should be included or excluded based on the four criteria. The final selection procedure resulted in a set of 27 papers (i.e., 4 from 2018, 13 from 2019, and 10 from 2020) that we included in this analysis. We do not report inter-rater reliability, as disagreement between the researchers was resolved through detailed discussions and critical reflection [224].

6.3.2. Method

Each of the three experts who also co-decided on the inclusion criteria subsequently analyzed each of the 27 selected papers in a two-step process. After reading the paper, including the described task design, task instructions, and crowd selection criteria, they first went through each item in the checklist and marked which cognitive bias could have affected the results. Here, the expert would consider the textual description of the task as well as additional available materials (e.g., screenshots). Each bias could be marked with either “yes” (i.e., if there was good reason to assume that the bias may have occurred) or “no” (i.e., if it was impossible or unlikely that the bias had occurred). Therefore, note that a “yes” here did not necessarily mean that crowd workers were indeed affected by the bias, but merely that such an influence could not be ruled out based on the provided task description and additional materials.

As a second step, the expert stated whether the paper at hand discussed the potential influence of cognitive biases on the results. The options here were “yes” (i.e., if the paper identified and at least discussed *all* possible cognitive biases that may have taken place), “partly” (i.e., if the paper at least discussed a subset of the potential cognitive biases), or “no” (i.e., if the paper did not consider any cognitive biases as a potential influence on the results). Thus, if a paper discussed the potential influence of cognitive biases on the collected data at all, it would receive a “yes” or “partly” label, depending on whether it mentioned all or just a subset of the potential biases identified by the expert. We included this additional label to gauge the degree to which requesters are considerate of such influences on data quality. While it may be difficult to rule out or fully mitigate the influence of cognitive biases on crowdsourcing tasks, discussing potential influences is important information for anyone who may want to use or build on the data set or the published research.

We used majority voting to aggregate the judgments corresponding to the three independent experts. For example, if two of the three experts judged “no” for a particular bias in a particular paper, we would adopt this label for this data point.[10] This resulted in a set of 13 labels per paper (i.e., one for each of the 12 cognitive biases from the checklist as well as the overall judgment on whether the paper considered cognitive biases). Note that this analysis did not concern the methods or evaluations presented in those papers but merely the task design they described.

6.3.3. Results

Table 6.2 shows the results of our retrospective analysis of crowdsourcing papers at the AAAI HCOMP conference from the last three years. We identified each cognitive bias from the checklist in at least some of the papers we analyzed. Whereas the *salience bias* (93%), *anchoring effect* (81%), and *halo effect* (78%) were marked rather often, biases such as the self-interest bias (30%), loss aversion (22%), or groupthink (15%) were identified comparatively seldom. We also found that some biases often co-occurred in our analysis. Specifically, the *confirmation bias* and *availability bias* as well as *overconfidence* and *disaster neglect* were most often identified for the same papers.

Eight out of the 27 analyzed papers at least partly considered cognitive biases in their task design or discussion. For instance, Otterbacher et al. [256] show how cognitive biases and stereotypes can affect image labeling and Peng et al. [266] discuss at length how cognitive biases may affect the hiring process. Mohanty et al. [235] and Kemmer et al. [178] acknowledge that a variety of biases, such as the confirmation bias, can lead to low-quality data labels and propose methods to mitigate these effects.

Note that we intentionally do not disclose which potential cognitive biases had been identified per paper. We wish to point out that this retrospective analysis is not meant to discredit the work of others. Instead, we performed this analysis to show (a) that cognitive biases can occur in a variety of crowdsourcing tasks, (b) that the influence of cognitive biases in crowdsourcing is rarely considered, and (c) that the checklist we propose in this

---

[10] No conflicts arose for the last (3-option) label as there was a majority judgment for all 27 selected papers.

Table 6.2: Results of the retrospective analysis of cognitive biases in crowdsourcing papers from HCOMP proceedings in 2018, 2019, 2020. Here, *biases considered* refers to papers that discussed the identified cognitive biases at least partly.

Bias	**2018**	**2019**	**2020**	**Total**
Self-interest Bias	0 (0%)	5 (38%)	3 (30%)	8 (30%)
Affect Heuristic	2 (50%)	8 (62%)	5 (50%)	15 (56%)
Groupthink	1 (25%)	2 (15%)	1 (10%)	4 (15%)
Salience Bias	4 (100%)	11 (85%)	10 (100%)	25 (93%)
Confirmation Bias	3 (75%)	8 (62%)	5 (50%)	16 (59%)
Availability Bias	4 (100%)	10 (77%)	5 (50%)	19 (70%)
Anchoring Effect	4 (100%)	9 (69%)	9 (90%)	22 (81%)
Halo Effect	4 (100%)	11 (85%)	6 (60%)	21 (78%)
Sunk Cost Fallacy	3 (75%)	6 (46%)	2 (20%)	11 (41%)
Overconfidence	3 (75%)	9 (69%)	3 (30%)	15 (56%)
Disaster Neglect	1 (25%)	6 (46%)	2 (20%)	9 (33%)
Loss Aversion	0 (0%)	5 (38%)	1 (10%)	6 (22%)
Biases Considered?	1 (25%)	5 (38%)	2 (20%)	8 (30%)

chapter is widely applicable and could assist researchers in identifying these potential biases.

6.4. Discussion

In this chapter, we have proposed a 12-item checklist to combat the negative influence of crowd workers' cognitive biases on the quality of crowdsourced data labels. Each item in this checklist refers to a different, commonly occurring cognitive bias that may affect crowd workers' judgments and thereby reduce data quality. Requesters may use our proposed checklist before or after data collection to identify, mitigate, and describe cognitive biases that may influence crowd workers in the tasks they design. To clarify the intended use of the checklist, we have demonstrated its practical application at the hand of a case study on viewpoint annotations for search results. We further showed in a retrospective analysis of recently published crowdsourcing studies that our proposed checklist is widely applicable and that most crowdsourcing studies currently do not consider the influence of cognitive biases on the data labels they obtain.

6.4.1. Limitations

Requesters (i.e., those who design and publish crowdsourcing annotation tasks) may apply the checklist we propose to their crowdsourcing tasks but should be aware of at least three important limitations. First, the checklist is unlikely to be exhaustive: several cognitive biases relevant to crowdsourcing may still be missing from it. That is why we set up an online repository that will always host the latest version of the checklist and provide an opportunity for contributors to suggest edits. The repository is available at `https://osf.io/rbucj`. Second, although our proposed checklist can help requesters identify potential cognitive biases that may affect the crowd workers they employ, it does not (yet) give direct recommendations regarding the measurement and mitigation of these biases. We give some pointers in this chapter on how the influence of cognitive biases could be assessed or mitigated in certain situations, and previous work has already proposed some further mitigation strategies [96, 156]. However, more research is needed to develop robust procedures that can deal with all the different cognitive biases in our checklist. Third, requesters should be aware that cognitive biases can, in some cases, be beneficial. The *bandwagon effect*, for instance, is often harnessed to promote collaboration between crowd workers, which can indeed increase data quality [186].

6.4.2. Implications

The checklist we propose in this chapter has implications for task design as well as data documentation in the crowdsourcing context. As we have discussed, requesters may use this checklist to assess and mitigate cognitive biases. This predominantly concerns adaptations to the task design itself (e.g., adding the collection of contextual metadata) but can also involve item selection or adapting the worker requirements. Furthermore, requesters can use our proposed checklist to document (the limitations of) the data they collect. The checklist is applicable to a wide range of crowdsourcing task types, including (but not limited to) validation tasks such as data matching, interpretation and analysis tasks such as relevance judgments, and surveys (e.g., opinion gathering).[11] Although following the procedures we suggest in this chapter may increase costs (e.g., due to elongating tasks) and deployment time (e.g., due to prolonged time needed to fine-tune the tasks), we believe that high data quality and reliability should be any requester's primary aim – especially when the data has a potentially high impact on individuals and society. This is particularly important to facilitate the appropriate reuse of data collections.

Initial steps have been taken towards defining a taxonomy of relevant attributes to report on crowdsourcing studies, such as the employed crowd, the task shown to the workers, the applied quality control mechanisms, and the experimental design [288, 289]. We believe that cognitive biases are an additional factor to consider in reports on crowdsourcing studies. Our retrospective analysis suggests that requesters should also clarify such aspects to the crowd if they aim to mitigate cognitive biases effectively. In particular, the *sunk cost fallacy* could be mitigated by providing the estimated duration of the task in the description of the task. Rejection criteria are also essential for crowd workers in deciding whether they continue to work on a given task (i.e., *loss aversion* and *disaster neglect*). Thus, our retrospective analysis suggests that some aspects that are recommended for reporting on crowdsourcing studies should be included in the actual task design and instructions. This would lead to increased requester-crowd transparency while mitigating several cognitive biases.

---

[11] We here refer to the taxonomy of microtasks on the web proposed by Gadiraju, Kawase, and Dietze [112].

6.5. Conclusion

Viewpoint bias evaluations and studies examining user interactions with debated topics in web search require high-quality viewpoint labels for search results. Such viewpoint labels often have to be collected in crowdsourcing annotation procedures, where cognitive biases of crowd workers can limit the resulting data quality. However, cognitive biases in the crowdsourcing context are typically hard to identify, assess, and mitigate. This chapter has taken a step toward tackling cognitive biases in crowdsourcing by proposing a simple, 12-item checklist for deciding whether (and how) some of the most commonly occurring cognitive biases may have an undesired influence in crowdsourcing tasks. Requesters and researchers can use this tool to improve their task designs and acknowledge cognitive biases as potential sources of sub-optimal data quality where necessary. This is particularly important when aiming to collect high-quality viewpoint labels for search results, which may be difficult to assign depending on factors such as document length and viewpoint salience. In the next chapter, at the hand of a similar use case (i.e., truthfulness judgments for politician statements), we dig deeper into specific worker traits and cognitive biases that can reduce annotation quality.

7

Identifying Crowd Worker Biases in the Context of Debated Content

---

This chapter is based on a published, full conference paper: Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. "The Effects of Crowd Worker Biases in Fact-Checking Tasks". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 2114–2124. DOI: `10.1145/3531146.3534629`.

Tim Draws collaborated with David La Barbera, Michael Soprano, and Kevin Roitero in planning and carrying out the conceptualization, investigation, statistical analyses, project administration, visualization, and write-up of the work described in the paper referenced above. David La Barbera and Michael Soprano carried out the crowdsourcing study. The remaining co-authors supervised the project and made edits to the writing.

Our proposed 12-item checklist to combat cognitive biases in crowdsourcing enables requesters to independently assess the tasks they design and thereby collect higher-quality viewpoint labels for search results and similar documents (see Chapter 6). However, it is currently still unclear what specific worker traits or cognitive biases generally play a role in this context. There is a need to identify these factors and create guidelines for requesters who wish to crowdsource reliable annotations for debated content (e.g., opinionated news articles or social media posts). To support these efforts, this chapter explores what specific worker traits and cognitive biases may reduce annotation quality in such scenarios. Our use case for this chapter is crowdsourcing *truthfulness judgments*: identifying to what degree a document contains misinformation. This task is slightly different from annotating viewpoint labels but similarly involves debated content as well as some extent of crowd worker subjectivity (e.g., due to worker opinions).

Although experts are considered the most reliable truthfulness annotators, recent research has shown that crowd workers can also reliably perform such fact-checking tasks [197, 300, 301, 302, 303] and assess information quality across multiple truthfulness dimensions or quality aspects [216, 334, 347]. Crowdsourced fact-checking is now widely used in academic research [267, 273, 321, 322, 367] and has already found applications in industry [12, 279]. However, because crowdsourcing often relies on contributions from large groups of laypeople with different backgrounds, expertise, and skills, (subconscious) biases among those workers may reduce the quality of their annotations [96, 156]. For example, in fact-checking tasks, factors such as workers' political affiliation or general trust in politics may affect their ability to recognize misinformation.

Identifying systematic biases in crowdsourced fact-checking is an important issue. Because expert-provided assessments are expensive and slow to gather, crowdsourced truthfulness judgments are often used as training sets for supervised machine learning methods. The presence of bias in training data may lead to bias in the classification performed by these systems. Moreover, such biases might affect the accuracy (or even question the feasibility) of human-in-the-loop hybrid systems that try to identify misinformation at scale by combining experts, crowd, and automatic machine learning systems [73]. Unveiling these systematic biases would support a more reliable collection of crowdsourced training data and enable bias mitigation methods for existing data sets.

In this chapter, we investigate systematic biases that may decrease the quality of crowdsourced truthfulness judgments for politician statements. Three research questions guide our work:

$\mathbf{RQ_{II.1}}$ What individual characteristics of crowd workers and statements may lead to systematic biases in crowd workers' truthfulness judgments?

$\mathbf{RQ_{II.2}}$ What cognitive biases can affect crowd workers' truthfulness judgments?

$\mathbf{RQ_{II.3}}$ Are different truthfulness dimensions affected by different biases?

To address these research questions, we first conducted an exploratory study on an earlier collected data set containing crowdsourced truthfulness judgments for political statements (Section 7.2). These data also contain information on the political leaning of statements as well as individual worker characteristics (e.g., workers' level of education and political leaning). We used the findings from these exploratory analyses to formulate specific hypotheses concerning which individual worker and statement characteristics (**RQ**$_{\mathbf{II.1}}$) and what cognitive worker biases (**RQ**$_{\mathbf{II.2}}$) may affect the accuracy of crowd workers' truthfulness judgments. To test these hypotheses, we subsequently conduct a new, preregistered crowdsourcing study (Section 7.3). Our findings suggest that crowd workers' degree of belief in science matters in this context (**RQ**$_{\mathbf{II.1}}$), that workers generally overestimate truthfulness, and that cognitive biases such as the *affect heuristic* and *overconfidence* can reduce their annotation quality (**RQ**$_{\mathbf{II.2}}$). We also find exploratory evidence that different truthfulness dimensions may be affected by these biases to different degrees (**RQ**$_{\mathbf{II.3}}$; Section 7.4).

Supplementary material such as data sets, task screenshots, and analysis code related to this chapter is available at `https://osf.io/8yu5z`.

7.1. Crowdsourced Fact-Checking in Earlier Work

To allow fact-checking tasks to scale and keep up with the large amounts of information posted online, previous research has studied methods to address the misinformation issue using crowd-powered systems. Many of those studies employed crowdsourcing to collect *truthfulness* judgments [273, 321, 357]. For example, La Barbera et al. [197], extending earlier work [300], studied the effect of both judgment scale and assessor bias when fact-checking political statements. Their work demonstrated that coarse-grained scales are preferred by workers and that workers' political background is the main bias influencing workers' ability to effectively assess misinformation statements.

Roitero et al. [301] used crowdsourcing to collect thousands of truthfulness labels on multiple data sets for political fact-checking, employing different scales. They found that adjacent categories in the assessment scale can be grouped together to increase both worker effectiveness and agreement and that different scales lead to similar agreement levels. More recently, Soprano et al. [334] re-assessed Roitero et al.'s [301] statements. Breaking down truthfulness on a multidimensional scale, they found that using multiple dimensions measures different aspects of the misinformation statement evaluated by the crowd workers. Roitero et al. [303] focused on fact-checking statements related to the COVID-19 pandemic. Besides reporting results on crowd effectiveness and agreement, they performed a longitudinal study and presented an in-depth study on how the crowd's effectiveness changes when it is asked to perform fact-checking over different time spans. They also provide a failure analysis to investigate the statements that are mislabeled by crowd workers. Epstein, Pennycook, and Rand [101] deployed a survey to 1000 Americans to study their perceived trust in popular news websites, finding that mainstream sources are usually more trusted than fact-checking websites or hyper-partisan sources. Bhuiyan et al. [34] adopted a similar approach by surveying students enrolled in journalism or media programs about information dealing with climate change. Ghenai and Mejova [122] used crowdsourcing and machine learning to track misinformation on Twitter. Pennycook and Rand [268] crowdsourced news source quality labels. Giachanou and Rosso [124] developed a tutorial on online misinformation and fact-checking, with a focus on social media data.

7.2. Exploratory Study

To identify specific hypotheses concerning our research questions, we conducted an exploratory study using a publicly available data set. This section details this exploratory study and describes the hypotheses we formulated as a result.

7.2.1. Data

We conducted our exploratory study on a data set collected and published by Soprano et al. [334].[1] The data set is composed of crowdsourced truthfulness judgments for 180 statements from two political fact-checking websites: *Politifact* [367] and *ABC*.[2] *Politifact* is a collection of more than 10000 statements from mainly US politicians, labeled by experts on a six-level truthfulness scale containing the categories *pants-on-fire*, *false*, *barely-true*, *half-true*, *mostly-true*, and *true*. *ABC* is a collection of more than 500 statements on Australian politics that are first labeled by experts on a fine-grained semantic scale with more than 30 levels and then mapped into a three-level scale with the labels *negative*, *in-between*, and *positive*. The 180 statements in the data set had been selected by sampling, per truthfulness level, 10 statements for each of the main two political parties present in the *Politifact* and *ABC* data sets (i.e., Republican and Democrat for *Politifact*; Liberal and Labor for *ABC*). This resulted in $10 * 2$ (political parties) $* 6$ (truthfulness levels) = 120 *Politifact* statements, and $10 * 2$ (political parties) $* 3$ (truthfulness levels) = 60 *ABC* statements.

Soprano et al. [334] asked crowd workers to reassess the 180 statements in a set of Human Intelligence Tasks (HITs). In addition to the *overall truthfulness* of the statement, they employed a multidimensional truthfulness scale composed of the following dimensions: *correctness*, *neutrality*, *comprehensibility*, *precision*, *completeness*, *speaker's trustworthiness*, and *informativeness*.[3] They recruited 200 US-based crowd workers from *Amazon Mechanical Turk* (MTurk)[4]. Each HIT required workers to assess the truthfulness of 11 statements: six from Politifact (i.e., one for each truthfulness level), three from ABC (i.e., one for each truthfulness level), and two additional hand-crafted statements that the authors used for quality control (i.e., to identify malicious or low-quality workers). The statement sets were also balanced in terms of political parties (i.e., all political parties were equally represented). Workers assessed the truthfulness of statements using a set of five-point Likert scales ranging from "strong disagreement" (-2) to "strong agreement" (2). Each statement was evaluated by 10 distinct workers. Before judging the truthfulness of statements, each worker completed a mandatory questionnaire (i.e., to record their age group, level of education, income, general political view, favored political party, opinion on the US southern border, and opinion on US environmental regulations) and a *Cognitive Reflection Test* (CRT) [107] to assess their cognitive reasoning abilities. We used these bits of worker-specific information as independent variables for our exploratory study.

---

[1] The data set is publicly available at `https://github.com/KevinRoitero/crowdsourcingTruthfulness`.

[2] See `https://www.abc.net.au/news/factcheck/`.

[3] For the rationales behind and detailed discussion of these extra dimensions, we refer to [334, Section 4.3].

[4] `https://www.mturk.com/`

7.2.2. Data preprocessing

We performed several preprocessing steps on the data described in Section 7.2.1 so that they fit our purposes. Specifically, we transformed several scales and computed worker-related bias metrics.

Scale Transformations

Each statement in the data set (see Section 7.2.1) contains a truthfulness judgment from either *Politifact* or *ABC*, as well as several truthfulness judgments from crowd workers. However, these different types of judgments all adhere to different (ordinal) scales: whereas *Politifact* judgments are made on a six-level scale, *ABC* judgments are made on a three-level scale and worker judgments are made on a five-level (Likert) scale. Comparing the different assessments required that we align all of those scales. Assuming that all the *Politifact*, *ABC*, and Likert scales are linear equally spaced scales,[5] we converted the *Politifact* and Likert scales to the three-level scale used by *ABC*. This meant transforming each judgment to one of three labels: *negative* (-1), *neutral* (0), and *positive* (1):

- Politifact: we mapped *pants-on-fire* and *false* into *negative* (-1), *barely-true* and *half-true* into *neutral* (0), and *mostly-true* and *true* into *positive* (1).
- ABC: *negative* and *positive* maintain the same semantic meaning, while *in-between* was mapped into *neutral* (0).
- Likert scale: we mapped -2 and -1 into *negative* (-1), 0 into *neutral* (0), and $+1$ and $+2$ into *positive* (1).

Annotation Bias Metrics

We computed three different metrics to quantify and evaluate annotation bias. We considered both *external* errors (i.e., when comparing crowd annotations with the ground truth) and *internal* errors (i.e., when comparing crowd annotations with other crowd annotations for the same set of items).

- **External Error** (`eE`): the difference between a worker's overall truthfulness judgment and the respective item's ground truth label as assessed by the expert. This metric assesses the degree to which a crowd worker overestimates or underestimates the overall truthfulness of a particular statement. Its values range in $[-2, 2]$: for example, if the ground truth label (i.e., from *Politifact* or *ABC*) for an item is positive (1) but the crowd worker's annotation is negative (-1), `eE` for this particular annotation is equal to -2.
- **External Absolute Error** (`eAE`): the *absolute* difference between a crowd worker's overall truthfulness judgment and the respective item's ground truth label. Its values range in $[0, 2]$. In contrast to `eE`, this metric quantifies the *magnitude* of bias. It is the absolute value of `eE`.[6]

---

[5] The same assumption has been made in previous studies and discussed in more detail by Roitero et al. [301, Section 3.3].

[6] We did not use the mean squared error here to avoid penalizing larger errors (e.g., an error of 2 should not be more than the double the error of 1).

- **Internal Error** (`iE`): the difference between a worker's judgment and the average judgment of other crowd workers for the same statement. Its values range in $[-2, 2]$. We computed nine such metrics in total, i.e., one for overall truthfulness, one for workers' confidence, and one for each of the seven truthfulness dimensions. These nine metrics quantify the degree to which a specific annotation was above or below other crowd workers' judgments on a particular dimension.

Worker Bias Metrics

We computed aggregate bias metrics that evaluate each worker's individual degree of bias based on the annotation bias metrics described in Section 7.2.2. Specifically, we compute each worker's mean `eE` (`eME`), mean `eAE` (`eMAE`), and – for overall truthfulness, confidence, and each of the seven dimensions – mean `iE` (`iME`). These 11 worker-specific metrics are used as dependent variables for the exploratory study.

7.2.3. Exploratory Analyses

We performed a series of exploratory analyses on the public data set (see Section 7.2.1) to identify potential systematic biases in crowd workers' truthfulness judgments. Specifically, we used different worker-related attributes (e.g., political views and average time per judgment) as independent variables and the aggregate worker bias metrics as dependent variables. We found the workers in the data set to be quite balanced in terms of demographics (e.g., age group and income) and political views (e.g., conservative versus liberal orientation). Note that the results we report in this subsection (e.g., p-values from hypothesis tests) are exploratory. We only conducted these analyses to identify concrete hypotheses that we would test on novel data (see Section 7.2.4).

Exploring Worker's `eME`

We began our exploratory analysis by computing workers' `eME`, corresponding to the average difference between a crowd worker's judgment and the respective item's ground truth label. We found that workers overall tended to overestimate truthfulness (mean `eME` = 0.32, sd = 0.42, $t = 10.93$, $p < 0.001$; result from a one-sample t-test; test value = 0). Looking at specific worker characteristics using linear regression and ANOVA models (incl. posthoc tests), we found that workers who identified as *very conservative* and/or *Republican* tended to overestimate truthfulness more than other worker groups (i.e., Tukey-adjusted $p = [0.006, 0.050]$ compared to other political views for *very conservative* workers; Tukey-adjusted $p = [0.012, 0.089]$ compared to other party affiliations for Republican workers). The results further showed that workers who agreed to the southern border question (see the full survey on our repository) overestimated truthfulness more than workers who disagreed (Tukey-adjusted $p = 0.004$); although this effect seemed to be explained by workers' political affiliation, as 78% of those workers also identified as Republicans.

When looking for explanations for the aforementioned systematic biases, we found a slight trend that workers (especially those who identified as Republicans) particularly overestimated the truthfulness of those statements that confirmed their political views (see the left-hand panel of Figure 7.1). Ironically, due to the general trend toward overestimating truthfulness, this led the average worker to judge the truthfulness of statements

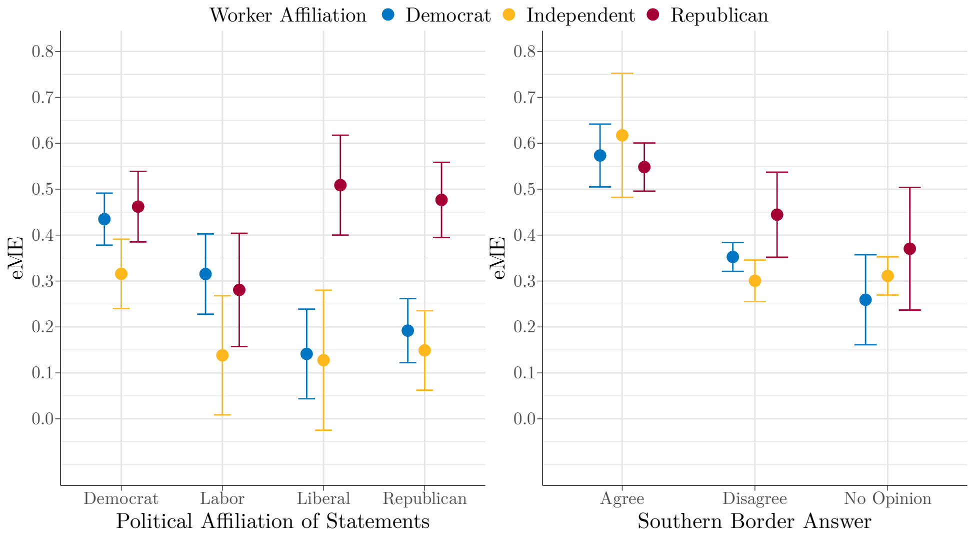


Figure 7.1: Mean `eME` per political affiliations of statements and workers (left) and mean `eMAE` per southern border answer and political affiliations of workers (right) in the public data set (see Section 7.2). Here, we excluded four workers who did not consider themselves a Democrat, independent, or Republican.

affiliated with other parties more accurately than their own. This phenomenon could be explained by different cognitive biases (see Chapter 6), i.e., the *affect heuristic* (crowd workers may overestimate truthfulness when they like the statement speaker) or the *confirmation bias* (crowd workers may overestimate truthfulness when they support the underlying political message).

Exploring Worker's `eMAE`

We also considered `eMAE`, which corresponds to the mean absolute difference between a crowd worker's judgment and the respective item's ground truth label. The mean `eMAE` in the data is 0.42 (sd = 0.31), reiterating that the average worker was somewhat biased in their annotations (i.e., `eAE` ranged from 0 to 1.11). Moreover, in line with the findings above, we found that workers who identified as *very conservative* (Tukey-adjusted $p = [0.012, 0.200]$), Republican (Tukey-adjusted $p = [0.031, 0.129]$), or agreed on the southern border question (Tukey-adjusted $p < 0.001$) were more biased than others (i.e., had a higher `eMAE`; see the right-hand panel of Figure 7.1).

We also found that the more biased worker groups mentioned above generally took less time for their judgments compared to other workers. Although we did not find an effect of cognitive reasoning on `eMAE` when considering all independent variables at the same time, workers with lower cognitive reasoning also tended to do the task quicker. It could thus be that cognitive reasoning abilities explain some of the variance between worker groups but that the effect was too small to be detected in this exploratory study. Another explanation could be workers' *belief in science*: we found that 78% of the "disagree" answers regarding additional environmental regulations came from (very) conservative workers. Given the clear scientific stance regarding the environment, some workers may simply not trust scientific results and therefore distrust statements in which scientific results are brought up as evidence. Although there are too few of these “disagree” answers overall to detect a direct effect here, *belief in science* may be an underlying variable that influences the accuracy of crowd workers’ truthfulness judgments.

Interestingly, the analyses also revealed a positive relationship between workers’ average confidence in their judgments and `eMAE` ($\beta = 0.14$, $p < 0.001$), which might be an indication of *overconfidence*, a cognitive bias whereby workers with too much confidence in their abilities make more inaccurate judgments than others (see Chapter 6).

Exploring Worker’s `iME`

Finally, we investigated `iME`, which corresponds to the mean difference between crowd workers’ judgments and other crowd workers’ judgments on the same statements. We found that workers with some postgraduate or professional schooling (no postgraduate degree) had higher confidence in their abilities to judge truthfulness compared to most workers with lower or higher education status (Tukey-adjusted $p = [< 0.001, 0.018]$). Our analyses also revealed that the more a worker identified as being conservative, the higher their self-reported confidence compared to other workers who annotated the same items. In general, confidence was higher in worker groups with greater bias, which further pointed to a potential *overconfidence* bias in some workers. This could also indicate that the confidence dimension acts as a proxy for explaining the political skewness of the results.

By far the strongest predictor of `eME` among the `iME` measures was the correctness dimension ($\beta = 0.51, p < 0.001$). This suggests that workers might see the correctness dimension as commensurable to overall truthfulness (as previously identified by Soprano et al. [334]) and indicates that workers who judge correctness higher than others likely also overestimate overall truthfulness.

Furthermore, we found that workers who identified as Democrats or Republicans judged truthfulness higher on most dimensions than workers who identified as independent or something else, which usually led to more accurate judgments for the latter group due to the general tendency toward overestimation of truthfulness. Even though these differences were small, this might be an indication that workers with higher *trust in politics* (as here represented by Republicans and Democrats) exhibit more overall bias because they overestimate truthfulness to a greater degree than workers with lower *trust in politics* (as here represented by other workers). This suspicion is underlined by the finding that workers who answered with “no opinion” to the southern border question tended to judge the speaker’s trustworthiness lower than other workers (see the right-hand panel of Figure 7.1).

Our analyses also revealed that `iME` for speaker’s trustworthiness was the strongest predictor among the `iME` measures for `eMAE` ($\beta = 0.16, p = 0.040$). This again could point to a potential affect heuristic (see Section 7.2.3).

7.2.4. Hypotheses for the Novel Data Collection

From our exploratory study (see Section 7.2), we derived seven different hypotheses that we planned to test on novel data. We differentiated our hypotheses based on whether they refer to general worker traits (e.g., their *trust in politics*) or task-related cognitive biases (e.g., the *affect heuristic*).

General Worker Traits

These hypotheses refer to expectations about which worker groups may be more prone to biased judgments compared to others (**RQ$_{\textbf{II.1}}$**).

H$_{\textbf{II.1a}}$ Workers with stronger *trust in politics* are less accurate in judging the overall truthfulness of statements compared to other workers.

Rationale: Workers who considered themselves Democrats or Republicans (i.e., the most "traditional" political parties) were less accurate in their truthfulness judgments than other workers in our exploratory study. Overly high *trust in politics* (i.e., the conviction that politicians and governmental bodies are trustworthy and aim to do the right thing) may lead some workers to strongly identify with political parties and could be the underlying reason for this bias. Such workers may not be skeptical enough when considering politicians' statements and therefore overestimate the likelihood of statements being true.

H$_{\textbf{II.1b}}$ Workers with stronger *belief in science* are more accurate in judging the overall truthfulness of statements compared to other workers.

Rationale: Workers who answered with "disagree" to the environmental regulations question (see the full questionnaire on our repository) tended to be more biased than others in our exploratory study. We hypothesize that the underlying responsible variable could be workers' *belief in science* (i.e., the conviction that scientific results are trustworthy and important for societal development). Workers with low belief in science may automatically doubt the truthfulness of statements that refer to scientific findings, e.g., related to climate change. This may undermine workers' ability to give accurate truthfulness judgments.

H$_{\textbf{II.1c}}$ Workers with better cognitive reasoning abilities are more accurate in judging the overall truthfulness of statements compared to other workers.

Rationale: In our exploratory study, we found that workers with lower cognitive reasoning abilities tended to perform the task quicker, which was generally associated with greater bias. Although we did not find a direct association of workers' cognitive reasoning abilities with their bias, we hypothesize that such a relationship could exist but that it might be hard to detect; especially given that many study participants have been exposed to the CRT before [134].

Cognitive Biases

These hypotheses are predictions about cognitive biases that may affect crowd workers (**RQ$_{\textbf{II.2}}$**).

H$_{\textbf{II.2a}}$ Workers generally overestimate truthfulness.

Rationale: We found that workers overestimated truthfulness in our exploratory study, so we expect to find the same in novel data.

$H_{II.2b}$ Workers' tendency to over- or underestimate the overall truthfulness of a statement is related to the degree to which they like the statement claimant.

Rationale: Our exploratory study revealed several relationships that hint at a potential *affect heuristic*. As detailed in Chapter 6, this bias occurs when workers' judgments are affected by the degree to which they like the document they annotate.

$H_{II.2c}$ Workers' tendency to overestimate or underestimate the overall truthfulness of a statement is related to the degree to which they personally support the goal of the statement.

Rationale: Some relationships we found as part of our exploratory study hint at a potential *confirmation bias*, which occurs when workers' judgments are affected by their pre-existing opinions (see Chapter 6).

$H_{II.2d}$ Workers with higher confidence in their ability to correctly judge the truthfulness of items exhibit more bias compared to other workers.

Rationale: We found that workers' confidence in their judgments is directly related to their degree of bias in our exploratory study. We thus expect to find similar *overconfidence* in novel data that we collect.

7.3. Methods

To test the hypotheses detailed in Section 7.2.4, we conducted a further crowdsourcing study. Note that we preregistered our hypotheses, research design, and data analysis plan before data collection.[7]

7.3.1. Procedure

For the data collection, we relied on the same experimental design as Soprano et al. [334]. Specifically, we used the same interface and the same HITs, to keep the new task as similar as possible. We also relied on the same code and framework used in Soprano et al. [334], discussed in Soprano et al. [333].

To investigate our hypotheses (see Section 7.2.4), we identified three additional variables (i.e., *trust in politics*, *belief in science*, and *affect for statement claimant*; see Section 7.2.4) that required modifications to the original task. We used a generalized version of the *Citizen Trust in Government Organizations* (CTGO) questionnaire [132] to measure workers' trust in politics and the *Belief in Science Scale* (BISS) [67] to record workers' belief in science.[8] These two surveys were placed in the task right after the original initial questionnaire. Finally, we added a single, five-point Likert scale item to capture the degree to which the workers like the claimant of the statement. This item also included an additional answer option that allowed the worker to state that they do not know the claimant.

7.3.2. Variables

Our task recorded the following *Independent Variables*:

---

[7] The preregistration is available at `https://osf.io/5jyu4`.

[8] All questionnaires can be found on our repository (see Section 1 for a link).

- *Trust in politics* (continuous; $[-2,2]$): the degree to which workers trust in media and politics as measured by the CTGO questionnaire (i.e., averaging all responses). Higher scores mean greater trust in politics.
- *Belief in science* (continuous; $[-2,2]$): the degree to which workers believe in science as measured by the BISS questionnaire (i.e., averaging all responses). Higher scores mean greater belief in science.
- *Cognitive reasoning* (ordinal; $[0,4]$): worker's cognitive reasoning abilities as measured by the CRT; we also measure the time spent on CRT as a proxy for cognitive effort. Higher scores mean greater cognitive reflection.
- *Political party affiliation* (categorical): whether workers consider themselves as Republican, Democrat, independent or something else (i.e., not represented by any of the three previous political parties). We here relied on workers' responses to Q5 of the initial questionnaire (see our repository for the full questionnaire).
- *Affect for the statement claimant* (ordinal; $[-3,3]$): each worker rated on a five-point Likert scale the degree to which they like each statement claimant; we also included the option "I don't know the claimant."
- *Mean confidence* (ordinal; $[-2,2]$): workers' average self-reported confidence regarding the accuracy across their truthfulness judgments (on a five-point Likert scale).
- *Statement support* (categorical): we approximate the degree to which workers support the cause of the statement (whether true or false) with their personal political orientation.

We considered the `eE`, `eME`, and `eMAE` as *Dependent Variables* (see Sections 7.2.2 and 7.2.2). Finally, we considered `iE`, `iME`, age group, gender, level of education, income, political views, opinion on US southern border and about US environmental regulation of the workers as *descriptive and exploratory variables* (i.e., we do not conduct any conclusive hypothesis tests using those variables). We collected data on these variables using a survey (see our repository).

7.3.3. Crowd Workers

We planned to collect data from at least 255 crowd workers. We computed this required sample size in a power analysis for a Between-Subjects ANOVA (see Section "Analysis Plan") using the software *G*Power* [104]. Here, based on our findings in the exploratory study, we specified a small effect size of $f = 0.10$, a significance threshold $\alpha = 0.05/7 = 0.007$ (due to testing multiple hypotheses), a statistical power of $(1-\beta) = 0.8$, and that we have three between-subjects groups (i.e., Republican, Democrat, independent/else) and four within-subjects groups (i.e., Republican, Democrat, Liberal, Labor). We computed the required sample size for each of our hypotheses using their respective degrees of freedom.

We deployed 200 MTurk HITs to evaluate the set of 180 statements outlined in Section 7.2.1. We collected 2200 judgments in total. We recruited crowd workers who were based in the United States. Each crowd worker was rewarded $2 for completing the task. This amount was based on the minimum time required to complete the task and the United States minimum wage of $7.25 per hour.

7.3.4. Statistical Analyses

To test our hypotheses, we conducted several statistical analyses. We performed a multiple linear regression to predict `eMAE` from *trust in politics* ($\mathbf{H_{II.1a}}$), *belief in science* ($\mathbf{H_{II.1b}}$), and *cognitive reflection* ($\mathbf{H_{II.1c}}$), and *mean confidence* ($\mathbf{H_{II.2d}}$). We conducted a one-sample t-test to assess $\mathbf{H_{II.2a}}$ (i.e., comparing `eME` to a test value of 0) and a Spearman correlation analysis to test $\mathbf{H_{II.2b}}$ (i.e., computing a correlation between affect for the statement claimant and `eE`). Finally, we tested $\mathbf{H_{II.2c}}$ by conducting a factorial mixed ANOVA with `eE` as the dependent variable, workers' political party affiliation as a between-subjects factor, and statement's political affiliation as a within-subjects factor (i.e., $\mathbf{H_{II.2c}}$ describes an interaction effect between these two variables).

7.4. Results

This section describes the results of the crowdsourcing study outlined in Section 7.3.

7.4.1. Descriptive Statistics

Abandonment

We measured the abandonment rate of the crowdsourcing task using the definition provided by Han et al. [137] (i.e., how many workers voluntarily terminated the task before completion). Overall, 2742 workers participated. About 302 (11%) workers completed the task, while 2065 workers (75%) voluntarily abandoned it. Furthermore, 375 workers (14%) failed at least one quality check at the end of the task. Each worker had up to 10 tries to complete the task. We compared abandonment and failure distributions with those of Soprano et al. [334] (see Figure 7.2).

The left-hand panel of Figure 7.2 shows how many workers abandoned the task per number of statements annotated. The vast majority of workers (98%) abandoned the task when reaching the first statement. The number of workers who abandoned the task after the first statement is negligible. There is an 18% increase in abandonment rate when comparing our values with those of Soprano et al. [334], compared to which our task adds two additional questionnaires and an evaluation dimension. Thus, our task required somewhat more effort from workers. A higher number of workers may have become bored or frustrated sooner. Indeed, when considering the task described by Soprano et al. [334], it can be seen that a fraction of workers abandoned the task even after reaching the fourth statement. Despite this difference, the general trend was that workers abandoned the task when reaching the first statement.

The right-hand panel of Figure 7.2 shows how many workers failed at least one quality check after submitting their work within their current try. The majority of workers who failed the task performed it only once (216, 58%), with 103 (27%) workers doing it a second time. The remaining 15% of workers who failed the task performed it from three up to 10 times. The failure rate drops from 18% to 14% compared to the task by Soprano et al. [334], meaning that those who submitted their work were less likely to fail. However, the

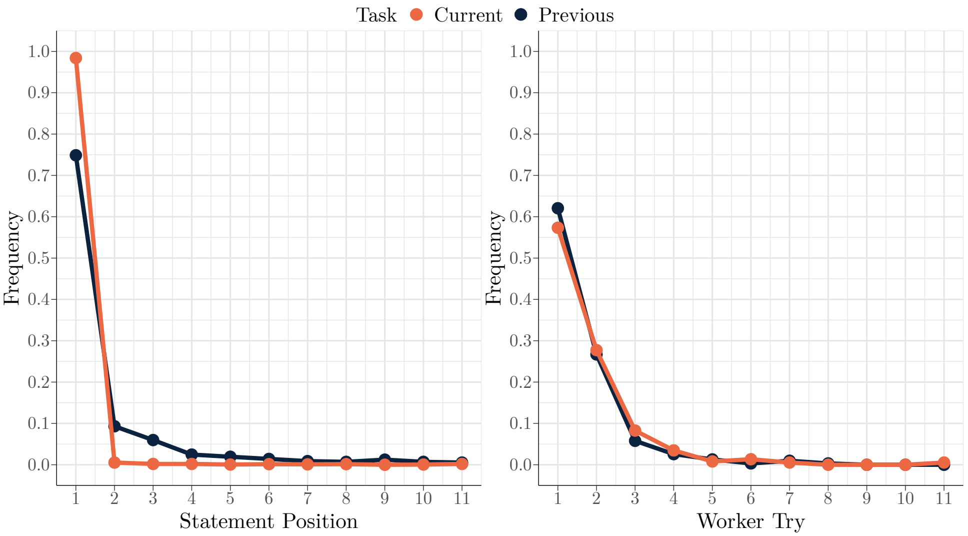


Figure 7.2: Comparison of workers' abandonment distribution (left) and workers' failure distribution (right). The orange line represents our task. The blue lines represent the task by Soprano et al. [334].

failure distribution of our task is in line with the one of Soprano et al. [334].

Demographics

We derived the following demographic statistics considering the 302 workers who completed the crowdsourcing task. Nearly 36% of workers were between 26 and 35 years old, while the 34% were between 35 and 50 years old. The majority of workers (52%) had a college/bachelor's degree. Concerning the total income before taxes, 25% of workers earned \$50k to less than \$75k, while 19% earned \$75k to less than \$100k. When considering workers' political views, 27% identified as moderate, 27% as conservative, and 26% as liberal. The majority of workers (53%) considered themselves Democrats, while the 27% as Republicans and the 17% as independent. The majority of workers (53%) agreed with building a wall at US southern border, with 25% of them disagreeing. Finally, the vast majority of workers (84%) thought that the government should increase environmental regulations to prevent climate change, while only 9% disagreed. In general, our sample was well balanced apart from a few categories and similar to the one of Soprano et al. [334], except that most workers in that study disagreed with building a wall at the US southern border.

Agreement

We measured the internal agreement among workers using Krippendorff's α [190] on the unit level. The use of this metric is motivated by earlier work [301, 334] and theoretical reasons [334]. We found a low level of agreement overall between the workers for each considered truthfulness dimension, which is in line with previous research [301, 334].

We also measured the external agreement between workers' aggregated scores for the overall truthfulness and corresponding experts' values. We recall that experts and workers used different scales (see Section 7.2.1). Whereas the experts used six- (*Politifact*) or three-level scales (*ABC*), the workers in our study evaluated the statements using a five-level scale. We observed that workers were generally in agreement with the experts, as workers tended to judge more truthful documents higher across truthfulness dimensions. Note here that although overall truthfulness is directly correlated with the ground truth, all other truthfulness dimensions capture orthogonal and independent information not directly measured by the experts.

7.4.2. Hypothesis Tests

Our multiple linear regression analysis revealed no evidence for a relationship between `eMAE` and *trust in politics* ($\mathbf{H_{II.1a}}$; $\beta = -0.04, p = 0.020$) or *cognitive reflection* ($\mathbf{H_{II.1c}}$; $\beta = 0.02, p = 0.152$). However, *belief in science* ($\mathbf{H_{II.1b}}$; $\beta = 0.07, p = 0.003$) and *mean confidence* ($\mathbf{H_{II.2d}}$; $\beta = 0.06$, $p =< 0.001$) were both significant predictors of `eMAE`. Partly in contrast to what we expected, workers with stronger *belief in science* and those with greater mean confidence were *more* biased in their truthfulness judgments compared to others. We also found that workers generally overestimated truthfulness, as their mean `eME` (i.e., 0.33, sd = 0.46) lay significantly above 0 in the one-sample t-test we performed ($\mathbf{H_{II.2a}}$; $t = 12.18, p < 0.001$). Our Pearson correlation analysis revealed a significant positive relationship between affect for the statement claimant and `eE` ($\mathbf{H_{II.2b}}$; $r = 0.25, p < 0.001$). Thus, the more the workers liked the statement claimant, the more they overestimated truthfulness; and the more workers disliked the statement claimant, the more they underestimated truthfulness. Our final analysis was an ANOVA with the statement's affiliated party and worker's affiliated party as independent variables and `eE` as dependent variable. This analysis revealed no evidence in favor of an interaction effect between the two independent variables ($\mathbf{H_{II.2c}}$; $F = 1.59, p = 0.112$), which means that we can make no conclusion about whether workers had different degrees of over- or underestimating truthfulness depending on whether the statement party matched their personally favored party or political direction.

In sum, we found evidence in favor of some of our hypotheses (i.e., $\mathbf{H_{II.2a}}$, $\mathbf{H_{II.2b}}$, and $\mathbf{H_{II.2d}}$), suggesting that workers with greater confidence were more biased in their truthfulness judgments, workers generally overestimated truthfulness, and workers' truthfulness judgments were affected by the degree to which they liked the statement claimant. We also found evidence for a relationship between *belief in science* and bias in truthfulness judgments; however, in contrast to $\mathbf{H_{II.1b}}$, our results show that workers with a stronger belief in science were more biased than others.

7.4.3. Exploratory Analyses

Next to the descriptive analyses and hypothesis tests detailed above, we also performed several exploratory analyses on the data we collected. In doing so, we aimed to explain some of the outcomes from the hypothesis tests and identify interesting trends that had not been covered by those planned analyses. Note that the results we report in this subsection are indeed of an exploratory nature, as we did not preregister these analyses.

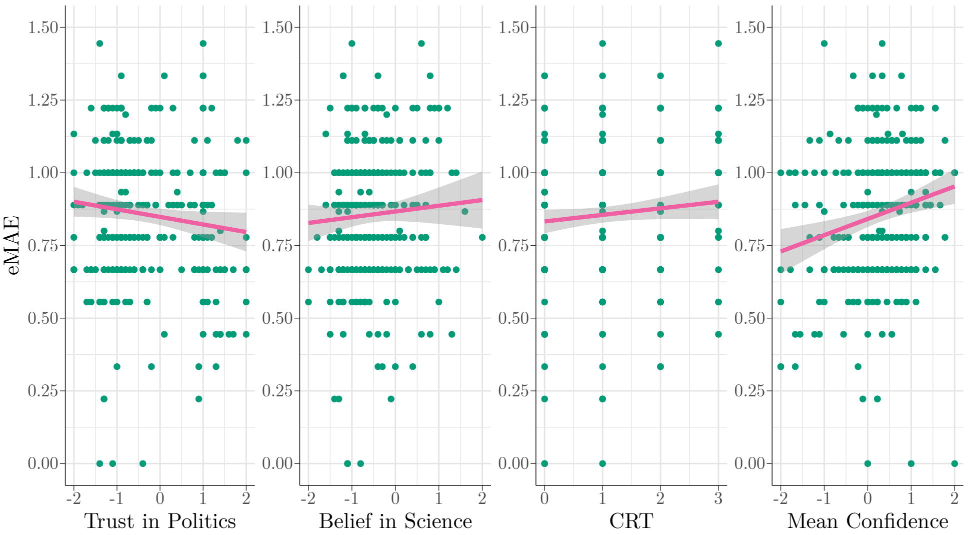


Figure 7.3: Scatter plots showing the relationships between workers' `eMAE` and their *trust in politics* (**$H_{II.1a}$**, left-hand plot), *belief in science* (**$H_{II.1b}$**, center-left plot), Cognitive Reflection Test (CRT; **$H_{II.1c}$**, center-right plot), and mean confidence (**$H_{II.2d}$**, right-hand plot). Our multiple linear regression analysis identified *belief in science* as well as mean confidence as significant predictors of `eMAE` (see Section 7.2.4).

Predicting `eMAE`

Our multiple linear regression identified workers' *belief in science* and *mean confidence* as significant predictors of `eMAE`. Interestingly, we found that, when conducting individual Pearson correlation analyses, only *mean confidence* correlates considerably with `eMAE` ($r = 0.20$, $p < 0.001$), whereas *belief in science* does not (see also Figure 7.3). This suggests that *belief in science* only becomes a relevant predictor of `eMAE` when also taking *trust in politics* and/or *cognitive reasoning* into account, as we did in our multiple linear regression analysis. These two variables might thus still play an important role in predicting workers' `eMAE`, although we did not find such evidence.

The Role of Workers' and Statements' Political Affiliations

The ANOVA we conducted shows no evidence for an interaction effect between workers' and statements' political affiliations in predicting `eE` (**$H_{II.2c}$**). This suggests that workers may not overestimate or underestimate truthfulness systematically based on whether they support the political party that the statement is affiliated with. The same model also contains no evidence for the main effect of workers' political affiliation on `eE` ($F = 1.43$, $p = 0.232$), thus suggesting that workers' political affiliation may not matter at all here. However, there is a significant main effect for statements' political affiliation ($F = 10.55$, $p < 0.001$). Comparing the different statement affiliations shows that workers overestimated the truthfulness of statements relevant to the Australian Labor party significantly more than those relevant to other parties (mean `eE` = 0.51, Tukey-adjusted $p = [< 0.001, 0.018]$). Workers also judged the truthfulness of statements affiliated with the Australian Liberal party significantly lower than those affiliated with other parties (mean `eE` = 0.08, Tukey-adjusted $p = [< 0.001, 0.014]$). Republican and Democrat statements were rated roughly equally on average. This suggests that the political parties connected to the statements may matter for predicting bias in crowd workers' truthfulness judgments, even –or perhaps especially– when those parties are not well-known among the crowd worker population (i.e., the crowd workers in our study were all US-based).

Looking at Individual Truthfulness Dimensions

$\mathbf{RQ_{II.3}}$ concerns whether different truthfulness dimensions are affected by different biases. Next to an overall tendency towards overestimation of truthfulness, our hypothesis tests revealed that workers' *belief in science*, mean confidence, and the degree to which they like the statement claimant may be related to bias in their truthfulness judgments. We thus looked at which specific truthfulness dimensions were particularly affected by these biases to get some more insight into the nature of these biases.

We found that the best `iME` predictors of `eMAE` were neutrality and comprehensibility. Workers thus exhibited more bias when they judged neutrality higher ($\beta = 0.10, p = 0.001$) or comprehensibility lower than others ($\beta = -0.08, p = 0.013$). Moreover, we found that workers' *belief in science* affected no other truthfulness dimensions except overall truthfulness, while the mean confidence of a worker was a significant predictor for all `iME` measures. We also found other interesting relationships, i.e., between workers' *trust in politics* and lower scores on neutrality ($\beta = -0.09, p = 0.028$), and between cognitive reasoning and higher scores on comprehensibility ($\beta = 0.08, p = 0.027$). Finally, affect for the statement claimant was positively related to all considered truthfulness dimensions.

7.5. Discussion

In this section, we report a summary of the key findings derived in this work, list their practical implications, and sketch possible directions for future research.

7.5.1. Key Findings

We have presented a study on the impact of worker biases in crowdsourced fact-checking. To perform our analyses, we conducted an exploratory study using a publicly available data set from which we derived several hypotheses. We then tested these hypotheses in a novel crowdsourcing study. Below, we summarize our findings.

$\mathbf{RQ_{II.1}}$. Our first research question concerned what *individual characteristics* of crowd workers may lead to systematic biases in crowd workers' truthfulness judgments. In this context, we found no evidence for any influence of workers' *trust in politics* (**$\mathbf{H_{II.1a}}$**) or cognitive reasoning abilities (**$\mathbf{H_{II.1c}}$**). Our results do indicate a relationship between workers' degree of *belief in science* (**$\mathbf{H_{II.1b}}$**). However, in contrast to what we expected, we found that workers who reported a stronger belief in science were *less accurate* in their truthfulness judgments.

$\mathbf{RQ_{II.2}}$. The second research question that guided this chapter concerned what *cognitive biases* can affect crowd workers' truthfulness judgments. Our results indicate that several cognitive biases can affect crowd workers' truthfulness judgments. Although we found no evidence for a *confirmation bias* [245] in this context (i.e., there was no interaction effect between workers' and statement's party affiliation on truthfulness judgments;

$\mathbf{H_{II.2c}}$), we found that workers generally overestimate truthfulness ($\mathbf{H_{II.2a}}$). Our findings also suggest an influence of the *affect heuristic* [327]: the more workers like the claimant of a statement, the more they overestimate the statement's truthfulness (and vice versa; $\mathbf{H_{II.2b}}$). Finally, we found evidence for *overconfidence* in crowd workers: the higher workers' self-reported confidence in their ability to judge the truthfulness of statements, the less accurate their judgments generally were ($\mathbf{H_{II.2d}}$).

$\mathbf{RQ_{II.1}}$. Our final research question concerned whether different truthfulness dimensions are affected by different biases. Our study returned exploratory evidence that more biased workers judged the neutrality of statements higher, and the comprehensibility of statements lower than others. Moreover, workers' trust in politics was negatively correlated with their neutrality judgments.

7.5.2. Practical Implications

Following the results of our study, we note several practical implications for crowdsourcing truthfulness judgments as well as adjacent domains such as the collection of document viewpoint annotations [233] (see also Chapters 5 and 6):

- Although crowd workers generally seem to be reliable when judging the truthfulness of statements, individual characteristics (e.g., their belief in science) or cognitive biases (e.g., the affect heuristic or overconfidence) can negatively affect the accuracy of their judgments. We therefore recommend assessing, documenting, and –where possible– mitigating these biases [96, 156]; either by adapting the task design or corrective post-processing of the collected data.

- Where applicable, we recommend that requesters measure relevant concepts such as workers' belief in science [67] to enable effective assessment of systematic biases. Requesters could also consider prioritizing workers with moderate political affiliation, *belief in science*, and confidence in their judgment abilities, as our study suggests that overly strong convictions in these contexts can lead to worse quality in truthfulness judgments.

- Related to the above point, we also recommend avoiding the employment of instruments that may *not* be strictly necessary, e.g., the cognitive reasoning test (CRT) for which we found no relationship to the quality of truthfulness judgments. Requesters should be aware that each such test may reduce the cognitive capacity of crowd workers to eventually perform the actual task. Thus, although we recommend assessment and mitigation of systematic biases, we note that requesters should also not overdo it in this respect.

- Where possible, we recommend that requesters hide unnecessary information (e.g., statement claimant identities or political affiliations) to mitigate the influence of cognitive biases such as the affect heuristic.

- Judgments coming from workers with high self-reported confidence in their ability to identify misinformation should be carefully adjusted, as we found that such workers tend to be more biased than others.

7.5.3. Limitations

We acknowledge that the work presented in this chapter is limited in several different ways. First, we asked crowd workers to self-report their political affiliations and views. Self-reports can be vulnerable to social desirability and other biases [53], so our results are conditional on the assumption that workers' assessments were accurate here. Second, recent research has demonstrated that crowd workers on the MTurk platform do not always accurately represent study populations [50, 365]. This could mean that our results do not generalize to crowd workers on other platforms or people in general. Similarly, we here focused on a narrow set of political statements relevant to specific countries (i.e., the United States and Australia). Our results do not necessarily generalize to other countries or cultural backgrounds, where political systems and discussions may be different. Third, a high percentage of workers abandoned our task early on (see Section 7.4 and Figure 7.2). Although comparable to the study conducted by Soprano et al. [334], this may point to unclarity or other difficulties in our task design, which may have impeded workers' participation in the study [113].

7.6. Conclusion

After introducing a checklist that allows requesters to independently assess crowdsourcing tasks for the potential influence of cognitive worker biases in Chapter 6, this chapter has identified several worker biases that can reduce annotation quality for truthfulness judgments. These biases (e.g., the *affect heuristic* and *overconfidence*) can skew workers' perceptions of the documents they annotate and thus lead them to assign incorrect labels. Although we focused on truthfulness judgments rather than viewpoint label annotations here, crowd workers may behave similarly across annotation scenarios that involve subjective assessments of debated content. It is important for requesters to consider these potential influences in the crowdsourcing tasks they design, e.g., by implementing ways to measure or mitigate cognitive biases such as the affect heuristic or overconfidence.

III

Viewpoint Bias Metrics for Search Results

Assigning the comprehensive viewpoint label we have developed in Part I at scale while applying the tools and insights concerning crowd worker biases from Part II of this dissertation, researchers and practitioners can begin to measure viewpoint bias in search results. Such viewpoint bias assessments are essential in scoping and understanding the general problem of viewpoint bias in current search engines, linking specific degrees of viewpoint bias to user behavior, and exploring how such bias could potentially be reduced. Much work has been devoted to measuring *diversity* concerning query subtopics (i.e., evaluating how well a ranked search result list covers all potential user intents given a query) and *fairness* toward protected groups of search results (i.e., evaluating whether documents that express a particular viewpoint are represented equally across a ranking; see Section 2.2.4). These methods consider the rank and particular characteristics (e.g., the subtopic relevance or protected attribute) of each search result. However, there is currently no notion of ideal viewpoint diversity or protected viewpoints for search results, and measuring viewpoint bias specifically has received comparatively little attention [195]. Recent research has looked at defining and measuring viewpoint bias for the domain of news recommender systems [147, 148, 361], but more work is needed to translate these concepts into the web search paradigm and develop viewpoint bias metrics for search results. In this third part of the dissertation, we thus ask the following research question:

RQ$_{\text{III}}$ What methods can evaluate viewpoint bias in search results?

We begin addressing **RQ$_{\text{III}}$** by exploring how existing ranking fairness metrics could be used to measure viewpoint bias in search results (Chapter 8). Specifically, we take several practical considerations and design choices to adapt existing ranking fairness metrics to the search result viewpoint bias use case and show in simulation studies how different metrics (including one novel metric that we propose) behave for different degrees of viewpoint (ranking) bias. From these simulations, we derive guidelines for measuring viewpoint bias in search results using ranking fairness metrics. We conclude that existing ranking fairness metrics can be used to measure viewpoint bias when search results are labeled using binary taxonomies (e.g., *against*/*in favor*), and the novel ranking fairness metric we propose can also accommodate multicategorical viewpoint labels (e.g., a seven-point stance taxonomy; see Table 2.1). However, ranking fairness metrics cannot incorporate multidimensional viewpoint representations such as the one we propose in Chapter 5. That is why Chapter 9 proposes *normalized discounted viewpoint bias* (nDVB), a rank-aware viewpoint bias metric for search results that considers our more comprehensive viewpoint label. This metric, which measures bias as a deviation from viewpoint plurality, is founded upon a clear notion of viewpoint diversity and can be adapted to fit different topics or viewpoint structures. We then also use nDVB to evaluate viewpoint bias in real search results from popular search engines and show how to increase the viewpoint diversity in such search result lists.

8

Assessing Viewpoint Bias in Search Results Using Ranking Fairness Metrics

---

This chapter is based on a published workshop paper: Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. “Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics”. In: *ACM SIGKDD Explorations Newsletter* 23.1 (May 2021), pp. 50–58. DOI: `10.1145/3468507.3468515`.

Tim Draws primarily planned and carried out the conceptualization, investigation, methodology, project administration, visualization, and write-up of the work described in the paper referenced above. His co-authors supervised him during the project and made edits to the writing.

Viewpoint biases in search results on debated topics can strongly influence user attitudes, preferences, and behavior [10, 99, 274] (see Sections 2.2.4 and 2.2.5). Although previous research has found that web search results may indeed be biased for particular debated topics (e.g., in the health [370, 371] and politics [284] domains), measuring and monitoring these biases at scale and across topics is currently impeded by the lack of available viewpoint bias metrics. The tools, resources, and insights we have introduced in Parts I and II of this dissertation allow researchers and practitioners to create high-quality data sets of search results with corresponding viewpoint labels, but it is currently unclear how to evaluate viewpoint bias in such data sets. To fill this research gap, research has to explore existing metrics from related domains and – if necessary – propose novel metrics that can comprehensively assess search result viewpoint bias. This chapter takes a first step in developing standard metrics for evaluating search result viewpoint bias by using *ranking fairness metrics* for this purpose.

Fair ranking metrics evaluate ranked lists in terms of their *fairness* concerning a given characteristic [35, 315, 381]. For example, a ranked list of candidates on a job seeking platform could be evaluated with respect to gender fairness. A fair ranking is then considered to be one in which gender does not affect the ranking of candidates. Analogously in this chapter, a ranked list of search results is evaluated with respect to *viewpoint* – to the best of our knowledge, a novel application of ranking fairness. Such a viewpoint can, for example, convey different stances on a topic or different underlying reasons for a given stance. A search result ranking that is fair (or unbiased) with respect to viewpoints would give each viewpoint its fair share of coverage, contributing to viewpoint diversity in the search results.[1] Several metrics have been developed that assess fairness in rankings [315, 381] (see Section 2.2.4). These metrics evaluate fairness in terms of *statistical parity*, which is satisfied in a ranking if a given variable of interest – here, the expressed viewpoint – does not influence how documents are ranked.

In this chapter, we generate a range of synthetic search result rankings with varying degrees of ranking bias and explore the behavior of existing and novel ranking fairness metrics on these rankings. Our core assumption here is that viewpoint bias is the degree of deviation from an ideal viewpoint-*diverse* or *fair* search result ranking (see Section 2.1.3). We consider two fundamental scenarios: *binomial viewpoint fairness*, in which the task is to measure viewpoint bias with respect to one specific *protected viewpoint*, and *multinomial viewpoint fairness*, where the aim is to protect all available viewpoint categories simultaneously. We make the following contributions:

1. We present a simulation study that illustrates how existing ranking fairness metrics can be used to assess viewpoint bias in search result rankings. We show how these metrics behave under varying conditions of viewpoint bias and provide a guide for their use (Section 8.2.2).

2. We propose a novel ranking fairness metric for assessing multinomial viewpoint fairness (Section 8.1.3) and also analyze its behavior (Section 8.2.2).

We find that all the considered ranking fairness metrics can distinguish well between different levels of viewpoint bias in search results. However, which specific metric is

---

[1] Note that here we thus look at fairness in the *outcome* of a ranking algorithm; i.e., not at procedural fairness.

Table 8.1: Notation used throughout this chapter.

Notation	**Description**
d	document
D	set of documents
s_d	viewpoint/stance label of document d
S	set of viewpoint/stance labels
S^p	number of items in set S that belong to subset p
τ	ranked list of set D
$S^p_{1...i}$	S^p in the top i ranked documents
N	number of elements in D, S, and τ

most sensitive to viewpoint bias (or a lack of viewpoint diversity) depends on how many viewpoint categories there are, the distribution of advantaged and disadvantaged items in the ranking, and bias severity.

Supplementary material (e.g., data sets and code) related to this chapter is available at `https://osf.io/nkj4g`.

8.1. Measuring Fairness in Rankings

Consider a user who wants to form an educated opinion on the topic *school uniforms* and turns to web search to gather information. Let us assume that each document the user encounters in the search result list will express a viewpoint concerning school uniforms or be neutral towards the topic.[2] These viewpoints can be represented as ordinal *stance* categories, i.e., by placing them on a seven-point scale ranging from *strongly opposing* to *strongly supporting* school uniforms (see Table 2.1).[3]

The notation we consider in this chapter is displayed in Table 8.1. We are given a set of documents D, and a set of viewpoint labels S. Both sets contain the same number of elements N. Each document $d \in D$ is uniquely associated with one label $s_d \in S$. Here, s_d reflects the viewpoint (or stance) of document d towards a given disputed topic, rated on a seven-point scale ranging from *strongly opposing* to *strongly supporting*. The viewpoint labels in S are integers ranging from -3 to 3, where negative values indicate an *opposing stance*, 0 indicates a *neutral stance*, and positive values indicate a *supporting stance* toward the debated topic (see Table 2.1 for an example). A ranked list of D is denoted as τ. We denote the number of items that belong to a subset p of S as S^p, which becomes $S^p_{1...i}$ when constrained to the top i ranked documents.

---

[2] Here, *neutral* could mean that a document is not opinionated, provides a balanced overview of the different viewpoints, or is irrelevant to the topic.

[3] Note that this is just one possible way to categorize existing viewpoints on a topic.

8.1.1. Defining Fairness and Viewpoint Bias

There are many definitions of fairness, so before describing fairness metrics, we first identify which type of fairness to handle. In this chapter, we focus on the notion of *statistical parity* (also commonly referred to as *group fairness*; see Section 2.2). This notion allows us to define several fairness aims for assessing viewpoint bias. We consider two such aims, which we call *binomial viewpoint fairness* and *multinomial viewpoint fairness*. Below we describe these aims and align them with the notion of statistical parity in rankings.

Binomial Viewpoint Fairness

One aim concerning web search on debated topics may be to treat one specific viewpoint, e.g., a minority viewpoint, fairly. For instance, if a search result ranking on the query `school uniforms pros and cons` is dominated by arguments *supporting* school uniforms, the ranking assessor may want to evaluate whether the minority viewpoint (i.e., *opposing* school uniforms) gets its fair share of coverage. The assessor may consider a binary classification of documents into (1) expressing the minority viewpoint or (2) not expressing the minority viewpoint. Here, expressing the minority viewpoint is analogous to a protected group. Statistical parity in a ranking of such documents is satisfied when expressing the minority viewpoint does not affect a document's position in the ranking.

Multinomial Viewpoint Fairness

Another aim may be to cover *all* viewpoints fairly. For example, a search result ranking on the query `school uniforms pros and cons` could be assessed without explicitly defining a specific viewpoint as the protected group but instead considering the distribution over several existing viewpoints. Here the assessor thus considers a multinomial classification of documents into some viewpoint taxonomy (e.g., into seven stance categories; see Table 2.1). In this case, we say that statistical parity is satisfied when expressing any viewpoint does not influence a document's position in the ranking. Multinomial viewpoint fairness is thus more fine-grained than binomial viewpoint fairness: whereas binomial viewpoint fairness focuses on fairness towards one protected viewpoint, multinomial viewpoint fairness requires being fair to all viewpoints simultaneously.

8.1.2. Desiderata and Practical Considerations for Metrics

Evaluating statistical parity. In this chapter, we use ranking fairness metrics to assess viewpoint bias in search result rankings. These are based on the notion of statistical parity, which is present in a ranking when the viewpoints that documents express do not affect their position in the ranking. However, we are only given the ranking and viewpoint per document and cannot assess the ranking algorithm directly. Statistical parity thus needs to be approximated. We choose to approximate statistical parity in the same way as previously developed ranking fairness metrics [381]. These metrics measure the extent to which the document distribution over groups (e.g., the protected and non-protected group) is the same in different top-i portions of the ranking compared to the overall ranking (see Section 2.2.4). The more dissimilar the distribution at different top-i portions is from the overall distribution, the less fair the ranking.

Discounting the ranking fairness computation. User attention depletes rapidly as the ranks go up, with many users not even exploring search results beyond the first page (i.e., the top 10 results) [168, 258]. This means that fairness is more important at higher ranks. A measure of viewpoint bias thus needs to consider the rank of documents and not just whether viewpoints are present. A practical way to incorporate this notion into a ranking fairness metric is to include a discount factor. Sapiezynski et al. [315] point out that such a discount depends on the user model related to the particular ranking one is assessing. Similar to the ranking fairness metrics introduced by Yang and Stoyanovich [381], we choose the commonly used $\log_2$ discount for each metric we introduce below. Yang and Stoyanovich [381] suggest discounting in steps of 10 (see Section 2.2.4). Such a binned discount nicely incorporates the notion that ranking fairness is more important in the top 10 documents than in the top 20 documents. However, especially on the first page of search results, individual ranks matter a lot [168, 258]. We therefore decide to discount by individual rank and consider the top $1, 2, \ldots N$ documents at each step of the aggregation.

Normalization. When evaluating and comparing metrics, it is useful if they all operate on the same scale. We thus only consider normalized ranking fairness metrics.

8.1.3. Ranking Fairness Metrics

In this section, we describe the metrics we use to assess viewpoint bias in search result rankings. These metrics are partly based on existing ranking fairness metrics and partly novel. We adapt each metric we use to fit the practical considerations outlined in Section 8.1.2. Taking these practical considerations into account, we define a template that each normalized ranking bias (nRB) metric that we use will follow:

$$\mathrm{nRB}(\tau) = \frac{1}{Z} \sum_{i=1}^{N} \frac{F(i)}{\log_2(i+1)}. \tag{8.1}$$

Here, F is a function that quantifies the ranking bias in the ranked list τ. All metrics that we describe in the following subsections will only differ in terms of how they define F. The function F is iteratively computed for the top i documents and subsequently aggregated using a $\log_2$ discount. Finally, Z is a normalizing constant that takes on the value for F given the maximally unfair permutation of τ.[4]

Metrics to Assess Binomial Viewpoint Fairness

Yang and Stoyanovich [381] propose three ranking fairness metrics to assess statistical parity in rankings (see Section 2.2.4). We interpret these metrics to fit binomial viewpoint fairness and adapt them to fit the considerations outlined in Section 8.1.2. Note that although we define a protected and a non-protected viewpoint before using any of these metrics, the metrics are in principle agnostic as to which of the two viewpoint categories (i.e., “protected” and “unprotected”) is advantaged in the ranking. That is, they do not only measure when the protected viewpoint is treated unfairly but also capture if a ranking is biased *towards* the protected viewpoint. The categorization into protected and non-protected viewpoints should thus be viewed as a binary classification of documents that – in a fair scenario – does not affect how documents are ranked.

---

[4] A description of how we normalize each metric can be found at `https://osf.io/nkj4g`.

Normalized Discounted Difference (rND). This metric computes the difference between the proportion of items that belong to the protected group at different top-i subsets of the ranking with the overall proportion:

$$\mathrm{rND}(\tau) = \frac{1}{Z}\sum_{i=1}^{N}\frac{1}{\log_2(i+1)}\left|\frac{S^p_{1\ldots i}}{i} - \frac{S^p}{N}\right|. \tag{8.2}$$

Here, S^p is the number of documents in the protected group and N is the total number of ranked documents.

Normalized Discounted Ratio (rRD). This metric measures the difference between the ratio of documents that express the protected viewpoint and those who do not, at different top-i portions of the ranking with the overall ratio:

$$\mathrm{rRD}(\tau) = \frac{1}{Z}\sum_{i=1}^{N}\frac{1}{\log_2(i+1)}\left|\frac{S^p_{1\ldots i}}{S^u_{1\ldots i}} - \frac{S^p}{S^u}\right|. \tag{8.3}$$

Here, S^u refers to the number of documents that do not express the protected viewpoint. We set the value of fractions to 0 if their denominator is 0 [381].

Normalized Discounted Kullback-Leibler Divergence (rKL). This metric makes use of the *Kullback-Leibler divergence* (KLD), an asymmetric measure of the difference between probability distributions [194]. For two discrete probability distributions P and Q that are defined on the same probability space $\mathcal{X}$, KLD is given by

$$\mathrm{KLD}(P||Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right). \tag{8.4}$$

To measure binomial viewpoint fairness in a ranking, P and Q can be defined as

$$P = \left(\frac{S^p_{1\ldots i}}{i}, \frac{S^u_{1\ldots i}}{i}\right), Q = \left(\frac{S^p}{N}, \frac{S^u}{N}\right).$$

This way, KLD measures the divergence between the proportion of protected items at rank i and in the ranking overall.[5] We can insert KLD in Equation 8.1:

$$\mathrm{rKL}(\tau) = \frac{1}{Z}\sum_{i=1}^{N}\frac{\mathrm{KLD}(P||Q)}{\log_2(i+1)}. \tag{8.5}$$

Metric to Assess Multinomial Viewpoint Fairness

To the best of our knowledge, no metrics have so far been proposed that explicitly assess ranking fairness for multiple categories at once. The previously introduced rKL metric can in principle be expanded to assess multinomial viewpoint fairness. KLD measures the distance between two discrete probability distributions P and Q. In the multinomial case, we can define P and Q as multinomial distributions over the available viewpoint categories. For instance, in our use case of viewpoints rated on a seven-point stance scale, P and Q may be given by:

---

[5] Note that KLD is not defined for $P = (0,1)$. In this case, we smooth to $P = (0.001, 0.999)$.

$$P = \Big(\frac{S^{-3}_{1...i}}{i}, \frac{S^{-2}_{1...i}}{i}, \frac{S^{-1}_{1...i}}{i}, \frac{S^{0}_{1...i}}{i}, \frac{S^{+1}_{1...i}}{i}, \frac{S^{+2}_{1...i}}{i}, \frac{S^{+3}_{1...i}}{i}\Big),$$
$$Q = \Big(\frac{S^{-3}}{N}, \frac{S^{-2}}{N}, \frac{S^{-1}}{N}, \frac{S^{0}}{N}, \frac{S^{+1}}{N}, \frac{S^{+2}}{N}, \frac{S^{+3}}{N}\Big),$$

where $S^{-3,-2,...,3}$ refer to the number of items in each viewpoint category.

A problem with using KLD for multinomial distributions is that its normalization becomes extremely complex. To normalize KLD, the maximally divergent distribution of items needs to be computed at each step. Whereas this is rather straightforward in the binomial case,finding the maximally divergent distribution becomes extremely expensive when more categories are added.

To resolve the normalization issue that comes with KLD, we propose a new metric that uses the Jensen-Shannon divergence (JSD) as an alternative distance function. Similarly to KLD, JSD measures the distance between two discrete probability distributions P and Q that are defined on the same sample space $\mathscr{X}$ [108]. JSD can, in fact, be expressed using KLD:

$$\mathrm{JSD}(P||Q) = 0.5 * \Big(KLD(P||R) + KLD(Q||R)\Big).$$

Here, $R = 0.5 * (P + Q)$ is the mid-point between P and Q. In contrast to KLD (which can go to infinity), JSD is bound by 1 as long as one uses a base 2 logarithm in its computation [206]. Knowing this maximally possible value for JSD, also an aggregated, discounted version of JSD is easily normalized. We thus propose *Normalized Discounted Jensen-Shannon Divergence* (nDJS) as given by

$$\mathrm{nDJS}(\tau) = \frac{1}{Z}\sum_{i=1}^{N} \frac{\mathrm{JSD}(P||Q)}{\log_2(i+1)}, \tag{8.6}$$

where $\mathrm{JSD}(P||Q)$ is the JSD between P and Q. Although we here propose nDJS specifically for assessing multinomial viewpoint fairness, note that it can be used to assess binomial viewpoint fairness as well.

8.2. Simulation Study

In this section, we show how the metrics introduced in Section 8.1.3 behave in different ranking scenarios. The code to implement the metrics and simulation is available on our repository.

8.2.1. Generating Synthetic Rankings

To simulate different ranking scenarios, we first generate three synthetic sets $S1$, $S2$, and $S3$ to represent different stance distributions (i.e., considering the seven-point stance taxonomy we use to represent viewpoints in this chapter; see Table 2.1). The items in each set simulate stance labels for 700 documents (i.e., to enable a simple balanced distribution over the seven stance categories) and are distributed as shown in Table 8.2. Whereas $S1$ has a balanced stance distribution, $S2$ and $S3$ are skewed towards supporting stances.[6] We use $S1$, $S2$, and $S3$ to simulate both binomial and multinomial viewpoint fairness.[7]

Table 8.2: Viewpoint (i.e., stance) distributions of the sets $S1$, $S2$, and $S3$.

	-3	**-2**	**-1**	**0**	**+1**	**+2**	**+3**
$S1$	100	100	100	100	100	100	100
$S2$	80	80	80	115	115	115	115
$S3$	60	60	60	130	130	130	130

Sampling. We create rankings of the stance labels in $S1$, $S2$, and $S3$ by conducting a weighted sampling procedure. To create a ranking, stance labels are gradually sampled from one of the three sets without replacement to fill the individual ranks. Each stance label in the set is assigned one of two different sample weights that determine the labels' probability of being drawn. These two sample weights are controlled by the ranking bias parameter α and given by: $w_1 = 1.0001 - 1 \times \alpha$; $w_2 = 1.0001 + 1 \times \alpha$.

Alpha. For our simulation of binomial and multinomial viewpoint fairness, ranking bias is controlled by the continuous parameter $\alpha = [-1, 1]$. More specifically, α controls the sample weights w_1 and w_2 that are used to create the rankings. Whereas a negative α will result in higher ranks for stances that are assigned w_1, a positive α will advantage stances that are assigned w_2. The further away α is from 0, the more extreme the ranking bias. If α is set to exactly 0, no ranking bias is present: here, it does not matter whether a stance label is assigned w_1 or w_2; the sample weights are the same. In each simulation, we try 21 degrees of ranking bias for $\alpha = -1$ to $\alpha = 1$ in steps of 0.1.

Simulating Binomial Viewpoint Fairness

To simulate binomial viewpoint fairness, we create ranked lists from $S1$, $S2$, and $S3$ with different degrees of ranking bias. Ranking bias – controlled by α – in this scenario refers to the degree to which expressing a protected viewpoint influences a document's position in the ranking. We define all *opposing* stances (i.e., -3, -2, and -1) together as the protected viewpoint and assign them the sample weight w_1. All other stances (i.e., 0, 1, 2, and 3) are thus non-protected and assigned the other sample weight w_2 when generating the rankings. Table 8.3 (left-hand table) shows an example of this sample weight allocation for $\alpha = 0.5$. In this example, the non-protected viewpoint is more likely to be drawn compared to the protected viewpoint.

---

[6] Due to symmetry, we do not include similar distributions for opposing stances.

[7] Because we are only interested in rankings with respect to viewpoint/stance labels, we do not generate any actual documents here. Instead, we rank the labels themselves.

Table 8.3: Examples of sample weight allocations for the simulation of binomial (left-hand table, $\alpha = 0.5$) and multinomial viewpoint fairness (right-hand table; $\alpha = -0.8$).

stance	-3	-2	-1	0	+1	+2	+3
weight	w_1	w_1	w_1	w_2	w_2	w_2	w_2
	0.5	0.5	0.5	1.5	1.5	1.5	1.5

stance	-3	-2	-1	0	+1	+2	+3
weight	w_2	w_2	w_1	w_2	w_2	w_2	w_2
	0.2	0.2	1.8	0.2	0.2	0.2	0.2

Our weighted sampling procedure (see above) will produce slightly different rankings even when the same α is used. To get reliable results, we therefore create 1000 ranked lists for each α and aggregate the results.

Simulating Multinomial Viewpoint Fairness

We simulate multinomial viewpoint fairness by again sampling rankings from $S1$, $S2$, and $S3$ with different degrees of ranking bias. This time, the ranking bias α is defined as the degree to which the expressed viewpoint generally affects a document's position in the ranking.

Since there are many scenarios in which one (or more) of several viewpoint (or stance) categories could be preferred over others in a ranking, we focus on just one specific case: our simulation prefers *one* of the seven stances over the other six. For example, this could be the case if a search result list is biased toward an extremely opposing stance. We randomly assign the sample weight w_1 to one of the opposing stances (i.e., -3, -2, or -1) and the sample weight w_2 to all remaining stances for each ranking we create. This means that each ranked list we create prefers a different stance, reflecting the idea that we do not know which stance might be preferred before evaluating the ranking, and we have no specific, pre-defined protected viewpoint. Table 8.3 (right-hand table) shows an example of this sample weight allocation for $\alpha = -0.8$. In this example, the ranked list will prefer the *somewhat opposing* stance (-1) over all other stance categories. We again compute 1000 ranked lists for each α and aggregate the results.

8.2.2. Metric Behavior

Here, we explore the behavior of the ranking fairness metrics introduced in Section 8.1.3 using the synthetic rankings from Section 8.2.1.

Binomial Viewpoint Fairness

Binomial viewpoint fairness can be assessed using rND, rRD, or rKL. Each of these metrics measures the degree to which expressing a protected viewpoint affects the ranking of documents. The ranking in our running example is considered fair if documents opposing school uniforms (i.e., -3, -2, and -1) get a similar coverage throughout the ranking compared to other stances (i.e., 0, +1, +2, and +3). A fair scenario should lead to a low score on each of the three metrics.

Figure 8.1 shows the mean outcome of rND, rRD, and rKL from 1000 ranked lists per data set (i.e., $S1$, $S2$, and $S3$) and α (i.e., ranking bias) setting. Each set represents a different overall stance distribution (see Table 8.2).

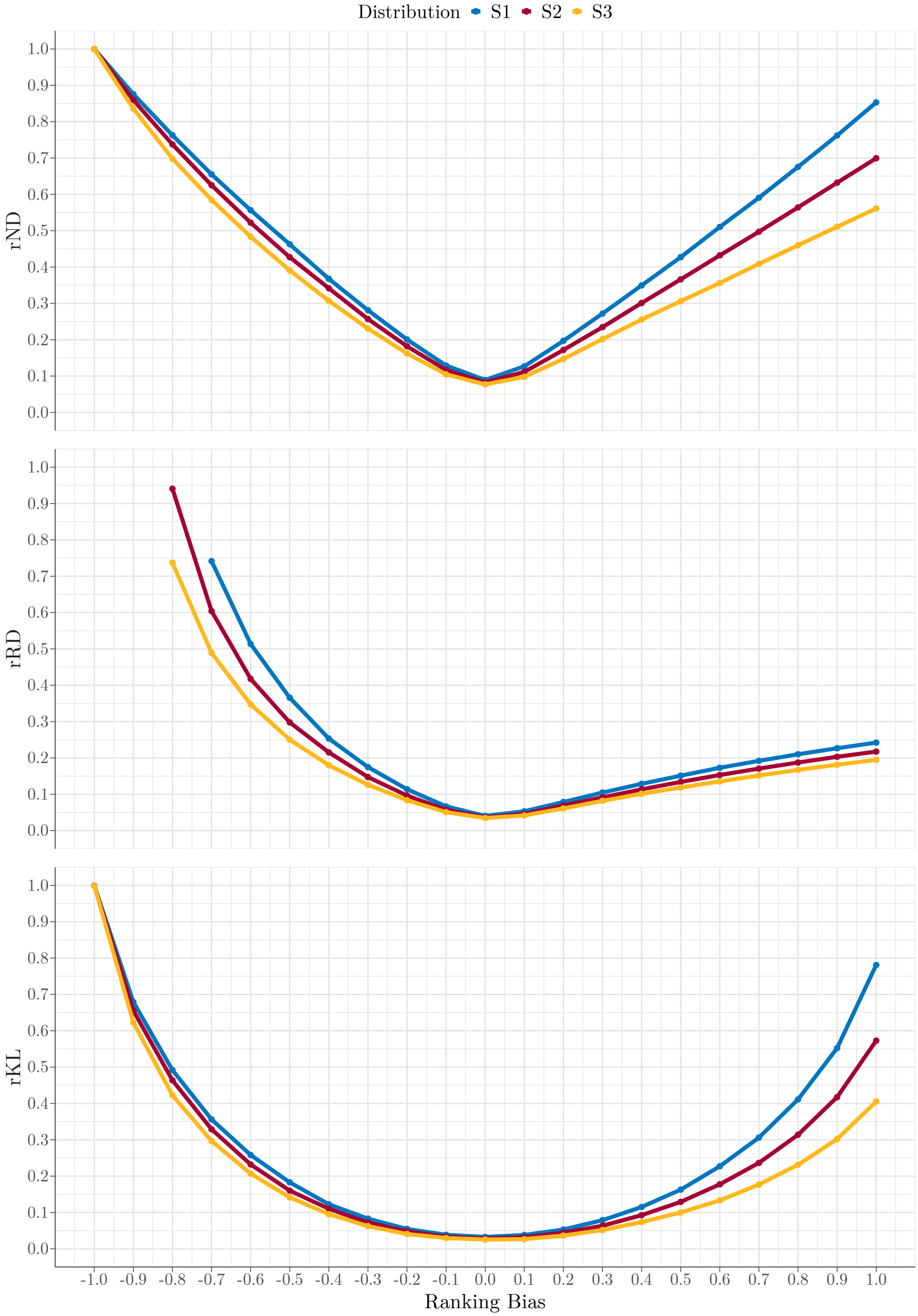


Figure 8.1: Behavior of the metrics rND (top plot), rRD (center plot), and rKL (bottom plot) on the sets $S1$ ($S^p = 300$), $S2$ ($S^p = 240$), and $S3$ ($S^p = 180$) across different α (ranking bias) settings.

We note three characteristics that all three metrics share. First, each of the three metrics is lowest for low bias ($\alpha = 0$) and increases from there as the absolute value of α increases. This means that all three metrics function as expected: they produce higher values as ranking bias becomes more extreme. Second, each metric shows a steeper curve as the data sets contained fewer items that express the minority viewpoint (here, the protected opposing stances) increases, i.e., $S1 > S2 > S3$. Different levels of ranking bias thus become easier to detect when the distribution of protected and non-protected items is more balanced. Third, each metric produces higher values for $\alpha = -1$ (protected viewpoint is *advantaged*) than for $\alpha = 1$ (protected viewpoint is *disadvantaged*). The reason behind this is that unfair treatment becomes increasingly harder to detect as the number of items in the disadvantaged group shrinks: if one group only encompasses around 25% of items (e.g., such as in $S3$), it is less odd to see several items of the other group ranked first than if the distribution is more balanced. That is also why each metric produces higher values at $\alpha = 1$ as the number of protected items increases.

Next to these general characteristics shared by all metrics, below we discuss differences that distinguish the metrics in terms of their behavior.

Normalized Discounted Difference. For each of the three data sets, rND reaches its maximum value of 1 when $\alpha = -1$ and is at its lowest with mean values of approximately 0.08 when $\alpha = 0$. Depending on the number of items that express the protected viewpoint, rND reaches mean values between 0.55 and 0.85 when $\alpha = 1$ for the three data sets in our simulation. The curves for rND in Figure 8.1 are also comparatively steep. This indicates that rND is especially useful for distinguishing low levels of ranking bias.

Normalized Discounted Ratio. The lowest mean rRD values in our simulation (reached at $\alpha = 0$ for each of the three data sets) all approximate 0.04. Even more so than rND, rRD reaches mean values far below 1 when the protected viewpoint is disadvantaged in the ranking. The mean values for this form of extreme ranking unfairness range from approximately 0.19 to 0.24 in our simulation, depending on the number of protected viewpoint items. In comparison to the other two metrics, rRD is less steep than rND but steeper than rKL. It could thus be useful for detecting medium levels of ranking bias. However, if a ranking is unfair towards the minority viewpoint, rRD does not distinguish different levels of ranking bias well. We also find that our normalization procedure (i.e., dividing each metric outcome by the outcome for a maximally unfair ranking) does not normalize rRD correctly. Thus, the maximal mean values for rRD (which it reaches at $\alpha = -1$) lie above 1 and are therefore not displayed in Figure 8.1 (which has 1 as its upper limit).[8]

Normalized Kullback-Leibler Divergence. Similar to the other metrics, rKL reaches its maximum value of 1 at $\alpha = -1$. In our simulation, the lowest mean values for rKL (reached at $\alpha = 0$) approximated 0.03. Large α settings (i.e., disadvantaging the minority viewpoint) produce mean rKL values between 0.40 and 0.78, depending on the number of items that express the minority viewpoint. Furthermore, rKL has a more parabolic shape compared

---

[8] We explore the reason behind this (including an alternative way to normalize rRD) in a supplementary document on our normalization procedures; see `https://osf.io/nkj4g`.

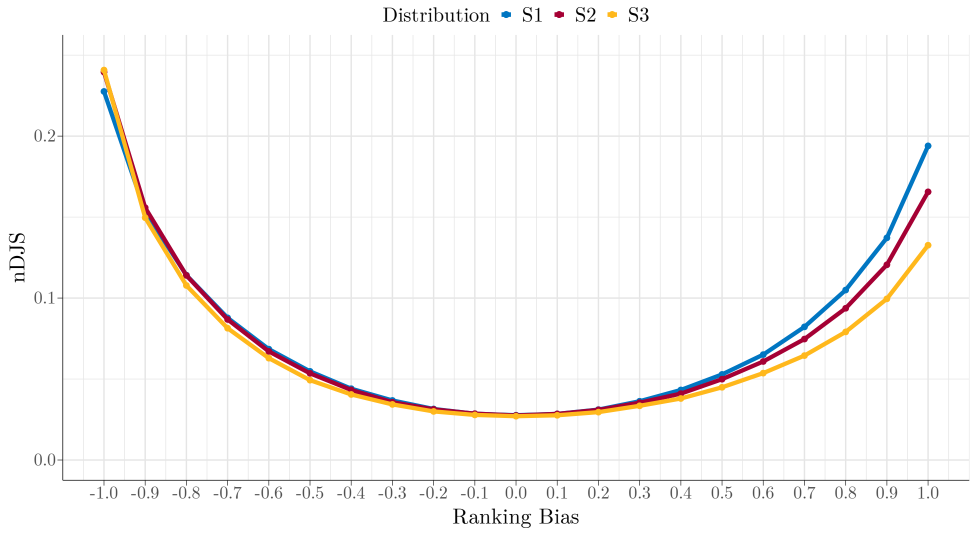


Figure 8.2: Behavior of nDJS on the sets $S1$, $S2$, and $S3$ across different α (ranking bias) settings. The number of items with sample weight w_1 for rankings from the sets $S1$, $S2$, and $S3$ are 100, 80, and 60, respectively. Note that we have zoomed in here compared to Figure 8.1 to better show nuanced differences between the lines.

to rND and rRD. Whereas rKL can thus not distinguish low values of ranking bias well, it is useful for differentiating between high levels of ranking bias.

Multinomial Viewpoint Fairness

To assess multinomial viewpoint fairness, we use nDJS. This metric measures the degree to which the viewpoint that documents express is a factor for a ranking in general. For example, in a search result ranking related to the topic *school uniforms*, a range of stance categories may exist, some of which may be advantaged in the ranking over other stances. That is why we cannot use binomial ranking fairness metrics here: we do not have a specific viewpoint to protect but instead wish to equally protect all viewpoints (e.g., represented across seven stance categories). A maximally fair ranking scenario would give all viewpoints coverage across the ranking proportional to their share in the overall distribution. For (approximately) fair rankings, nDJS should return a low value.

We test nDJS on synthetic rankings that simulate varying degrees of bias on three different sets of items ($S1$, $S2$, and $S3$, see Section 8.2.1). Figure 8.2 shows the mean outcome of nDJS from 1000 ranked lists per set and α (i.e., ranking bias) setting. Similar to the binomial ranking fairness metrics, nDJS does what it is expected to do: it produces its highest values at extreme α (ranking bias) settings and its lowest values at $\alpha = 0$. This means that nDJS can pick up the nuanced multinomial viewpoint fairness in our synthetic rankings. We observe, however, that due to its normalization, the maximum values for nDJS are much lower than for the metrics that assess binomial viewpoint fairness. When $\alpha = -1$ (i.e., when one random stance category is *disadvantaged* compared to others), nDJS produces mean values between approximately 0.18 and 0.21. Due to the different normalization, it is therefore impossible to compare results from nDJS directly to results from the binomial ranking fairness metrics. For low values of ranking bias, the mean nDJS values approximate 0.03 on all three data sets. The mean nDJS value lies between approximately 0.07 and 0.09 when $\alpha = 1$ (i.e., when one stance is *advantaged* compared to others).

Similar to the binomial fairness metrics, the values nDJS produces are again influenced by the proportion of advantaged items in the ranking. The more balanced this ratio, the easier it is to detect a ranking bias (i.e., the higher nDJS). Note that in this simulation, the distribution of advantaged and disadvantaged items was far from balanced, as we only treated one stance label differently per ranking.

8.3. Discussion

In this section, we summarize our findings, provide a guide to using the metrics we examined, and discuss the limitations and implications of this research.

8.3.1. Binomial Viewpoint Fairness

Each of the three metrics we tested in our simulation can measure binomial viewpoint fairness (rND, rRD, rKL; see Section 8.2.2). However, depending on the distribution of protected and non-protected items, as well as the direction and level of ranking bias, a different metric might be suitable. Table 8.4 shows which metric we recommend using in which scenario. In sum, we suggest taking the following considerations when assessing binomial viewpoint fairness:

1. Generally, the metrics are better able to distinguish different levels of ranking bias when the overall distribution of protected and non-protected items in the ranking is more balanced. When ranking bias is disadvantaging a protected group that only contains a small number of items, rRD appears to be the most suitable metric because it is the least vulnerable in this case.

2. Which metric is most suitable also depends on how severe the bias in the ranking is estimated to be. Whereas rND outputs the most divergent values for mild cases of ranking bias, rKL distinguishes more severe cases of ranking bias better. Although rRD is slightly better in distinguishing medium levels of negative ranking bias, we do not recommend using it at all due to its normalization issues and weak performance when ranking bias is positive.

3. If the minority viewpoint is preferred in the ranking, ranking bias is well detected by all three metrics. However, when the minority group is *disadvantaged*, all metrics show a decrease in performance. In this case, we suggest using either rND or rKL, depending on how strong the ranking bias is.

8.3.2. Multinomial Viewpoint Fairness

We find that our novel metric nDJS can assess multinomial ranking fairness. Similarly to the binomial fairness metrics, nDJS can distinguish different levels of ranking bias best when the overall distribution of advantaged and disadvantaged viewpoints is balanced. A

Table 8.4: Recommended metrics for different scenarios of ranking bias and overall viewpoint distribution (i.e., protected and non-protected items) in a ranked list.

		Ranking Bias		
		Low	Medium	High
Distribution	Low balance	rND	rND	rND
	Medium balance	rND	rND	rKL
	High balance	rND	rKL	rKL

weakness of nDJS is that its normalization causes its outcome values to be much lower in general compared to binomial fairness metrics. We note that nDJS cannot be directly compared to rND, rRD, or rKL and recommend interpreting nDJS carefully when ranking bias is mild.

8.3.3. Caveats and Limitations

We note that our simulation study is limited in at least three important ways. First, our results are indeed based on mere *simulations*. Although we believe that such simulations are essential for considering a wide array of possible scenarios, we may have missed realistic ranking bias settings in which the metrics we tested perform differently. For instance, our simulation of multinomial viewpoint fairness included only one specific case in which one viewpoint is treated differently compared to the other six (see Section 8.2). There are other scenarios where multinomial viewpoint fairness could become relevant. These scenarios differ in how many viewpoint categories there are, how many items are advantaged in the ranking, and to what degree. Simulating all of these potential scenarios was beyond the scope of this chapter.

Second, we consider a scenario in which documents have correctly been assigned multinomial viewpoint labels. This allows us to study their behavior in a controlled setting. In reality, existing viewpoint labeling methods are prone to biases and issues of accuracy. Current opinion mining techniques are still limited in their ability to assign such labels [366], and crowdsourcing viewpoint annotations from human annotators can be costly and also prone to biases and variance [368].

Third, we assume that any document in a search result ranking can be assigned some viewpoint label concerning a disputed topic. It is realistically possible for a document to express several or even all available viewpoints (e.g., a debate forum page). In these cases, assigning an overarching viewpoint label might oversimplify the nuances in viewpoints that exist *within* rankings and thereby lead to a skewed assessment of viewpoint bias in the search result ranking.

8.4. Conclusion

In this chapter, we adapted existing ranking fairness metrics to measure binomial viewpoint fairness and proposed a novel metric that evaluates multinomial viewpoint fairness. We found that, despite some limitations, these metrics reliably detect viewpoint bias in search results in our controlled scenarios. Our simulations further show the relative strengths of these metrics and how they can be interpreted. A crucial challenge that remains, however, is to accommodate more comprehensive viewpoint representations (see Part I), e.g., to consider not only stances but also logics of evaluation when evaluating search result viewpoint bias. Based on the findings from this chapter, the upcoming Chapter 9 will propose a novel viewpoint bias metric for search results that overcomes this limitation of current metrics. We compare this novel metric to some of the ranking fairness metrics applied in this chapter and use it to measure viewpoint bias in real search results from popular search engines.

9

Comprehensive Viewpoint Bias Evaluation and Viewpoint Diversification for Search Results

---

This chapter is based on a published, full conference paper: Tim Draws, Nirmal Roy, Oana Inel, Alisa Rieger, Rishav Hada, Mehmet Orcun Yalcin, Benjamin Timmermans, and Nava Tintarev. "Viewpoint Diversity in Search Results". In: *Advances in Information Retrieval*. Ed. by Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo. Vol. 13980. Cham: Springer Nature Switzerland, 2023, pp. 279–297. DOI: `10.1007/978-3-031-28244-7_18`.

Tim Draws primarily planned and carried out the conceptualization, investigation, project administration, visualization, and write-up of the work described in the paper referenced above. Tim collaborated with Nirmal Roy, Oana Inel, Alisa Rieger, and Mehmet Orcun Yalcin to implement the methodology. Nirmal Roy implemented the re-ranking algorithms we describe in Section 9.2.3. Benjamin Timmermans and Nava Tintarev supervised the project and made edits to the writing.

Measuring and reducing bias in search results has been studied extensively in recent years, e.g., to satisfy pluralities of search intents [4, 64, 308] or ensure fairness toward particular document classes [35, 381, 391, 393] (see Section 2.2.4). We have made a first attempt at measuring *viewpoint bias* specifically (using ranking fairness metrics) in Chapter 8, and recent work has already explored fostering *viewpoint diversity* in ranked outputs [240, 345]. However, two essential aspects have not been sufficiently addressed yet by previous research: (1) current methods only allow for limited viewpoint representations, i.e., one-dimensional, often binary labels, and (2) there is no clear conceptualization of viewpoint diversity or what constitutes viewpoint bias in search results. Current methods often assume that any top k portion of a ranked list should represent all available (viewpoint) categories proportionally to their overall distribution, i.e., analogous to the notion of *statistical parity* (see Chapter 8), without considering other notions of diversity [361]. This impedes efforts to meaningfully assess viewpoint bias in search results or measure improvements made by diversification algorithms. Hence, the current chapter focuses on the following research questions:

$\mathbf{RQ_{III.1}}$ What metric can thoroughly measure viewpoint bias in search results?

$\mathbf{RQ_{III.2}}$ What is the degree of viewpoint bias in actual search results?

$\mathbf{RQ_{III.3}}$ What method can foster viewpoint diversity in search results?

We address **$\mathbf{RQ_{III.1}}$** by proposing a metric that evaluates viewpoint bias (i.e., deviation from viewpoint diversity) considering the two-dimensional viewpoint representation we propose in Chapter 5. We show that this metric assesses viewpoint bias in a more comprehensive fashion than current methods and apply it in a case study of search results from two popular search engines (**$\mathbf{RQ_{III.2}}$**; Section 9.2). We find notable differences in search result viewpoint bias between queries, topics, and search engines and show that applying existing diversification methods can increase viewpoint diversity (**$\mathbf{RQ_{III.3}}$**; Section 9.2.3).

Supplementary material (e.g., data sets and code) related to this chapter is available at `https://osf.io/kz3je`.

9.1. Evaluating Viewpoint Bias in Search Results

This section introduces a novel metric for measuring viewpoint bias in ranked lists such as search results. To comprehensively capture documents' viewpoints, we adopt the two-dimensional viewpoint representation we propose in Chapter 5. Each document thus receives a single *stance* label on a seven-point ordinal scale from strongly opposing (-3) to strongly supporting (3) a topic and anywhere from no to seven *logic of evaluation* labels that reflect the underlying reason(s) behind the stance (i.e., *inspired, popular, moral, civic, economic, functional, ecological*). Although other viewpoint representations could be modeled, this two-dimensional representation supports more nuanced viewpoint bias analyses than current approaches, and it is still computationally tractable (i.e., only seven topic-independent categories per dimension).

We consider a set of documents retrieved in response to a query (e.g., `school uniforms well-being`) related to a particular debated topic (e.g., mandatory school uniforms). R is a ranked list of N retrieved documents (i.e., by the search engine), $R_{1...k}$ is the top-k portion of R, and R_k refers to the k^{th}-ranked document. We refer to the sets of stance and logic labels of the documents in R as $\mathscr{S}$ and $\mathscr{L}$, respectively, and use $\mathscr{S}_k$ or $\mathscr{L}_k$ to refer to the labels of the particular document at rank k. For instance, a document at rank k may receive the label [$\mathscr{S}_k = 2$; $\mathscr{L}_k =$ (*popular, functional*)] if the article *supports* (stance) school uniforms because they supposedly are popular among students (i.e., *popular* logic) and lead to better grades (i.e., *functional* logic). S and L, respectively, are the (multinomial) stance and logic distributions of the documents in R.

Defining Viewpoint Diversity

Undesired effects such as the *search engine manipulation effect*, whereby users change their opinion following viewpoint biases in search results (see Section 2.2.5), typically occur when search results are one-sided and unbalanced in terms of viewpoints [21, 99, 274]. To overcome this, we follow the normative values of *deliberative democracy* [147], and counteract these problems through viewpoint plurality and balance. We put these notions into practice by following three intuitions:

1. *Neutrality.* A set of documents should feature both sides of a debate equally and not take any particular side when aggregated. We consider a search result list as neutral if averaging its stance labels results in 0 (a neutral stance score).

2. *Stance Diversity.* A set of documents should have a balanced stance distribution so that different stance strengths (e.g., 1, 2, and 3) are covered. For example, we consider a search result list as stance-diverse if it contains equal proportions of all seven different stance categories, but not if it contains only the stance categories −3 and 3 (albeit satisfying *neutrality* here).

3. *Logic Diversity.* A set of documents should include a plurality of reasons for different stances (i.e., balanced logic distribution *within* each stance category). For example, a search result list may not satisfy *logic diversity* if documents containing few reasons (here, logics) are over-represented.

Our metric *normalized discounted viewpoint bias* (nDVB) measures the degree to which a ranked list *diverges* from a pre-defined scenario of ideal viewpoint diversity. It combines the three sub-metrics *normalized discounted polarity bias* (nDPN), *normalized discounted stance bias* (nDSB), and *normalized discounted logic bias* (nDLB), which respectively assess the three characteristics of a viewpoint-diverse search result list (i.e., *neutrality*, *stance diversity*, and *logic diversity*).

9.1.1. Measuring Polarity, Stance, and Logic Bias

We propose three sub-metrics that contribute to nDVB by considering different document aspects. They all ignore irrelevant during their computation and – like other IR evaluation metrics [286] – apply a discount factor for rank-awareness.

Normalized Discounted Polarity Bias (nDPB)

Polarity bias considers the mean stance label balance. *Neutrality*, the first trait in our viewpoint diversity notion, posits that the stance labels for documents in any top k portion should balance each other out (mean stance = 0). We assess how much a top k

search result list *diverges* from this ideal scenario (i.e., *polarity bias*; PB; see Eq. 9.1) by computing the average normalized stance label. Here, $\mathscr{S}_{1...k}$ is the set of stance labels for all documents in the top k portion of the ranking. PB normalizes all stance labels $\mathscr{S}_i$ in the top k to a score between -1 and 1 (by dividing it by its absolute maximum, i.e., 3) and takes their average. To evaluate the neutrality of an entire search result list τ with N documents, we compute PB iteratively for the top $1, 2, \ldots, N$ ranking portions, aggregate the results in a discounted fashion, and apply min-max normalization to produce nDPB (see Eq. 9.2). Here, Z is a normalizer equal to the highest possible value for the aggregated and discounted absolute PB values and I is an indicator variable equal to -1 if $\sum_{k=1}^{N} \frac{\text{PB}(\mathscr{S},k)}{log_2(k+1)} < 0$ and 1 otherwise. nDPB quantifies a search result list's bias toward opposing or supporting a topic and ranges from -1 to 1 (more extreme values indicate greater bias, values closer to 0 indicate neutrality).

$$\text{PB}(\mathscr{S}, k) = \frac{\sum_{i=1}^{k} \frac{\mathscr{S}_i}{3}}{|\mathscr{S}_{1...k}|} \qquad (9.1) \qquad \text{nDPB}(\tau) = \frac{1}{Z} I \sum_{k=1}^{N} \frac{|\text{PB}(\mathscr{S}, k)|}{\log_2(k+1)} \qquad (9.2)$$

Normalized Discounted Stance Bias (nDSB)

Stance bias evaluates how much the stance distribution diverges from the viewpoint-diverse scenario. *Stance diversity,* the second trait of our viewpoint diversity notion, suggests that all stance categories are equally covered in any top k portion. We capture this ideal scenario of a balanced stance distribution in the uniform target distribution $T = \left(\frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}\right)$. The stance distribution of the top k-ranked documents is given by $S_{1...k} = \left(\frac{|\mathscr{S}^{-3}_{1...k}|}{k}, \ldots, \frac{|\mathscr{S}^{3}_{1...k}|}{k}\right)$, where each numerator refers to the number of top-k search results in a stance category. We assess how much $S_{1...k}$ diverges from T by computing their *Jensen-Shannon divergence* (JSD), a symmetric distance metric for discrete probability distributions [108]. This approach is inspired by work suggesting divergence metrics to measure viewpoint bias [360, 361] (see also Chapter 8). We then normalize JSD between $S_{1...k}$ and T by dividing it by the maximal divergence, i.e., $\text{JSD}(U \| T)$ where $U = (1, 0, 0, 0, 0, 0, 0)$ and call the result *stance bias* (SB; see Eq. 9.3). SB ranges from 0 (desired scenario of stance diversity) to 1 (maximal stance bias). Notably, SB will deliberately *always* return high values for the very top portions (e.g., top one or two) of any search result list, as it is impossible to get a balanced distribution of the seven stance categories in just a few documents. We evaluate an entire search result list using nDSB (see Eq. 9.4), by computing SB iteratively for the top $1, 2, \ldots, N$ ranking portions, aggregating the results in a discounted fashion, and normalizing.

$$\text{SB}(S, k) = \frac{JSD(S_{1...k} || T)}{JSD(U || T)} \qquad (9.3) \qquad \text{nDSB}(\tau) = \frac{1}{Z} \sum_{k=1}^{N} \frac{\text{SB}(S, k)}{\log_2(k+1)} \qquad (9.4)$$

Normalized Discounted Logic Bias (nDLB)

Logic bias measures how balanced documents in each stance category are in terms of logics. *Logic diversity* suggests that all logics are equally covered in each document group when splitting documents by stance category. Thus, when a search result list contains documents, e.g., with stances -1, 0, and 1, the logic distributions of each of those three groups should be balanced. The logic distribution of all top k results belonging to a particular stance category s is given by $L^s_{1\ldots k} = \big(\frac{|\mathscr{L}^{s,l_1}_{1\ldots k}|}{|\mathscr{L}^{s}_{1\ldots k}|}, \ldots, \frac{|\mathscr{L}^{s,l_7}_{1\ldots k}|}{|\mathscr{L}^{s}_{1\ldots k}|}\big)$, where each numerator $|\mathscr{L}^{s,l}_{1\ldots k}|$ refers to the number of times logic l (e.g., *inspired*) appears in the top k documents with stance category s. Each denominator $|\mathscr{L}^{s}_{1\ldots k}|$ is the total number of logics that appear in the top k documents with stance category s. $L^s_{1\ldots k}$ reflects the relative frequency of each logic in the top k documents in a specific stance category. Similar to SB, we evaluate the degree to which $L^s_{1\ldots k}$ diverges from T by computing the normalized JSD for the logic distributions of each available stance category and then produce *logic bias* (LB) by averaging the results (Eq. 9.5). Here, $\mathscr{S}^*_k$ is the set of unique stance categories among the top k-ranked documents. LB thus quantifies, on a scale from 0 to 1, the average degree to which the logic distributions diverge from the ideal, viewpoint-diverse scenario where all logics are equally present within each stance category. We produce nDLB by computing LB iteratively for the top $1, 2, \ldots, N$ documents and applying our discounted aggregation and normalization procedures (Eq. 9.6).

$$\mathrm{LB}(\mathscr{S}, L, k) = \frac{1}{|\mathscr{S}^*_k|} \sum_{s \in \mathscr{S}^*_k} \frac{JSD(L^s_{1\ldots k}||T)}{JSD(U||T)} \qquad (9.5) \qquad \mathrm{nDLB}(\tau) = \frac{1}{Z} \sum_{k=1}^{N} \frac{\mathrm{LB}(\mathscr{S}, L, k)}{\log_2(k+1)} \qquad (9.6)$$

9.1.2. Normalized Discounted Viewpoint Bias

To evaluate overall viewpoint bias, we combine nDPB, nDSB, and nDLB into a single metric, called *normalized discounted viewpoint bias* (nDVB):

$$\mathrm{nDVB}(\tau) = I \frac{\alpha|\mathrm{nDPB}(\tau)| + \beta \mathrm{nDSB}(\tau) + \gamma \mathrm{nDLB}(\tau)}{\alpha + \beta + \gamma}.$$

Here, I is an indicator variable that equals -1 when $\mathrm{nDPB}(\tau) < 0$ and 1 otherwise. The parameters α, β, and γ are weights that control the relative importance of the three submetrics. Thus, nDVB measures the degree to which a ranked list of documents diverges from an ideal, viewpoint-diverse scenario. It ranges from -1 to 1, indicating the direction and severity with which such a ranked list (e.g., search results) is biased (values closer to 0 imply greater viewpoint diversity). Our proposed metric nDVB allows for a more comprehensive assessment of viewpoint bias in search results compared to metrics such as rND or RB (see Section 2.2.4 and Chapter 8). It does so by allowing for comprehensive viewpoint representations of search results, simultaneously considering *neutrality*, *stance diversity*, and *logic diversity*.

9.2. Case Study: Evaluating Viewpoint Bias and Fostering Viewpoint Diversity

This section presents a case study in which we show how to practically apply the viewpoint bias metric we propose (nDVB; see Section 9.1.2). We examine viewpoint biases in real search results from commonly used search engines, using relevant queries for currently debated topics (i.e., *atheism*, *school uniforms*, and *intellectual property*). Finally, we demonstrate how viewpoint diversity in search results can be enhanced using existing diversification algorithms. More details on the materials and results (incl. figures) are available in our repository.

9.2.1. Materials

Topics. We aimed to include in our case study three topics that are not scientifically answerable (i.e., with legitimate arguments in both the opposing and supporting directions) and cover a broad range of search outcomes (i.e., consequences for the individual user, a business, or society). To find such topics, we considered the *IBM-ArgQ-Rank-30kArgs* data set [130], which contains arguments on controversial issues. The three topics we (manually) selected from this data set were *atheism* (where attitude change may primarily affect the user themselves, e.g., they become an atheist), *intellectual property rights* (where attitude change may affect a business, e.g., the user decides to capitalize on intellectual property they own), and *school uniforms* (where attitude change may affect society, e.g., the user votes to abolish school uniforms in their municipality).

Queries. We conducted a user study to find, per topic, five different queries that users might enter into a web search engine if they were wondering whether one should be an atheist (individual use case), intellectual property rights should exist (business use case), or students should have to wear school uniforms (societal use case). In a survey, we asked participants to imagine the three search scenarios and select, for each, three "neutral" and four "biased" queries from a pre-defined list. The neutral queries did not specify a particular debate side (e.g., `school uniforms opinions`), while the biased queries prompted opposing (e.g., `school uniforms disadvantages`) or supporting results (e.g., `school uniforms pros`).

We recruited 100 participants from *Prolific* (`https://prolific.co`) who completed our survey for a reward of $0.75 (i.e., $8.09 per hour). All participants were fluent English speakers older than 18. For our analysis, we excluded data from two participants who had failed at least one of two attention checks. The remaining 98 participants were gender-balanced (49% female, 50% male, 1% non-binary) and rather young (50% were between 18 and 24). We selected five queries per topic: the three most commonly selected neutral queries and the single most commonly selected opposing- and supporting-biased queries (see Table 9.1).[1]

Search Results. We retrieved the top 50-ranked search results for each of the $3 \times 5 = 15$ queries listed in Table 9.1 from two of the most commonly used search engines through web crawling or an API.[2] This resulted in a data set of $15 \times 2 \times 50 = 1500$ search results, 25 of which (mostly the last one or two results) were not successfully retrieved. The remaining 1475 (i.e., 973 unique) search results were recorded, including their query, URL, title, and snippet.

---

[1] Due to error, we used the 2^{nd} most common supporting query for the *IPR* topic.

[2] The retrieval took place on December 12th, 2021 in the Netherlands.

Table 9.1: Viewpoint diversity evaluation for all 30 search result lists from Engine 1 and 2: rND, RB, and nDVB (including its sub-metrics nDPB, nDSB, and nDLB). Queries were designed to retrieve neutral (neu), opposing (opp), or supporting (sup) results (↔).

		Engine 1						**Engine 2**					
Query	↔	rND	RB	nDPB	nDSB	nDLB	nDVB	rND	RB	nDPB	nDSB	nDLB	nDVB
`why people become atheists or theists`	neu	.70	.27	.32	.33	.38	.34	.69	.14	.21	.36	.33	.30
`should I be atheist or theist`	neu	.68	.13	.24	.39	.44	.35	.80	.04	.05	.51	.40	.32
`atheism vs theism`	neu	.58	-.06	-.07	.52	.37	-.32	.77	.01	.03	.53	.39	.32
`why theism is better than atheism`	opp	.47	.19	.22	.28	.35	.29	.53	-.04	-.15	.45	.30	-.30
`why atheism is better than theism`	sup	.35	.05	.15	.23	.43	.27	.68	.10	.15	.45	.34	.31
`why companies maintain or give away IPRs`	neu	.77	.46	.49	.41	.45	.45	.97	.61	.60	.48	.51	.53
`should we have IPRs or not`	neu	.80	.34	.34	.35	.33	.34	.93	.47	.44	.42	.41	.43
`IPRs vs open source`	neu	.80	.10	.09	.45	.43	.32	.92	.18	.19	.57	.53	.43
`why IPRs don’t work`	opp	.69	.30	.33	.42	.40	.38	.54	.18	.19	.40	.35	.31
`should we respect IPRs`	sup	.90	.48	.49	.41	.36	.42	.95	.60	.59	.50	.35	.48
`why countries adopt or ban school unif.`	neu	.59	-.01	.14	.37	.25	.26	.54	-.10	-.11	.37	.20	-.23
`should students wear school unif. or not`	neu	.62	-.10	-.10	.45	.20	-.25	.85	.14	.15	.42	.19	.26
`school unif. well-being`	neu	.55	.07	.09	.28	.25	.21	.54	.13	.23	.31	.35	.30
`why school unif. don’t work`	opp	.30	-.22	-.31	.33	.18	-.27	.59	-.01	-.03	.37	.21	-.20
`why school unif. work`	sup	.89	.43	.49	.38	.27	.38	.92	.45	.03	.50	.39	.36
Overall mean absolute bias		.65	.21	.26	.37	.34	.32	.75	.21	.24	.44	.34	.34

Note. In contrast to the actual queries, we here abbreviate *intellectual property rights* (IPRs) and *uniforms* (unif.).

Viewpoint Annotations. To assign each search result the two-dimensional viewpoint label (see Section 9.1), we employed six experts, familiar with the three topics, the annotation task, and the viewpoint labels. This is more than the one to three annotators typically employed for information retrieval (IR) annotation practices [128, 359]. The viewpoint label consists of *stance* (i.e., position on the debated topic on an ordinal scale ranging from -3; strongly opposing; to 3; strongly supporting) and *logics of evaluation* (i.e., arguments or reasons behind the stance).[3] First, the experts discussed annotation guidelines and examples before individually annotating the same set of 30 search results (i.e., two results randomly chosen per query). Then, they discussed their disagreements, created an improved, more consistent set of annotation guidelines, and revised their annotations. Following discussions, their overall agreement increased to satisfactory levels for stance (Krippendorff's $\alpha = .90$) and the seven logics ($\alpha = \{.79, .66, .73, .86, .77, .36, .57\}$). Such agreement values represent common ground in the communication sciences, where, e.g., two trained annotators got $\alpha = \{.21, .58\}$ when annotating *morality* and *economical* frames in news [51]. Each expert finally annotated an equal and topic-balanced share of the remaining 943 unique search results.

9.2.2. Viewpoint Bias Evaluation Results

We conducted viewpoint bias analyses per topic, search engine, and query. Specifically, we examined the overall viewpoint distributions and then measured viewpoint bias in each of the ($15 \times 2 =$) 30 different top 50 search result lists retrieved from the two search engines, by computing the existing metrics rND and RB (see Section 2.2.4 and Chapter 8) as well as our proposed metric including its sub-metrics (see Section 9.1).

---

[3] Note that viewpoint labels do not refer to specific web search queries, but always to the topic (or claim) at hand. For example, a search result supporting the idea that students should have to wear school uniforms always receives a positive stance label (i.e., 1, 2, or 3), no matter what query was used to retrieve it.

Overall Viewpoint Distributions

Among the 973 unique URLs in our search results data set, 306, 334, and 263 respectively related to the topics *atheism*, *intellectual property rights* (IPRs), and *school uniforms*. A total of 70 unique search results were judged irrelevant to their topic and excluded from the analysis. Search Engine 1 (SE_1) provided a somewhat greater proportion of unique results for the 15 queries (77%) than Search Engine 2 (SE_2, 69%). For all three topics, supporting stances were more common. Regarding logics, the *school uniforms* topic was overall considerably more balanced than the others. Atheism-related documents often focused on *inspired*, *moral*, and *functional* logics (e.g., religious people have higher moral standards, atheism explains the world better). Documents related to IPRs often referred to *civic, economic*, and *functional* logics (e.g., IPRs are an important legal concept, IPRs harm the economy).

Viewpoint Bias per Query, Topic, and Search Engine

We analyzed the viewpoint bias of search results using the existing metrics rND, RB, and our proposed (combined) metric nDVB. We slightly adapted rND and RB to make their outcomes more comparable, aggregating both in steps of one and measuring viewpoint *imbalance* (or bias) rather than ranking fairness. Our rND implementation considered all documents with negative stance labels as *protected*, all documents with positive stance labels as *non-protected*, and ignored neutral documents. Computing RB required standardizing all stance labels to scores ranging from -1 to 1. To compute nDVB, we set the parameters to $\alpha = \beta = \gamma = 1$, i.e., giving all sub-metrics equal weights. Table 9.1 shows the evaluation results for all metrics across the 30 different search result lists from the two search engines. Scores closer to 0 suggest greater diversity (i.e., less distance to the ideal scenario), whereas scores further away from 0 suggest greater bias.

Divergence from Neutrality. As we note in Section 9.1, viewpoint-diverse search result lists should feature both sides of debates equally. While rND does not indicate whether a search result list is biased against or in favor of the protected group (see Chapter 8), the RB and nDPB outcomes suggest that most of the search result lists we analyzed are biased towards *supporting* viewpoints. We observed that results on IPRs tended to be more biased than results on the other topics but, interestingly, we did not observe clear differences between query types. Moreover, except for the *school uniforms* topic, supposedly neutral queries generally returned results that were just as biased as queries targeted specifically at opposing or supporting results.

Divergence from Stance Diversity. Another trait of viewpoint-diverse search result lists is a balanced stance distribution. Since rND, RB, and nDPB cannot clarify whether all stances (i.e., all categories ranging from -3 to 3) are uniformly represented, we here only inspect the nDSB outcomes. While we did not observe a noteworthy difference between topics or queries, we found that SE_2 returned somewhat more biased results than SE_1. Closer examination of queries where the two engines differed most in terms of nDSB (e.g., *why theism is better than atheism*) revealed that SE_2 was biased in the sense that it often returned fewer opinionated (and more neutral) results than SE_1. Regarding their balance between mildly and extremely opinionated results, both engines behaved similarly.

Divergence from Logic Diversity. The final characteristic of viewpoint-diverse search result lists concerns their distribution of logics, i.e., the diversity of reasons brought forward to oppose or support topics. When inspecting the nDLB outcomes, we found that logic distributions in the search result lists were overall more balanced than stance distributions (see nDSB results) and similar across search engines and queries. However, we did observe that nDLB on the *school uniforms* topic tended to be lower than for other topics, suggesting that greater diversities of reasons opposing or supporting school uniforms were brought forward.

Overall Viewpoint Bias. To evaluate overall viewpoint bias in the search result lists, we examined nDVB, the only metric that simultaneously evaluates divergence from neutrality, stance diversity, and logic diversity. Bias *magnitude* per nDVB ranged from .20 to .53 across results from search engines, with only four out of 30 search result lists being biased against the topic. Regarding topics, search results for neutral queries were somewhat less biased on *school uniforms* compared to *atheism* or *intellectual property rights.*

Interestingly, search results for neutral queries on all topics were often just as viewpoint-biased as those from directed queries. Some queries returned search results with different bias magnitudes (e.g., *school uniforms well-being*) or bias directions (e.g., *atheism vs theism*) depending on the search engine. Moreover, whereas search results for supporting-biased queries were indeed always biased in the supporting direction (i.e., positive nDVB score), results for opposing-biased queries were often also biased towards supporting viewpoints.

Figure 9.1 shows, per topic and search engine, how the absolute nDVB developed on average when evaluated at each rank. It illustrates that nDVB tended to decrease over the ranks across engines, topics, and queries but highlights that the top, say 10, search results that users typically examine are often much more viewpoint-biased than even the top 30 (i.e., more search results could offer more viewpoints).

9.2.3. Viewpoint Diversification

To improve viewpoint diversity, we build on earlier work on diversifying search results concerning *user intents* [2, 4, 91, 175, 314]. *xQuAD* [314] and *HxQuAD* [155] are two such models that re-rank search results with the aim of fulfilling diverse ranges of information needs at high ranks. Whereas xQuAD diversifies for single dimensions of (multi-categorical) subtopics, HxQuAD adapts xQuAD to accommodate multiple dimensions of subtopics and diversifies in a multi-level hierarchical fashion. For example, for the query *java*, two first-level subtopics may be *java island* and *java programming*. For the former, queries such as *java island restaurant* and *java island beach* may then be second-level subtopics. To the best of our knowledge, such methods have so far not been used to foster viewpoint diversity in ranked lists.

We implemented four diversification algorithms to foster viewpoint diversity in search results by (1) re-ranking and (2) creating viewpoint-diverse top 50 search result lists using

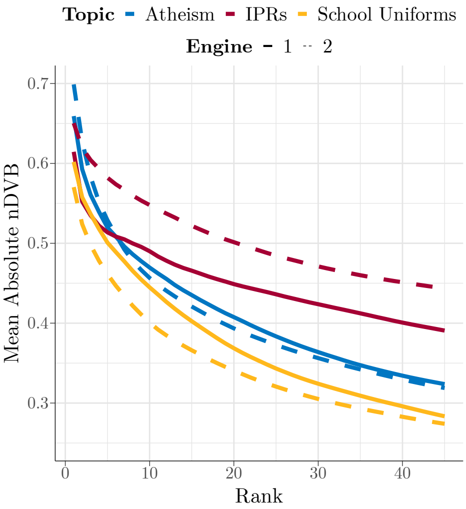


Figure 9.1: Development of mean absolute nDVB@k across search result ranks, split by topic and search engine.

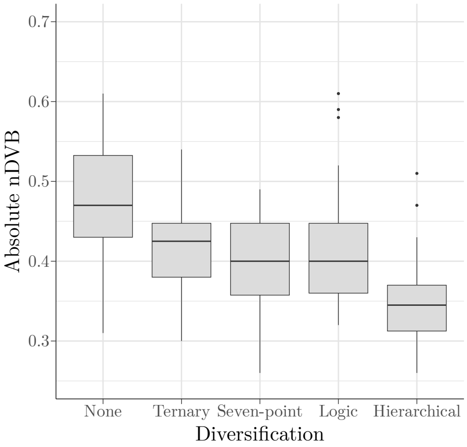


Figure 9.2: Mean absolute viewpoint diversity (nDVB@10) per diversification algorithm across the 30 search result lists.

all unique results from each topic. Specifically, we performed *ternary stance diversification*, *seven-point stance diversification*, *logic diversification* (all based on xQuAD; i.e., diversifying search results according to ternary stance labels, a seven-point stance taxonomy, or logic labels, respectively), and *hierarchical viewpoint diversification* (based on HxQuAD; i.e., diversifying search results hierarchically: first for seven-point ordinal stance labels and then, within each stance category, for logic labels; giving both dimensions equal weights). We evaluated the resulting search result lists using nDVB.

Re-ranked Top 50 Search Result Lists

Figure 9.2 compares absolute nDVB between the original top 50 search result lists and the four diversification strategies. All strategies improved the viewpoint diversity of our lists. Whereas the ternary stance diversification only showed marginal improvements (mean abs. nDVB@10 = .42, nDVB@50 = .35) compared to the original search result lists (mean abs. nDVB@10 = .47, nDVB@50 = .33), the hierarchical viewpoint diversification based on stances and logics was the most effective in fostering viewpoint diversity (mean abs. nDVB@10 = .35, nDVB@50 = .27). Viewpoint bias for the seven-point stance diversification (mean abs. nDVB@10 = .39, nDVB@50 = .29) and logic diversification (mean abs. nDVB@10 = .42, nDVB@50 = .31) were comparable, and in between the ternary stance and hierarchical diversification.

“Best-case” Comparison

Despite the promising re-ranking results, diversification methods can only work with the specific sets of documents they are given. To show a “best-case” scenario for comparison, we employed our diversification algorithms to create, per topic, one maximally viewpoint-diverse search result list using all topic-relevant search results (i.e., from across queries and search engines). We found that all four diversification algorithms yielded search result lists with much less bias when given more documents compared to when they only re-ranked top 50 search results lists. Here, the hierarchical diversification was again most effective (mean abs. nDVB@10 = .29, nDVB@50 = .20); improving by a magnitude of .07 on average over the re-ranked top 50 search result lists. Compared to the average search result list we had retrieved from the two search engines, the "best-case" hierarchical diversification improved viewpoint diversity by margins of .17 (nDVB@10) and .13 (nDVB@50), reflecting a mean improvement of 39%. The other diversification algorithms showed similar improvements, albeit not as impactful as the hierarchical method (i.e., mean abs. nDVB@10 was .37, .37, .34 and mean abs. nDVB@50 was .31, .24, .24 for the ternary stance, seven-point stance, and logic diversifications, respectively).

9.3. Discussion

In this chapter, we identified that viewpoint diversity in search results can be conceptualized based on the deliberative notion of diversity by looking at *neutrality*, *stance diversity*, and *logic diversity*. Although we were able to adapt existing metrics to partly assess these aspects, a novel metric was needed to incorporate all of them simultaneously. We thus proposed the metric *normalized discounted viewpoint bias* (nDVB), which considers two important viewpoint dimensions (i.e., *stances* and *logics of evaluation*) and measures viewpoint bias, i.e., the deviation of a search result list from an ideal, viewpoint-diverse scenario (**RQ$_{\mathbf{III.1}}$**). Findings from our case study suggest that nDVB is sensitive to expected data properties, e.g., in aligning with the query polarity and decreasing bias for larger lists of search results. Although further refinement and investigation of the metric are required (e.g., to find the most practical and suitable balance between the three notions of diversity or outline interpretation guidelines), our results indicate that the metric is a good foundation for measuring viewpoint bias.

The degree of viewpoint bias across search engines in our case study was comparable: neither engine was consistently more biased than the other (**RQ$_{\mathbf{III.2}}$**). However, we found notable differences in bias magnitude and even bias direction between search engines *regarding the same query* and queries related to the same topic. This lends credibility to the idea that nDVB indeed measures viewpoint bias and is able to detect different kinds of biases. Similar to previous research [370], we found that search results were mostly biased in the *supporting* direction. This suggests that actual search results on debated topics may often not reflect a satisfactory degree of viewpoint diversity and instead be systemically biased in terms of viewpoints. More worryingly, depending on where (which search engine) or how (which query) users search for information, they may not only be exposed to different viewpoints but ones representing a different bias than their peers. We also found that neutrally formulated queries often returned similarly biased search results as queries calling for specific viewpoints. In light of findings surrounding the search engine manipulation effect (SEME; see Section 2.2.5) and similar effects, this could have serious ramifications for individual users' well-being, business decision-making, and societal polarization.

Our case study further showed that diversification approaches based on xQuAD and HxQuAD can improve the viewpoint diversity in search results. Here, the hierarchical viewpoint diversification (based on HxQuAD, and able to consider both documents' *stances* and *logics of evaluation*) was most effective ($\mathbf{RQ_{III.3}}$).

Limitations

Although our case study covered debated topics with consequences for individuals, businesses, and society, it is important to note that our results may not generalize to all search engines and controversial issues. We carefully selected the deliberative notion of diversity to guide our work as we believe it suits many debated topics, especially those with legitimate arguments on all sides of the viewpoint spectrum. However, we note that some scenarios may require applying other diversity notions and that presenting search results according to the deliberative notion of diversity (i.e., representing all viewpoints equally) may even cause harm to individual users or help spread fake news. For example, this could be the case for topics such as medical treatment or climate change, where only one viewpoint is scientifically correct [10, 37, 274, 372]. Recent work has already begun to address harm prevention in web search [395, 396]. In line with this emerging body of work, assessing and increasing the viewpoint diversity search results is a contentious issue in itself and should always be done with care.

Another limitation of our work is that, despite providing a diverse range of queries to choose from, queries may not have represented all users adequately. Moreover, despite efforts to represent a general user during the search result retrieval, the search results we received (and how they were ranked) may have been different had we entered queries from a different location or at a different time.

Finally, our proposed metric nDVB is still limited in several ways, e.g., it does not yet incorporate document relevance, other viewpoint diversity notions, or the personal preferences and beliefs of users. Implementing such factors may be necessary depending on the use case at hand. Finally, annotating viewpoints is a difficult, time-consuming task even for expert annotators [51] (see also Chapter 5 and Part II). We have already applied automatic stance detection methods to search results in Chapter 3 but, to the best of our knowledge, no earlier work has so far not attempted to identify logics of evaluation. However, once such automatic systems have become more comprehensive, researchers and practitioners could easily combine them with existing methods for extracting arguments [40, 336] and visualize viewpoints [6, 57] in search results.

9.4. Conclusion

Although the tools and methods we propose in Parts I and II allow researchers to create search result data sets with high-quality viewpoint labels, existing metrics can measure viewpoint bias in search results only to a limited extent (see Chapter 8). This chapter proposed a metric for more comprehensive evaluations of search result viewpoint bias, measuring the divergence from an ideal scenario of equal viewpoint representation. Our novel metric nDVB overcomes the limitations of existing methods that cannot handle multi-dimensional viewpoint representations (e.g., incorporating both stances and logics of evaluation). In a case study evaluating search results on three different debated topics from two popular search engines, we found that search results may often not be viewpoint-diverse, even if queries are formulated neutrally. We also saw notable differences between search engines concerning bias *magnitude* and *direction*. Our hierarchical viewpoint diversification, based on HxQuAD, consistently improved the viewpoint diversity of search results. In sum, our results suggest that, while viewpoint bias in search results is not pervasive, users may unknowingly be exposed to high levels of viewpoint bias, depending on the query, topic, or search engine. These factors may influence (especially vulnerable and undecided) users' behavior and opinions by means of recently demonstrated search engine manipulation effects and thereby affect individuals, businesses, and society. In the fourth and final part of this dissertation, we examine user behavior in the context of search result viewpoint biases more closely.

IV

How Search Result Viewpoint Biases Affect User Behavior

In Parts I, II, and III, we have presented empirical results and proposed tools that not only allow for comprehensive measurement of viewpoint bias in search results but also the investigation of user behavior in this context. Recent research has already demonstrated that severe viewpoint bias in search results can lead to phenomena such as the *search engine manipulation effect* (SEME), where users without strong pre-search opinions change their opinions following search result viewpoint biases [10, 36, 37, 99, 274]. For example, viewpoint-biased search results concerning *school uniforms* could lead users to vote in favor of a school uniform mandate without having properly considered the opposing side. Understanding the underlying mechanisms of such user behavior is vital in developing systems that can support users in their web search for debated topics. However, it is still unclear what underlying mechanisms cause such user behavior. This makes it difficult to predict, e.g., at what *degree* viewpoint biases begin to cause systematic user tendencies and whether they occur across search topics. Part IV addresses this research gap by investigating the following research question:

$\mathbf{RQ_{IV}}$ What cognitive processes underlie the effect of search result viewpoint bias on users' opinion formation?

We address **$\mathbf{RQ_{IV}}$** by conducting a user study investigating whether lower-degree viewpoint biases in search results can lead users to adopt particular viewpoints (Chapter 10). Whereas previous research in this area has largely presented users with strongly biased search results, we expose users to single, top 10 search engine results pages (SERPs) that are overall viewpoint-balanced (i.e., five opposing and five supporting results) but are ranked with different degrees of bias (e.g., ranking all opposing results above the supporting results or ranking them in alternating fashion). We find no differences between these rankings of overall viewpoint-balanced top 10 SERPs concerning users' attitude change across topics (e.g., whether zoos should exist). Further analyses provide exploratory evidence that, rather than *order effects* (i.e., posing that users assign stronger weights to search results at higher ranks in their opinion formation), *exposure effects* (i.e., posing that users' opinion formation is affected by the majority viewpoint among the search results they engage with) may guide user behavior in this context.

10

Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics

---

This chapter is based on a published, full conference paper: Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. "This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '21. New York, NY, USA: Association for Computing Machinery, July 2021, pp. 295–305. DOI: `10.1145/3404835.3462851`.

Tim Draws primarily planned and carried out the conceptualization, investigation, methodology, project administration, visualization, and write-up of the work described in the paper referenced above. His co-authors supervised him during the project and made edits to the writing.

In Parts I, II, and III of this dissertation, we have explored how to best represent viewpoints expressed in search results, collect high-quality viewpoint annotations, and measure search result viewpoint biases. This has culminated in an analysis demonstrating that viewpoint bias can occur in web search results across debated topics, queries, and search engines. However, web search interactions can also be biased in other ways [26], e.g., cognitive user biases such as the *confirmation bias* strongly affect how users consume and process information from search results [21, 298]. A well-established finding is that users typically pay more attention to higher-ranked items when consuming ranked search result lists [168]. This phenomenon – known as *position bias* – leads users to primarily engage with the first *search engine results page* (SERP) [43] and click on results at higher ranks with greater probability [168, 177, 258]. Consequently, the ranking of search results greatly influences users' post-search opinions when exploring debated topics: recent work has demonstrated that when a search result ranking is biased towards any particular viewpoint (i.e., assigning higher ranks to documents that express it), users tend to change their opinion accordingly [10, 99, 274] (see Section 2.2.5). This type of opinion change induced by a viewpoint-biased search result list has been called the *search engine manipulation effect* (SEME) [99]. It can occur even after single search sessions, for a variety of topics (e.g., political elections and medical treatment) [10, 99, 274], and without users' awareness [123]. In this chapter, we seek to understand the mechanisms that underlie such user behavior.

Why do users fall prey to SEME? Although position bias can explain how users *select* search results to engage with, it does not explain how users *process* the search results they have picked for consumption. Two different cognitive biases have been suspected to drive the information processing that leads to SEME: *exposure effects* and *order effects* [21]. Exposure effects imply that being exposed to messages pertaining to a particular viewpoint increases an individuals' favorability towards that viewpoint [21, 231, 388]. In the context of web search, this would mean that users' tendency to adopt a particular viewpoint increases with the proportion of consumed documents that express this viewpoint. Order effects occur when users assign *more weight* to information drawn from higher-ranked results [21]. This would mean that the influence of a document's expressed viewpoint is weighted by its rank.

Despite these considerations, there is a lack of empirical evidence as to whether exposure effects, order effects, or both are responsible for SEME. Previous studies have demonstrated SEME using search result lists that had a majority viewpoint among documents on the first SERP [10, 99, 274, 371], which makes both exposure effects and order effects plausible (but not necessary) explanations. For instance, suppose a user queries `should zoos exist?` and sees eight documents supporting zoos and two documents opposing zoos on the first SERP. Assuming that the user engages with only these first ten search results, the user may change their opinion towards favorability for zoos because, among the results they consumed, zoo-supporting documents were in the majority (i.e., exposure effects) and ranked higher (i.e., order effects). More recent research indicates that users may look for majority viewpoints but are unaware of any order effects when they search the web [123]. However, humans are often unaware of their biases [280], so it is currently unclear which cognitive processes truly contribute to SEME.

Mitigating SEME requires a thorough understanding of its underlying mechanisms.

This chapter investigates whether cognitive order effects contribute to SEME by studying the influence of algorithmic ranking bias for *overall viewpoint-balanced* top 10 search results (i.e., the first SERP) on user opinions towards debated topics. Unlike previous research, this method exposes users to SERPs that contain equal proportions of opposing and supporting documents, thereby mitigating potential exposure effects and isolating potential order effects. To explore which users might be particularly vulnerable to SEME, we additionally study whether factors such as *actively open-minded thinking, user engagement,* and *perceived diversity* play a role here. We have four research questions:

$\text{RQ}_{\text{IV.1}}$ Does the ranking of overall viewpoint-balanced top 10 search results affect opinion change concerning debated topics in users with mild pre-existing opinions?

$\text{RQ}_{\text{IV.2}}$ Are individual user characteristics such as *actively open-minded thinking* and *user engagement* related to opinion change?

$\text{RQ}_{\text{IV.3}}$ Do factors such as *actively open-minded thinking, user engagement,* or *perceived diversity* interact with search result rankings to cause opinion change?

$\text{RQ}_{\text{IV.4}}$ Are users aware of varying degrees of viewpoint diversity in the search results they consume?

Results from our preregistered user study show that most users changed their opinion after viewing the search results. However, opinion change did not differ across different result rankings, a finding that contravenes the predictions of order effects. We similarly find no evidence that individual differences (i.e., actively open-minded thinking, user engagement) or perceived diversity affect opinion change directly or in interaction with search result rankings. Exploratory analyses suggest that *exposure effects* may explain opinion change.

Supplementary material related to this chapter, including the data sets, preregistration, results, and data analysis code, are available at `https://osf.io/6tbvw`.

10.1. Hypotheses

Azzopardi [21] argues that two well-known cognitive biases may contribute to SEME: *exposure effects* and *order effects*. Exposure effects occur when exposure to messages pertaining to a particular viewpoint increases an individual's favorability towards that viewpoint [231, 388]. In the context of web search, this would mean that the higher the proportion of consumed documents expressing a particular viewpoint, the greater user's tendency to adopt that viewpoint. Order effects imply a *weighting* of consumed information according to its position in a given order (e.g., assigning more weight to information encountered first) [153]. Order effects may nudge users to adopt the viewpoint that is expressed by highly-ranked documents – even when they consume an overall viewpoint-balanced set of search results that has no majority viewpoint.[1]

---

[1] Order effects could also mean weighting in favor of information encountered *last*. However, in situations such as web search, where the number of consumed items is typically low [258], assigning more weight to information seen *first* is more likely [21, 153].

Both exposure effects and order effects are plausible explanations for SEME, but there is currently a lack of empirical understanding as to whether either of them (or both) are responsible for SEME. In previous studies that explored SEME, the first SERP that participants saw reflected a viewpoint imbalance so that one viewpoint was in the majority among the results on the first page [10, 99, 274, 371]. This means that either exposure effects (i.e., users adopting the majority viewpoint among the results they consume), order effects (i.e, users adopting the viewpoint of higher-ranked results), or a combination of both could have been responsible for SEME. For example, White and Horvitz [371] demonstrated SEME by comparing imbalanced, ranking-biased SERPs (i.e., allowing for both exposure and order effects to take place) to a controlled setting in which the shown SERPs were viewpoint-balanced *and* ranked in random order (i.e., ruling out both exposure and order effects). Recent research indicates that exposure effects may indeed underlie SEME [123], but the evidence surrounding order effects in this context is inconclusive. Users put more *trust* in higher-ranked results [168] and it has been argued that the viewpoint expressed by the first search result acts as an *anchor* in users' exploration of search results [248]. However, users do not consciously experience order effects [123], which have – to the best of our knowledge – not been demonstrated in situations where potential exposure effects are mitigated or ruled out.

If order effects underlie SEME, users will tend to adopt the viewpoint expressed by higher-ranked results, at least when they have mild pre-existing opinions [99, 371]. SEME should then occur even when users consume a ranking-biased but overall viewpoint-balanced set of documents. Our hypothesis follows earlier work arguing that order effects are (partly) responsible for SEME [21, 248].

$H_{III.1a}$ The ranking of an overall viewpoint-balanced list of 10 search results affects opinion change towards debated topics in users with mild pre-existing opinions.

Although SEME has been demonstrated for a variety of topics, to the best of our knowledge, no previous research has assessed topical differences concerning SEME. Given that some topics are more polarizing than others, we expect that the effect of search result rankings on opinion change is moderated by topic.

$H_{III.1b}$ Topic moderates the effect of search result rankings on opinion change.

10.1.1. Vulnerability of Users to Opinion Change

Users with mild pre-existing opinions are more susceptible to opinion change when searching the web compared to users with strong opinions [371]. If biased search result rankings can cause opinion change in such users (i.e., elicit SEME), an important step towards developing mitigation strategies is to understand which other factors (aside from having a mild pre-existing opinion) characterize users who are particularly affected. Psychological research has identified that willingness to process (counter-attitudinal) information [346], and engagement with the topic at hand [271] may increase an individual's vulnerability to opinion change.

We thereby defined two distinct user-specific factors that we expected to (1) predict opinion change directly and (2) affect opinion change in interaction with search result rankings. First, we predicted such a role for *actively open-minded thinking* (AOT). AOT is a style of thinking that involves considering counter-attitudinal information and opinions of others when forming one's own opinion [139]. Consequently, AOT predicts information acquisition [139] and reasoning independently from pre-existing opinion [338].

$\mathbf{H_{III.2a}}$ Actively open-minded thinking (AOT) predicts opinion change in web search.

$\mathbf{H_{III.3a}}$ Actively open-minded thinking (AOT) moderates the effect of search result rankings on opinion change.

Second, we hypothesized that *user engagement* will act as a direct and moderating factor in this context. High interaction with a search result list may be analogous to strong engagement with a topic. Moreover, depending on the degree of ranking bias present in a search result list, engaged users may be exposed to a growing diversity of viewpoints as they move down the search results list.

$\mathbf{H_{III.2b}}$ User engagement predicts opinion change in web search.

$\mathbf{H_{III.3b}}$ *User engagement moderates the effect of search result rankings on opinion change.*

It has been shown that higher engagement with presented information mediates the relationship between AOT and task performance [139], which is why we expected the same in our study.

$\mathbf{H_{III.2c}}$ User engagement mediates the relationship between AOT and opinion change.

Additionally, we expected perceived diversity to moderate the effect of search result rankings on opinion change. Perceiving search result lists as more or less diverse could reflect the degree to which users have recognized potential biases or considered the different viewpoints present on the topic. Merely perceiving that a search result list has a high diversity could therefore change how a search result ranking affects opinion change.

$\mathbf{H_{III.3c}}$ Perceived diversity in search result lists moderates the effect of search result rankings on opinion change.

10.1.2. User Perception of Viewpoint Diversity

To the best of our knowledge, no previous work has explored whether users perceive an existing (lack of) viewpoint diversity in sets of search results. We investigate the effect of search result rankings on perceived diversity to better understand how and why user opinions might be affected. Because previous research has shown that users truly engage with only the top few results on a SERP [168, 250, 258], we expected that the ranking of search results (i.e., different degrees of bias) would skew their perception of search result list diversity.

$\mathbf{H_{III.4a}}$ Search result rankings affect perceived diversity in search result lists.

10.2. Data

Debated Topics. We first conducted a preliminary study to identify a set of disputed topics that most people hold undecided or mild opinions on (i.e., because we aimed to test users without strong pre-search opinions). We picked 18 different topics from *ProCon* (`https://procon.org`), a website that lists controversial issues.[2] We then created a survey in which participants could state their opinion on each of the 18 topics using a seven-point Likert scale ranging from “strongly agree” to “strongly disagree”. Each topic was phrased as a question (e.g., “Should zoos exist?”). A total of 100 participants completed the survey for a $0.60 reward after being recruited from the online participant pool *Prolific* (`https://prolific.co`). We excluded seven responses from data analysis due to failing at least one of two attention checks we had included.

We defined two inclusion criteria for topics. First, we aimed to include topics for which opinions were generally not skewed towards a particular side. We evaluated this by transforming all survey responses to integers ranging from -3 (i.e., “strongly disagree”) to 3 (i.e., “strongly agree”) and subsequently conducting one-sample Wilcoxon tests against a test value of 0 for each topic. A significant result in this test suggested that the mean opinion on the topic at hand is not undecided (i.e., not equal to 0). We thus included only topics for which the Wilcoxon test had a *non-significant* result.[3] Second, we desired topics that a majority of people held a mild (i.e., uncertain) opinion towards. We implemented this criterion by classifying all survey responses into *mild* and *strong* viewpoints: responses among the three central options from the Likert scale (i.e., ranging from “somewhat disagree” to “somewhat agree”) were mapped onto the *mild* class, all other responses were mapped to the *strong* class. We included a topic in our study only if the proportion of mild opinions was above 0.5. Five topics met both criteria and were therefore included in our study:

1. *Are social networking sites good for our society?*
2. *Should zoos exist?*
3. *Is cell phone radiation safe?*
4. *Should bottled water be banned?*
5. *Is obesity a disease?*

Search Results. Per topic, we created a set of 14 queries according to a pre-defined template. This template included neutrally-formulated queries (e.g., `zoos opinions`, `zoos arguments`) as well as viewpoint-biased queries (e.g., `opinions supporting zoos`, `arguments opposing zoos`).[4] We then retrieved the top 50 search results for each of these queries using the API of the search engine *Bing* (`https://bing.com`).

From the search results we retrieved on each topic, we handpicked 56 opinionated search result items (i.e., items that expressed some viewpoint on the topic) and had them annotated by crowd workers on *Amazon Mechanical Turk* (`https://www.mturk.com`). We collected at least three annotations per item, both for relevance (binary) and expressed stance (i.e., representing viewpoints on a seven-point scale ranging from “strongly opposing” to “strongly supporting”). We paid crowd workers \$2 per task in which they annotated 14 different search results. Additionally, workers could earn a \$0.50 bonus if they passed two attention checks. Data from participants who did not pass at least one attention check were excluded from the analysis. According to Krippendorff’s α, inter-rater reliability for the viewpoint judgments was satisfactory ($\alpha = 0.79$) [191]. Qualitative feedback from crowd workers revealed that the task was understandable and could be performed without issues. We assigned each search result item the median annotation for both of these measurements.

---

[2] We excluded topics that were highly politicized (e.g., *gun control* or *abortion*).

[3] We corrected for multiple testing by applying a Bonferroni correction; i.e., only p-values below $\frac{0.05}{18} = 0.003$ were considered significant.

[4] The full list of queries we used is available on our repository.

Our final data set thus consisted of 280 search result items (including title, snippet, and URL for each document) that were annotated concerning their relevance and stance with respect to the five debated topics.

10.3. Method and Experimental Setup

This section describes the materials, procedure, participants, and statistical analysis related to our user study. Next to constructs introduced in Chapter 2, we here describe several additional measurements that we included in our study (i.e., topical interest, gender, and age). We used these measurements for descriptive and exploratory analyses; more specifically, to obtain a clearer image of our sample (e.g., whether participants had a realistic level of topical interest) and to explore directions for future research.

10.3.1. Materials

Search Result Rankings. Using the data set described in Section 10.2, we assembled one set of 10 search results for each of the five topics. We did that by randomly sampling three “opposing”, two “somewhat opposing”, two “somewhat supporting”, and three “supporting” documents from the search results that were deemed relevant to a given topic by crowd workers.[5] Thus, although the search result lists were different in terms of topic and content, they were consistent with respect to the representation of viewpoints, as each topic-specific search result set had the same stance distribution.

We ranked the search result sets to reflect three levels of bias (*little*, *moderate*, and *extreme*) by computing viewpoint diversity for all possible ranking orders using the three metrics rND, rKL, and RB (see Section 2.2.4 and Part III), and summing up the metric outcomes for each ranking. We selected ranking permutations for each level based on their score; *little bias*– lowest combined score, *moderate bias*– closest to the mean, and *extreme bias*– highest combined score. Table 10.1 illustrates the three conditions with a bias toward opposing viewpoints. In practice, we counter-balanced the search result rankings so that half contained bias for the opposing viewpoint and half the supporting viewpoint.[6]

---

[5] We did not include “strongly opposing” and “strongly supporting” items in the search result sets because they were non-existent for several topics.

[6] Note that the metrics have the same output for symmetrical search result rankings (e.g., ranking all opposing documents before any supporting documents and vice versa).

Table 10.1: Three conditions representing the three levels of ranking bias. Here, all rankings are biased toward opposing stances, but our study also included their symmetrical opposites (favoring supporting stances).

	Stance Label		
Rank	Little Bias	Moderate Bias	Extreme Bias
1	-1	-2	-2
2	2	-2	-2
3	1	1	-2
4	- 2	2	-1
5	-1	-1	-1
6	2	2	1
7	-2	-2	1
8	2	2	2
9	-2	-1	2
10	1	1	2

Actively Open-Minded Thinking (AOT) Scale. A seven-item scale that measures the degree to which a person is willing to consider opposing viewpoints and change their mind about topics [139]. Responses were recorded on a seven-point Likert scale ranging from “strongly agree” to “strongly disagree” and later aggregated by taking their mean. To ensure reliability of responses, we added an attention check item to the AOT scale.

User Engagement Scale - Short Form (UES-SF). A 12-item scale that measures the degree to which a person was involved and satisfied with a given experience [249]. Responses were recorded on a seven-point Likert scale and averaged.

Perceived Diversity Scale. We measured *perceived diversity* using adapted versions of items from a scale for measuring recommendation variety in a list of recommended items [184]. Responses were recorded on a seven-point Likert scale and averaged.

10.3.2. Variables

Independent Variables

- *Topic* (categorical; between-subjects). Each participant saw search results that relate to one of five different debated topics that we included in this study (see Section 10.2).

- *Condition* (categorical; between-subjects). Each participant was randomly assigned to one of three conditions that each involve a different ranking of search results. These search result rankings reflected (1) little, (2) moderate, and (3) extreme ranking bias (see Section 10.3.1 and Table 10.1). This variable was nested within *topic*.

Dependent Variable

- *Opinion change* (continuous). We measured each participant's opinion towards their assigned topic twice (i.e., once before and once after exposing them to a ranked list of search results related to a debated topic). Specifically, we asked them to respond to a topic statement (e.g., "Zoos should exist") on a seven-point Likert scale ranging from "strongly disagree" to "strongly agree". Similar to previous research [99], we computed *opinion change* by subtracting the first measurement from the second: [-6,6]. We additionally computed the *absolute opinion change*: [0,6].

Covariates

- *Actively open-minded thinking* (continuous). Measured using the AOT scale: [0;6].
- *User engagement* (continuous). We quantified user engagement by aggregating three different metrics: UES-SF, total time spent examining the search results, and the number of links a user clicked. We normalized these metrics and then aggregated them by taking their mean: [0;1].
- *Perceived diversity* (continuous). We measured the degree of viewpoint diversity that participants perceived in the search results using the perceived diversity scale: [0;6].

Descriptive and Exploratory Measurements

- *Gender*. Participants could select their self-identified gender.
- *Age*. Participants could type their ages in an open text field.
- *Topical interest* (continuous). We measured interest in the topic at hand using the item "I was interested in learning more about this topic," which could be answered by selecting the appropriate option from a seven-point Likert scale: [0,6].

10.3.3. Procedure

We conducted our study on the online task platform *Qualtrics* (`https://www.qualtrics.com`). Participants went through three subsequent steps:

Step 1. Participants received a short introduction to the task and subsequently stated their gender, age, and opinion toward each of the five debated topics. The introduction read: *Imagine the government is seeking informed opinions from the population about a number of debated topics. In order to decide on future policies, they would like to know what the public thinks. You happen to be one of the randomly selected individuals the government is asking for such an informed opinion.*

Step 2. Participants were assigned to one of the topics they had a *mild* opinion on (i.e., responding with "somewhat agree", "neither agree nor disagree", or "somewhat disagree").[7] They learned which topic they had been assigned to and were instructed to pick one of 14 different queries for their web search.[8] Here, the sole purpose of selecting a query was to make the task more realistic. Participants did not get different treatments based on the query they picked.

---

[7] Participants without a mild opinion on any topic were ejected from the study.

Step 3. Participants were randomly assigned to one of the three conditions and were presented with a list of search results, i.e., including search result title, snippet, and URL, similar to most popular search engines.[9] This list contained search result items that were relevant to the assigned topic and ranked according to the assigned condition. For example, if a participant was assigned the topic “Should zoos exist?” and the condition *extreme bias*, that participant saw a search result list in which all documents supporting zoos were ranked above all documents opposing zoos (or vice versa). At the bottom of the page was a “more” button that participants could click, but that would not provide more search results. This allowed us to study the number of participants who might have explored further results if they were available. Participants could explore the search results by reading the names and snippets or by directly examining the web pages they found most interesting. They had to spend at least two minutes exploring the search results but otherwise could take as much time as they need for this part of the study.

Step 4. Participants stated their (updated) opinion and interest concerning their assigned topic.

Step 5. Participants filled in a post-questionnaire that consisted of the AOT scale, the UES-SF, and the perceived diversity scale.

10.3.4. Statistical Analyses

We performed an ANCOVA using *absolute opinion change* as the dependent variable, *condition* and *topic* as between-subjects factors and *AOT*, *user engagement*, and *perceived diversity* as covariates.[10] We looked at the main effects of *condition* ($\mathbf{H_{III.1a}}$), *AOT* ($\mathbf{H_{III.2a}}$), and *user engagement* ($\mathbf{H_{III.2b}}$) on *opinion change* as well as the interaction effects of *condition* and *topic* ($\mathbf{H_{III.1b}}$), *condition* and *AOT* ($\mathbf{H_{III.3a}}$), *condition* and *user engagement* ($\mathbf{H_{III.3b}}$), and *condition* and *perceived diversity* ($\mathbf{H_{III.3c}}$). Additionally, we conducted a One-Way ANOVA to analyze the main effect of *condition* on *perceived diversity* ($\mathbf{H_{III.4a}}$). We decided to conduct AN(C)OVAs despite the anticipation that our data may not be normally distributed because these analyses have been shown to be robust to Likert-type ordinal data [246]. To correct for testing nine hypotheses, we applied a Bonferroni correction so that the significance threshold decreased to $\frac{0.05}{9} = 0.006$.

We also conducted two Bayesian ANOVAs according to the protocol proposed by van den Bergh et al. [351]. In contrast to classical (frequentist) analyses, Bayesian hypothesis tests quantify evidence that the data provide *in favor of the null hypothesis* as opposed to the alternative hypothesis [363]. This is especially useful when trying to interpret non-significant results from classical hypothesis tests because such results do not mean that the null hypothesis is true [129]. Practically, performing Bayesian hypothesis tests allowed us to weigh the evidence in favor of some of the null hypotheses opposing the hypotheses laid out in Section 10.1. We performed these analyses using the software JASP [165] with default settings. We computed *Bayes Factors* (BFs) by comparing the model of interest to a *null model*[11] and interpret them in adherence to the guide proposed by Lee and Wagenmakers [201], who adopted it from Jeffreys [167].

---

[8] The queries that participants could choose from were the same 14 queries that we used for retrieving the search results per topic (see Section 10.2).

[9] Screenshots of the task are available on our repository.

[10] We made two adjustments here compared to our preregistration. First, our preregistration stated that we would perform an ANOVA using these variables; however, ANCOVA is the suitable analysis. Second, we used *absolute* (instead of raw) *opinion change* as the dependent variable. SERPs favored either the supporting or opposing viewpoint, which means that raw scores could have balanced each other out. We were mainly interested in the *magnitude* of opinion change here.

10.3.5. Participants

Before recruiting participants, we computed the required sample size in a power analysis for a *Between-Subjects ANOVA* using the software *G*Power* [104]. We specified the default effect size $f = 0.25$ (i.e., indicating a moderate effect), a significance threshold $\alpha = \frac{0.05}{9} = 0.006$ (i.e., due to testing multiple hypotheses; see Section 10.3.4), a statistical power of $(1\text{-}\beta) = 0.8$, and that we will test $5 \times 3 = 15$ groups (i.e., three conditions, five topics). We computed the required sample size for each of our hypotheses using their respective degrees of freedom. This resulted in a required sample size of 368 participants.

We thus recruited 391 participants from *Prolific* (reward: $2). All participants were proficient English speakers above the age of 18. We excluded participants from data analysis if they did not hold a mild opinion toward any of the five topics, failed at least one attention check, or represented an outlier in terms of the amount of time they spent exploring the SERP. Outliers were participants (seven in total) who spent more or less time on the SERP than two standard deviations from the mean time spent.[12] The resulting sample of 364 participants had an average age of 37 (sd = 13) and a balanced gender distribution (59% female, 41% male, < 1% other).

10.4. Results

In this section, we present the results of our study. We discuss descriptive statistics, the outcomes of the hypothesis tests we conducted, and exploratory findings.

10.4.1. Descriptive Statistics

Participants were distributed over the five topics as follows: 25 (*social networking sites*), 9 (*zoos*), 48 (*cell phone radiation*), 73 (*bottled water*), and 209 (*obesity*).[13] Whereas most participants (81.6%) chose neutral queries (e.g., `is obesity a disease?`) for their task, some picked either opposing queries (6.6%; e.g., `why cell phone radiation is unsafe`) or supporting queries (11.8%; e.g., `arguments supporting zoos`). The number of participants was balanced between conditions: there were 125, 119, and 120 participants in the *little bias*, *moderate bias*, and *extreme bias* conditions, respectively.

---

[11] Here, the *null model* contained nothing but an intercept.

[12] This exclusion criterion was not mentioned in our preregistration, but we felt it was necessary as some participants spent excessive amounts of time on the SERP.

[13] Note that random (balanced) allocation of participants over topics was not possible because we specifically targeted users with mild pre-existing opinions.

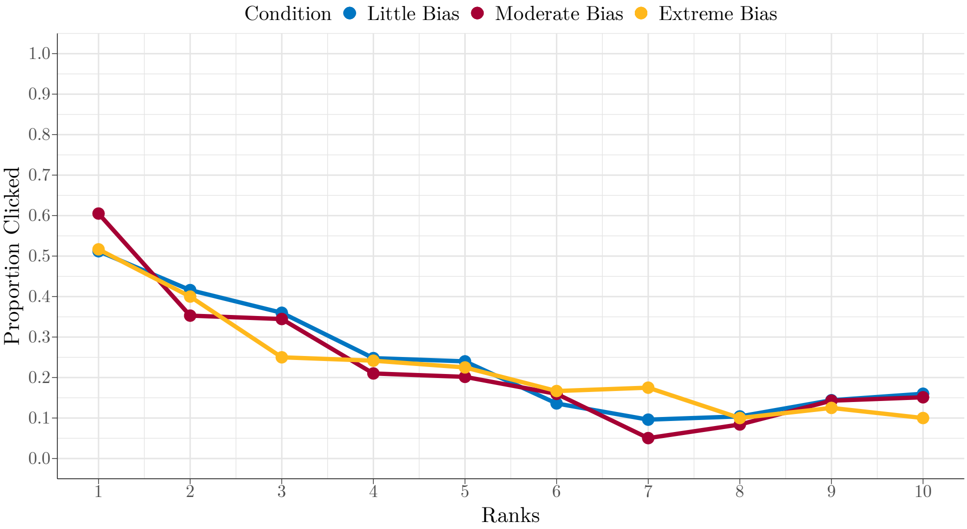


Figure 10.1: Click proportions over the ranks. Users exhibited a weak position bias across conditions.

Most participants (i.e., 83%) were at least somewhat interested in their assigned topic. Overall AOT (mean = 3.89, sd = 0.41), user engagement (mean = 0.33, sd = 0.10), and perceived diversity (mean = 3.40, sd = 0.94) were moderate.

Figure 10.1 shows the click proportions over the ranks for each of the three conditions. In all conditions, we observe that click proportions decrease from roughly 0.55 at the first rank to roughly 0.15 at the sixth rank and below. This reflects a weaker position bias compared to what previous research has found, where click proportions decrease much more severely (i.e., from similar proportions at the first rank down to approximately 0.03 at the tenth rank) [168, 250, 258]. As expected, it thus seemed that participants in our study distributed their attention across the ranks to a satisfactory degree (i.e., as opposed to just focusing on the first few results). This meant that if order effects were strong, we should have found an effect of search result rankings on opinion change.

Two other interesting metrics to look at were (1) the time participants spent exploring the SERP and (2) the number of URLs they clicked. These two metrics – that both contributed towards our *user engagement* measure (see Section 10.3.2) – could tell us more about participant's sincerity in doing the task as well as their motivation and behavior related to informing themselves on the debated topic. First, participants spent an average of 3.32 minutes on the SERP (sd = 1.89). Participants thus spent considerably more time here than the required minimum (i.e., two minutes). This indicated that participants took the task seriously and were motivated to inform themselves. Second, participants clicked a mean of 2.34 URLs (sd = 1.49), and 36% of them clicked the "more" button at the bottom of the SERP. The participants who clicked the "more" button did not, however, engage more in terms of their click behavior (mean = 2.45). These findings suggest that participants used the search result titles and snippets to obtain an overview of the topic and then clicked on particular URLs that interested them.

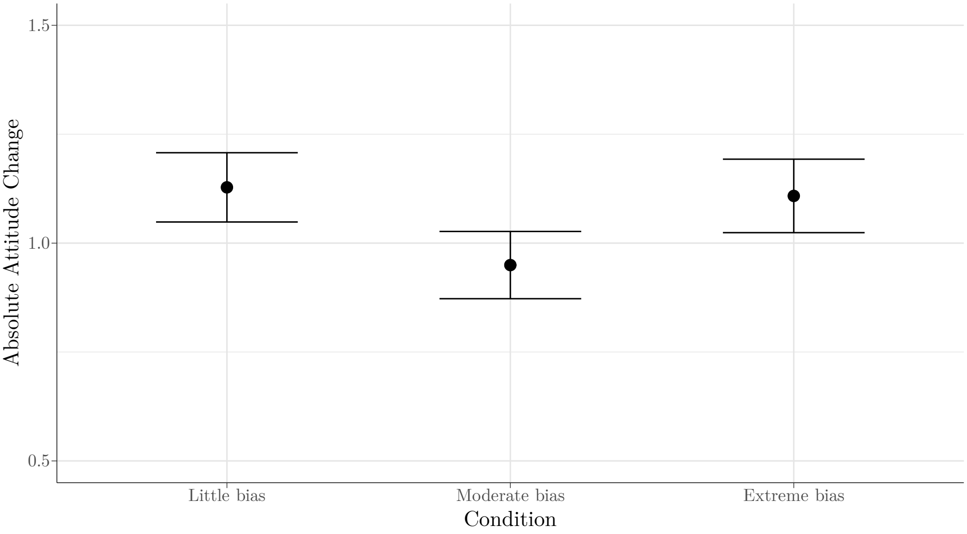


Figure 10.2: Mean absolute opinion change over the three conditions. Error bars represent the standard error.

Overall, 70% of participants expressed an opinion change after viewing the SERP. That is, they moved at least one point on the Likert scale in their post-search opinion compared to their pre-existing opinion towards their assigned debated topic. Mean absolute opinion change (over all conditions) was 1.06 (sd = 0.89), with 30% of participants experiencing an opinion change of two points or more on the Likert scale. This indicates that the search results we showed to participants had the potential to cause opinion change. In line with previous research [371], we find that most participants who reported opinion change (57%) became more *supportive* (rather than more opposing).

10.4.2. Hypothesis Tests

Table 10.2 shows the ANCOVA results. There was no significant difference between conditions (i.e., levels of ranking bias) in terms of opinion change ($F = 1.67$, $p = 0.19$, $\eta^2 = 0.01$; $\mathbf{H_{III.1a}}$; see Figure 10.2). We thus found no evidence in favor of $\mathbf{H_{III.1a}}$. In contrast, as a result of conducting a Bayesian ANOVA, we found moderate evidence in favor of the null hypothesis *opposing* $\mathbf{H_{III.1a}}$, namely that condition had *no influence* on opinion change ($BF_{01} = 8.56$).

The ANCOVA also revealed no direct effects of AOT ($F = 0.90$, $p = 0.34$, $\eta^2 = 0.00$; $\mathbf{H_{III.2a}}$) or user engagement ($F = 0.01$, $p = 0.94$, $\eta^2 = 0.00$; $\mathbf{H_{III.2b}}$). Similarly, there were no significant interaction effects between *condition* and *topic* ($F = 0.74$, $p = 0.66$, $\eta^2 = 0.01$; $\mathbf{H_{III.1b}}$), *AOT* ($F = 0.23$, $p = 0.80$, $\eta^2 = 0.00$; $\mathbf{H_{III.3a}}$), *user engagement* ($F = 3.81$, $p = 0.02$, $\eta^2 = 0.02$; $\mathbf{H_{III.3b}}$),[14] or *perceived diversity* ($F = 2.93$, $p = 0.06$, $\eta^2 = 0.01$; $\mathbf{H_{III.3c}}$). We did not perform a mediation analysis to test $\mathbf{H_{III.2c}}$ because we did not find *AOT* to be significantly related to *opinion change* (i.e., there was no effect to be mediated.)

---

[14] Note that we corrected our significance threshold to 0.006 (see Section 10.3.4), which also rendered this p-value of 0.02 insignificant.

Table 10.2: ANCOVA (absolute opinion change as dependent variable). Colons represent interaction effects.

Hyp.	**Variables**	F	p	η^2
$\mathbf{H_{III.1a}}$	condition	1.67	0.19	0.01
$\mathbf{H_{III.1b}}$	condition:topic	0.74	0.66	0.01
$\mathbf{H_{III.2a}}$	AOT	0.90	0.34	0.00
$\mathbf{H_{III.2b}}$	user engagement	0.01	0.94	0.00
$\mathbf{H_{III.3a}}$	condition:AOT	0.23	0.80	0.00
$\mathbf{H_{III.3b}}$	condition:user eng.	3.81	0.02	0.02
$\mathbf{H_{III.3c}}$	condition:perc. div.	2.93	0.06	0.01

The One-Way ANOVA showed no significant effect of *condition* on *perceived diversity* ($F = 0.07$, $p = 0.94$, $\eta^2 = 0.00$; $\mathbf{H_{III.4a}}$). Conversely, a Bayesian ANOVA revealed strong evidence in favor of the opposing null hypothesis ($BF_{01} = 31.23$), indicating that participants did *not* perceive different levels of diversity.

In sum, we cannot reject any of the null hypotheses opposing the hypotheses we specified in Section 10.1. Bayesian analyses reveal moderate to strong evidence that there are no effects of search result rankings on user opinions and perceived diversity.

10.4.3. Exploratory Findings

Our results suggest that although most users experienced opinion change due to viewing SERPs on debated topics, opinions may not be affected by top 10 search result rankings or the individual differences we measured. We aim to further understand these findings in this subsection and first probe the data for order effects that our previous analyses may not have picked up. However, if there were no order effects, what else drove participants' opinion change? We discuss the roles of exposure effects, confirmation bias, and position bias. Note that the statistical analyses presented here were *not* preregistered: they are of exploratory nature.

A Closer Look at Order Effects

Before turning to alternative explanations for opinion change in our study, we examined the order effects hypothesis more closely. A total of 77 users in our study consumed exactly as many opposing as supporting documents. At least these users should have changed their opinion in accordance with higher-ranked results if order effects occurred. However, we found no evidence for a difference in opinion change between users in this group who saw opposing-biased SERPs and those who saw supporting-biased SERPs ($t = -0.20$, $df = 75$, $p = 0.85$).

Exposure Effects

If order effects are not responsible for SEME, exposure effects may play a bigger role in search results-driven opinion change than we previously anticipated. Exposure effects suggest that the higher the consumed proportion of documents pertaining to a particular viewpoint, the stronger the tendency to adopt that viewpoint. Our study aimed to mitigate exposure effects by letting users explore viewpoint-balanced SERPs for a minimum of two minutes. Nevertheless, some users consumed much higher proportions of supporting or opposing documents than others. We indeed found a relationship between the proportion of supporting documents among the results a user clicked on and opinion change ($r = 0.34$, $df = 342$, $p < 0.001$; the result of a Pearson correlation analysis).[15] As exposure effects would predict, participants thus changed their opinion in accordance with the majority viewpoint (i.e., stance) among the documents they had consumed.

No Evidence for Confirmation Bias

Previous research has demonstrated that *confirmation bias* (i.e., a tendency to engage with pro-attitudinal information) can occur in web search [185, 385]. An explanation for opinion change in our study could thus be that users engaged with mainly pro-attitudinal search results, which subsequently caused the exposure effects. However, we found no evidence for a difference between pre-existing opinions (i.e., somewhat opposing, neutral, somewhat supporting) concerning the proportion of supporting documents that participants clicked on ($F = 0.01$, $p = 0.93$, $\eta^2 = 0.00$, result of a one-way ANOVA).[16]

When Position Bias Meets Ranking Bias

We show in Section 10.2 that, despite our efforts to make users engage with the full SERP, there was some position bias: click proportions decreased from the first to the sixth result (i.e., from 0.55 to 0.15). This means that in the *extreme bias* condition, the average user consumed documents that largely promoted one particular stance because the first five results all pertained to the same stance in this condition (see Table 10.3). If exposure effects took place, there should thus be a difference in opinion change between users that saw the supporting-biased SERPs and those who saw the opposing-biased SERPs. We looked at opinion change per condition, split by the ranking bias on the SERPs (see Table 10.4). We indeed observed a tendency for values of opinion change to drift apart as conditions became more extreme. Here, the *directions* that these values drifted towards corresponded to the direction of the ranking bias on the SERP (e.g., opinion change was more positive in more extreme conditions when the SERP was supporting-biased). A t-test comparing opinion change between the two bias directions in the *extreme bias* condition revealed a potential difference ($t = 2.61$, $df = 113$, $p = 0.01$).

10.5. Discussion

We expected to find that user behavior is guided by order effects and predicted that *changing the order* of items on an overall viewpoint-balanced SERP would lead to varying degrees of opinion change. However, we found *no evidence for order effects*; conversely, we found moderate evidence that there is no effect of search result order on opinion change in this context (i.e., a conclusion drawn from a Bayesian analysis that quantified evidence in favor of the null hypothesis; $\mathbf{RQ_{IV.1}}$). Our results further do not contain evidence for

---

[15] We did not conduct the same analysis for the proportion of opposing results due to symmetry. Twenty participants were excluded from this analysis because they did not click on any URLs (i.e., no proportion of supporting documents could be calculated).

[16] See Footnote 15.

Table 10.3: Proportions of supporting documents among the search results that users clicked on (± standard deviation) in each condition, split by in what direction the SERP was biased.

	Prop. of supporting documents among clicked results	
Condition	**Supporting-biased SERP**	**Opposing-biased SERP**
Little bias	0.53 (±0.30)	0.49 (±0.34)
Moderate bias	0.66 (±0.33)	0.44 (±0.33)
Extreme bias	0.74 (±0.30)	0.32 (±0.37)

Table 10.4: Mean opinion change (± std. dev.) in each condition, split by SERP bias direction.

	Mean opinion change	
Condition	**Supporting-biased SERP**	**Opposing-biased SERP**
Little bias	0.13 (±1.37)	0.11 (±1.5)
Moderate bias	0.21 (±1.34)	0.04 (±1.18)
Extreme bias	0.50 (±1.30)	-0.17 (±1.50)

an interaction effect of topic and the order of search results. We similarly found no evidence for direct effects (**RQ$_{\mathbf{IV.2}}$**) or interaction effects (**RQ$_{\mathbf{IV.3}}$**) concerning other factors we measured (i.e., AOT, user engagement, and perceived diversity). Moreover, our results suggest that participants did not perceive the varying degrees of viewpoint diversity (i.e., different levels of ranking bias) in the SERP we presented to them (**RQ$_{\mathbf{IV.4}}$**). Our findings therefore imply that order effects – if they exist in this context – may contribute less strongly to SEME than previously anticipated.

10.5.1. Explaining SEME?

Exploratory analyses that we conducted indicate that exposure effects as a result of viewing search results may cause opinion change. As exposure effects predict, we found that the more search results pertaining to a particular viewpoint users consumed, the more they tended to adopt that viewpoint. Our results suggest that users did not have confirmation bias when engaging with search results but instead selected documents with a position bias (i.e., they were likely to consume higher-ranked results). This selection then led some users to engage with more documents pertaining to a particular viewpoint, which in turn guided their opinion change.

How do all these results fit together? If participants were affected by position bias in selecting documents and exposure effects regarding their opinions, why did this not result in different levels of opinion change across conditions in our study? A potential explanation is that our manipulation (i.e., presenting overall viewpoint-balanced but ranking-biased SERPs) was too weak for SEME to occur. Previous studies that investigated SEME exposed users to SERPs where one viewpoint was in the majority [10, 99, 274, 371].

This allowed for much more reliable exposure effects as most users would have consumed a great proportion of one particular viewpoint.

It should be pointed out that many different cognitive biases and other external factors play a role in web search [21]. Our study highlights that explaining SEME is a complex problem that requires at least a thorough understanding of (1) how users select documents from SERPs when searching for debated topics and (2) how the selected results affect them. After several studies have shown contexts in which SEME can occur [10, 99, 274, 371], we show that it does not occur in all cases of viewpoint-related ranking bias. Our results suggest that users may not exhibit strong order effects when consuming search results but that exposure effects can contribute to opinion change as a result of viewing search results.

10.5.2. Implications

Our findings have implications for the measurement and mitigation of ranking bias and SEME. First, if order effects do not contribute to SEME, the top-k portion of the ranking does not need to reflect optimal viewpoint diversity at every rank k. This means that the discount function in ranking bias metrics should be chosen according to a good estimate of at which ranks a lack of diversity could cause SEME. For example, it might be suitable to apply the log-discount in steps of ten [381] or to apply an alternative discount function [315]. Second, if exposure effects are the main contributor to SEME, it seems plausible that it can (in part) be mitigated by addressing the ranking bias so that there is a viewpoint balance on the first SERP. Several re-ranking algorithms have already been proposed for similar purposes [19, 35, 55, 238, 391]. Third, applying an (interface) intervention that makes users consider a more diverse selection of documents could also mitigate SEME. Previous research has already investigated this option and found that SEME could be mitigated by alerting users to an existing ranking bias [100]. This alert led users to examine more (and thereby a more viewpoint-balanced set of) search results. Similarly, interventions that nudge users to engage with more search results (e.g., by displaying search results in a different format than a list [171]), increase cognitive reasoning [269], provide additional information about the search topic or the ranking [215, 379, 382], visualize bias among search results [57], or recommend counter-attitudinal substitutes for selected documents [62, 385] could prove fruitful here.

10.5.3. Caveats and Limitations

This work has several limitations. First, we only measured opinion change twice (before and after users interacted with SERPs) and did not collect data on the order in which users clicked on the different documents they engaged with. We thereby cannot deduce the *point at which* opinion change occurred. Second, we cannot ascertain whether phenomena such as confirmation bias affect users on a more nuanced level, e.g., only early or late in the search. Third, we only investigated user behavior in exploring single SERPs in single search sessions that lasted a minimum amount of time (i.e., two minutes), which may not be realistic as users may wish to search more extensively or come back to the search after some reflection. Fourth, we did not control for users' perceived credibility of search results from different sources. Being familiar with a particular media outlet may have led some users to click on corresponding (rather than other) search results.

Fifth, participants' distribution over topics in our study was not balanced, which might have affected the results. Sixth, asking users to self-report their opinions towards debated topics could have prompted them to evaluate their opinions, a process that otherwise might not have occurred.

10.6. Conclusions

While Parts I, II, and III of this dissertation have focused on obtaining high-quality viewpoint labels and measuring viewpoint bias in search results, we finally studied user behavior in the context of biased search results. This chapter presented a user study investigating the effect of light viewpoint biases in search results on user opinions. We found that viewing a viewpoint-balanced SERP containing 10 search results related to a debated topic led to opinion change in most users. However, neither the *order* in which these search results were ranked nor the individual differences we measured affected opinion change. These findings imply that order effects are not a likely explanation for SEME in users with mild pre-search opinions. Instead, our exploratory analyses suggest that exposure effects could be responsible in this context (i.e., users adopting the majority viewpoint among the results they examine). We propose that simple interventions merit further study as user bias mitigation strategies.

11

Conclusion

Web search engines have become ubiquitous tools in the lives of many users, who are now employing web search even for highly subjective tasks such as forming opinions and seeking advice on debated topics (e.g., whether to support school uniform mandates) [119, 120, 369]. Users generally trust search engines to be accurate and impartial [54, 283, 328] but may be unaware that search results can be biased toward particular viewpoints [114, 284, 370, 371] and prompt biased user behavior [21, 168]. Concerning debated topics specifically, recent research has shown that users exposed to viewpoint-biased search results can experience bias-corresponding opinion change [10, 99, 100, 274]. This necessitates a thorough assessment of viewpoint biases in search results and calls for the development of web search applications that support users in their search for debated topics. However, measuring search result viewpoint biases and assisting users in this context has faced four crucial limitations: (1) earlier work represented viewpoints using simple viewpoint taxonomies (e.g., *against*/*neutral*/*in favor*) [121, 385] that are comparatively easy to obtain and handle but ignore important nuances between viewpoints; (2) viewpoint labels were often gathered using crowdsourced annotations without considering the cognitive biases of crowd workers that can reduce data quality [96, 156]; (3) there was a lack of comprehensive viewpoint bias metrics for search results; and (4) the underlying mechanisms of user behavior in web search on debated topics were unclear. Our work has focused on addressing these four key research gaps. In this final chapter, we revisit our research questions to summarize, interpret, and contextualize our findings, discuss the implications and limitations of the dissertation, and describe potentially fruitful avenues for future research.

11.1. Summary of Findings

In this section, we return to our main research questions and summarize our findings.

Part I: Representing Viewpoints

Choosing a taxonomy by which to *represent* viewpoints expressed in search results is an important practical consideration when measuring viewpoint bias or examining user interactions with debated topics. Depending on the level of nuance in these representations, assessors can label search results and perform subsequent analyses with varying degrees of comprehensiveness. Our first research question, which we addressed in Part I, guided our investigation into different ways of representing viewpoints:

$\mathbf{RQ_I}$ What label taxonomy can accurately represent viewpoints on debated topics?

We began addressing **$\mathbf{RQ_I}$** by examining the potential and limitations of currently available methods. Specifically, we automatically assigned ternary stance labels (i.e., *against*/ *neutral*/ *in favor*) to search results, generated explanations for those labels, and evaluated the explanations with users (Chapter 3). Our results showed that, although users found some explanations helpful, users were often unsatisfied with the quality and amount of viewpoint-related information they received. This indicated that considering more comprehensive viewpoint representations for search results could enable more meaningful viewpoint bias evaluations and better assist users in their web search for debated topics. In Chapter 4, we introduced *perspectives* (i.e., reasons for opposing or supporting a topic) as an alternative viewpoint representation format and showed how to automatically discover them in text using unsupervised topic models. Users being able to identify these perspectives as valid reasons for opposing or supporting a debated topic suggested that including reasons behind stances in viewpoint representations could be feasible and useful. Chapter 5 then proposed a comprehensive viewpoint representation for human information interaction. This novel label taxonomy consists of two dimensions: *stance* (i.e., a viewpoint's position regarding a debated topic, measured on a seven-point ordinal scale ranging from *strongly opposing* to *strongly supporting*) and *logic of evaluation* (i.e., the stance's underlying reasons or perspectives categorized into seven topic-independent categories). We showed in a case study how our proposed viewpoint label could be obtained via crowdsourcing with acceptable reliability. By analyzing the resulting data set and conducting a user study, we further demonstrated that the two-dimensional viewpoint representation we proposed allows for more meaningful analyses and diversification interventions compared to current approaches.

Part II: Crowdsourcing Viewpoint Annotations

Our multi-dimensional viewpoint representation (see Part I) enables researchers to create search result data sets with nuanced viewpoint labels (e.g., to train automatic methods or evaluate viewpoint bias). However, obtaining such labels at scale typically requires the input of crowd workers, whose *cognitive biases* can strongly reduce data quality. It is vital to reduce such cognitive worker biases when collecting viewpoint annotations for search results to prevent data biases and ensure high-quality research and practical applications. Our second research question, addressed in Part II, aimed at examining these cognitive crowd worker biases:

$\mathbf{RQ_{II}}$ What cognitive biases reduce crowd workers' abilities to correctly annotate web search results with viewpoint labels?

To address **$\mathbf{RQ_{II}}$**, we proposed a checklist for combating cognitive biases in crowdsourcing (Chapter 6). Our checklist comprises 12 items referring to particularly common or problematic (groups of) cognitive biases that may reduce the quality of crowdsourced data labels. We presented a retrospective analysis of past crowdsourcing papers, showing that cognitive biases are rarely considered but may affect data quality for most tasks. Furthermore, we demonstrated how to use our checklist for crowdsourcing viewpoint labels. Crowdsourcing task requesters can use the checklist we propose to inform their task design (e.g., to mitigate cognitive biases) and document potential influences of cognitive biases on the data they collect. Chapter 7 contains another demonstration of our proposed checklist for the use case of truthfulness judgments. We found that several cognitive biases, such as the affect heuristic and overconfidence, may reduce the quality of such (potentially subjective) crowdsourced annotations.

Part III: Viewpoint Bias Metrics for Search Results

Using the frameworks, tools, and guidelines we provided in Parts I and II, researchers can crowdsource high-quality, comprehensive viewpoint labels for search results and begin to evaluate viewpoint bias in these search result lists. Viewpoint bias assessments of web search results are essential in scoping and understanding the general problem of viewpoint biases in current search engines, linking specific degrees of viewpoint bias to user behavior, and exploring how search result viewpoint diversity could potentially be improved. So far, however, little work had been devoted to developing viewpoint bias metrics for ranked lists of documents, and it was unclear how to incorporate comprehensive viewpoint representations into such evaluations. Our third research question, which we addressed in Part III, focused on this gap:

RQ$_{\mathbf{III}}$ What methods can evaluate viewpoint bias in search results?

We addressed **RQ$_{\mathbf{III}}$** by first exploring how existing ranking fairness metrics could be used to measure viewpoint bias in search results (Chapter 8). Based on simulation studies we conducted with these metrics, we derived guidelines for measuring viewpoint bias in search results. We concluded that existing ranking fairness metrics could be used to measure viewpoint bias when search results are labeled using binary taxonomies (e.g., *against*/*in favor*). A novel ranking fairness metric we proposed in this chapter can also accommodate multi-categorical viewpoint labels (e.g., a seven-point stance taxonomy). However, ranking fairness metrics cannot handle multi-*dimensional* viewpoint representations such as the one we introduced in Part I. That is why we proposed *normalized discounted viewpoint bias* (nDVB), a rank-aware viewpoint bias metric for search results that considers our two-dimensional viewpoint label (Chapter 9). This metric, which measures bias as a deviation from viewpoint plurality, is founded upon a clear notion of viewpoint diversity and can be adapted to fit different topics or viewpoint structures. We further found considerable viewpoint bias (as measured by nDVB and other metrics) in search results from popular search engines and showed how to increase the viewpoint diversity in such search result lists.

Part IV: How Search Result Viewpoint Biases Affect User Behavior

Our contributions from the first three parts of the dissertation allow researchers and practitioners to assign comprehensive viewpoint labels to search results at scale and evaluate viewpoint bias in those search result lists. However, next to search result biases, cognitive user biases can also skew their online viewpoint exposure and drive opinion change. Earlier work had shown that viewpoint biases in search results can lead to phenomena such as the *search engine manipulation effect* (SEME), whereby users without strong pre-search attitudes change their opinions following the most prominent viewpoints among high-ranking search results. Understanding the underlying mechanisms of such user behavior is essential in developing systems that support users in their web search for debated topics. Our fourth research question, addressed in Part IV, guided our investigations into the underlying mechanisms of user behavior in this context:

$\mathbf{RQ_{IV}}$ What cognitive processes underlie the effect of search result viewpoint bias on users' opinion formation?

To address **$\mathbf{RQ_{IV}}$**, Chapter 10 presented a user study investigating whether lower-degree viewpoint biases in search results can also lead users to adopt particular viewpoints. We found no differences between rankings of overall viewpoint-balanced top 10 SERPs concerning users' opinion formation across topics. Further analyses provided exploratory evidence that, rather than *order effects* (i.e., posing that users weigh information importance according to the ranking of search results), *exposure effects* (i.e., posing that users adopt the majority viewpoint among the search results they engage with) seem to guide user behavior in web search on debated topics.

11.2. Implications

Our work has several important implications for the fields of human information interaction and information retrieval. On a general note, we followed open science principles by preregistering all user studies before data collection and openly sharing relevant supplementary materials such as task screenshots, data sets, and code. Doing so has allowed us to deliver rigorous work and made it easy for others to scrutinize and build on our work. We believe that especially the practice of preregistration [247], which is still uncommon in academic computer science, is a promising way toward high-quality research and reproducible findings. To that end, we encourage researchers to implement such open science practices in their own work.

In Part I of this dissertation, we have examined the suitability of different viewpoint representations for search results, including ternary stance labels (i.e., *against*/*neutral*/*in favor*; see Chapter 3), perspectives (i.e., stances' underlying reasons or arguments; see Chapter 4), and a two-dimensional viewpoint label representing stances (i.e., on a seven-point scale) and logics of evaluation (i.e., seven topic-independent perspective categories; see Chapter 5). Although basic viewpoint bias analyses and user support systems are feasible with simple stance labels, we conclude that the nuanced viewpoint representation we propose in Chapter 5 allows for more comprehensive viewpoint bias evaluations and web search applications. We encourage researchers to consider our novel viewpoint label in the studies and systems they design. As we have shown, crowd workers are able to assign the label reliably, and the topic-independent nature of *stances* and *logics* should moreover allow for the development of automatic classification methods.

Part II focused on combating cognitive biases in crowdsourcing viewpoint labels and similar subjective tasks. Our retrospective analysis of past crowdsourcing papers revealed that cognitive crowd worker biases may often reduce the quality of crowdsourced data annotations but are rarely considered by requesters. To improve data quality and reliability, we thus recommend that crowdsourcing task requesters use our proposed 12-item checklist (see Chapter 6) to assess, mitigate, and document the influence of such cognitive biases in the data sets they create. Specific cognitive biases that may pose problems in the context of subjective tasks, such as annotating viewpoints or truthfulness, include the *affect heuristic* and *overconfidence* (see Chapter 7).

In Part III, we turned to the measurement of viewpoint bias in search results, using both existing ranking fairness metrics and novel viewpoint bias metrics we proposed. We conclude that ranking fairness metrics can be used for search result viewpoint bias assessments when there is a particular viewpoint category of concern (e.g., when the aim is to ensure fair treatment for viewpoints opposing school uniforms; see Chapter 8). To measure viewpoint bias in the most comprehensive way, however, we suggest applying nDVB, the viewpoint bias metric we propose in Chapter 9. This metric considers search results' stances as well as logics of evaluation and is, to the best of our knowledge, the most advanced metric developed for this purpose to date. Moreover, as we show, search result viewpoint biases likely occur across search engines, topics, and queries. This means that users of web search engines may often be confronted with viewpoint-biased search results and calls for strategies that can mitigate such biases or support users in their web search on debated topics. Furthermore, we have also demonstrated that search results can be diversified to reduce viewpoint bias with relative ease using existing diversification algorithms. Practitioners who wish to increase the viewpoint diversity in search results (or similar ranked lists of documents) may follow our protocol.

Part IV concerned the underlying mechanisms of user behavior in the context of viewpoint biases and opinion change during web search. We found no evidence of biased opinion change for overall viewpoint-balanced top 10 search result lists (i.e., regardless of how they were ranked; see Chapter 10). This indicates that users may not be vulnerable to subtle search result viewpoint biases and that there may be a search result viewpoint bias threshold after which a reliable user tendency toward biased opinion change sets in – however, further studies are needed to confirm these indications. Our exploratory findings suggest that user behavior during web search for debated topics may not be guided by *order effects* (i.e., higher-ranked search results influence user opinions more than lower-ranked search results) but rather by *exposure effects* (i.e., users are primarily affected by the majority viewpoint among the results they consume, irrespective of where those search results were ranked). This would mean that web search engines may not have to reduce search result viewpoint bias to an extreme extent but can trade off bias reduction with ranking utility. Furthermore, next to adapting the ranking, web search engines could control exposure effects by adapting the interface to support users in their diverse search result consumption. Approaches such as nudging users to engage with more search results (e.g., by displaying search results in a different format than a list [171]), increasing users' cognitive reasoning abilities [213, 269], providing additional information about the search topic or the ranking [100, 215, 379, 382], visualizing bias among search results [57], or recommending counter-attitudinal substitutes for selected documents [62, 385] have already been proposed and could prove fruitful here.

11.3. Limitations and Future Work

Despite the aforementioned contributions and implications of our work, it is important to acknowledge its limitations. This section details these limitations and outlines promising directions for future research in this context. Specifically, we believe there are interesting research opportunities in all four areas this dissertation has covered, i.e., viewpoint representations, crowdsourcing viewpoint labels, measuring search result viewpoint bias, and investigating user interactions with viewpoints during web search.

Representing viewpoints expressed in search results and explaining viewpoint labels to users (see Part I) is still novel. Although we have explored several current techniques and proposed a comprehensive two-dimensional viewpoint representation, we did not consider all possible methods and features in this context. Viewpoint label explanations for search results (see Chapter 3) could take other shapes than the ones we tested (i.e., salience-based and bar plot explanations), such as textual explanations in natural language. Future work could build on our work by proposing novel ways to explain viewpoint labels in the web search context and to explain debated topics to users in general (e.g., in visual or interactive fashions) [57, 215, 298, 379]. Moreover, it is worth investigating what role external and individual user-related factors (e.g., users' trust in different web page sources [123, 226, 269]) play in how users perceive such explanations.

Some of the data sets we publish as part of this dissertation (e.g., see Chapters 5 and 9) provide a starting point for researchers to work with more comprehensive viewpoint labels but are comparatively small. To enable the development of reliable, automatic viewpoint detection models that predict our two-dimensional viewpoint representation, we recommend creating larger data sets of search results (i.e., including upwards of 10000 documents, similar to earlier stance detection data sets [233]) with high-quality (likely crowdsourced) viewpoint annotations. Part II of this dissertation explored how to combat the negative influence of cognitive crowd worker biases on data quality. However, other factors such as high costs and task difficulty can pose obstacles to the efficient and effective crowdsourcing of viewpoint labels for search results: assigning viewpoint labels to search results may require six or more annotations per search result, and each annotation may take considerable time as viewpoint labeling is no easy task (see Chapter 5). To lower the cost and assist crowd workers, automatic methods such as stance detection (see Chapter 3) or topic modeling (see Chapter 4) could be used as preprocessing methods that assist crowd workers in their tasks. Earlier work has already experimented with such hybrid (human-AI) viewpoint labeling procedures [23]. There is also extensive literature concerning subjective tasks in general, how to assist crowd workers in this context, and how to best aggregate such crowdsourced [16, 380]. For example, future work could incorporate collaborative workflows [59, 182] or ask crowd workers to provide rationales for their annotations, either in a free-text fashion or by highlighting the words in the text that support their decision. A second annotator could then approve or reject these. Finally, asking crowd annotators to provide rationales for their annotations proved useful for increasing quality and informing automatic methods [166, 196].

Part III investigated how to measure search result viewpoint bias, which resulted in the development of a novel viewpoint bias metric (nDVB) that considers both stances and logics of evaluation (see Chapter 9). Although we applied nDVB to real search results from popular search engines, its outcomes may differ depending on factors such as viewpoint distribution, topic, or query, e.g., our viewpoint bias evaluations in Chapter 9 may have resulted in different conclusions had we used other queries. More research is necessary to find interpretation guidelines for viewpoint bias metrics such as nDVB and determine when viewpoint bias becomes problematic or affects user behavior. Similarly, future work should investigate how automatic viewpoint diversification of search results can influence users and whether such strategies can contribute to more general efforts of supporting users in their web search for debated topics [26, 100]. Researchers and practitioners could also use our proposed metric to build on our viewpoint bias evaluations (see Chapter 9) and conduct large-scale analyses of real search results across different types of debated topics and queries. Interesting opportunities in this area include developing methods to automatically identify debated topics in realistic search scenarios (see Section 2.1.1), examining viewpoint biases for scientifically answerable topics [44, 241] and studying whether user perceptions align with the outcomes of viewpoint bias metrics [136, 227, 298]. Finally, our proposed metric nDVB is still limited in several ways, e.g., it does not yet incorporate document relevance, other viewpoint diversity notions, or the personal preferences and beliefs of users. We encourage researchers to build on our work to help improve the measurement of viewpoint bias in search results.

In Part IV, we have taken first steps to understand user behavior and opinion change in the context of search result viewpoint biases. Our findings in this area clearly indicate particular user behavior patterns (e.g., exposure effects) but have to be studied in more detail to form a nuanced picture. Specifically, identifying the circumstances (e.g., search scenario, topic, degree of viewpoint bias) under which user behavior and opinions are influenced presents an exciting challenge for future research. This could help identify particularly vulnerable user groups and inform user support strategies. Future work should also conduct longitudinal studies to investigate how robust opinion changes from search result viewpoint biases are. Moreover, similar to our work on representing viewpoints expressed in search results (see Part I), there is a need to measure user opinions more comprehensively, i.e., to develop accurate assessments of what logics of evaluation a user subscribes to, whether they adopt logics due to search result viewpoint biases [36], and how robust such opinion changes are.

A general limitation of this dissertation is that our work was conducted before the recent arrival of advanced generative AI systems such as ChatGPT [290]. Although it is currently still unclear how exactly the availability of such systems will affect web search going forward, it is fair to assume that users' web search interactions will be different from those considered in this dissertation. Future research should examine this shift in how users retrieve and interact with online information and revisit our conclusions in light of this novel search context, e.g., regarding the viewpoint diversity in automatically generated answers and how biases in such answers can affect user behavior and opinions.

11.4. Ethical Considerations

As we have discussed extensively throughout this dissertation, highly-ranked web search results can strongly affect users' behavior and opinions [10, 36, 37, 99, 274] (see Section 2.2.5). This warrants a general note of care with respect to our work. Researchers and practitioners should be aware that – if applied carelessly or maliciously – the methods we have proposed could harm individual users or society. To give a few examples, search result stance and logic classification errors could unintentionally lead to biased search results; malicious actors could harness crowd worker biases to achieve low-relevance judgments for documents from particular news outlets so that web search engines rank those documents lower; and search results could be diversified for scientifically answerable topics to create the impression of a debate, instilling “balance as bias” [44], e.g., causing users to adopt a harmful medical procedure. We do not wish to recommend specific topics or scenarios in which to (not) apply our results and approaches, as doing so comprehensively is beyond the scope of this dissertation. However, we encourage readers to carefully consider the ethical implications of dealing with viewpoint biases in web search when building on our work.

11.5. Concluding Remarks

This dissertation has contributed empirical evidence, tools, and resources that foster a greater understanding of viewpoint biases in search results and their effects on user behavior. In Part I, we identified that two-dimensional viewpoint labels (i.e., representing *stances* and *logics of evaluation*) can represent viewpoints on debated topics better than previously handled binary or ternary stance labels ($\mathbf{RQ_{I}}$). Part II resulted in a comprehensive list of cognitive biases (e.g., the *affect heuristic*) that may reduce crowd workers’ abilities to annotate web search results with viewpoint labels ($\mathbf{RQ_{II}}$). We described in Part III how to measure search result viewpoint bias more comprehensively than current methods using our proposed metric nDVB ($\mathbf{RQ_{III}}$). Finally, our work in Part IV provided exploratory evidence that *exposure effects* (i.e., rather than *order effects*) underlie the effect of search result viewpoint bias on users’ opinion formation ($\mathbf{RQ_{IV}}$).

Empirical findings surrounding viewpoint biases and their effects on user behavior call for taking a socio-technical perspective on search for debated topics and revisiting the role of search engines in this context. The way in which users form opinions while searching the web can make an important qualitative difference [189, 265]: Kornblith [189] posits that *responsible beliefs* are the product of actively gathering evidence and critically evaluating it. Search engines could assist users in forming opinions responsibly by (1) providing diverse resources that offer unbiased and complete information and (2) accommodating or encouraging thorough information-seeking strategies. Such strategies include exploring different resources, making unbiased comparisons, and objectively assessing the provided information for sense-making and learning [230, 323, 328]. In sum, web search engines can and should be platforms for users to explore debated topics in all their nuances without cognitive overwhelm. We hope this dissertation can make a meaningful contribution to the development of such more comprehensive web search engines by providing a better understanding of search result viewpoint biases and their effects on users.

Bibliography

[1] Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. “Internet Argument Corpus 2.0: An SQL Schema for Dialogic Social Media and the Corpora to Go with It”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 4445–4452. URL: `https://aclanthology.org/L16-1704`.

[2] Adnan Abid, Naveed Hussain, Kamran Abid, Farooq Ahmad, Muhammad Shoaib Farooq, Uzma Farooq, Sher Afzal Khan, Yaser Daanial Khan, Muhammad Azhar Naeem, and Nabeel Sabir. “A Survey on Search Results Diversification Techniques”. In: *Neural Computing and Applications* 27.5 (July 2016), pp. 1207–1229. DOI: `10.1007/s00521-015-1945-5`.

[3] Aseel Addawood, Jodi Schneider, and Masooda Bashir. “Stance Classification of Twitter Debates: The Encryption Debate as A Use Case”. In: *Proceedings of the 8th International Conference on Social Media & Society*. New York, NY, USA: Association for Computing Machinery, 2017. DOI: `10.1145/3097286.3097288`.

[4] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. “Diversifying Search Results”. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. WSDM '09. New York, NY, USA: Association for Computing Machinery, 2009, pp. 5–14. DOI: `10.1145/1498759.1498766`.

[5] Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. “Modeling Frames in Argumentation”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2922–2932. DOI: `10.18653/v1/D19-1290`.

[6] Yamen Ajjour, Henning Wachsmuth, Dora Kiesel, Patrick Riehmann, Fan Fan, Giuliano Castiglia, Rosemary Adejoh, Bernd Fröhlich, and Benno Stein. “Visualization of the Topic Space of Argument Search Results in Args.Me”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 60–65. DOI: `10.18653/v1/D18-2011`.

[7] Marwah Alaofi, Luke Gallagher, Dana Mckay, Lauren L. Saling, Mark Sanderson, Falk Scholer, Damiano Spina, and Ryen W. White. “Where Do Queries Come From?” In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 2850–2862. DOI: `10.1145/3477495.3531711`.

[8] Abeer Aldayel and Walid Magdy. “Stance Detection on Social Media: State of the Art and Trends”. In: *Information Processing & Management* 58.4 (July 2021), p. 102597. DOI: `10.1016/j.ipm.2021.102597`.

[9] Abeer Aldayel and Walid Magdy. “Your Stance Is Exposed! Analysing Possible Factors for Stance Detection on Social Media”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: `10.1145/3359307`.

[10] Ahmed Allam, Peter Johannes Schulz, and Kent Nakamoto. “The Impact of Search Engine Selection and Sorting Criteria on Vaccination Beliefs and Attitudes: Two Experiments Manipulating Google Output”. In: *Journal of Medical Internet Research* 16.4 (Apr. 2014), e100. DOI: `10.2196/jmir.2642`.

[11] Emily Allaway and Kathleen McKeown. “Zero-Shot Stance Detection: A Dataset and Model Using Generalized Topic Representations”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 8913–8931. DOI: `10.18653/v1/2020.emnlp-main.717`.

[12] Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. “Scaling up Fact-Checking Using the Wisdom of Crowds”. In: *Science Advances* 7.36 (Sept. 2021), eabf4393. DOI: `10.1126/sciadv.abf4393`.

[13] Alessia Antelmi, Delfina Malandrino, and Vittorio Scarano. “Characterizing the Behavioral Evolution of Twitter Users and the Truth behind the 90-9-1 Rule”. In: *Companion Proceedings of the 2019 World Wide Web Conference*. WWW ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1035–1038. DOI: `10.1145/3308560.3316705`.

[14] Judd Antin and Aaron Shaw. “Social Desirability Bias and Self-Reports of Motivation: A Study of Amazon Mechanical Turk in the US and India”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’12. New York, NY, USA: Association for Computing Machinery, 2012, pp. 2925–2934. DOI: `10.1145/2207676.2208699`.

[15] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney. “FactSheets: Increasing Trust in AI Services through Supplier’s Declarations of Conformity”. In: *arXiv:1808.07261 [cs]* (Feb. 2019). arXiv: `1808.07261 [cs]`. URL: `http://arxiv.org/abs/1808.07261` (visited on 07/13/2021).

[16] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. “Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions”. In: *Companion Proceedings of the 2019 World Wide Web Conference*. WWW ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1100–1105. DOI: `10.1145/3308560.3317083`.

[17] Charles Arthur. *What Is the 1% Rule?* July 2006. URL: `https://www.theguardian.com/technology/2006/jul/20/guardianweeklytechnologysection2`.

[18] Victor Asal and Paul Harwood. "Search Engines: Terrorism's Killer App." In: *Studies in Conflict & Terrorism* 31.7 (June 2008), pp. 641–654. DOI: `10.1080/10576100802149675`.

[19] Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. "Designing Fair Ranking Schemes". In: *Proceedings of the 2019 International Conference on Management of Data.* SIGMOD '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1259–1276. DOI: `10.1145/3299869.3300079`.

[20] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. "Stance Detection with Bidirectional Conditional Encoding". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 876–885. DOI: `10.18653/v1/D16-1084`.

[21] Leif Azzopardi. "Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval". In: *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval.* CHIIR '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 27–37. DOI: `10.1145/3406522.3446023`.

[22] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." In: *LREC.* Vol. 10. Proceedings of LREC. Jan. 2010. URL: `http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf`.

[23] Christian Baden, Neta Kligler-Vilenchik, and Moran Yarchi. "Hybrid Content Analysis: Toward a Strategy for the Theory-Driven, Computer-Assisted Classification of Large Text Corpora". In: *Communication Methods and Measures* 14.3 (July 2020), pp. 165–183. DOI: `10.1080/19312458.2020.1803247`.

[24] Christian Baden and Nina Springer. "Com(Ple)Menting the News on the Financial Crisis: The Contribution of News Users' Commentary to the Diversity of Viewpoints in the Public Debate". In: *European Journal of Communication* 29.5 (Oct. 2014), pp. 529–548. DOI: `10.1177/0267323114538724`.

[25] Christian Baden and Nina Springer. "Conceptualizing Viewpoint Diversity in News Discourse". In: *Journalism* 18.2 (Feb. 2017), pp. 176–194. DOI: `10.1177/1464884915605028`.

[26] Ricardo Baeza-Yates. "Bias on the Web". In: *Communications of the ACM* 61.6 (May 2018), pp. 54–61. DOI: `10.1145/3209581`.

[27] Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. "Integrating Stance Detection and Fact Checking in a Unified Corpus". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).* New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 21–27. DOI: `10.18653/v1/N18-2004`.

[28] Judit Bar-Ilan, Kevin Keenoy, Mark Levene, and Eti Yaari. "Presentation Bias Is Significant in Determining User Preference for Search Results—A User Study". In: *Journal of the American Society for Information Science and Technology* 60.1 (2009), pp. 135–149.

[29] Natã M. Barbosa and Monchu Chen. “Rehumanized Crowdsourcing: A Labeling Framework Addressing Bias and Ethics in Machine Learning”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–12. DOI: `10.1145/3290605.3300773`.

[30] Francesco Barile, Tim Draws, Oana Inel, Alisa Rieger, Shabnam Najafian, Amir Ebrahimi Fard, Rishav Hada, and Nava Tintarev. “Evaluating Explainable Social Choice-Based Aggregation Strategies for Group Recommendation”. In: *User Modeling and User-Adapted Interaction* (June 2023). DOI: `10.1007/s11257-023-09363-0`.

[31] Francesco Barile, Shabnam Najafian, Tim Draws, Oana Inel, Alisa Rieger, Rishav Hada, and Nava Tintarev. “Toward Benchmarking Group Explanations: Evaluating the Effect of Aggregation Strategies versus Explanation”. In: *Proceedings of the Perspectives on the Evaluation of Recommender Systems Workshop (PERSPECTIVES 2021)* (2021). URL: `http://ceur-ws.org/Vol-2955/paper11.pdf`.

[32] Iz Beltagy, Matthew E. Peters, and Arman Cohan. *Longformer: The Long-Document Transformer*. 2020. DOI: `10.48550/ARXIV.2004.05150`.

[33] Emily M Bender and Batya Friedman. “Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science”. In: *Transactions of the Association for Computational Linguistics* 6 (2018), pp. 587–604.

[34] Momen Bhuiyan, Amy Zhang, Connie Sehat, and Tanushree Mitra. “Investigating “Who” in the Crowdsourcing of News Credibility”. In: *Computational Journalism Symposium*. 2020.

[35] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. “Equity of Attention: Amortizing Individual Fairness in Rankings”. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR ’18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 405–414. DOI: `10.1145/3209978.3210063`.

[36] Markus Bink, Sebastian Schwarz, Tim Draws, and David Elsweiler. “Investigating the Influence of Featured Snippets on User Attitudes”. In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. CHIIR ’23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 211–220. DOI: `10.1145/3576840.3578323`.

[37] Markus Bink, Steven Zimmerman, and David Elsweiler. “Featured Snippets and Their Influence on Users’ Credibility Judgements”. In: *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*. CHIIR ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 113–122. DOI: `10.1145/3498366.3505766`.

[38] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. “Latent Dirichlet Allocation”. In: *Journal of Machine Learning Research* 3.null (Mar. 2003), pp. 993–1022.

[39] Luc Boltanski and Laurent Thévenot. *On Justification: Economies of Worth*. Vol. 27. Princeton University Press, 2006.

[40] Alexander Bondarenko, Yamen Ajjour, Valentin Dittmar, Niklas Homann, Pavel Braslavski, and Matthias Hagen. “Towards Understanding and Answering Comparative Questions”. In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining.* 2022, pp. 66–74.

[41] Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. “Overview of Touché 2022: Argument Retrieval”. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction.* Ed. by Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro. Cham: Springer International Publishing, 2022, pp. 311–336.

[42] Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. “Overview of Touché 2021: Argument Retrieval”. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction.* Ed. by K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Birger Larsen, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro. Cham: Springer International Publishing, 2021, pp. 450–467.

[43] Danah Boyd and Michael Golebiewski. *Data Voids: Where Missing Data Can Easily Be Exploited.* May 2018. URL: `https://datasociety.net/library/data-voids-where-missing-data-can-easily-be-exploited/` (visited on 10/26/2023).

[44] Maxwell T Boykoff and Jules M Boykoff. “Balance as Bias: Global Warming and the US Prestige Press”. In: *Global Environmental Change* 14.2 (2004), pp. 125–136.

[45] Aras Bozkurt. “Generative Artificial Intelligence (AI) Powered Conversational Educational Agents: The Inevitable Paradigm Shift”. In: *Asian Journal of Distance Education* 18.1 (Mar. 2023). URL: `https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/718` (visited on 10/25/2023).

[46] Dean Brady, Michael Meany, Janet Fulton, and Phillip McIntyre. “Search Engines as Opinion Leaders”. In: *Creating Space in the Fifth Estate.* 2017, pp. 129–143.

[47] Virginia Braun and Victoria Clarke. “Using Thematic Analysis in Psychology”. In: *Qualitative Research in Psychology* 3.2 (Jan. 2006), pp. 77–101. DOI: `10.1191/1478088706qp063oa`.

[48] Katarzyna Budzynska and Chris Reed. “Advances in Argument Mining”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts.* Florence, Italy: Association for Computational Linguistics, 2019, pp. 39–42. DOI: `10.18653/v1/P19-4008`.

[49] Alec Burmania, Srinivas Parthasarathy, and Carlos Busso. “Increasing the Reliability of Crowdsourcing Evaluations Using Online Quality Assessment”. In: *IEEE Transactions on Affective Computing* 7.4 (2016), pp. 374–388. DOI: `10.1109/TAFFC.2015.2493525`.

[50] Martin J. Burnham, Yen K. Le, and Ralph L. Piedmont. “Who Is Mturk? Personal Characteristics and Sample Consistency of These Online Workers”. In: *Mental Health, Religion & Culture* 21.9-10 (Nov. 2018), pp. 934–944. DOI: `10.1080/13674676.2018.1486394`.

[51] Björn Burscher, Daan Odijk, Rens Vliegenthart, Maarten De Rijke, and Claes H De Vreese. “Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis”. In: *Communication Methods and Measures* 8.3 (2014), pp. 190–206.

[52] Berfu Büyüköz, Ali Hürriyetoğlu, and Arzucan Özgür. “Analyzing ELMo and DistilBERT on Socio-Political News Classification”. In: *Proceedings of the Workshop on Automated Extraction of Socio-Political Events from News 2020*. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 9–18. URL: `https://aclanthology.org/2020.aespen-1.4`.

[53] Andrea Caputo. “Social Desirability Bias in Self-Reported Well-Being Measures: Evidence from an Online Survey”. In: *Universitas Psychologica* 16.2 (June 2017), pp. 245–255. DOI: `10.11144/javeriana.upsy16-2.sdsw`.

[54] Noel Carroll. “In Search We Trust: Exploring How Search Engines Are Shaping Society”. In: *International Journal of Knowledge Society Research* 5.1 (Jan. 2014), pp. 12–27. DOI: `10.4018/ijksr.2014010102`.

[55] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. “Ranking with Fairness Constraints”. In: *arXiv:1704.06840 [cs]* (July 2018). arXiv: `1704.06840 [cs]`. URL: `http://arxiv.org/abs/1704.06840` (visited on 07/13/2021).

[56] Andres Chacoma and Damian H Zanette. “Opinion Formation by Social Influence: From Experiments to Modeling”. In: *PloS one* 10.10 (2015), e0140406.

[57] Jon Chamberlain, Udo Kruschwitz, and Orland Hoeber. “Scalable Visualisation of Sentiment and Stance”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. LREC 2018. 2018, p. 5.

[58] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. “Do Explanations Make VQA Models More Predictable to a Human?” In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 10, pp. 1036–1042. DOI: `10.18653/v1/D18-1128`.

[59] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. “Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI ’17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 2334–2346. DOI: `10.1145/3025453.3026044`.

[60] Yi-Chin Chen, Zhao-Yang Liu, and Hung-Yu Kao. “IKM at SemEval-2017 Task 8: Convolutional Neural Networks for Stance Detection and Rumor Verification”. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 465–469. DOI: `10.18653/v1/S17-2081`.

[61] Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. “Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims”. In: *arXiv:1906.03538 [cs]* (June 2019). arXiv: `1906.03538 [cs]`. URL: `http://arxiv.org/abs/1906.03538` (visited on 07/13/2021).

[62] Tong Chen, Hongzhi Yin, Guanhua Ye, Zi Huang, Yang Wang, and Meng Wang. “Try This Instead: Personalized and Interpretable Substitute Recommendation”. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 891–900. DOI: `10.1145/3397271.3401042`.

[63] Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, and Rosso Paolo. “Sardistance@ Evalita2020: Overview of the Task on Stance Detection in Italian Tweets”. In: *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. Ceur, 2020, pp. 1–10. URL: `https://ceur-ws.org/Vol-2765/paper159.pdf`.

[64] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. “Novelty and Diversity in Information Retrieval Evaluation”. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’08. New York, NY, USA: Association for Computing Machinery, 2008, pp. 659–666. DOI: `10.1145/1390334.1390446`.

[65] Kevyn Collins-Thompson, Preben Hansen, and Claudia Hauff. “Search as Learning (Dagstuhl Seminar 17092)”. In: *Dagstuhl Reports*. Vol. 7. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2017.

[66] Costanza Conforti, Mohammad Taher Pilehvar, and Nigel Collier. “Towards Automatic Fake News Detection: Cross-Level Stance Detection in News Articles”. In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 40–49. DOI: `10.18653/v1/W18-5507`.

[67] Neil Dagnall, Andrew Denovan, Kenneth Graham Drinkwater, and Andrew Parker. “An Evaluation of the Belief in Science Scale”. In: *Frontiers in Psychology* 10 (2019), p. 861.

[68] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. “Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions”. In: *ACM Computing Surveys (CSUR)* 51.1 (2018), pp. 1–40.

[69] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. “A Survey of the State of Explainable AI for Natural Language Processing”. In: *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*. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 447–459. URL: `https://aclanthology.org/2020.aacl-main.46`.

[70] Kareem Darwish, Walid Magdy, and Tahar Zanouda. “Improved Stance Prediction in a User Similarity Feature Space”. In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017.* ASONAM ’17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 145–148. DOI: `10.1145/3110025.3110112`.

[71] Anubrata Das and Matthew Lease. “A Conceptual Framework for Evaluating Fairness in Search”. In: *arXiv:1907.09328 [cs]* (July 2019). arXiv: `1907.09328 [cs]`. URL: `http://arxiv.org/abs/1907.09328` (visited on 07/13/2021).

[72] Gianluca Demartini. “Implicit Bias in Crowdsourced Knowledge Graphs”. In: *Companion Proceedings of the 2019 World Wide Web Conference.* WWW ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 624–630. DOI: `10.1145/3308560.3317307`.

[73] Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. “Human-in-the-Loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities”. In: *Bulletin of IEEE Computer Society* 43.3 (2020), pp. 65–74.

[74] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: `10.18653/v1/N19-1423`.

[75] Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. “Topical Stance Detection for Twitter: A Two-Phase LSTM Model Using Attention”. In: *Advances in Information Retrieval.* Ed. by Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury. Cham: Springer International Publishing, 2018, pp. 529–536.

[76] Marcelo Dias and Karin Becker. “Inf-Ufrgs-Opinion-Mining at Semeval-2016 Task 6: Automatic Generation of a Training Corpus for Unsupervised Identification of Stance in Tweets”. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).* 2016, pp. 378–383.

[77] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. “Mechanical Cheat: 1st International Workshop on Crowdsourcing Web Search, CrowdSearch 2012 - Workshop Held in Conjunction with WWW 2012 Conference”. In: *CEUR Workshop Proceedings* 842 (2012), pp. 20–25. URL: `http://www.scopus.com/inward/record.url?scp=84892543307&partnerID=8YFLogxK` (visited on 10/26/2023).

[78] Shuoyang Ding and Philipp Koehn. “Evaluating Saliency Methods for Neural Language Models”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Online: Association for Computational Linguistics, 2021, pp. 5034–5052. DOI: `10.18653/v1/2021.naacl-main.399`.

[79] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning.* Mar. 2017. arXiv: `1702.08608 [cs, stat]`. URL: `http://arxiv.org/abs/1702.08608` (visited on 08/10/2022).

[80] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. “Shepherding the Crowd Yields Better Work”. In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work.* 2012, pp. 1013–1022.

[81] Tim Draws. “Understanding How Algorithmic and Cognitive Biases in Web Search Affect User Attitudes on Debated Topics”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* SIGIR ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2709. DOI: `10.1145/3404835.3463273`.

[82] Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. “Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions with Debated Topics”. In: *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval.* CHIIR ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 135–145. DOI: `10.1145/3498366.3505812`.

[83] Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. “The Effects of Crowd Worker Biases in Fact-Checking Tasks”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* FAccT ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 2114–2124. DOI: `10.1145/3531146.3534629`.

[84] Tim Draws, Jody Liu, and Nava Tintarev. “Helping Users Discover Perspectives: Enhancing Opinion Mining with Joint Topic Models”. In: *2020 International Conference on Data Mining Workshops (ICDMW).* Sorrento, Italy: IEEE, Nov. 2020, pp. 23–30. DOI: `10.1109/ICDMW51313.2020.00013`.

[85] Tim Draws, Karthikeyan Natesan Ramamurthy, Ioana Baldini, Amit Dhurandhar, Inkit Padhi, Benjamin Timmermans, and Nava Tintarev. “Explainable Cross-Topic Stance Detection for Search Results”. In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval.* CHIIR ’23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 221–235. DOI: `10.1145/3576840.3578296`.

[86] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. “A Checklist to Combat Cognitive Biases in Crowdsourcing”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing.* Vol. 9. HCOMP ’21. 2021, pp. 48–59. DOI: `10.1609/hcomp.v9i1.18939`.

[87] Tim Draws, Nirmal Roy, Oana Inel, Alisa Rieger, Rishav Hada, Mehmet Orcun Yalcin, Benjamin Timmermans, and Nava Tintarev. “Viewpoint Diversity in Search Results”. In: *Advances in Information Retrieval.* Ed. by Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo. Vol. 13980. Cham: Springer Nature Switzerland, 2023, pp. 279–297. DOI: `10.1007/978-3-031-28244-7_18`.

[88] Tim Draws, Zoltán Szlávik, Benjamin Timmermans, Nava Tintarev, Kush R. Varshney, and Michael Hind. “Disparate Impact Diminishes Consumer Trust Even for Advantaged Users”. In: *Persuasive Technology.* Ed. by Raian Ali, Birgit Lugrin,

and Fred Charles. Vol. 12684. Cham: Springer International Publishing, 2021, pp. 135–149. DOI: `10.1007/978-3-030-79460-6_11`.

[89] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. "Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics". In: *ACM SIGKDD Explorations Newsletter* 23.1 (May 2021), pp. 50–58. DOI: `10.1145/3468507.3468515`.

[90] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. "This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* SIGIR '21. New York, NY, USA: Association for Computing Machinery, July 2021, pp. 295–305. DOI: `10.1145/3404835.3462851`.

[91] Marina Drosou and Evaggelia Pitoura. "Search Result Diversification". In: *SIGMOD Record* 39.1 (2010), p. 7.

[92] Lorik Dumani, Patrick J. Neumann, and Ralf Schenkel. "A Framework for Argument Retrieval". In: *Advances in Information Retrieval.* Ed. by Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins. Cham: Springer International Publishing, 2020, pp. 431–445.

[93] Anca Dumitrache, Lora Aroyo, and Chris Welty. "Capturing Ambiguity in Crowdsourcing Frame Disambiguation". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 6 (June 2018), pp. 12–20. DOI: `10.1609/hcomp.v6i1.13330`.

[94] Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. "CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement". In: *arXiv preprint arXiv:1808.06080* (2018). arXiv: `1808.06080`.

[95] Tom Edixhoven, Sihang Qiu, Lucie Kuiper, Olivier Dikken, Gwennan Smitskamp, and Ujwal Gadiraju. "Improving Reactions to Rejection in Crowdsourcing through Self-Reflection". In: *13th ACM Web Science Conference 2021.* 2021, pp. 74–83.

[96] Carsten Eickhoff. "Cognitive Biases in Crowdsourcing". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining.* WSDM '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 162–170. DOI: `10.1145/3159652.3159654`.

[97] Robert M Entman. "Cascading Activation: Contesting the White House's Frame after 9/11". In: *Political Communication,* 20.4 (2003), pp. 415–432.

[98] Sacha Epskamp and Eiko I Fried. "A Tutorial on Regularized Partial Correlation Networks." In: *Psychological Methods* 23.4 (2018), p. 617.

[99] Robert Epstein and Ronald E. Robertson. "The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcomes of Elections". In: *Proceedings of the National Academy of Sciences* 112.33 (Aug. 2015), E4512–E4521. DOI: `10.1073/pnas.1419828112`.

[100] Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. “Suppressing the Search Engine Manipulation Effect (SEME)”. In: *Proceedings of the ACM on Human-Computer Interaction* 1.CSCW (Dec. 2017), pp. 1–22. DOI: `10.1145/3134677`.

[101] Ziv Epstein, Gordon Pennycook, and David Rand. “Will the Crowd Game the Algorithm? Using Layperson Judgments to Combat Misinformation on Social Media by Downranking Distrusted Sources”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* CHI ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–11. DOI: `10.1145/3313831.3376232`.

[102] Boi Faltings, Radu Jurca, Pearl Pu, and Bao Duy Tran. “Incentives to Counter Bias in Human Computation”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 2.1 (Sept. 2014), pp. 59–66. URL: `https://ojs.aaai.org/index.php/HCOMP/article/view/13145`.

[103] Siamak Faradani, Björn Hartmann, and Panagiotis G Ipeirotis. “What’s the Right Price? Pricing Tasks for Finishing on Time”. In: *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence.* 2011.

[104] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. “Statistical Power Analyses Using G* Power 3.1: Tests for Correlation and Regression Analyses”. In: *Behavior Research Methods* 41.4 (2009), pp. 1149–1160.

[105] Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. “Pathologies of Neural Models Make Interpretations Difficult”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 3719–3728. DOI: `10.18653/v1/D18-1407`.

[106] William Ferreira and Andreas Vlachos. “Emergent: A Novel Data-Set for Stance Classification”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* San Diego, California: Association for Computational Linguistics, June 2016, pp. 1163–1168. DOI: `10.18653/v1/N16-1138`.

[107] Shane Frederick. “Cognitive Reflection and Decision Making”. In: *Journal of Economic Perspectives* 19.4 (Dec. 2005), pp. 25–42.

[108] Bent Fuglede and Flemming Topsøe. “Jensen-Shannon Divergence and Hubert Space Embedding”. In: *IEEE Int. Symp. Inf. Theory - Proc.* 2004, p. 31. DOI: `10.1109/isit.2004.1365067`.

[109] Norbert Fuhr, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Jarvelin, Rosie Jones, YiquN Liu, Josiane Mothe, Wolfgang Nejdl, Isabella Peters, and Benno Stein. “An Information Nutritional Label for Online Documents”. In: *ACM SIGIR Forum* 51.3 (Feb. 2018), pp. 46–66. DOI: `10.1145/3190580.3190588`.

[110] Adrian Furnham and Hua Chu Boo. “A Literature Review of the Anchoring Effect”. In: *The Journal of Socio-Economics* 40.1 (Feb. 2011), pp. 35–42. DOI: `10.1016/j.socec.2010.10.008`.

[111] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. “Using Worker Self-Assessments for Competence-Based Pre-Selection in Crowdsourcing Microtasks”. In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 24.4 (2017), pp. 1–26.

[112] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. “A Taxonomy of Microtasks on the Web”. In: *Proceedings of the 25th ACM Conference on Hypertext and Social Media*. HT ’14. New York, NY, USA: Association for Computing Machinery, 2014, pp. 218–223. DOI: `10.1145/2631775.2631819`.

[113] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. “Clarity Is a Worthwhile Quality: On the Role of Task Clarity in Microtask Crowdsourcing”. In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. 2017, pp. 5–14.

[114] Ruoyuan Gao and Chirag Shah. “Toward Creating a Fairer Ranking in Search Engine Results”. In: *Information Processing & Management* 57.1 (Jan. 2020), p. 102138. DOI: `10.1016/j.ipm.2019.102138`.

[115] Atul Gawande. *The Checklist Manifesto: How to Get Things Right*. New York: Picador, 2010.

[116] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. “Datasheets for Datasets”. In: *arXiv:1803.09010 [cs]* (Mar. 2020). arXiv: `1803.09010 [cs]`. URL: `http://arxiv.org/abs/1803.09010` (visited on 07/13/2021).

[117] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. “Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?” In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 325–336. DOI: `10.1145/3351095.3372862`.

[118] Mor Geva, Yoav Goldberg, and Jonathan Berant. “Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1161–1166. DOI: `10.18653/v1/D19-1107`.

[119] Lisa Gevelber. “How Mobile Has Changed How People Get Things Done: New Consumer Behavior Data”. In: *Think with Google* (2016). URL: `https://think.storage.googleapis.com/docs/mobile-search-consumer-behavior-data.pdf`.

[120] Lisa Gevelber. *It’s All about ‘Me’—How People Are Taking Search Personally*. Tech. rep. 2018. URL: `https://www.thinkwithgoogle.com/intl/en-145/marketing-strategies/search/personal-needs-search-trends/`.

[121] Gizem Gezici, Aldo Lipani, Yucel Saygin, and Emine Yilmaz. “Evaluation Metrics for Measuring Bias in Search Engine Results”. In: *Information Retrieval Journal* 24.2 (Apr. 2021), pp. 85–113. DOI: `10.1007/s10791-020-09386-w`.

[122] Amira Ghenai and Yelena Mejova. “Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter”. In: *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. 2017, pp. 518–518.

[123] Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. “A Think-Aloud Study to Understand Factors Affecting Online Health Search”. In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. CHIIR ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 273–282. DOI: `10.1145/3343413.3377961`.

[124] Anastasia Giachanou and Paolo Rosso. “The Battle against Online Harmful Information: The Cases of Fake News and Hate Speech”. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 3503–3504.

[125] Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. “Overview of the Transformer-Based Models for NLP Tasks”. In: *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2020, pp. 179–183. DOI: `10.15439/2020F20`.

[126] Fausto Giunchiglia, Styliani Kleanthous, Jahna Otterbacher, and Tim Draws. “Transparency Paths - Documenting the Diversity of User Perceptions”. In: *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 415–420. DOI: `10.1145/3450614.3463292`.

[127] Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. “Stance Detection in COVID-19 Tweets”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1596–1611. DOI: `10.18653/v1/2021.acl-long.127`.

[128] Catherine Grady and Matthew Lease. “Crowdsourcing Document Relevance Assessment with Mechanical Turk”. In: *NAACL HLT Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*. 2010, pp. 172–179.

[129] Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. “Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations”. In: *European Journal of Epidemiology* 31.4 (Apr. 2016), pp. 337–350. DOI: `10.1007/s10654-016-0149-3`.

[130] Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. “A Large-Scale Dataset for Argument Quality Ranking: Construction and Analysis”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (Apr. 2020), pp. 7805–7813. DOI: `10.1609/aaai.v34i05.6285`.

[131] Thomas L Griffiths and Mark Steyvers. “Finding Scientific Topics”. In: *Proceedings of the National Academy of Sciences* 101.suppl 1 (2004), pp. 5228–5235. DOI: `10.1073/pnas.0307752101`.

[132] Stephan Grimmelikhuijsen and Eva Knies. “Validating a Scale for Citizen Trust in Government Organizations”. In: *International Review of Administrative Sciences* 83.3 (2017), pp. 583–601.

[133] Alexander Haas and Julian Unkel. “Ranking versus Reputation: Perception and Effects of Search Result Credibility”. In: *Behaviour & Information Technology* 36.12 (Dec. 2017), pp. 1285–1298. DOI: `10.1080/0144929X.2017.1381166`.

[134] Matthew Haigh. “Has the Standard Cognitive Reflection Test Become a Victim of Its Own Success?” In: *Advances in Cognitive Psychology* 12.3 (Sept. 2016), pp. 145–149.

[135] Alexander Halavais. *Search Engine Society.* John Wiley & Sons, 2017.

[136] Bin Han, Chirag Shah, and Daniel Saelid. “Users’ Perception of Search-Engine Biases and Satisfaction”. In: *Advances in Bias and Fairness in Information Retrieval.* Ed. by Ludovico Boratto, Stefano Faralli, Mirko Marras, and Giovanni Stilo. Communications in Computer and Information Science. Cham: Springer International Publishing, 2021, pp. 14–24. DOI: `10.1007/978-3-030-78818-6_3`.

[137] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. “The Impact of Task Abandonment in Crowdsourcing”. In: *IEEE Transactions on Knowledge and Data Engineering* (2019).

[138] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. “A Retrospective Analysis of the Fake News Challenge Stance-Detection Task”. In: *Proceedings of the 27th International Conference on Computational Linguistics.* Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1859–1874. URL: `https://aclanthology.org/C18-1158`.

[139] Uriel Haran, Ilana Ritov, and Barbara A Mellers. “The Role of Actively Open-Minded Thinking in Information Acquisition, Accuracy, and Calibration”. In: *Judgment and Decision Making* 8.3 (2013), p. 16.

[140] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. “A Survey on Stance Detection for Mis- and Disinformation Identification”. In: *Findings of the Association for Computational Linguistics: NAACL 2022.* Seattle, United States: Association for Computational Linguistics, July 2022, pp. 1259–1277. DOI: `10.18653/v1/2022.findings-naacl.94`.

[141] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. “Cross-Domain Label-Adaptive Stance Detection”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 9011–9028. DOI: `10.18653/v1/2021.emnlp-main.710`.

[142] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. “Few-Shot Cross-Lingual Stance Detection with Sentiment-Based Pre-Training”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.10 (June 2022), pp. 10729–10737. DOI: `10.1609/aaai.v36i10.21318`.

[143] Eszter Hargittai, Lindsay Fullerton, Ericka Menchen-Trevino, and Kristin Yates Thomas. "Trust Online: Young Adults' Evaluation of Web Content". In: *International Journal of Communication* 4.0 (Apr. 2010), p. 27. URL: `https://ijoc.org/index.php/ijoc/article/view/636` (visited on 10/26/2023).

[144] Christopher G Harris. "Detecting Cognitive Bias in a Relevance Assessment Task Using an Eye Tracker". In: *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. 2019, pp. 1–5.

[145] Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. "Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?" In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4351–4367. DOI: `10.18653/v1/2020.findings-emnlp.390`.

[146] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. *DeBERTa: Decoding-Enhanced BERT with Disentangled Attention*. 2020. DOI: `10.48550/ARXIV.2006.03654`.

[147] Natali Helberger. "On the Democratic Role of News Recommenders". In: *Digital Journalism* 7.8 (Sept. 2019), pp. 993–1012. DOI: `10.1080/21670811.2019.1623700`.

[148] Natali Helberger, Kari Karppinen, and Lucia D'Acunto. "Exposure Diversity as a Design Principle for Recommender Systems". In: *Information, Communication & Society* 21.2 (Feb. 2018), pp. 191–207. DOI: `10.1080/1369118X.2016.1271900`.

[149] Tomás Hercig, Peter Krejzl, Barbora Hourová, Josef Steinberger, and Ladislav Lenc. "Detecting Stance in Czech News Commentaries." In: *ITAT* 176 (2017), p. 180.

[150] Martin Hilbert. "Toward a Synthesis of Cognitive Biases: How Noisy Information Processing Can Bias Human Decision Making." In: *Psychological Bulletin* 138.2 (2012), p. 211.

[151] L. M. Hinman. "Searching Ethics: The Role of Search Engines in the Construction and Distribution of Knowledge". In: *Web Search: Multidisciplinary Perspectives*. Ed. by Amanda Spink and Michael Zimmer. Information Science and Knowledge Management. Berlin, Heidelberg: Springer, 2008, pp. 67–76. DOI: `10.1007/978-3-540-75829-7_5`.

[152] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. "Incentivizing High Quality Crowdwork". In: *Proceedings of the 24th International Conference on World Wide Web*. 2015, pp. 419–429.

[153] Robin M. Hogarth and Hillel J. Einhorn. "Order Effects in Belief Updating: The Belief-Adjustment Model". In: *Cognitive Psychology* 24.1 (1992), pp. 1–55. DOI: `10.1016/0010-0285(92)90002-J`.

[154] Suzanne Hoogeveen, Alexandra Sarafoglou, Balazs Aczel, Yonathan Aditya, Alexandra J. Alayan, Peter J. Allen, Sacha Altay, Shilaan Alzahawi, Yulmaida Amir, Francis-Vincent Anthony, Obed Kwame Appiah, Quentin D. Atkinson, Adam Baimel, Merve Balkaya-Ince, Michela Balsamo, Sachin Banker, František Bartoš, Mario Becerra, Bertrand Beffara, Julia Beitner, Theiss Bendixen, Jana B. Berkessel, Renatas Berniūnas, Matthew I. Billet, Joseph Billingsley, Tiago Bortolini, Heiko Breitsohl, Amélie Bret, Faith L. Brown, Jennifer Brown, Claudia C. Brumbaugh, Jacek Buczny, Joseph Bulbulia, Saúl Caballero, Leonardo Carlucci, Cheryl L. Carmichael, Marco E. G. V. Cattaneo, Sarah J. Charles, Scott Claessens, Maxinne C. Panagopoulos, Angelo Brandelli Costa, Damien L. Crone, Stefan Czoschke, Christian Czymara, E. Damiano D'Urso, Örjan Dahlström, Anna Dalla Rosa, Henrik Danielsson, Jill De Ron, Ymkje Anna de Vries, Kristy K. Dean, Bryan J. Dik, David J. Disabato, Jaclyn K. Doherty, Tim Draws, et al. "A Many-Analysts Approach to the Relation between Religiosity and Well-Being". In: *Religion, Brain & Behavior* 0.0 (2022), pp. 1–47. DOI: `10.1080/2153599X.2022.2070255`.

[155] Sha Hu, Zhicheng Dou, Xiaojie Wang, Tetsuya Sakai, and Ji-Rong Wen. "Search Result Diversification Based on Hierarchical Intents". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. CIKM '15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 63–72. DOI: `10.1145/2806416.2806455`.

[156] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. "Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–12. DOI: `10.1145/3290605.3300637`.

[157] Nicholas C. Hunt and Andrea M. Scheetz. "Using MTurk to Distribute a Survey or Experiment: Methodological Considerations". In: *Journal of Information Systems* 33.1 (Mar. 2019), pp. 43–65. DOI: `10.2308/isys-52021`.

[158] Yuki Igarashi, Hiroya Komatsu, Sosuke Kobayashi, Naoaki Okazaki, and Kentaro Inui. "Tohoku at SemEval-2016 Task 6: Feature-Based Model versus Convolutional Neural Network for Stance Detection". In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 2016, pp. 401–407.

[159] Ross Ihaka and Robert Gentleman. "R: A Language for Data Analysis and Graphics". In: *Journal of Computational and Graphical Statistics* 5.3 (1996), pp. 299–314. DOI: `10.1080/10618600.1996.10474713`. eprint: `https://www.tandfonline.com/doi/pdf/10.1080/10618600.1996.10474713`.

[160] Oana Inel, Tim Draws, and Lora Aroyo. "Collect, Measure, Repeat: Reliability Factors for Responsible AI Data Collection". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 11.1 (Nov. 2023), pp. 51–64. DOI: `10.1609/hcomp.v11i1.27547`.

[161] Oana Inel, Giannis Haralabopoulos, Dan Li, Christophe Van Gysel, Zoltán Szlávik, Elena Simperl, Evangelos Kanoulas, and Lora Aroyo. “Studying Topical Relevance with Evidence-Based Crowdsourcing”. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2018, pp. 1253–1262.

[162] Peter Ingwersen and Kalervo Järvelin. “Information Retrieval in Context: IRiX”. In: *Proceedings of The Twelfth Annual International Acm Sigir Conference On Research and Development in Information Retrieval* 39.2 (Dec. 2005), pp. 31–39. DOI: `10.1145/1113343.1113351`.

[163] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. “Quality Management on Amazon Mechanical Turk”. In: *Proceedings of the ACM SIGKDD Workshop on Human Computation*. 2010, pp. 64–67.

[164] Alon Jacovi and Yoav Goldberg. “Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4198–4205. DOI: `10.18653/v1/2020.acl-main.386`.

[165] JASP Team. *JASP (Version 0.13)[Computer Software]*. 2020. URL: `https://jasp-stats.org/`.

[166] Sahil Jayaram and Emily Allaway. “Human Rationales as Attribution Priors for Explainable Stance Detection”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 5540–5554. DOI: `10.18653/v1/2021.emnlp-main.450`.

[167] Harold Jeffreys. *The Theory of Probability*. OUP Oxford, 1998.

[168] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. “Accurately Interpreting Clickthrough Data as Implicit Feedback”. In: *ACM SIGIR Forum* 51.1 (2016), p. 8.

[169] Daniel Kahneman, Dan Lovallo, and Olivier Sibony. “Before You Make That Big Decision...” In: *Harvard Business Review* 89.6 (2011).

[170] Sriram Kalyanaraman and James D. Ivory. “Enhanced Information Scent, Selective Discounting, or Consummate Breakdown: The Psychological Effects of Web-Based Search Results”. In: *Media Psychology* 12.3 (2009), pp. 295–319. DOI: `10.1080/15213260903052232`.

[171] Yvonne Kammerer and Peter Gerjets. “Chapter 10 How Search Engine Users Evaluate and Select Web Search Results: The Impact of the Search Engine Interface on Credibility Assessments”. In: *Library and Information Science*. Ed. by Dirk Lewandowski. Emerald Group Publishing Limited, Feb. 2012, pp. 251–279. DOI: `10.1108/S1876-0562(2012)002012a012`.

[172] Adam Kapelner and Dana Chandler. “Preventing Satisficing in Online Surveys”. In: *Proceedings of CrowdConf* (2010).

[173] Hema Karande, Rahee Walambe, Victor Benjamin, Ketan Kotecha, and TS Raghu. "Stance Detection with BERT Embeddings for Credibility Analysis of Information on Social Media". In: *PeerJ Computer Science* 7 (2021), e467.

[174] Kornraphop Kawintiranon and Lisa Singh. "Knowledge Enhanced Masked Language Model for Stance Detection". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 2021.

[175] Mesut Kaya and Derek Bridge. "Subprofile-Aware Diversification of Recommendations". In: *User Modeling and User-Adapted Interaction* 29.3 (July 2019), pp. 661–700. DOI: `10.1007/s11257-019-09235-6`.

[176] Przemysław Kazienko, Julita Bielaniewicz, Marcin Gruza, Kamil Kanclerz, Konrad Karanowski, Piotr Miłkowski, and Jan Kocoń. "Human-Centered Neural Reasoning for Subjective Content Processing: Hate Speech, Emotions, and Humor". In: *Information Fusion* 94 (June 2023), pp. 43–65. DOI: `10.1016/j.inffus.2023.01.010`.

[177] Mark T Keane and Maeve O'Brien. "Are People Biased in Their Use of Search Engines?" In: *Communications of the ACM* 51.2 (2008), pp. 49–52.

[178] Ryan Kemmer, Yeawon Yoo, Adolfo Escobedo, and Ross Maciejewski. "Enhancing Collective Estimates by Aggregating Cardinal and Ordinal Inputs". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8.1 (Oct. 2020), pp. 73–82. DOI: `10.1609/hcomp.v8i1.7465`.

[179] Anant Khandelwal. "Fine-Tune Longformer for Jointly Predicting Rumor Stance and Veracity". In: *8th ACM IKDD CODS and 26th COMAD.* CODS COMAD 2021. New York, NY, USA: Association for Computing Machinery, 2021, pp. 10–19. DOI: `10.1145/3430984.3431007`.

[180] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. "Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems.* NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, pp. 2288–2296.

[181] Jaewon Kim, Paul Thomas, Ramesh Sankaranarayana, Tom Gedeon, and Hwan-Jin Yoon. "Understanding Eye Movements on Mobile Devices for Better Presentation of Search Results". In: *Journal of the Association for Information Science and Technology* 67.11 (2016), pp. 2607–2619.

[182] Joy Kim, Sarah Sterman, Allegra Argent Beal Cohen, and Michael S. Bernstein. "Mechanical Novel: Crowdsourcing Complex Work through Reflection and Revision". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.* CSCW '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 233–245. DOI: `10.1145/2998181.2998196`.

[183] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. "The Future of Crowd Work". In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work.* 2013, pp. 1301–1318.

[184] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. “Explaining the User Experience of Recommender Systems”. In: *User Modeling and User-Adapted Interaction* 22.4-5 (Oct. 2012), pp. 441–504. DOI: `10.1007/s11257-011-9118-4`.

[185] Silvia Knobloch-Westerwick, Benjamin K. Johnson, and Axel Westerwick. “Confirmation Bias in Online Searches: Impacts of Selective Exposure Before an Election on Political Attitude Strength and Shifts”. In: *Journal of Computer-Mediated Communication* 20.2 (Mar. 2015), pp. 171–187. DOI: `10.1111/jcc4.12105`.

[186] Masaki Kobayashi, Hiromi Morita, Masaki Matsubara, Nobuyuki Shimizu, and Atsuyuki Morishima. “An Empirical Study on Short- and Long-Term Effects of Self-Correction in Crowdsourced Microtasks”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 6.1 (June 2018), pp. 79–87. DOI: `10.1609/hcomp.v6i1.13324`.

[187] Pang Wei Koh and Percy Liang. “Understanding Black-Box Predictions via Influence Functions”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70.* ICML’17. Sydney, NSW, Australia: JMLR.org, 2017, pp. 1885–1894.

[188] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. *Captum: A Unified and Generic Model Interpretability Library for PyTorch.* 2020. DOI: `10.48550/ARXIV.2009.07896`.

[189] Hilary Kornblith. “Justified Belief and Epistemically Responsible Action”. In: *The Philosophical Review* 92 (1983), pp. 33–48. DOI: `10.2307/2184520`.

[190] Klaus Krippendorff. “Agreement and Information in the Reliability of Coding”. In: *Communication Methods and Measures* 5.2 (Apr. 2011), pp. 93–112. DOI: `10.1080/19312458.2011.568376`.

[191] Klaus Krippendorff. “Reliability in Content Analysis.: Some Common Misconceptions and Recommendations”. In: *Human Communication Research* 30.3 (July 2004), pp. 411–433. DOI: `10.1111/j.1468-2958.2004.tb00738.x`.

[192] Justin Kruger and David Dunning. “Unskilled and Unaware of It: How Difficulties in Recognizing One’s Own Incompetence Lead to Inflated Self-Assessments.” In: *Journal of Personality and Social Psychology* 77.6 (1999), p. 1121.

[193] Dilek Küçük and Fazli Can. “Stance Detection: A Survey”. In: *ACM Computing Surveys* 53.1 (Jan. 2021), pp. 1–37. DOI: `10.1145/3369026`.

[194] S. Kullback and R. A. Leibler. “On Information and Sufficiency”. In: *The Annals of Mathematical Statistics* 22.1 (Mar. 1951), pp. 79–86. DOI: `10.1214/aoms/1177729694`.

[195] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. “Search Bias Quantification: Investigating Political Bias in Social Media and Web Search”. In: *Information Retrieval Journal* 22.1-2 (Apr. 2019), pp. 188–227. DOI: `10.1007/s10791-018-9341-2`.

[196] Mucahid Kutlu, Tyler McDonnell, Tamer Elsayed, and Matthew Lease. “Annotator Rationales for Labeling Tasks in Crowdsourcing”. In: *Journal of Artificial Intelligence Research* 69 (2020), pp. 143–189.

[197] David La Barbera, Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. “Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias”. In: *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II.* Berlin, Heidelberg: Springer-Verlag, 2020, pp. 207–214. DOI: `10.1007/978-3-030-45442-5_26`.

[198] Mirko Lai, Alessandra Teresa Cignarella, Delia Irazu Hernandez Farias, et al. “Itacos at Ibereval2017: Detecting Stance in Catalan and Spanish Tweets”. In: *IberEval 2017.* Vol. 1881. CEUR-WS. org. 2017, pp. 185–192.

[199] Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. “Stance Evolution and Twitter Interactions in an Italian Political Debate”. In: *Natural Language Processing and Information Systems.* Ed. by Max Silberztein, Faten Atigui, Elena Kornyshova, Elisabeth Métais, and Farid Meziane. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 15–27. DOI: `10.1007/978-3-319-91947-8_2`.

[200] John Lawrence and Chris Reed. “Argument Mining: A Survey”. In: *Computational Linguistics* 45.4 (Jan. 2020), pp. 765–818. DOI: `10.1162/coli_a_00364`.

[201] Michael D Lee and Eric-Jan Wagenmakers. *Bayesian Cognitive Modeling: A Practical Course.* Cambridge University Press, 2014.

[202] Dirk Lewandowski. “The Retrieval Effectiveness of Search Engines on Navigational Queries”. In: *Aslib Proceedings.* Emerald Group Publishing Limited. 2011.

[203] Dirk Lewandowski. *Understanding Search Engines.* Cham: Springer International Publishing, 2023. DOI: `10.1007/978-3-031-22789-9`.

[204] Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. “P-Stance: A Large Dataset for Stance Detection in Political Domain”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.* Online: Association for Computational Linguistics, Aug. 2021, pp. 2355–2365. DOI: `10.18653/v1/2021.findings-acl.208`.

[205] Chenghua Lin and Yulan He. “Joint Sentiment/Topic Model for Sentiment Analysis”. In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management.* CIKM ’09. New York, NY, USA: Association for Computing Machinery, 2009, pp. 375–384. DOI: `10.1145/1645953.1646003`.

[206] Jianhua Lin. “Divergence Measures Based on the Shannon Entropy”. In: *IEEE Transactions on Information Theory* 37.1 (1991), pp. 145–151. DOI: `10.1109/18.61115`.

[207] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. “Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies”. In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 521–535.

[208] Can Liu, Wen Li, Bradford Demarest, Yue Chen, Sara Couture, Daniel Dakota, Nikita Haduong, Noah Kaufman, Andrew Lamont, Manan Pancholi, et al. “Iucl at Semeval-2016 Task 6: An Ensemble Model for Stance Detection in Twitter”. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 2016, pp. 394–400.

[209] Liran Liu, Shi Feng, Daling Wang, and Yifei Zhang. “An Empirical Study on Chinese Microblog Stance Detection Using Supervised and Semi-Supervised Machine Learning Methods”. In: *Natural Language Understanding and Intelligent Applications*. Ed. by Chin-Yew Lin, Nianwen Xue, Dongyan Zhao, Xuanjing Huang, and Yansong Feng. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 753–765. DOI: `10.1007/978-3-319-50496-4_68`.

[210] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. DOI: `10.48550/ARXIV.1907.11692`.

[211] Felicia Loecherbach, Judith Moeller, Damian Trilling, and Wouter van Atteveldt. “The Unified Framework of Media Diversity: A Systematic Literature Review”. In: *Digital Journalism* 8.5 (May 2020), pp. 605–642. DOI: `10.1080/21670811.2020.1764374`.

[212] Edward Loper and Steven Bird. “NLTK: The Natural Language Toolkit”. In: *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*. ETMTNLP ’02. USA: Association for Computational Linguistics, 2002, pp. 63–70. DOI: `10.3115/1118108.1118117`.

[213] Philipp Lorenz-Spreen, Michael Geers, Thorsten Pachur, Ralph Hertwig, Stephan Lewandowsky, and Stefan M Herzog. “Boosting People’s Ability to Detect Microtargeted Advertising”. In: *Scientific Reports* 11.1 (2021), pp. 1–9.

[214] Nikita Lozhnikov, Leon Derczynski, and Manuel Mazzara. “Stance Prediction for Russian: Data and Analysis”. In: *Proceedings of 6th International Conference in Software Engineering for Defence Applications*. Ed. by Paolo Ciancarini, Manuel Mazzara, Angelo Messina, Alberto Sillitti, and Giancarlo Succi. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2020, pp. 176–186. DOI: `10.1007/978-3-030-14687-0_16`.

[215] Ramona Ludolph, Ahmed Allam, and Peter J Schulz. “Manipulating Google’s Knowledge Graph Box to Counter Biased Information Processing During an Online Search on Vaccination: Application of a Technological Debiasing Strategy”. In: *Journal of Medical Internet Research* 18.6 (June 2016), e137. DOI: `10.2196/jmir.5430`.

[216] Eddy Maddalena, Davide Ceolin, and Stefano Mizzaro. “Multidimensional News Quality: A Comparison of Crowdsourcing and Nichesourcing”. In: *Proceedings of the CIKM 2018 Workshops Co-Located with 27th ACM International Conference on Information and Knowledge Management (CIKM 2018), Torino, Italy, October 22, 2018*. Ed. by Alfredo Cuzzocrea, Francesco Bonchi, and Dimitrios Gunopulos.

Vol. 2482. CEUR Workshop Proceedings. CEUR-WS.org, 2018. URL: `https://ceur-ws.org/Vol-2482/paper17.pdf`.

[217] Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. "Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining". In: *arXiv:2110.08412 [cs]* (Oct. 2021). arXiv: `2110.08412 [cs]`. URL: `http://arxiv.org/abs/2110.08412` (visited on 05/06/2022).

[218] Andreas Madsen, Siva Reddy, and Sarath Chandar. "Post-Hoc Interpretability for Neural NLP: A Survey". In: *arXiv:2108.04840 [cs]* (Apr. 2022). arXiv: `2108.04840 [cs]`. URL: `http://arxiv.org/abs/2108.04840` (visited on 05/09/2022).

[219] Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan Schwamb, Chris Lintott, and Arfon Smith. "Volunteering versus Work for Pay: Incentives and Tradeoffs in Crowdsourcing". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 1.1 (Nov. 2013), pp. 94–102. URL: `https://ojs.aaai.org/index.php/HCOMP/article/view/13075`.

[220] Andrea Masini and Peter Van Aelst. "Actor Diversity and Viewpoint Diversity: Two of a Kind?" In: *Communications* 42.2 (Jan. 2017). DOI: `10.1515/commun-2017-0017`.

[221] Andrea Masini, Peter Van Aelst, Thomas Zerback, Carsten Reinemann, Paolo Mancini, Marco Mazzoni, Marco Damiani, and Sharon Coen. "Measuring and Explaining the Diversity of Voices and Viewpoints in the News: A Comparative Study on the Determinants of Content Diversity of Immigration News". In: *Journalism Studies* 19.15 (Nov. 2018), pp. 2324–2343. DOI: `10.1080/1461670X.2017.1343650`.

[222] Matthew Matero, Nikita Soni, Niranjan Balasubramanian, and H. Andrew Schwartz. "MeLT: Message-Level Transformer with Masked Document Representations as Pre-Training for Stance Detection". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2959–2966. DOI: `10.18653/v1/2021.findings-emnlp.253`.

[223] Katja Mayer. "On the Sociometry of Search Engines". In: *Deep Search: The Politics of Search Beyond Google* (2009), pp. 54–72.

[224] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. "Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–23.

[225] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. "Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4.1 (Sept. 2016), pp. 139–148. URL: `https://ojs.aaai.org/index.php/HCOMP/article/view/13287`.

[226] Dana Mckay, Stephann Makri, Marisela Gutierrez-Lopez, Andrew MacFarlane, Sondess Missaoui, Colin Porlezza, and Glenda Cooper. “We Are the Change That We Seek: Information Interactions during a Change of Viewpoint”. In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval.* CHIIR ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 173–182. DOI: `10.1145/3343413.3377975`.

[227] Dana McKay, Kaipin Owyong, Stephann Makri, and Marisela Gutierrez Lopez. “Turn and Face the Strange: Investigating Filter Bubble Bursting Information Interactions”. In: *ACM SIGIR Conference on Human Information Interaction and Retrieval.* 2022, pp. 233–242.

[228] Florian Meier and David Elsweiler. “Studying Politicians’ Information Sharing on Social Media”. In: *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval.* CHIIR ’19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 237–241. DOI: `10.1145/3295750.3298944`.

[229] Vincent Menger, Floor Scheepers, and Marco Spruit. “Comparing Deep Learning and Classical Machine Learning Approaches for Predicting Inpatient Violence Incidents from Clinical Text”. In: *Applied Sciences* 8.6 (2018), p. 981.

[230] Boaz Miller and Isaac Record. “Justified Belief in the Digital Age: On the Epistemic Implications of Secret Internet Technologies”. In: *Episteme; rivista critica di storia delle scienze mediche e biologiche* 10 (2013), pp. 117–134. DOI: `10.1017/epi.2013.11`.

[231] Richard L. Miller. “Mere Exposure, Psychological Reactance and Attitude Change”. In: *The Public Opinion Quarterly* 40.2 (1976), pp. 229–233.

[232] Amita Misra, Brian Ecker, Theodore Handleman, Nicolas Hahn, and Marilyn Walker. “NLDS-UCSC at SemEval-2016 Task 6: A Semi-Supervised Approach to Detecting Stance in Tweets”. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).* 2016, pp. 420–427.

[233] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. “SemEval-2016 Task 6: Detecting Stance in Tweets”. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).* San Diego, California: Association for Computational Linguistics, 2016, pp. 31–41. DOI: `10.18653/v1/S16-1003`.

[234] Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. “Stance and Sentiment in Tweets”. In: *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media* 17.3 (2017).

[235] Vikram Mohanty, Kareem Abdol-Hamid, Courtney Ebersohl, and Kurt Luther. “Second Opinion: Supporting Last-Mile Person Identification with Crowdsourcing and Face Recognition”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7.1 (Oct. 2019), pp. 86–96. DOI: `10.1609/hcomp.v7i1.5272`.

[236] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. “Do Not Blame It on the Algorithm: An Empirical Assessment of Multiple Recommender Systems and Their Impact on Content Diversity”. In: *Information, Communication & Society* 21.7 (July 2018), pp. 959–977. DOI: `10.1080/1369118X.2018.1444076`.

[237] Richard D Morey, Jeffrey N Rouder, Tahira Jamil, and Maintainer Richard D Morey. “Package ‘BayesFactor’”. In: (2015). URL: `http://cran/r-projectorg/web/packages/BayesFactor/BayesFactor.pdf`.

[238] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. “Controlling Fairness and Bias in Dynamic Learning-to-Rank”. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* SIGIR ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 429–438. DOI: `10.1145/3397271.3401100`.

[239] Robert R Morris, Mira Dontcheva, and Elizabeth M Gerber. “Priming for Better Performance in Microtask Crowdsourcing Environments”. In: *IEEE Internet Computing* 16.5 (2012), pp. 13–19.

[240] Mats Mulder, Oana Inel, Jasper Oosterman, and Nava Tintarev. “Operationalizing Framing to Support Multiperspective Recommendations of Opinion Pieces”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* FAccT ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 478–488. DOI: `10.1145/3442188.3445911`.

[241] Sean A. Munson and Paul Resnick. “Presenting Diverse Political Opinions: How and How Much”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* CHI ’10. New York, NY, USA: Association for Computing Machinery, 2010, pp. 1457–1466. DOI: `10.1145/1753326.1753543`.

[242] Akiko Murakami and Rudy Raymond. “Support or Oppose? Classifying Positions in Online Debates from Reply Activities and Opinion Expressions”. In: *Coling 2010: Posters.* 2010, pp. 869–875.

[243] Shabnam Najafian, Tim Draws, Francesco Barile, Marko Tkalcic, Jie Yang, and Nava Tintarev. “Exploring User Concerns about Disclosing Location and Emotion Information in Group Recommendations”. In: *Proceedings of the 32nd ACM Conference on Hypertext and Social Media.* HT ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 155–164. DOI: `10.1145/3465336.3475104`.

[244] Dong Nguyen. “Comparing Automatic and Human Evaluation of Local Explanations for Text Classification”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).* New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1069–1078. DOI: `10.18653/v1/N18-1097`.

[245] Raymond S Nickerson. “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises”. In: *Review of General Psychology* 2.2 (1998), pp. 175–220.

[246] Geoff Norman. “Likert Scales, Levels of Measurement and the “Laws” of Statistics”. In: *Advances in Health Sciences Education* 15.5 (Dec. 2010), pp. 625–632. DOI: `10.1007/s10459-010-9222-y`.

[247] Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. “The Preregistration Revolution”. In: *Proceedings of the National Academy of Sciences* 115.11 (2018), pp. 2600–2606.

[248] Alamir Novin and Eric Meyers. “Making Sense of Conflicting Science Information: Exploring Bias in the Search Engine Result Page”. In: *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval.* CHIIR ’17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 175–184. DOI: `10.1145/3020165.3020185`.

[249] Heather L. O’Brien, Paul Cairns, and Mark Hall. “A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form”. In: *International Journal of Human-Computer Studies* 112 (Apr. 2018), pp. 28–39. DOI: `10.1016/j.ijhcs.2018.01.004`.

[250] Maeve O’Brien and Mark T Keane. “Modeling Result-List Searching in the World Wide Web: The Role of Relevance Topologies and Trust Bias”. In: *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Vol. 28. Citeseer. 2006, pp. 1881–1886.

[251] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. “Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries”. In: *Frontiers in Big Data* 2 (July 2019), p. 13. DOI: `10.3389/fdata.2019.00013`.

[252] *Opinion.* URL: `https://dictionary.cambridge.org/dictionary/english/opinion`.

[253] Jason W Osborne and Elaine Waters. “Four Assumptions of Multiple Regression That Researchers Should Always Test”. In: *Practical Assessment, Research & Evaluation* 8.1 (2002), p. 2.

[254] Jahna Otterbacher. “Addressing Social Bias in Information Retrieval”. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction.* Ed. by Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric SanJuan, Linda Cappellato, and Nicola Ferro. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 121–127. DOI: `10.1007/978-3-319-98932-7_11`.

[255] Jahna Otterbacher. “Crowdsourcing Stereotypes: Linguistic Bias in Metadata Generated via Gwap”. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* 2015, pp. 1955–1964.

[256] Jahna Otterbacher, Pınar Barlas, Styliani Kleanthous, and Kyriakos Kyriakou. “How Do We Talk about Other People? Group (Un)Fairness in Natural Language Image Descriptions”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7.1 (Oct. 2019), pp. 106–114. DOI: `10.1609/hcomp.v7i1.5267`.

[257] Jahna Otterbacher, Jo Bates, and Paul Clough. “Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* CHI ’17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 6620–6631. DOI: `10.1145/3025453.3025727`.

[258] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. “In Google We Trust: Users’ Decisions on Rank, Position, and Relevance”. In: *Journal of Computer-Mediated Communication* 12.3 (Apr. 2007), pp. 801–823. DOI: `10.1111/j.1083-6101.2007.00351.x`.

[259] Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You.* Penguin UK, 2011.

[260] Sachin Pathiyan Cherumanal, Damiano Spina, Falk Scholer, and W. Bruce Croft. "Evaluating Fairness in Argument Retrieval". In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management.* CIKM '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 3363–3367. DOI: `10.1145/3459637.3482099`.

[261] Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. "JU_NLP at SemEval-2016 Task 6: Detecting Stance in Tweets Using Support Vector Machines". In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).* 2016, pp. 440–444.

[262] Michael Paul and Roxana Girju. "A Two-Dimensional Topic-Aspect Model for Discovering Multi-Faceted Topics". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 24.1 (July 2010), pp. 545–550. DOI: `10.1609/aaai.v24i1.7669`.

[263] Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. "Comparing Bayesian Models of Annotation". In: *Transactions of the Association for Computational Linguistics* 6 (Dec. 2018), pp. 571–585. DOI: `10.1162/tacl_a_00040`.

[264] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[265] Rik Peels. *Responsible Belief: A Theory in Ethics and Epistemology.* Oxford University Press, 2016.

[266] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. "What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7.1 (Oct. 2019), pp. 125–134. DOI: `10.1609/hcomp.v7i1.5281`.

[267] Gordon Pennycook and David G Rand. "Crowdsourcing Judgments of News Source Quality". In: *SSRN. com* (2018).

[268] Gordon Pennycook and David G Rand. "Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality". In: *Proceedings of the National Academy of Sciences* 116.7 (2019), pp. 2521–2526.

[269] Gordon Pennycook and David G. Rand. "Lazy, Not Biased: Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning than by Motivated Reasoning". In: *Cognition* 188 (July 2019), pp. 39–50. DOI: `10.1016/j.cognition.2018.06.011`.

[270] Rafael B Pereira, Alexandre Plastino, Bianca Zadrozny, and Luiz HC Merschmann. "Correlation Analysis of Performance Measures for Multi-Label Classification". In: *Information Processing & Management* 54.3 (2018), pp. 359–369.

[271] Richard E Petty and Pablo Brinol. “Attitude Change”. In: *Advanced Social Psychology: The State of the Science.* Ed. by R. F. Baumeister and E. J. Finkel. Oxford University Press, 2010. Chap. 7, pp. 217–259. DOI: `10.1016/B978-0-12-375000-6.00040-9`.

[272] Richard E Petty, Duane T Wegener, and Leandre R Fabrigar. “Attitudes and Attitude Change”. In: *Annual review of psychology* 48.1 (1997), pp. 609–647.

[273] Marcos Rodrigues Pinto, Yuri Oliveira de Lima, Carlos Eduardo Barbosa, and Jano Moreira de Souza. “Towards Fact-Checking through Crowdsourcing”. In: *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD).* IEEE. 2019, pp. 494–499.

[274] Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. “The Positive and Negative Influence of Search Results on People’s Decisions about the Efficacy of Medical Treatments”. In: *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval.* ICTIR ’17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 209–216. DOI: `10.1145/3121050.3121074`.

[275] Dean Pomerleau and Delip Rao. *Fake News Challenge Stage 1 (FNC-I): Stance Detection.* 2017. URL: `https://fakenewschallenge.org`.

[276] Mauro P. Porto. “Frame Diversity and Citizen Competence: Towards a Critical Approach to News Quality”. In: *Critical Studies in Media Communication* 24.4 (Oct. 2007), pp. 303–321. DOI: `10.1080/07393180701560864`.

[277] John Pougué-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, and Doyne Farmer. “DEBAGREEMENT: A Comment-Reply Dataset for (Dis) Agreement Detection in Online Debates”. In: *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).* 2021.

[278] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. “Manipulating and Measuring Model Interpretability”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 2021, pp. 1–52.

[279] Nicolas Pröllochs. “Community-Based Fact-Checking on Twitter’s Birdwatch Platform”. In: *arXiv preprint arXiv:2104.07175* (2021). arXiv: `2104.07175`.

[280] Emily Pronin, Daniel Y Lin, and Lee Ross. “The Bias Blind Spot: Perceptions of Bias in Self versus Others”. In: *Personality and Social Psychology Bulletin* 28.3 (2002), pp. 369–381.

[281] Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C. Lipton, Graham Neubig, and William W. Cohen. “Evaluating Explanations: How Much Do Explanations from the Teacher Aid Students?” In: *arXiv:2012.00893 [cs]* (Dec. 2021). arXiv: `2012.00893 [cs]`. URL: `http://arxiv.org/abs/2012.00893` (visited on 03/22/2022).

[282] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. “Estimating Training Data Influence by Tracing Gradient Descent”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 19920–19930.

[283] Kristen Purcell, Lee Rainie, and Joanna Brenner. *Search Engine Use 2012.* 2012. URL: `https://www.pewresearch.org/internet/2012/03/09/search-engine-use-2012-2/` (visited on 01/18/2023).

[284] Cornelius Puschmann. “Beyond the Bubble: Assessing the Diversity of Political Search Results”. In: *Digital Journalism* 7.6 (July 2019), pp. 824–843. DOI: `10.1080/21670811.2018.1539626`.

[285] Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. “Short Text Topic Modeling Techniques, Applications, and Performance: A Survey”. In: *IEEE Transactions on Knowledge and Data Engineering* (2020).

[286] Filip Radlinski and Nick Craswell. “Comparing the Sensitivity of Information Retrieval Metrics”. In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2010, pp. 667–674.

[287] Pavani Rajula, Chia-Chien Hung, and Simone Paolo Ponzetto. “Stacked Model Based Argument Extraction and Stance Detection Using Embedded LSTM Model”. In: *Working Notes Papers of the CLEF* (2022).

[288] Jorge Ramírez, Marcos Baez, Fabio Casati, Luca Cernuzzi, and Boualem Benatallah. “DREC: Towards a Datasheet for Reporting Experiments in Crowdsourcing”. In: *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing.* CSCW ’20 Companion. New York, NY, USA: Association for Computing Machinery, 2020, pp. 377–382. DOI: `10.1145/3406865.3418318`.

[289] Jorge Ramírez, Burcu Sayin, Marcos Baez, Fabio Casati, Luca Cernuzzi, Boualem Benatallah, and Gianluca Demartini. “On the State of Reporting in Crowdsourcing Experiments and a Checklist to Aid Current Practices”. In: *Proceedings of the ACM on Human-Computer Interaction (PACM HCI), Presented at the 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2021). October 2021.* 2021.

[290] Partha Pratim Ray. “ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope”. In: *Internet of Things and Cyber-Physical Systems* 3 (Jan. 2023), pp. 121–154. DOI: `10.1016/j.iotcps.2023.04.003`.

[291] Radim Řehůřek and Petr Sojka. “Software Framework for Topic Modelling with Large Corpora”. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.* Valletta, Malta: ELRA, May 2010, pp. 45–50.

[292] Jan Heinrich Reimer, Johannes Huck, and Alexander Bondarenko. “Grimjack at Touché 2022: Axiomatic Re-Ranking and Query Reformulation”. In: *Working Notes Papers of the CLEF* (2022).

[293] Myrthe Reuver, Suzan Verberne, Roser Morante, and Antske Fokkens. “Is Stance Detection Topic-Independent and Cross-Topic Generalizable? - A Reproduction Study”. In: *Proceedings of the 8th Workshop on Argument Mining*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 46–56. DOI: `10.18653/v1/2021.argmining-1.5`.

[294] Rezvaneh Rezapour, Ly Dinh, and Jana Diesner. “Incorporating the Measurement of Moral Foundations Theory into Analyzing Stances on Controversial Topics”. In: *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*. HT ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 177–188. DOI: `10.1145/3465336.3475112`.

[295] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “" Why Should i Trust You?" Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144.

[296] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Semantically Equivalent Adversarial Rules for Debugging NLP Models”. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. 2018.

[297] Alisa Rieger, Tim Draws, Mariët Theune, and Nava Tintarev. “Nudges to Mitigate Confirmation Bias during Web Search for Opinion Formation: Support vs. Manipulation”. In: *ACM Transactions on the Web* (in press).

[298] Alisa Rieger, Tim Draws, Nava Tintarev, and Mariet Theune. “This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias”. In: *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*. HT ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 189–199. DOI: `10.1145/3465336.3475101`.

[299] Kevin Roitero, Cristian Bozzato, Vincenzo Della Mea, Stefano Mizzaro, and Giuseppe Serra. “Twitter Goes to the Doctor: Detecting Medical Tweets Using Machine Learning and BERT.” In: *SIIRH@ ECIR*. 2020.

[300] Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. “How Many Truth Levels? Six? One Hundred? Even More? Validating Truthfulness of Statements via Crowdsourcing”. In: *Proceedings of the CIKM 2018 Workshops Co-Located with 27th ACM International Conference on Information and Knowledge Management (CIKM 2018)*. New York, NY, USA: Association for Computing Machinery, 2018. URL: `https://ceur-ws.org/Vol-2482/paper38.pdf`.

[301] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. “Can the Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor’s Background”. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 439–448. DOI: `10.1145/3397271.3401112`.

[302] Kevin Roitero, Michael Soprano, Beatrice Portelli, Massimiliano De Luise, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. "Can the Crowd Judge Truthfulness? A Longitudinal Study on Recent Misinformation about COVID-19". In: *Personal and Ubiquitous Computing* 1.1 (2021), pp. 1–31.

[303] Kevin Roitero, Michael Soprano, Beatrice Portelli, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. "The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively?" In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 1305–1314.

[304] Mir Rosenberg. *Toward a More Intelligent Search: Bing Multi-Perspective Answers*. 2018. URL: `https://blogs.bing.com/search-quality-insights/february-2018/Toward-a-More-Intelligent-Search-Bing-Multi-Perspective-Answers` (visited on 01/02/2023).

[305] Alexis Ross, Ana Marasović, and Matthew Peters. "Explaining NLP Models via Minimal Contrastive Editing (MiCE)". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 3840–3852. DOI: `10.18653/v1/2021.findings-acl.336`.

[306] Arjun Roy, Pavlos Fafalios, Asif Ekbal, Xiaofei Zhu, and Stefan Dietze. "Exploiting Stance Hierarchies for Cost-Sensitive Stance Detection of Web Documents". In: *Journal of Intelligent Information Systems* 58.1 (Feb. 2022), pp. 1–19. DOI: `10.1007/s10844-021-00642-z`.

[307] Farah Saab, Imad H Elhajj, Ayman Kayssi, and Ali Chehab. "Modelling Cognitive Bias in Crowdsourcing Systems". In: *Cognitive Systems Research* 58 (2019), pp. 1–18.

[308] Tetsuya Sakai, Nick Craswell, Ruihua Song, Stephen Robertson, Zhicheng Dou, and Chin-Yew Lin. "Simple Evaluation Metrics for Diversified Search Results". In: (2010), p. 9.

[309] Tetsuya Sakai and Ruihua Song. "Evaluating Diversified Search Results Using Per-Intent Graded Relevance". In: *SIGIR'11 - Proc. 34th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.* (2011), pp. 1043–1052. DOI: `10.1145/2009916.2010055`.

[310] Tetsuya Sakai and Zhaohao Zeng. "Which Diversity Evaluation Measures Are "Good"?" In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 595–604. DOI: `10.1145/3331184.3331215`.

[311] Younes Samih and Kareem Darwish. "A Few Topical Tweets Are Enough for Effective User Stance Detection". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, 2021, pp. 2637–2646. DOI: `10.18653/v1/2021.eacl-main.227`.

[312] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. *DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter.* 2019. DOI: `10.48550/ARXIV.1910.01108`.

[313] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. "Search Result Diversification". In: *Foundations and Trends in Information Retrieval* 9.1 (2015), pp. 1–90. DOI: `10.1561/1500000040`.

[314] Rodrygo L.T. Santos, Craig Macdonald, and Iadh Ounis. "Exploiting Query Reformulations for Web Search Result Diversification". In: *Proceedings of the 19th International Conference on World Wide Web.* WWW '10. New York, NY, USA: Association for Computing Machinery, 2010, pp. 881–890. DOI: `10.1145/1772690.1772780`.

[315] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. "Quantifying the Impact of User Attentionon Fair Group Representation in Ranked Lists". In: *Companion Proceedings of the 2019 World Wide Web Conference.* WWW '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 553–562. DOI: `10.1145/3308560.3317595`.

[316] Alexandra Sarafoglou, Anna van der Heijden, Tim Draws, Joran Cornelisse, Eric-Jan Wagenmakers, and Maarten Marsman. "Combine Statistical Thinking With Open Scientific Practice: A Protocol of a Bayesian Research Project". In: *Psychology Learning & Teaching* (Feb. 2022), pp. 1–13. DOI: `10.1177/14757257221077307`.

[317] Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. "Stance Detection Benchmark: How Robust Is Your Stance Detection?" In: *KI-Künstliche Intelligenz* 35.3 (2021), pp. 329–341.

[318] Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. "Human Interpretation of Saliency-Based Explanation over Text". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 611–636. DOI: `10.1145/3531146.3533127`.

[319] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *Proceedings of the IEEE International Conference on Computer Vision.* 2017, pp. 618–626.

[320] Robiert Sepúlveda-Torres, Marta Vicente, Estela Saquete, Elena Lloret, and Manuel Palomar. "Exploring Summarization to Enhance Headline Stance Detection". In: *Natural Language Processing and Information Systems.* Ed. by Elisabeth Métais, Farid Meziane, Helmut Horacek, and Epaminondas Kapetanios. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 243–254. DOI: `10.1007/978-3-030-80599-9_22`.

[321] Ricky J. Sethi. "Crowdsourcing the Verification of Fake News and Alternative Facts". In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media.* HT '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 315–316. DOI: `10.1145/3078714.3078746`.

[322] Shaban Shabani and Maria Sokhn. "Hybrid Machine-Crowd Approach for Fake News Detection". In: *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC).* IEEE. 2018, pp. 299–306.

[323] Chirag Shah and Emily M. Bender. "Situating Search". In: *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval.* CHIIR '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 221–232. DOI: `10.1145/3498366.3505816`.

[324] Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. "Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Online: Association for Computational Linguistics, July 2020, pp. 5248–5264. DOI: `10.18653/v1/2020.acl-main.468`.

[325] S. M. Sadiq-Ur-Rahman Shifath, Mohammad Faiyaz Khan, and Md. Saiful Islam. *A Transformer Based Approach for Fighting COVID-19 Fake News.* 2021. DOI: `10.48550/ARXIV.2101.12027`.

[326] Ashudeep Singh and Thorsten Joachims. "Fairness of Exposure in Rankings". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* KDD '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 2219–2228. DOI: `10.1145/3219819.3220088`.

[327] Paul Slovic, Melissa L Finucane, Ellen Peters, and Donald G MacGregor. "The Affect Heuristic". In: *European journal of operational research* 177.3 (2007), pp. 1333–1352.

[328] Catherine L. Smith and Soo Young Rieh. "Knowledge-Context in Search Systems: Toward Information-Literate Actions". In: *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval.* CHIIR '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 55–62. DOI: `10.1145/3295750.3298940`.

[329] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. "Cheap and Fast – but Is It Good? Evaluating Non-Expert Annotations for Natural Language Tasks". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing.* Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 254–263. URL: `https://aclanthology.org/D08-1027`.

[330] Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. "A Dataset for Multi-Target Stance Detection". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers.* 2017, pp. 551–557.

[331] Parinaz Sobhani, Saif M. Mohammad, and Svetlana Kiritchenko. "Detecting Stance in Tweets and Analyzing Its Interaction with Sentiment". In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*Sem).* Berlin, Germany, 2016.

[332] Swapna Somasundaran and Janyce Wiebe. “Recognizing Stances in Ideological On-Line Debates”. In: *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.* 2010, pp. 116–124.

[333] Michael Soprano, Kevin Roitero, Francesco Bombassei De Bona, and Stefano Mizzaro. “Crowd_Frame: A Simple and Complete Framework to Deploy Complex Crowdsourcing Tasks off-the-Shelf”. In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining.* WSDM ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 1605–1608. DOI: `10.1145/3488560.3502182`.

[334] Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. “The Many Dimensions of Truthfulness: Crowdsourcing Misinformation Assessments on a Multidimensional Scale”. In: *Information Processing & Management* 58.6 (Nov. 2021), p. 102710. DOI: `10.1016/j.ipm.2021.102710`.

[335] “Spearman Rank Correlation Coefficient”. In: *The Concise Encyclopedia of Statistics.* New York, NY: Springer New York, 2008, pp. 502–505. DOI: `10.1007/978-0-387-32833-1_379`.

[336] Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. “Argumentext: Searching for Arguments in Heterogeneous Sources”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations.* 2018, pp. 21–25.

[337] *Stance.* URL: `https://dictionary.cambridge.org/de/thesaurus/stance`.

[338] Keith E. Stanovich and Richard F. West. “Reasoning Independently of Prior Belief and Individual Differences in Actively Open-Minded Thinking”. In: *Journal of Educational Psychology* 89.2 (1997), pp. 342–357. DOI: `10.1037/0022-0663.89.2.342`.

[339] Mukund Sundararajan and Amir Najmi. “The Many Shapley Values for Model Explanation”. In: *Proceedings of the 37th International Conference on Machine Learning.* ICML’20. JMLR.org, 2020.

[340] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic Attribution for Deep Networks”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70.* ICML’17. Sydney, NSW, Australia: JMLR.org, 2017, pp. 3319–3328.

[341] James Surowiecki. *The Wisdom of Crowds.* Knopf Doubleday Publishing Group, Aug. 2005.

[342] Mariona Taulé, M Antonia Martí, Francisco M Rangel, Paolo Rosso, Cristina Bosco, Viviana Patti, et al. “Overview of the Task on Stance and Gender Detection in Tweets on Catalan Independence at IberEval 2017”. In: *2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017.* Vol. 1881. CEUR-WS. 2017, pp. 157–177.

[343] Matt Thomas, Bo Pang, and Lillian Lee. “Get out the Vote: Determining Support or Opposition from Congressional Floor-Debate Transcripts”. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.* EMNLP ’06. USA: Association for Computational Linguistics, 2006, pp. 327–335.

[344] Thibaut Thonet, Guillaume Cabanac, Mohand Boughanem, and Karen Pinel-Sauvagnat. “VODUM: A Topic Model Unifying Viewpoint, Topic and Opinion Discovery”. In: *Advances in Information Retrieval.* Ed. by Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello. Vol. 9626. Cham: Springer International Publishing, 2016, pp. 533–545. DOI: `10.1007/978-3-319-30671-1_39`.

[345] Nava Tintarev, Emily Sullivan, Dror Guldin, Sihang Qiu, and Daan Odjik. “Same, Same, but Different: Algorithmic Diversification of Viewpoints in News”. In: *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization.* UMAP ’18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 7–13. DOI: `10.1145/3213586.3226203`.

[346] Zakary L. Tormala and Derek D. Rucker. “Attitude Certainty: Antecedents, Consequences, and New Directions”. In: *Consumer Psychology Review* 1.1 (Jan. 2018), pp. 72–89. DOI: `10.1002/arcp.1004`.

[347] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. “Detecting Fake News in Social Networks via Crowdsourcing”. In: *arXiv preprint arXiv:1711.09025* (2017). arXiv: `1711.09025`.

[348] Martin Tutek, Ivan Sekulić, Paula Gombar, Ivan Paljak, Filip Čulinović, Filip Boltužić, Mladen Karan, Domagoj Alagić, and Jan Šnajder. “Takelab at Semeval-2016 Task 6: Stance Classification in Tweets Using a Genetic Algorithm Based Ensemble”. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).* 2016, pp. 464–468.

[349] Amos Tversky and Daniel Kahneman. “Judgment under Uncertainty: Heuristics and Biases”. In: *Science (New York, N.Y.)* 185.4157 (1974), pp. 1124–1131. DOI: `10.1021/ac60205a032`.

[350] Jodie B Ullman and Peter M Bentler. “Structural Equation Modeling”. In: *Handbook of Psychology* (2003), pp. 607–634.

[351] Don van den Bergh, Johnny van Doorn, Maarten Marsman, Tim Draws, Erik-Jan van Kesteren, Koen Derks, Fabian Dablander, Quentin Frederik Gronau, Šimon Kucharský, Akash R. Komarlu Narendra Gupta, Alexandra Sarafoglou, Jan G. Voelkel, Angelika Stefan, Max Hinne, Dora Matzke, and Eric-Jan Wagenmakers. “A Tutorial on Conducting and Interpreting a Bayesian ANOVA in JASP”. In: *Annee Psychologique* 120.1 (Feb. 2020), pp. 73–96. DOI: `10.3917/anpsy1.201.0073`.

[352] Antal Van den Bosch, Toine Bogers, and Maurice De Kunder. “Estimating Search Engine Index Size Variability: A 9-Year Longitudinal Study”. In: *Scientometrics* 107.2 (2016), pp. 839–856.

[353] Johnny van Doorn, Don van den Bergh, Udo Böhm, Fabian Dablander, Koen Derks, Tim Draws, Alexander Etz, Nathan J. Evans, Quentin F. Gronau, Julia M. Haaf, Max Hinne, Šimon Kucharský, Alexander Ly, Maarten Marsman, Dora Matzke, Akash R. Komarlu Narendra Gupta, Alexandra Sarafoglou, Angelika Stefan, Jan G. Voelkel, and Eric-Jan Wagenmakers. “The JASP Guidelines for Conducting and Reporting a Bayesian Analysis”. In: *Psychonomic Bulletin & Review* 28.3 (June 2021), pp. 813–826. DOI: `10.3758/s13423-020-01798-5`.

[354] Marieke van Hoof, Corine S Meppelink, Judith Moeller, and Damian Trilling. “Searching Differently? How Political Attitudes Impact Search Queries about Political Issues”. In: *New Media & Society* (July 2022), p. 146144482211044. DOI: `10.1177/14614448221104405`.

[355] Caspar J. Van Lissa, Wolfgang Stroebe, Michelle R. vanDellen, N. Pontus Leander, Maximilian Agostini, Tim Draws, Andrii Grygoryshyn, Ben Gützgow, et al. “Using Machine Learning to Identify Important Predictors of COVID-19 Infection Prevention Behaviors during the Early Phase of the Pandemic”. In: *Patterns* 3.4 (Apr. 2022), p. 100482. DOI: `10.1016/j.patter.2022.100482`.

[356] David Vilares and Yulan He. “Detecting Perspectives in Political Debates”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 1573–1582. DOI: `10.18653/v1/D17-1165`.

[357] Jacky Visser, John Lawrence, and Chris Reed. “Reason-Checking Fake News”. In: *Communications of the ACM* 63.11 (Oct. 2020), pp. 38–40.

[358] Paul S Voakes, Jack Kapfer, David Kurpius, and David Shano-yeon Chern. “Diversity in the News: A Conceptual and Methodological Framework”. In: *Journalism & Mass Communication Quarterly* 73.3 (1996), pp. 582–593.

[359] Ellen M Voorhees. “Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness”. In: *Information processing & management* 36.5 (2000), pp. 697–716.

[360] Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, and Maarten De Rijke. “RADio – Rank-Aware Divergence Metrics to Measure Normative Diversity in News Recommendations”. In: *Proceedings of the 16th ACM Conference on Recommender Systems.* RecSys ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 208–219. DOI: `10.1145/3523227.3546780`.

[361] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. “Recommenders with a Mission: Assessing Diversity in Newsrecommendations”. In: *arXiv:2012.10185 [cs]* (Dec. 2020). arXiv: `2012.10185 [cs]`. URL: `http://arxiv.org/abs/2012.10185` (visited on 07/13/2021).

[362] Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. “Using Argument Mining to Assess the Argumentation Quality of Essays”. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers.* Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 1680–1691. URL: `https://aclanthology.org/C16-1158`.

[363] Eric-Jan Wagenmakers, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Jonathon Love, Ravi Selker, Quentin F. Gronau, Martin Šmíra, Sacha Epskamp, Dora Matzke, Jeffrey N. Rouder, and Richard D. Morey. "Bayesian Inference for Psychology. Part I: Theoretical Advantages and Practical Ramifications". In: *Psychonomic Bulletin & Review* 25.1 (Feb. 2018), pp. 35–57. DOI: `10.3758/s13423-017-1343-3`.

[364] Paul Wais, Shivaram Lingamneni, Duncan Cook, Jason Fennell, Benjamin Goldenberg, Daniel Lubarov, David Marin, and Hari Simons. "Towards Building a High-Quality Workforce with Mechanical Turk". In: *Proceedings of computational social science and the wisdom of crowds (NIPS)* (2010). URL: `https://people.cs.umass.edu/~wallach/workshops/nips2010css/papers/wais.pdf`.

[365] Kelly Walters, Dimitri A. Christakis, and Davene R. Wright. "Are Mechanical Turk Worker Samples Representative of Health Status and Health Behaviors in the U.S.?" In: *PLOS ONE* 13.6 (June 2018), e0198835. DOI: `10.1371/journal.pone.0198835`.

[366] Rui Wang, Deyu Zhou, Mingmin Jiang, Jiasheng Si, and Yang Yang. "A Survey on Opinion Mining: From Stance to Product Aspect". In: *IEEE Access* 7 (2019), pp. 41101–41124. DOI: `10.1109/ACCESS.2019.2906754`.

[367] William Yang Wang. ""Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 422–426. DOI: `10.18653/v1/P17-2067`.

[368] Fabian L. Wauthier and Michael I. Jordan. "Bayesian Bias Mitigation for Crowdsourcing". In: *Adv. Neural Inf. Process. Syst. 24 25th Annu. Conf. Neural Inf. Process. Syst. 2011, NIPS 2011.* 2011, pp. 1–9.

[369] Ingmar Weber and Alejandro Jaimes. "Who Uses Web Search for What: And How". In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining.* 2011, pp. 15–24.

[370] Ryen White. "Beliefs and Biases in Web Search". In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval.* SIGIR '13. New York, NY, USA: Association for Computing Machinery, 2013, pp. 3–12. DOI: `10.1145/2484028.2484053`.

[371] Ryen W. White and Ahmed Hassan. "Content Bias in Online Health Search". In: *ACM Transactions on the Web* 8.4 (Nov. 2014), pp. 1–33. DOI: `10.1145/2663355`.

[372] Ryen W. White and Eric Horvitz. "Belief Dynamics and Biases in Web Search". In: *ACM Transactions on Information Systems* 33.4 (May 2015), pp. 1–46. DOI: `10.1145/2746229`.

[373] Michael Wojatzki and Torsten Zesch. "Ltl. Uni-Due at Semeval-2016 Task 6: Stance Detection in Social Media Using Stacked Classifiers". In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).* 2016, pp. 428–433.

[374] Meng-Han Wu and Alexander Quinn. "Confusing the Crowd: Task Instruction Quality on Amazon Mechanical Turk". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 5.1 (Sept. 2017), pp. 206–215. URL: `https://ojs.aaai.org/index.php/HCOMP/article/view/13317`.

[375] Zhangyi Wu, Tim Draws, Federico Cau, Francesco Barile, Alisa Rieger, and Nava Tintarev. "Explaining Search Result Stances to Opinionated People". In: *Explainable Artificial Intelligence*. Ed. by Luca Longo. Communications in Computer and Information Science. Cham: Springer Nature Switzerland, 2023, pp. 573–596. DOI: `10.1007/978-3-031-44067-0_29`.

[376] Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. "Cross-Target Stance Classification with Self-Attention Networks". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 778–783. DOI: `10.18653/v1/P18-2123`.

[377] Luyan Xu, Mengdie Zhuang, and Ujwal Gadiraju. "How Do User Opinions Influence Their Interaction with Web Search Results?" In: *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 240–244. URL: `https://doi.org/10.1145/3450613.3456824`.

[378] Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. "Overview of NLPCC Shared Task 4: Stance Detection in Chinese Microblogs". In: *Natural Language Understanding and Intelligent Applications*. Ed. by Chin-Yew Lin, Nianwen Xue, Dongyan Zhao, Xuanjing Huang, and Yansong Feng. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 907–916. DOI: `10.1007/978-3-319-50496-4_85`.

[379] Yusuke Yamamoto and Satoshi Shimada. "Can Disputed Topic Suggestion Enhance User Consideration of Information Credibility in Web Search?" In: *Proceedings of the 27th ACM Conference on Hypertext and Social Media*. HT '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 169–177. DOI: `10.1145/2914586.2914592`.

[380] Jie Yang, Judith Redi, Gianluca Demartini, and Alessandro Bozzon. "Modeling Task Complexity in Crowdsourcing". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4.1 (Sept. 2016), pp. 249–258. DOI: `10.1609/hcomp.v4i1.13283`.

[381] Ke Yang and Julia Stoyanovich. "Measuring Fairness in Ranked Outputs". In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. SSDBM '17. New York, NY, USA: Association for Computing Machinery, 2017. DOI: `10.1145/3085504.3085526`.

[382] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, HV Jagadish, and Gerome Miklau. "A Nutritional Label for Rankings". In: *Proceedings of the 2018 International Conference on Management of Data*. SIGMOD '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1773–1776. DOI: `10.1145/3183713.3193568`.

[383] Scott Cheng-Hsin Yang, Tomas Folke, and Patrick Shafto. “A Psychological Theory of Explainability”. In: *Proceedings of the 39 Th International Conference on Machine Learning.* Vol. 162. Proceedings of Machine Learning Research. Baltimore, Maryland, USA: PMLR, 2022, pp. 25007–25021.

[384] Arnold Yeung, Shalmali Joshi, Joseph Jay Williams, and Frank Rudzicz. *Sequential Explanations with Mental Model-Based Policies.* 2020.

[385] Elad Yom-Tov, Susan Dumais, and Qi Guo. “Promoting Civil Discourse Through Search Engine Diversity”. In: *Social Science Computer Review* 32.2 (Apr. 2014), pp. 145–154. DOI: `10.1177/0894439313506838`.

[386] Yisong Yue, Rajan Patel, and Hein Roehrig. “Beyond Position Bias: Examining Result Attractiveness as a Source of Presentation Bias in Clickthrough Data”. In: *Proceedings of the 19th International Conference on World Wide Web.* WWW ’10. New York, NY, USA: Association for Computing Machinery, 2010, pp. 1011–1018. DOI: `10.1145/1772690.1772793`.

[387] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. “Disentangling Fairness Perceptions in Algorithmic Decision-Making: The Effects of Explanations, Human Oversight, and Contestability”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* CHI ’23. New York, NY, USA: Association for Computing Machinery, 2023. DOI: `10.1145/3544548.3581161`.

[388] Robert B. Zajonc. “Attitudinal Effects of Mere Exposure.” In: *Journal of Personality and Social Psychology* 9.2, Pt.2 (1968), pp. 1–27. DOI: `10.1037/h0025848`.

[389] Hamed Zamani, Michael Bendersky, Xuanhui Wang, and Mingyang Zhang. “Situational Context for Ranking in Personal Search”. In: *Proceedings of the 26th International Conference on World Wide Web.* WWW ’17. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017, pp. 1531–1540. DOI: `10.1145/3038912.3052648`.

[390] Antonia Zapf, Stefanie Castell, Lars Morawietz, and André Karch. “Measuring Inter-Rater Reliability for Nominal Data – Which Coefficients and Confidence Intervals Are Appropriate?” In: *BMC Medical Research Methodology* 16.1 (Dec. 2016), p. 93. DOI: `10.1186/s12874-016-0200-9`.

[391] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. “FA*IR: A Fair Top-k Ranking Algorithm”. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* CIKM ’17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 1569–1578. DOI: `10.1145/3132847.3132938`.

[392] Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. “Fair Top-k Ranking with Multiple Protected Groups”. In: *Information Processing & Management* 59.1 (Jan. 2022), p. 102707. DOI: `10.1016/j.ipm.2021.102707`.

[393] Meike Zehlike, Ke Yang, and Julia Stoyanovich. “Fairness in Ranking: A Survey”. In: *arXiv:2103.14000 [cs]* (May 2021). arXiv: `2103.14000 [cs]`. URL: `http://arxiv.org/abs/2103.14000` (visited on 10/28/2021).

[394] Guangzhen Zhao and Peng Yang. "Pretrained Embeddings for Stance Detection with Hierarchical Capsule Network on Social Media". In: *ACM Transactions on Information Systems* 39.1 (Sept. 2020). DOI: `10.1145/3412362`.

[395] Steven Zimmerman. "Exploring Strategies to Prevent Harm from Web Search". PhD thesis. University of Essex, Jan. 2021. URL: `https://repository.essex.ac.uk/29762/` (visited on 10/31/2023).

[396] Steven Zimmerman, Stefan M. Herzog, David Elsweiler, Jon Chamberlain, and Udo Kruschwitz. "Towards a Framework for Harm Prevention in Web Search". In: *Proceedings of the First Workshop on Bridging the Gap between Information Science, Information Retrieval and Data Science (BIRDS 2020), Co-Located with 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 202.* Ed. by Ingo Frommholz, Haiming Liu, and Massimo Melucci. Vol. 2741. Xi'an, China (online).: CEUR Workshop Proceedings, July 2020, pp. 30–46. URL: `http://ceur-ws.org/Vol-2741/paper-07.pdf` (visited on 10/31/2023).

List of Figures

3.1 Example of a salience-based explanation (using BERT-base and LIME) from our user study. 32

3.2 Example of a bar plot explanation (using BERT-base and LIME) from our user study. 32

3.3 Distributions of macro-f1 scores across stance detection models (see also Table 3.2). Box plots (white) show medians and interquartile ranges, while violin plots (green) show how macro-f1 scores were distributed over the 100 runs. 37

3.4 Mean simulation proportion per explanation content, split by explanation visualization (Coeff. = coefficients, LR = logistic regression, LSVM = linear SVM, IG = integrated gradients, GC = Grad-CAM). The dotted line reflects always selecting the true instead of (as instructed) the predicted stance label (i.e., 10 out of 20 explanations were for incorrect predictions). 43

4.1 Screenshot of the main task. Word groups 1 and 7 are the two honeypot topics. 57

4.2 Mean *nCor* (i.e., the mean number of correctly identified perspectives) per model. The error bars represent the standard error. 60

4.3 Normalized distribution of how often each available perspective was chosen (excluding the two honeypot checks). Whereas p_1-p_6 were actually present in the corpus (see Table 4.3), the remaining perspectives were not. The horizontal line at $\frac{1}{16} = 0.0625$ shows the expected proportion for random selection. 62

5.1 Proposed viewpoint representation at the example of a tweet from the *SemEval 2016 Stance Detection* data set [233]. A viewpoint is evaluated on two dimensions: *stance* (i.e., on a seven-point ordinal scale ranging from "strongly opposing" to "strongly supporting" a topic) and *logic of evaluation* (i.e., in a multi-categorical format to include all logics present; see Table 2.2). 67

5.2 Relative frequency of the seven different logics across the topics *atheism*, *Donald Trump*, and *feminist movement*. 74

5.3 Network plots of all tweets divided into the three topics *atheism* (left-hand panel), *Donald Trump* (central panel), and the *feminist movement* (right-hand panel). Each node is a single tweet, whereby its color indicates the stance. Edges indicate the Jaccard similarity between tweets based on the assigned logics (i.e., the stronger the edge, the greater the similarity). . . . 75

6.1 Example item to collect viewpoint (i.e., stance label) annotations for search results in our case study. 95

7.1 Mean `eME` per political affiliations of statements and workers (left) and mean `eMAE` per southern border answer and political affiliations of workers (right) in the public data set (see Section 7.2). Here, we excluded four workers who did not consider themselves a Democrat, independent, or Republican. . . 109
7.2 Comparison of workers' abandonment distribution (left) and workers' failure distribution (right). The orange line represents our task. The blue lines represent the task by Soprano et al. [334]. 115
7.3 Scatter plots showing the relationships between workers' `eMAE` and their *trust in politics* ($\mathbf{H_{II.1a}}$, left-hand plot), *belief in science* ($\mathbf{H_{II.1b}}$, center-left plot), Cognitive Reflection Test (CRT; $\mathbf{H_{II.1c}}$, center-right plot), and mean confidence ($\mathbf{H_{II.2d}}$, right-hand plot). Our multiple linear regression analysis identified *belief in science* as well as mean confidence as significant predictors of `eMAE` (see Section 7.2.4). 117

8.1 Behavior of the metrics rND (top plot), rRD (center plot), and rKL (bottom plot) on the sets $S1$ ($S^p = 300$), $S2$ ($S^p = 240$), and $S3$ ($S^p = 180$) across different α (ranking bias) settings. 134
8.2 Behavior of nDJS on the sets $S1$, $S2$, and $S3$ across different α (ranking bias) settings. The number of items with sample weight w_1 for rankings from the sets $S1$, $S2$, and $S3$ are 100, 80, and 60, respectively. Note that we have zoomed in here compared to Figure 8.1 to better show nuanced differences between the lines. 136

9.1 Development of mean absolute nDVB@k across search result ranks, split by topic and search engine. 150
9.2 Mean absolute viewpoint diversity (nDVB@10) per diversification algorithm across the 30 search result lists. 150

10.1 Click proportions over the ranks. Users exhibited a weak position bias across conditions. 170
10.2 Mean absolute opinion change over the three conditions. Error bars represent the standard error. 171

List of Tables

2.1 The seven ordinal stance categories we consider. 14

2.2 The seven logics of evaluation we consider, adapted from Baden and Springer [24]. Each logic represents a particular orientation of what is desired and can be used to either support or oppose a given claim. 14

3.1 The topic and stance distributions in our data set. 33

3.2 Mean test set performances (± standard error) of stance detection models over 100 trials (i.e., using 10 different seeds controlling any model randomness for each of 10 different data splits; best scores in each column are bold). 36

3.3 Mean test set performances (± standard error; except for deterministic models) over 10 random seeds on the data split where the four selected models performed best. 38

4.1 Abortion perspectives in the final data set. Whereas the perspectives p_1, p_2, and p_3 support the legalization of abortion, p_4, p_5, and p_6 oppose it. . . . 53

4.2 Models used in the user study. 55

4.3 Participant's pre-existing abortion stance. 59

4.4 Descriptive statistics of the user study. Here, n refers to the number of participants, mean *nCor* to the mean number of correctly identified perspectives per model (ranging from 0 to 6), and SE to the standard error. 61

4.5 Descriptive statistics (mean ± standard deviation) on the exploratory measurements. Responses are from five-point Likert scales with 1 = "strongly disagree" and 5 = "strongly agree." 62

4.6 The six topics computed by TAM. 63

6.1 The stance label taxonomy we considered as viewpoint representation in our case study. Crowd workers could assign a stance label to each search result by selecting one of the seven options ranging from "strongly opposing" ($l = -3$) to "strongly supporting" ($l = 3$). 95

6.2 Results of the retrospective analysis of cognitive biases in crowdsourcing papers from HCOMP proceedings in 2018, 2019, 2020. Here, *biases considered* refers to papers that discussed the identified cognitive biases at least partly. 100

8.1 Notation used throughout this chapter. 127

8.2 Viewpoint (i.e., stance) distributions of the sets $S1$, $S2$, and $S3$. 132

8.3 Examples of sample weight allocations for the simulation of binomial (left-hand table, $\alpha = 0.5$) and multinomial viewpoint fairness (right-hand table; $\alpha = -0.8$). 133

8.4 Recommended metrics for different scenarios of ranking bias and overall viewpoint distribution (i.e., protected and non-protected items) in a ranked list. 138

9.1 Viewpoint diversity evaluation for all 30 search result lists from Engine 1 and 2: rND, RB, and nDVB (including its sub-metrics nDPB, nDSB, and nDLB). Queries were designed to retrieve neutral (neu), opposing (opp), or supporting (sup) results (↔). 147

10.1 Three conditions representing the three levels of ranking bias. Here, all rankings are biased toward opposing stances, but our study also included their symmetrical opposites (favoring supporting stances). 166
10.2 ANCOVA (absolute opinion change as dependent variable). Colons represent interaction effects. 172
10.3 Proportions of supporting documents among the search results that users clicked on (± standard deviation) in each condition, split by in what direction the SERP was biased. 174
10.4 Mean opinion change (± std. dev.) in each condition, split by SERP bias direction. 174

Summaries

English Summary

Web search engines have a growing influence on individuals and society as people increasingly rely on online resources when forming opinions and seeking advice. Although users typically trust search engines to be impartial, recent research has shown that search results can be biased toward particular viewpoints. Moreover, interacting with such viewpoint-biased search results can cause biased user behavior and opinion formation. This calls for a thorough assessment of viewpoint biases in search results as well as the development of bias mitigation and user support strategies. However, such efforts currently face multiple limitations, which we address in this dissertation.

A fundamental decision when examining viewpoints expressed in search results is how to *represent* those viewpoints and label search results accordingly. In Part I, we study the possibilities and limitations of automatic methods that consider ternary stance labels (i.e., *against*/*neutral*/*in favor*) and explore how to enhance such methods by adding *perspectives* (i.e., underlying reasons). We then propose a novel viewpoint representation framework consisting of two dimensions (i.e., *stance* and *logic of evaluation*). As we demonstrate, the corresponding viewpoint labels are feasible to obtain via crowdsourcing and allow for more nuanced viewpoint bias analyses than current methods.

Crowdsourcing comprehensive viewpoint labels for search results can be impeded by *cognitive biases* of crowd workers. Part II proposes a checklist to combat such cognitive worker biases. This checklist, comprising 12 commonly occurring cognitive biases, can be utilized by practitioners to assess, mitigate, and document potential influences of cognitive worker biases in the tasks they design. We apply our checklist in two different case studies (i.e., crowdsourcing viewpoint labels and truthfulness judgments).

Collecting high-quality viewpoint labels allows for comprehensive viewpoint bias evaluations, but it is currently unclear how to conduct such assessments. In Part III, we probe existing *ranking fairness metrics* for this task and propose a novel viewpoint bias metric that considers our proposed viewpoint representation from Part I. We find considerable viewpoint bias across topics, queries, and search engines in real search results and show how to increase viewpoint diversity using existing diversification algorithms.

Aside from viewpoint biases in search results, cognitive user biases can also affect web search interactions. In Part IV, we seek to identify what underlying mechanisms lead users to change their opinions following search result viewpoint biases. Exploratory evidence suggests that rather than order effects, *exposure effect* (i.e., adopting the majority viewpoint among the results users consume) seem to guide user behavior here.

Web search engines can and should be platforms for users to explore debated topics in all their nuances without cognitive overwhelm. We hope that the empirical evidence, tools, and resources we contribute can support this vision by developing a greater understanding of viewpoint biases in search results and their effects on user behavior.

Nederlandse Samenvatting

Zoekmachines oefenen een groeiende invloed uit op gebruikers en de samenleving omdat mensen steeds vaker vertrouwen op online bronnen om meningen te vormen.Hoewel gebruikers doorgaans vertrouwen op de onpartijdigheid van zoekmachines, kunnen zoekresultaten vertekend zijn ten opzichte van bepaalde standpunten (*viewpoint bias*). Bovendien kan de interactie met dergelijk vertekende zoekresultaten het gedrag van gebruikers beïnvloeden en leiden tot bevooroordeelde meningsvorming. Dit vraagt om grondig onderzoek naar viewpoint bias in zoekresultaten en de ontwikkeling van strategieën om deze bias te verminderen en gebruikers te ondersteunen. Dit proefschrift behandelt de beperkingen waarmee dergelijke inspanningen momenteel te maken hebben.

Een cruciale beslissing bij het onderzoeken van standpunten in zoekresultaten is hoe deze standpunten moeten worden gerepresenteerd. In Deel I evalueren wij de mogelijkheden en beperkingen van automatische methoden die ternaire standpunten (*tegen*/*neutraal*/*voor*) in overweging nemen, en onderzoeken wij hoe deze methoden kunnen worden verbeterd door *perspectives* (d.w.z., onderliggende premissen) toe te voegen. Vervolgens introduceren wij een nieuw raamwerk voor het representeren van standpunten dat bestaat uit twee dimensies: *stance* en *logic of evaluation*. Wij laten zien dat de bijbehorende labels kunnen worden verkregen via crowdsourcing en een verfijndere analyse van viewpoint bias mogelijk maken dan de huidige methoden.

Het crowdsourcen van uitgebreide standpuntenlabels voor zoekresultaten kan bemoeilijkt worden door cognitieve heuristieken van crowdworkers. In Deel II presenteren wij een checklist om dergelijke cognitieve heuristieken tegen te gaan. Deze checklist omvat 12 veelvoorkomende heuristieken en kan worden gebruikt om mogelijke invloeden van cognitieve heuristieken van crowdworkers te beoordelen, te beperken, en vast te leggen. Wij passen onze checklist toe in twee verschillende case studies, namelijk het crowdsourcen van viewpoint labels en waarheidsbeoordelingen.

Het crowdsourcen van hoogwaardige standpuntenlabels maakt uitgebreide analyses van viewpoint bias mogelijk, maar het is nog onduidelijk hoe dergelijke analyses moeten worden uitgevoerd. In Deel III onderzoeken wij of bestaande *ranking fairness* metrieken deze taak kunnen vervullen en presenteren wij een nieuwe metriek voor viewpoint bias die het door ons voorgestelde raamwerk voor standpuntenlabels uit Deel I omvat. Wij ontdekken een aanzienlijke mate van viewpoint bias in echte zoekresultaten van diverse onderwerpen, zoekopdrachten, en zoekmachines, en laten zien hoe de de diversiteit van meningen in zoekresultaten kan worden vergroot met behulp van bestaande methodes.

Naast viewpoint bias in zoekresultaten kunnen ook cognitieve heuristieken van gebruikers zoekinteracties beïnvloeden. In Deel IV proberen wij de onderliggende mechanismen te identificeren die gebruikers ertoe aanzetten om hun mening door viewpoint bias in zoekresultaten te veranderen. Voorlopig bewijs suggereert dat het gedrag van gebruikers lijkt te worden gestuurd door het *exposure effect* (d.w.z., het overnemen van het meerderheidsstandpunt onder de resultaten die gebruikers consumeren).

Zoekmachines op het web zouden platformen moeten zijn die gebruikers in staat stellen om gedebatteerde onderwerpen in al hun nuances te verkennen zonder overweldigd te raken. Wij hopen dat het empirische bewijs, de tools, en de ressourcen die wij aanbieden deze visie kunnen ondersteunen door een beter begrip te ontwikkelen van de viewpoint bias in zoekresultaten en de impact daarvan op het gedrag van gebruikers.

Acknowledgements

This dissertation could not have been written without the many people who have supported me over the years. To everyone, please realize that this has been a group effort. I feel incredibly lucky and am grateful for all of you.

I would like to express my deepest gratitude to my promotor and supervisor Nava for her continuous guidance, encouragement, and patience throughout my PhD journey. Your attention to detail and scientific rigor inspired and pushed me to apply the same standards to my own work. I think we can look back on a highly successful four years together. Thank you for always showing up for both the highs and lows. I would also like to thank my promotor Geert-Jan for being ready to invest time and effort into helping PhD candidates whenever that was needed. Your guidance has helped me immensely on several occasions. Thank you to my co-supervisor Ben, who provided a highly relevant industry perspective to our work and supported me wherever he could, e.g., in finding an internship and communicating our research findings to broader audiences.

My gratitude goes out to the members of my dissertation defense committee: Professors Gerd Kortuem, Udo Kruschwitz, Carsten Eickhoff, Mark Sanderson, and Markus Specht. Thank you for reading my earlier dissertation draft and for your valuable feedback. I am honored to defend my work before such a prestigious academic committee.

I cannot thank my wonderful colleagues at TU Delft enough – all of you have made the last four years so much more successful and enjoyable. Thank you to Alessandro and Ujwal for your guidance, feedback, and research contributions. Both of you played key roles in the making of this dissertation. Thank you to my paranymphs, Alisa and Felipe, who were always there to discuss important work-related and non-work-related topics. Alisa was instrumental in much of the research described here and Felipe regularly took time out of his day to explain information retrieval concepts or give me valuable career advice. I would like to thank my Epsilon lab colleagues Shabnam, Oana, Jody, and Mesut, as well as my office mates and collaborators Arthur, Nirmal, Gustavo, Agathe, and Mireia for their research contributions and so much fun during office chats and team outings. Thank you to my remaining WIS colleagues Avishek, Christoph, Claudia, Sole, Asterios, Jie, Rihan, Andra, Christos, David, Gaole, Garrett, George, Petros, Kyriakos, Lijun, Lorenzo, Nirmal, Peide, Philip, Sara, Sepideh, Shahin, Sihang, Wenbo, and Ziyu for all the fun lunch breaks, exciting conversations, and mutual support. Finally, a huge thank you to our secretary Daphne, who was always there for us when we needed her. We are a big family.

I would like to thank IBM for sponsoring our work and allowing me to conduct some of our research in an industry setting. Thank you to Michael, Kush, Karthi, Ioana, Amit, and Inkit for their valuable guidance and feedback during our collaborative projects. A special thank you goes out to Zoltan, who was one of the first to encourage me to embark on this journey and has supported me throughout.

Several people from other institutions deserve my gratitude. Thank you to Francesco, Federico, Amir, Rishav, Orçun, and Zhangyi from the University of Maastricht; David, Markus, and Sebastian from the University of Regensburg; Michael, David, Kevin, and Stefano from the University of Udine; as well as my research collaborators Christian Baden, Davide Ceolin, and Alessandro Checco.

I would also like to thank my exceptional former teachers and colleagues at the University of Amsterdam. Thank you EJ for allowing me to contribute to JASP as a student and teaching me good research practices. I do not believe this dissertation could have been written without your early influence. My gratitude also goes out to the remaining JASP team, including Alexander, Joris, Johnny, Don, Quentin, and Bruno, who taught me valuable things about maths and software development. Additionally, I would like Robert and Abe from the Behavioral Data Science Master's program for giving us a fantastic career preparation. I had such a great time at UvA.

Although my professional colleagues and collaborators were instrumental in the making of this dissertation, I could not have achieved any of this without the continuous and extensive support of my family. Thank you to my parents Elvi and Joe for their unconditional love and support. You have given everything to raise me well and, despite my efforts to pursue a career as a professional DJ in my late teens,[1] never left a doubt in my mind that you would be there for me no matter what. I dedicate this dissertation to you both. My deepest gratitude goes out to my fiancé Carol, the love of my life, whose ridiculous support has included preparing food when I could barely sustain myself close to paper deadlines and listening to boring talks as I practiced for conference appearances. On top of that, Carol designed a beautiful cover for this dissertation. You have been my anchor for the better part of this PhD journey, celebrating my wins, consoling my losses, and distracting me when I was too deep into some meaningless research problem. Thank you for everything – I love you. I would also like to thank my siblings Kai and Laura for their friendship and encouragement. Our time together has often helped me get my mind off of work. Thank you to Oma for teaching me yoga, mindfulness, and gratitude. Thank you to Markus, Debbie, Leah, Jack, Christel, and Bernard for all the fun during family meetups. Finally, a big thank you to my dog Sepp, who has been with me almost from the start of my PhD candidacy. You forced me to take breaks when I really needed to and taught me that it is okay to be lazy and do nothing at all. You are such a good boy.

I want to thank my friends for all the happy moments we have shared over the years. To Sascha, Timbow, Sven, and Marv, my second family whom I have known almost my entire life, thank you for making me laugh and reminding me that there are more important things than work. Thank you to Irish, Mitos, Phil, Katie, and David for all the fun trips, hot pots, and barbecues. Thank you to Alexandra and Akash for inspiring me to dance(!) and pick up other random hobbies. Thank you to Davud, Ori, Raphael, Liam, Ferron, Marleen, and Andreas for the countless hours of fun conversations and laughs we have shared over the past few years.

Last but not least, I want to thank the Netherlands for giving me the opportunity to study what I wanted (psychology) and providing a fantastic ground for PhD candidates. My almost nine years in Amsterdam included meeting my future wife and many dear friends, interesting cultural encounters, and, ultimately, writing this dissertation. I am grateful for the many friendly Dutch people who have made me feel welcome and – despite my German accent – were willing to talk Dutch to me. Ik ga jullie missen, hoor.

---

[1] For those who are curious, that one did not work out at all.

Curriculum Vitæ

Tim Alexander DRAWS

12-04-1991	Born in Mannheim, Germany.

Professional Experience

2023–present	Data Scientist, Otto, Hamburg, Germany
2019–2023	PhD Candidate, Delft University of Technology, Netherlands
2022	Research Intern, IBM Watson Research Center, Yorktown Heights, USA
2019	Research Intern, IBM Benelux, Amsterdam, Netherlands
2017–2019	Research Assistant, University of Amsterdam, Netherlands

Education

2019–2023	Doctor of Philosophy (PhD), Computer Science Delft University of Technology, Netherlands
2018–2019	Master of Science (MSc), Psychology, cum laude University of Amsterdam, Netherlands
2015–2018	Bachelor of Science (BSc), Psychology, with distinction University of Amsterdam, Netherlands

Awards and Grants

2023	Best Paper Award at the ACM CHI conference
2022	Best Paper Award at the ACM CHIIR conference
2021	Best Paper Award at the ACM HCOMP conference
2021	Best Paper Award at the ACM Hypertext conference
2021	Delft Design for Values Open Subsidy

List of Publications

25. Alisa Rieger, Tim Draws, Mariët Theune, and Nava Tintarev. “Nudges to Mitigate Confirmation Bias during Web Search for Opinion Formation: Support vs. Manipulation”. In: *ACM Transactions on the Web* (in press)

24. Oana Inel, Tim Draws, and Lora Aroyo. “Collect, Measure, Repeat: Reliability Factors for Responsible AI Data Collection”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 11.1 (Nov. 2023), pp. 51–64. DOI: 10.1609/hcomp.v11i1.27547

23. Zhangyi Wu, Tim Draws, Federico Cau, Francesco Barile, Alisa Rieger, and Nava Tintarev. “Explaining Search Result Stances to Opinionated People”. In: *Explainable Artificial Intelligence.* Ed. by Luca Longo. Communications in Computer and Information Science. Cham: Springer Nature Switzerland, 2023, pp. 573–596. DOI: 10.1007/978-3-031-44067-0_29

22. Francesco Barile, Tim Draws, Oana Inel, Alisa Rieger, Shabnam Najafian, Amir Ebrahimi Fard, Rishav Hada, and Nava Tintarev. “Evaluating Explainable Social Choice-Based Aggregation Strategies for Group Recommendation”. In: *User Modeling and User-Adapted Interaction* (June 2023). DOI: 10.1007/s11257-023-09363-0

21. Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. “Disentangling Fairness Perceptions in Algorithmic Decision-Making: The Effects of Explanations, Human Oversight, and Contestability”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* CHI ’23. New York, NY, USA: Association for Computing Machinery, 2023. DOI: 10.1145/3544548.3581161

20. Markus Bink, Sebastian Schwarz, Tim Draws, and David Elsweiler. “Investigating the Influence of Featured Snippets on User Attitudes”. In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval.* CHIIR ’23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 211–220. DOI: 10.1145/3576840.3578323

19. Tim Draws, Karthikeyan Natesan Ramamurthy, Ioana Baldini, Amit Dhurandhar, Inkit Padhi, Benjamin Timmermans, and Nava Tintarev. “Explainable Cross-Topic Stance Detection for Search Results”. In: *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval.* CHIIR ’23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 221–235. DOI: 10.1145/3576840.3578296

18. Tim Draws, Nirmal Roy, Oana Inel, Alisa Rieger, Rishav Hada, Mehmet Orcun Yalcin, Benjamin Timmermans, and Nava Tintarev. “Viewpoint Diversity in Search Results”. In: *Advances in Information Retrieval.* Ed. by Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo. Vol. 13980. Cham: Springer Nature Switzerland, 2023, pp. 279–297. DOI: 10.1007/978-3-031-28244-7_18

17. Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. “The Effects of Crowd Worker Biases in Fact-Checking Tasks”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.*

FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 2114–2124. DOI: `10.1145/3531146.3534629`

16. Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. "Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions with Debated Topics". In: *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval.* CHIIR '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 135–145. DOI: `10.1145/3498366.3505812`

15. Suzanne Hoogeveen, Alexandra Sarafoglou, Balazs Aczel, Yonathan Aditya, Alexandra J. Alayan, Peter J. Allen, Sacha Altay, Shilaan Alzahawi, Yulmaida Amir, Francis-Vincent Anthony, Obed Kwame Appiah, Quentin D. Atkinson, Adam Baimel, Merve Balkaya-Ince, Michela Balsamo, Sachin Banker, František Bartoš, Mario Becerra, Bertrand Beffara, Julia Beitner, Theiss Bendixen, Jana B. Berkessel, Renatas Berniūnas, Matthew I. Billet, Joseph Billingsley, Tiago Bortolini, Heiko Breitsohl, Amélie Bret, Faith L. Brown, Jennifer Brown, Claudia C. Brumbaugh, Jacek Buczny, Joseph Bulbulia, Saúl Caballero, Leonardo Carlucci, Cheryl L. Carmichael, Marco E. G. V. Cattaneo, Sarah J. Charles, Scott Claessens, Maxinne C. Panagopoulos, Angelo Brandelli Costa, Damien L. Crone, Stefan Czoschke, Christian Czymara, E. Damiano D'Urso, Örjan Dahlström, Anna Dalla Rosa, Henrik Danielsson, Jill De Ron, Ymkje Anna de Vries, Kristy K. Dean, Bryan J. Dik, David J. Disabato, Jaclyn K. Doherty, Tim Draws, et al. "A Many-Analysts Approach to the Relation between Religiosity and Well-Being". In: *Religion, Brain & Behavior* 0.0 (2022), pp. 1–47. DOI: `10.1080/2153599X.2022.2070255`

14. Alexandra Sarafoglou, Anna van der Heijden, Tim Draws, Joran Cornelisse, Eric-Jan Wagenmakers, and Maarten Marsman. "Combine Statistical Thinking With Open Scientific Practice: A Protocol of a Bayesian Research Project". In: *Psychology Learning & Teaching* (Feb. 2022), pp. 1–13. DOI: `10.1177/14757257221077307`

13. Caspar J. Van Lissa, Wolfgang Stroebe, Michelle R. vanDellen, N. Pontus Leander, Maximilian Agostini, Tim Draws, Andrii Grygoryshyn, Ben Gützgow, et al. "Using Machine Learning to Identify Important Predictors of COVID-19 Infection Prevention Behaviors during the Early Phase of the Pandemic". In: *Patterns* 3.4 (Apr. 2022), p. 100482. DOI: `10.1016/j.patter.2022.100482`

12. Fausto Giunchiglia, Styliani Kleanthous, Jahna Otterbacher, and Tim Draws. "Transparency Paths - Documenting the Diversity of User Perceptions". In: *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization.* UMAP '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 415–420. DOI: `10.1145/3450614.3463292`

11. Johnny van Doorn, Don van den Bergh, Udo Böhm, Fabian Dablander, Koen Derks, Tim Draws, Alexander Etz, Nathan J. Evans, Quentin F. Gronau, Julia M. Haaf, Max Hinne, Šimon Kucharský, Alexander Ly, Maarten Marsman, Dora Matzke, Akash R. Komarlu Narendra Gupta, Alexandra Sarafoglou, Angelika Stefan, Jan G. Voelkel, and Eric-Jan Wagenmakers. "The JASP Guidelines for Conducting and Reporting a Bayesian Analysis". In: *Psychonomic Bulletin & Review* 28.3 (June 2021), pp. 813–826. DOI: `10.3758/s13423-020-01798-5`

10. Francesco Barile, Shabnam Najafian, Tim Draws, Oana Inel, Alisa Rieger, Rishav Hada, and Nava Tintarev. "Toward Benchmarking Group Explanations: Evaluating the Effect of Aggregation Strategies versus Explanation". In: *Proceedings of the Perspectives on the Evaluation of Recommender Systems Workshop (PERSPECTIVES 2021)* (2021). URL: `http://ceur-ws.org/Vol-2955/paper11.pdf`

9. Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. “Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics”. In: *ACM SIGKDD Explorations Newsletter* 23.1 (May 2021), pp. 50–58. DOI: 10.1145/3468507.3468515

8. Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. “A Checklist to Combat Cognitive Biases in Crowdsourcing”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing.* Vol. 9. HCOMP ’21. 2021, pp. 48–59. DOI: 10.1609/hcomp.v9i1.18939

7. Tim Draws, Zoltán Szlávik, Benjamin Timmermans, Nava Tintarev, Kush R. Varshney, and Michael Hind. “Disparate Impact Diminishes Consumer Trust Even for Advantaged Users”. In: *Persuasive Technology.* Ed. by Raian Ali, Birgit Lugrin, and Fred Charles. Vol. 12684. Cham: Springer International Publishing, 2021, pp. 135–149. DOI: 10.1007/978-3-030-79460-6_11

6. Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. “This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* SIGIR ’21. New York, NY, USA: Association for Computing Machinery, July 2021, pp. 295–305. DOI: 10.1145/3404835.3462851

5. Tim Draws. “Understanding How Algorithmic and Cognitive Biases in Web Search Affect User Attitudes on Debated Topics”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* SIGIR ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2709. DOI: 10.1145/3404835.3463273

4. Shabnam Najafian, Tim Draws, Francesco Barile, Marko Tkalcic, Jie Yang, and Nava Tintarev. “Exploring User Concerns about Disclosing Location and Emotion Information in Group Recommendations”. In: *Proceedings of the 32nd ACM Conference on Hypertext and Social Media.* HT ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 155–164. DOI: 10.1145/3465336.3475104

3. Alisa Rieger, Tim Draws, Nava Tintarev, and Mariet Theune. “This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias”. In: *Proceedings of the 32nd ACM Conference on Hypertext and Social Media.* HT ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 189–199. DOI: 10.1145/3465336.3475101

2. Don van den Bergh, Johnny van Doorn, Maarten Marsman, Tim Draws, Erik-Jan van Kesteren, Koen Derks, Fabian Dablander, Quentin Frederik Gronau, Šimon Kucharský, Akash R. Komarlu Narendra Gupta, Alexandra Sarafoglou, Jan G. Voelkel, Angelika Stefan, Max Hinne, Dora Matzke, and Eric-Jan Wagenmakers. “A Tutorial on Conducting and Interpreting a Bayesian ANOVA in JASP”. in: *Annee Psychologique* 120.1 (Feb. 2020), pp. 73–96. DOI: 10.3917/anpsy1.201.0073

1. Tim Draws, Jody Liu, and Nava Tintarev. “Helping Users Discover Perspectives: Enhancing Opinion Mining with Joint Topic Models”. In: *2020 International Conference on Data Mining Workshops (ICDMW).* Sorrento, Italy: IEEE, Nov. 2020, pp. 23–30. DOI: 10.1109/ICDMW51313.2020.00013

SIKS Dissertation Series

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines

02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data

05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

06 Michel Wilson (TUD), Robust scheduling in an uncertain environment

07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training

08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data

09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts

10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms

11 Anne Schuth (UvA), Search Engines that Learn from Their Users

12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems

13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach

14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization

15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward

17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms

18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web

19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data

20 Daan Odijk (UvA), Context & Semantics in News & Web Search

21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground

22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems

23 Fei Cai (UvA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (TiU), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy

47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks

48 Tanja Buttler (TUD), Collecting Lessons Learned

49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis

50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

---

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation

03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store

05 Mahdieh Shadi (UvA), Collaboration Behavior

06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search

07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly

08 Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery

09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text

10 Robby van Delden (UT), (Steering) Interactive Play Behavior

11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment

12 Sander Leemans (TU/e), Robust Process Mining with Guarantees

13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology

14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

15 Peter Berck (RUN), Memory-Based Text Correction

16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines

17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution

18 Ridho Reinanda (UvA), Entity Associations for Search

19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility

21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

22 Sara Magliacane (VUA), Logics for causal inference under uncertainty

23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces

24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning

25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch

27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors

28 John Klein (VUA), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (TiU), The Power of Facial Expressions

31 Ben Ruijl (UL), Advances in computational methods for QFT calculations

32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity

34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics

35 Martine de Vos (VUA), Interpreting natural science spreadsheets

36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging

37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

38 Alex Kayal (TUD), Normative Social Applications

39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems

41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets

43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement

46 Jan Schneider (OU), Sensor-based Learning Support

47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration

48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations

02 Felix Mannhardt (TU/e), Multi-perspective Process Mining

03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction

04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process

06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems

07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems

08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems

09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks

12 Xixi Lu (TU/e), Using behavioral context in process mining

13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future

14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters

15 Naser Davarzani (UM), Biomarker discovery in heart failure

16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children

17 Jianpeng Zhang (TU/e), On Graph Sample Clustering

18 Henriette Nakad (UL), De Notaris en Private Rechtspraak

19 Minh Duc Pham (VUA), Emergent relational schemas for RDF

20 Manxia Liu (RUN), Time and Bayesian Networks

21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games

22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks

23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis

24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots

25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections

26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology

27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis

28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel

29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech

30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty

03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources

04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data

05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data

06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction

11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs

12 Jacqueline Heinerman (VUA), Better Together

13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation

14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses

15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments

16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models

17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts

18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication

19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents

20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence

21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection

22 Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture

23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification

24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing

25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description

26 Prince Singh (UT), An Integration Platform for Synchromodal Transport

27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses

28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations

29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances

30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems

31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics

32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games

33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks

34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming

36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills

37 Jian Fang (TUD), Database Acceleration on FPGAs

38 Akos Kadar (OU), Learning visually grounded and multilingual representations

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour

02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models

03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding

04 Maarten van Gompel (RUN), Context as Linguistic Bridges

05 Yulong Pei (TU/e), On local and global structure mining

06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support

07 Wim van der Vegt (OU), Towards a software architecture for reusable game components

08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search

09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research

10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining

11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation-Methods for Long-Tail Entity Recognition Models

12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment

13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming

14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases

15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games

16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling

17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences

18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems

19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems

20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations

21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be

22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar

23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging

24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots

25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining

26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization

27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context

28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality

29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference

30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst

31 Gongjin Lan (VUA), Learning better – From Baby to Better

32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising

33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation

34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development

35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

---

2021 01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space

02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models

03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices

04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning

05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems

06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot

07 Armel Lefebvre (UU), Research data management for open science

08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking

09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play

10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning

11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision

12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs

13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning

14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support

15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm

16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues

17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks

18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks

19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management

20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems

21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes

22 Sihang Qiu (TUD), Conversational Crowdsourcing

23 Hugo Manuel Proença (UL), Robust rules for prediction and description
24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
14 Michiel Overeem (UU), Evolution of Low-Code Platforms
15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
16 Pieter Gijsbers (TU/e), Systems for AutoML Research
17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification

18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation

19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation

20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media

21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments

22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations

23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents

24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values

25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty

26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences

27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences

28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems

29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality

30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays

31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management

32 Cezara Pastrav (UU), Social simulation for socio-ecological systems

33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management

34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change

35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions

02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts

03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations

04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval

05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
14 Selma Čaušević (TUD), Energy resilience through self-organization
15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
22 Alireza Shojaifar (UU), Volitional Cybersecurity
23 Theo Theunissen (UU), Documentation in Continuous Software Development
24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions

28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts

29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results

2024 01 Daphne Miedema (TU/e), On SQL Learning: Disentangling concepts in data systems education

02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems

03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis

04 Mike Huisman (UL), Understanding Deep Meta-Learning

05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair