



**RPTU**

**Black-Box Analysis of  
Algorithmic Decision-Making Systems**

Thesis approved by  
the Department of Computer Science  
University of Kaiserslautern-Landau  
for the award of the Doctoral Degree  
Doctor of Engineering (Dr.-Ing.)

to

Tobias Krafft

Date of Defense: 09.10.2024  
Dean: Prof. Dr. Christoph Garth  
Reviewer: Prof. Dr.-Ing. Jörg Dörr  
Reviewer: Prof. Dr.-Ing. Florian Gallwitz

DE-386

“In our daily lives we are confronted at every turn with systems whose internal mechanisms are not fully open to inspection, and which must be treated by the methods appropriate to the Black Box.” (Ashby, [1956](#))

---

## Acknowledgment

I would like to express my heartfelt gratitude to my family, Silke, Bodo, Roman, and Shari, for their unwavering support throughout my PhD journey. Your encouragement has meant the world to me.

Special thanks go to my supervisor, Katharina Zweig, for her invaluable advice and assistance at every stage of my scientific growth. Your guidance has shaped my research and my development as a scholar.

I am grateful to the Algorithm Accountability Lab, especially Marc Hauer, for their support and contributions to my work. Your collaboration has been instrumental in the success of my research.

I would like to acknowledge the critical friends who have provided valuable insights and feedback throughout my journey. Anita Klingel and Philipp Bird, thank you for your constructive criticism and thought-provoking discussions. Your perspectives have played an essential role in shaping my research and enhancing its quality.

I would also like to extend my thanks to the experts involved in the various studies and research projects I have been a part of. A heartfelt thank you to Pascal König, Michael Gamer and Anna Couturier for their passionate participation and valuable input. Your contributions have been vital to the successful completion of our research.

Once again, I am truly grateful to all those mentioned above for their unwavering support, guidance, and collaboration. Without you, this journey would not have been possible. Thank you.

---

## Abstract

The advancing digitalization has led to the widespread use of algorithmic decision-making (ADM) systems in various sectors of our lives. While these systems have proven beneficial to optimizing workflows and decision-making processes, their use in assessing and classifying individuals can have significant impacts on the respective lives. Public stakeholders studying these adverse effects often face the challenge of examining ubiquitous, but opaque ADM systems from the outside. In critical sectors, like healthcare, consumer protection, or safety, the impact of algorithmically governed platforms is technologically complex to address within traditional research modalities, which poses a major obstacle to oversight. Therefore, (1) barely any analyses can be performed by researchers operating outside the field of computer science, because of the necessary technical knowledge. (2) It is also unclear whether an analysis of opaque ADM systems without knowledge of internal functions is even possible and sufficient, in order to investigate the cause of a negative outcome that raised suspicions. So how do we bridge the gaps of knowledge between those experts trying to analyze the potential dangers of ADM systems and those computer science researchers accustomed to employing methodologically grounded tools like black-box analyses? How can we empower non-computer science stakeholders like patient and consumer advocates, charities/NGOs, regulators and centres of social science and biomedical research to undertake more complex investigations into the ADM systems that shape their respective fields? Which information and access is required to be able to carry out appropriate analyses at all? This dissertation focuses on targeting those questions by presenting a process model for investigating ADM systems as a black box, building on traditional black-box analysis approaches. The model is both accessible and adaptable for non-computer science domain experts and has been developed, tested, and discussed in multiple studies, each based on a different use case in different disciplines. Depending on the possible access to an ADM system, the presented process model can be used to investigate black-box systems with regard to a variety of questions. Additionally, it shows regulating authorities indirectly how costly external analyses are without direct access provided by the platform operators. The model developed in this dissertation empowers expert actors within diverse fields of civil society to find answers to questions with far-reaching social and societal consequences, like whether Google's search results aim for political manipulation, or whether medical advertisements try to exploit the susceptible part of the population. The limitations of the process model are presented and discussed from social, technical and legal perspectives. Understanding and addressing these limitations is essential for conducting effective and reliable black-box analyses. Furthermore, the studies show, that there are limitations that cannot be solved by methods of black-box analysis alone and therefore the political implications of this research are significant, particularly in the context of increasing interest and regulatory efforts in enhancing the transparency and accountability of algorithmic systems. The process model aligns with these initiatives and provides concrete guidelines for promoting black-box analyses. Additionally, the recommendations for further research and political action highlight the need for strengthened rights of data subjects, establishment of suitable interfaces for investigation, legal certainty for black-box analyses, and a watchdog approach for continuous monitoring and evaluation of ADM systems. As society continues to grapple with the challenges and opportunities presented by ADM systems, the insights and methodologies presented in this dissertation contribute to a more comprehensive and critical understanding of these systems. By fostering interdisciplinary collaboration and promoting a distributed approach to analyzing ADM systems, this research aims to shape a society that can navigate the complexities of algorithmic decision-making while safeguarding fundamental rights and values.

# Contents

<b>Abstract</b>	<b>i</b>
<b>1. Introduction</b>	<b>2</b>
1.1. Research Questions . . . . .	7
1.2. Outline of the thesis . . . . .	7
<b>2. Definitions and related work</b>	<b>13</b>
2.1. Algorithmic decision-making systems . . . . .	13
2.2. Socioinformatics, emergence, and socioinformatics systems . . . . .	19
2.3. The black-box analysis . . . . .	21
2.4. Audit . . . . .	25
2.4.1. Original use of the term 'audit' in social science . . . . .	26
2.4.2. Use of the term 'audit' in the evaluation of software . . . . .	28
2.4.3. Algorithmic audit . . . . .	30
<b>3. Societal demands for transparency and verifiability of ADM systems</b>	<b>34</b>
3.1. Differentiated regulatory efforts of ADM systems . . . . .	36
3.2. Risk-based approach to assessing the damage potential of an ADM system in a socioinformatic system . . . . .	40
<b>4. Phenomenon-induced socioinformatic analysis</b>	<b>43</b>
4.1. Structure of a phenomenon-induced socioinformatic analysis . . . . .	44
4.1.1. Phase 1: Description of the technical foundation . . . . .	45
4.1.2. Phase 2: Identification of the relevant social agents and their mo- tivations . . . . .	45
4.1.3. Phase 3: Creating the effect structure . . . . .	46
4.1.4. Phase 4: Analysis of the effect structure . . . . .	48
4.1.5. Phase 5: Identification of possible countermeasures . . . . .	49
4.2. Phenomenon-induced socioinformatic analysis of the filter bubble theory on Google . . . . .	49
<b>5. Black-box analysis – Filter bubble on Google</b>	<b>62</b>
5.1. Conception of the black-box analysis . . . . .	64

5.2. Operationalization of personalization . . . . .	73
5.3. Data collection . . . . .	75
5.4. Examination of the results . . . . .	76
5.4.1. Data cleaning and preprocessing . . . . .	76
5.4.2. Data evaluation . . . . .	80
5.4.3. Threats to the validity of the results . . . . .	88
5.5. Interpretation of the results . . . . .	89
5.6. Lessons Learned . . . . .	91
5.6.1. General Learnings . . . . .	91
5.6.2. Technical Learnings . . . . .	91
5.7. Further use of the collected data . . . . .	92
<b>6. Black-box analysis – Stem cell advertising ban on Google</b>	<b>94</b>
6.1. On the problem of stem cell therapies . . . . .	95
6.2. Socioinformatic analysis . . . . .	100
6.3. Conception of the black-box analysis . . . . .	103
6.4. Data collection . . . . .	113
6.5. Examination of the results . . . . .	115
6.5.1. Captured Ads Analysis . . . . .	115
6.5.2. Roll-out behavior of the web displays . . . . .	117
6.6. Threats to the validity of the results . . . . .	118
6.7. Interpretation of the results . . . . .	120
6.8. Lessons Learned . . . . .	122
6.8.1. General Learnings . . . . .	122
6.8.2. Technical Learnings . . . . .	123
<b>7. Black-box analysis – Price differentiation in online retail</b>	<b>125</b>
7.1. Dynamic and personalized prices in online retailing . . . . .	128
7.1.1. Price differentiation . . . . .	128
7.1.2. Consumer reactions to personalized pricing . . . . .	132
7.1.3. Profiling . . . . .	134
7.2. Is it possible to capture personalized prices with a black-box analysis? . . . . .	135
7.2.1. Measuring price differentiation and personalization . . . . .	142
7.2.2. Challenges and state of the art in monitoring prices in online retailing	143
7.3. Findings of our black-box analysis of price differentiation in online retailing	145
7.3.1. Preliminary study with EU-Preis . . . . .	148
7.3.2. Study with EU-Preis . . . . .	149
7.3.3. Interpretation of the results . . . . .	149
7.3.4. Pre-training profiles with bots . . . . .	149
7.3.5. Lessons Learned . . . . .	150
7.3.6. General Learnings . . . . .	151

7.3.7. Technical Learnings . . . . .	151
<b>8. Process model</b>	<b>153</b>
8.1. From suspicion to falsifiable statement . . . . .	155
8.1.1. Find a suspicion . . . . .	155
8.1.2. Verify the suspicion . . . . .	156
8.1.3. Formulate a testable hypothesis . . . . .	157
8.2. Design Decisions . . . . .	159
8.2.1. Identify black-box scenario . . . . .	159
8.2.2. Design of the analysis method . . . . .	162
8.2.3. Determine access to the system . . . . .	170
8.3. Concept of the Black-Box Analysis . . . . .	178
8.3.1. Develop and document the study design . . . . .	178
8.3.2. Threats to validity I . . . . .	181
8.4. Preliminary study . . . . .	186
8.5. Data collection . . . . .	188
8.6. Analysis of the results . . . . .	190
<b>9. Limits, political conclusions and outlook</b>	<b>193</b>
9.1. Insights gained from & limits of the three studies . . . . .	193
9.2. General limitations of black-box analysis . . . . .	196
9.3. Further research . . . . .	198
9.4. Conclusion, value added and political implications . . . . .	199
<b>Bibliography</b>	<b>203</b>
<b>A. Curriculum Vitae Tobias Krafft, M.Sc.</b>	<b>234</b>
<b>B. Own Publications</b>	<b>238</b>

Algorithmic decision-making solves increasingly complex cognitive problems, partly due to advances in our capacity to process data. Modern society employs a growing array of such algorithmic decision-making systems (ADM systems for short) for many different tasks. The possible applications are diverse and range from decisions regarding the placement of online advertisements (Datta et al., 2015; Sweeney, 2013) to credit scoring (West, 2000) to the recidivism risk assessment of criminals (Angwin et al., 2016) to the classification of the unemployed (Niklas et al., 2015).

The selected examples illustrate the consequences these decisions can have for the lives of the respective individuals: ADM systems serve to simplify people's daily lives, but they also have a wide-ranging impact on society, especially as they take on more and more tasks. Although avoiding misjudgments and unintended side effects is an expected goal in the development of ADM systems, it is evident that these forms of errors are very diverse and therefore difficult to predict or avoid. For example, risk categories like functional, physical, social, and financial are known from consumer behavior research (Schiffman et al., 2012, p. 186), while other risks are of a legal nature and concern possible violations of intellectual property rights, privacy (Saurwein et al., 2015, p. 37), or discrimination (Sweeney, 2013, pp. 674-675; Barocas & Selbst, 2016).

People in a wide range of social and societal roles can be affected by ADM: be it bank customers denied credit (West, 2000), individuals convicted of crime prevented from being released on bail by an ADM system (Angwin et al., 2016), or welfare recipients falsely accused of benefiting from an overpayment by the state (Braithwaite, 2020).

Numerous journalistic and scientific studies on problems with ADM systems demonstrate that current control mechanisms are insufficient to prevent unfavorable outcomes for those affected by the outputs of the ADM systems (see, e.g. Angwin et al., 2016; Braithwaite, 2020). Errors in the system are only recognized later, and often too late, especially since necessary review and adjustment processes also tend to take a very long time.

In light of these potential risks, the political debate about the necessity for so-called algorithm audits has taken up significant speed: Both individual actors such as O'Neil, 2016 and political groups such as the German Data Ethics Commission (DEK) (Data Ethics Commission, 2019) demand the examination of ADM systems with respect to



---

a variety of risks in external, standardized ADM system audits. Both the DEK (Data Ethics Commission, 2019) and the German Parliament’s enquete commission “Künstliche Intelligenz - Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale” (Deutscher Bundestag, 2020) demand the opportunity for stakeholders prioritizing the common good to subject automated decision-making (ADM) systems operating in crucial domains such as medicine, healthcare, consumer protection, and security to an external audit. The primary aim is to ascertain the potential risks inherent in these systems.

On the European level, the General Data Protection Regulation (GDPR, European Parliament & Council of the European Union, 2016) is frequently referenced. However, a study conducted by the European Parliamentary Research Service suggests that GDPR alone may not be sufficient in guaranteeing the accountability of ADM systems (Koene et al., 2019).

While the latest EU regulatory efforts such as the AI Act (European Parliament & European Council, 2021) call for comprehensive transparency for ADM systems in critical application areas, the majority of AI systems that do not involve major risks will not be controlled. The scope of transparency requirements will be limited, as legal regulatory demands always require proportionality (see, for example European Parliament & European Council, 2021, Section 2.3). Although it is evident that systems with significant implications, such as the risk of misdiagnosing tumors in medical images, can be subjected to transparency requirements due to the severe consequences of incorrect evaluations, other ADM systems like Google Search might not require the same level of transparency. While there are scenarios where incorrect decisions may lead to individual or societal harm, such cases are infrequent. Therefore, it could be argued that imposing transparency requirements on such systems may not be proportionate.

To deal with the requirement of proportionality, the European Union has presented a preliminary categorization of the application areas where AI systems pose a “high risk” (European Parliament & European Council, 2021, Section 5.2.3.). However, while the AI Act establishes formal requirements for each class, it leaves room for interpretation and implementation. It does not provide explicit guidance on the specific methodologies and procedures that should be followed when auditing an AI system. This creates a gap in the legislation, as there is a lack of detailed instructions on how organizations should conduct comprehensive audits of AI systems to assess their compliance with the law and to ensure transparency and accountability.

Transparency is at the center of this debate: Due to the possibility of highly sensitive decisions being made in certain areas, which can harm people’s health, safety, and fundamental rights (Orwat et al., 2022, pp. 258-260)(Ben-Israel et al., 2020; HLEG on AI, 2019; Zuiderveen Borgesius, 2018), it is essential to make the design process of such ADM systems and the interpretation of their results as comprehensible as possible. Demands for transparency can go as far as requiring full disclosure of the software code (Leite & Cappelli, 2010; Portugal et al., 2017). However, this is countered by

## 1. Introduction

---

data protection concerns as well as fears on the part of operators that business secrets would not be kept if their systems were entirely transparent. Furthermore, there are also general problems that can arise if internal processes are published. For instance, the disclosure of Google’s “PageRank” algorithm (Page et al., 1998) demonstrated that complete transparency in socio-technical systems may have unintended consequences. After the exact sorting method used in the Google search engine was published by the developers (Page et al., 1998), both individuals and companies tried to influence it. By gaining access to the exact procedure for evaluating individual pages, people quickly determined how the ranking could be changed in their own favor. Link farms were used, among other methods, to trick the algorithm into believing that the desired page was more popular (Grimmelmann, 2008; Zweig et al., 2021, p. 139). However, scholars have questioned whether transparency provides sufficient insight for evaluating an ADM system, since even with full transparency, a complex system will remain opaque (Ananny & Crawford, 2018). Such a system is referred to as a black box (Ananny & Crawford, 2018).

Complete transparency on socially significant topics can therefore be more of a problem than a solution. In addition, the sheer size and complexity of the codes of today’s systems exceed what can be captured: In France, for instance, software was released that computed the 2014 tax payments of 37 million tax households. This code contained over 17,000 variables and approximately 1,000 functions<sup>1</sup>. Therefore, it must be stated that there is no universal answer to the issue of transparency of ADM systems. Rather, it must be handled with sensitivity in relation to each individual case, keeping in mind that the disclosure of an algorithm will not necessarily result in deeper insight or comprehension.

The second concept frequently discussed with regard to algorithmic audits is that of accountability. As per M. Bovens, 2007, accountability is understood as a relationship wherein an actor is obligated to explain and justify their conduct to a forum, which in turn has the right to ask questions, judge, and potentially impose consequences. However, in the context of algorithmic accountability (Wieringa, 2020), the focus shifts from the actor’s behavior to the behavior of the algorithmic system they employ, requiring justification and explanation. The interplay between the actor and the forum is then defined by the transparency, dialog, and evaluation of the explanation, culminating in the consequences decided upon by the forum.

In order to tackle both the complexity and the potential risks of these issues regarding ADM systems, a large and rapidly expanding scientific community from a variety of disciplines has formed to address the challenges of algorithmic accountability. Contributions to algorithm ethics (Ananny & Crawford, 2018; Binns, 2018; Hauer et al., 2023; Lanzing, 2019; Mittelstadt et al., 2016) and to the fields of law and regulation (Brauneis & Good-

---

<sup>1</sup>Direction Générale des Finances, Le blog d’Etalab, Hackathon “#CodeImpôt”, <https://www.etalab.gouv.fr/codeimpot-un-hackathon-autour-de-louverture-du-code-source-du-calculateur-impots>, Accessed 26.07.2021, Results: <https://github.com/etalab/calculateur-impots-m-source-code>.

---

man, 2018; Hildebrandt, 2016; Koops, 2013; Yeung, 2017a) identify fundamental ethical principles and procedures necessary to ensure that decisions made by ADM systems do not harm or violate the rights of those affected. Similar research exists in criminology (e.g., Wormith, 2017) and science and technology studies, respectively (e.g., Ananny & Crawford, 2018). Furthermore, more technically oriented work points to a variety of approaches for making ADM systems transparent and subject to scrutiny (Bryson & Theodorou, 2019; Diakopoulos, 2014a, 2014b; Guidotti et al., 2018; Kroll et al., 2017; Lepri et al., 2018; Sokol & Flach, 2020; Wieringa, 2020).

Here, the use of algorithm-driven platforms has brought many new challenges that can only be addressed to a limited extent within the framework of traditional research modalities (Koshiyama et al., 2021), since no standardized possibilities for investigating ADM systems are currently available.

Therefore, it is imperative to develop and implement methods for monitoring such ADM systems in areas deemed critical by the European Parliament’s AI Act (European Parliament & European Council, 2021, p. 5.2.3). In addition, it would be necessary to consider whether other areas could or should be deemed critical if people are directly impacted by decisions made by an ADM system (T. D. Krafft et al., 2022). In this line of thought, the U.S. Government’s “AI Bill of Rights” mandates a thorough examination of systems that interfere with civil rights and civil liberties (Lander & Nelson, 2011).

However, despite the widely recognized need for algorithmic audits, there is, to date, no widely recognized systematic procedure to perform such an audit.

In 2014, Nicholas Diakopoulos developed a method of examining algorithmic decision-making systems as a black box that has no knowledge of the precise operation of the integrated components. He applied techniques from cybernetics research and reverse engineering<sup>2</sup> to ADM systems (Diakopoulos, 2014a, 2014b) and proposed verifying quality statements about systems into which the general public has no or very limited insight. Here, the ADM system is regarded as a black box because one can only infer properties of the underlying implementation based on the relationship between input and output. There are ADM systems in which not even the training data or the type of instance data used, i.e., the algorithm’s input, is disclosed for reasons of privacy or trade secrecy. This is referred to by Diakopoulos as “varying degrees of observability” (Diakopoulos, 2014a, p. 22). Depending on the scope of the systems and the social issues involved, this level of observability can pose significant challenges for the black-box approach, as the evaluation of an ADM system’s output can be based on a variety of quality measures.

Essentially, it must be stated that many ADM systems are opaque “black boxes” that are not visible to society, making it difficult, if not impossible, to comprehend how they function. Jenna Burrell, 2016 distinguishes three forms of opacity: “(1) opacity

---

<sup>2</sup>Reverse engineering is the process of deconstructing and analyzing a product or system to understand its design, functionality, and underlying principles. In software testing, this method is used to test the correct operation of software components) (Eilam, 2011)

as intentional corporate or state secrecy; (2) opacity as technical illiteracy; and (3) opacity that arises from the characteristics of machine learning algorithms and the scale required to apply them usefully”. Opacity as intentional corporate or state secrecy (1) refers to the deliberate choice of corporations to keep their algorithms opaque in order to protect trade secrets and gain a competitive advantage. However, this opacity can also serve as a means to conceal regulatory violations, consumer manipulation, or discriminatory practices. To address this, the solution proposed by Burrell is to make the code available for scrutiny through regulatory measures or independent audits to ensure transparency and accountability. Opacity as technical illiteracy (2) refers to the lack of understanding and proficiency in writing and comprehending code and algorithms, which is currently limited to a specialized skillset inaccessible to the majority of people. This form of opacity arises due to the unique rules, precision, and formality required in coding languages, which differ from human languages. Increasing diversity in STEM fields and promoting computational thinking at all levels of education are crucial in addressing this opacity and enable the public to have the knowledge and skills necessary to evaluate and critique the mechanisms impacting their lives. Opacity as the way algorithms operate at the scale of application (3) refers to the challenges posed by the complexity and scale of machine learning algorithms. These algorithms, such as those used in ADM systems, involve numerous interlinked components, vast amounts of data, and inherent complexity, which make it difficult not only to read and comprehend the code but also to understand how the algorithm operates on the data. These challenges arise from factors like the “curse of dimensionality” and the need to handle diverse data properties, which can contribute to the opacity of the algorithm’s logic and operation.

Because of this lack of transparency, researchers outside the field of computer science frequently exclude studies on the impact of ADM systems from their analyses due to the technological complexity of such black boxes. Safiya Umoja Nobles, among others, outlines extensive findings on ADM systems in her book “Algorithms of Oppression” (Noble, 2018). However, an examination of the methodology reveals that the research is limited to normative user testing conducted by the author herself. This demonstrates that social science research is often methodologically constrained in its ability to examine the technical side of algorithmic black boxes; consequently, the lack of insight and access severely restricts efforts to expand control mechanisms by non-governmental organizations and/or public oversight. If social scientists are unable to comprehend the actual functioning and environment of these ADM systems, their ability to thoroughly investigate and analyze them is hindered. Consequently, the significance and effectiveness of social science research are diminished.

This highlights the urgent need for methods that are not only specifically designed for such situations, but can also be applied by scholars outside the field of computer science.

The literature provides numerous theoretical concepts and practical examples of black-box analyses and their implementation (for a comprehensive literature review, see Bandy, 2021). What is lacking, however, is a systematic approach to bridging the knowledge and

skill gaps between domain experts faced with the impact of ADM systems on societal challenges and computer scientists accustomed to interacting with technical systems without knowledge about the inside of the system.

Recognizing this gap, the work undertaken here focuses on addressing this precise issue. It aims to bridge the divide between the general requirements outlined in the AI Act and the practical steps necessary to perform an effective black-box analysis of ADM systems for researchers from different fields.

## 1.1. Research Questions

This dissertation presents a process model that is as generic as possible, allowing experts from a variety of fields, including those outside the field of computer science, to examine ADM systems as black boxes (see section 8). The goal of this work is to enable stakeholders such as patient and consumer advocates, charities/NGOs, regulators, social science and biomedical research centers, and so on - in short, non-computer scientists - to conduct complex investigations of ADM systems that pose challenges in their respective domains.

### Research questions

1. What kinds of questions can a black-box analysis address?
2. How is the structure of a black-box analysis of an ADM system determined?
3. Which accesses to an ADM system can be used to perform which forms of black-box analysis?
4. What are the methodological limitations of black-box analysis?

## 1.2. Outline of the thesis

In order to answer these questions, the next section introduces definitions and lays the foundations for this dissertation (Section 2). First, it will explained how algorithmic decision-making works and on which basis ADM systems are created. Following this conceptual foundation, we will introduce the study field of “socioinformatics”, which examines the interactions between technology and society. Concepts such as emergence and socioinformatic systems will be introduced to show how technology and society are interconnected and influence each other. Subsequently, the process of analyzing systems as black boxes will be elaborated and applied specifically to ADM systems. This form of analysis is particularly useful, as ADM systems are frequently opaque in their operations and can thus be regarded as black boxes. Scrutinizing ADM systems has antecedents

in audits conducted in social science since the 1940s as a research method for human decision-making processes (Gaddis, 2018, p. 9). At that time, researchers in social science faced the challenge of locating evidence for undesirable or forbidden behavior using only observational data (Gaddis, 2018, p. 3). As a result, a method of investigation was developed in which the behavior of individuals or institutions is examined to determine the presence of such behavioral or decision-making patterns. For this purpose, the process was either tested and evaluated on-site (so-called “field experiment”) (Gaddis, 2018, p. 3 et seqq.), or an attempt was made to conduct the audit via a communication channel such as letter, fax, and later also email (so-called “correspondence audit”). However, because these audits were quite expensive, they were typically conducted on a small scale. Even so, these more than eighty years of experience (Gaddis, 2018; Vecchione et al., 2021) can certainly still be applied in part to the black-box investigations of ADM systems. The section concludes with a classification of black-box analysis in the field of algorithmic audit and introduces, for this purpose, the necessary areas of the term “audit” in software development and the term “algorithm audit” according to Sandvig et al., 2014. Section 3 explains the demands resulting from the increasing use of ADM systems in areas with high societal impact and discusses excerpts from the development process of the regulatory framework for ADM systems. As it is of immense importance to society that the design process of systems with high societal impact and the interpretation of the results are comprehensible, a high degree of transparency is often demanded, even to the point of full disclosure of the code. Since 2013, a so-called “Algorithm TÜV” has been frequently demanded by society (Mayer-Schönberger & Cukier, 2013), which is asking for the technical component of an ADM system to be demonstrated to a verifying authority and to be certified by it. However, this approach has a significant flaw, since the same technical component of a recommendation system, for example, can be used both for suggesting products in online commerce and for disseminating political advertising based on existing preferences, which is associated with a different risk. Therefore, in early 2019, T. D. Krafft and K. A. Zweig published a risk-based regulatory proposal for ADM systems (T. D. Krafft & Zweig, 2019), which was further developed together with the political scientist Pascal König (T. D. Krafft et al., 2022). In this proposal, the overall damage potential of an ADM system is determined to be dependent on the embedding of the technical component into the overall socioinformatic system. In this context, we consider both the possible consequences of an erroneous judgment and the degree of exposure of the affected person in terms of whether it is possible to challenge a decision. Depending on the level of the overall damage potential determined in this way, we propose correspondingly increasing the transparency and examinability requirements that the technical component should fulfill in this area of application. Politicians on

both the national<sup>3</sup> and international<sup>4</sup> levels have embraced this proposal.

It is evident that current regulatory efforts seek to strike a balance between the interests of society (e.g., transparency, comprehension, education) and the interests of ADM system operators (e.g., trade secrecy, ...). Orwat et al., 2022 delve into the legal aspects of risk-based regulatory mechanisms for ADM systems. In their study, they point to the debate on whether the precautionary principle should be applied in this specific context. This principle fundamentally argues for proactive, protective regulatory measures when there remains scientific ambiguity regarding the potential harm associated with specific ADM systems. They assert that, under the precautionary principle, regulatory intercessions are justifiable and necessary until a comprehensive scientific understanding of the damage potential of these systems is conclusively established. This negotiation process will continue for a long time and will not result in widespread ADM system transparency. Therefore, it is necessary to address the fact that, for many ADM systems, society has little or no insight into their internal processes and decision-making rules. Nicholas Diakopoulos applied black-box analysis techniques derived from the field of cybernetics to further support current methods of assessing and analyzing ADM systems (Diakopoulos, 2014a). In this case, the ADM system is opaque, and the assessing entity can only infer the properties of the underlying implementation from the input-output relationship. There are ADM systems in which not even the training data or the type of instance data used, i.e., the input to the system, is disclosed due to privacy or trade secret concerns. Depending on the size of ADM systems and the social factors involved, the level of opacity can pose significant challenges for the black-box approach. The output of an ADM system can be evaluated using a variety of different quality metrics. The difficulty lies in determining what kinds of questions can be answered with a black-box approach, what variations of this approach exist, and what kinds of information and access to the ADM system are necessary for its implementation.

The discourse at the time of the proposal made by T. D. Krafft & Zweig, 2019 on the regulation and review of ADM systems lacked an appropriate method for analyzing whether, for example, an ADM system may actually lead to unfair discrimination. Without a well-founded justification, however, it is difficult to demand a cost-intensive review by means of a black box analysis. In this case, the phenomenon-induced socioinformatic analysis developed jointly with Katharina Anna Zweig, Anita Klingel, and Enno Park has proven to be a qualitative method for identifying the effects of technical components on specific socioinformatic phenomena. It is possible to use this method to determine whether the ADM system in question may actually be causally responsible for the alleged

---

<sup>3</sup>German Standardization Roadmap AI of DIN & DKE, 2020, Data Ethics Commission Data Ethics Commission, 2019, German Parliament's enquete commission "Künstliche Intelligenz - Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale" (Deutscher Bundestag, 2020)

<sup>4</sup>AI White Paper of the European Commission, 2020, the draft of the EU AI Act of the European Parliament & European Council, 2021.

phenomenon. Section 4 presents this method of analysis.

Sections 5 to 7 cover black-box analyses for various questions and present use cases. In section 5, the first black-box analysis is presented, which is dedicated to the filter bubble theory of Eli Pariser. According to this theory, personalized algorithms in social media tend to show individuals content that corresponds to their previous interests, allowing for the formation of distinct information spheres. Individual filtering of the information flow may result in groups or individuals being informed of different facts and thus may lead to them living in their own information universe.

Section 6 examines our black-box analysis of Google ads for unproven stem cell treatments. The starting point was the lack of oversight and the difficulty to detect errors in individual results with regard to personalization by ADM systems, because in the field of search engine ad distribution, the lack of transparency regarding who sees what ads can become problematic. The need became apparent given the discrepancy between the small number of scientifically proven stem cell therapies and the alarming increase in the number of direct-to-consumer marketing of therapies worldwide that offer such treatments for various diseases. Google issued a statement on 1 October 2019 prohibiting the advertising of stem cell and gene therapy treatments with questionable and unproven effects in order to protect its users from the proliferation of advertisements for unproven medical interventions. People with serious diseases such as Parkinson’s Disease or Multiple Sclerosis have reported seeing advertisements for this type of therapy, despite Google’s ban on this type of advertising. However, there was no procedure to collect and verify these anecdotal reports through a structured approach. Our project’s objective was to assess the enforcement and efficacy of this ban and monitor the impact of Google’s ADM-based advertising on end-user outcomes. To this end, the study investigated the potential risk posed to vulnerable patient groups when they search Google for health-related information. Using a black-box analysis with a browser plugin, we were able to identify the continued presence of banned and problematic advertisements in stem cell-related searches in the months following Google’s ban (Reber et al., 2020). We chose a combination of scraping audit and crowdsourced audit as our research method due to the smaller patient population relevance. The question of unequal treatment with regard to a parameter is a common one when investigating ADM systems, and Section 7 is devoted to addressing this issue. Typically, such an investigative question is prompted by an allegation of discrimination. Due to the fact that the outcome of an ADM system may be influenced by numerous, largely unidentified parameters, the investigation of such unequal treatment is complex. In online retail, algorithmic pricing may employ parameters that are not socially acceptable and should not affect pricing. In general, store owners can personalize prices for products to specific customers or consumer groups. This can also be accomplished through automated decision-making or consumer behavior profiling. Nonetheless, because such pricing could be perceived as a risk in purchasing decisions, the EU Commission requires vendors to inform consumers when the prices of goods or services have been personalized using these methods (European



Commission & Council of the European Union, 2019, paragraph 45).

Algorithmic pricing can result in unequal treatment of individuals with protected characteristics if these characteristics are directly or indirectly incorporated into the evaluation of purchasing behavior and personalized prices are offered based on this profile. To establish unjustified or even punishable discrimination, society must demonstrate that a particular population group is statistically and systemically disadvantaged. However, it is difficult to prove the presence of personalized prices, particularly in online retail, because individual consumers cannot compare prices by their very nature. Such pricing strategies could only be demonstrated by comparing the prices displayed to various customers in real time. In order to investigate this question, a black-box analysis consisting of a crowdsourced and a sock-puppet audit of dynamic/personalized pricing is presented. In order to test an ADM system for pricing with regard to the unequal treatment of a group of people, we made an attempt to determine how the system responds when a product is searched twice with only one characteristic, such as gender, differing between the two queries.

Setting up the so-called bot accounts required for the analysis was challenging. These bot accounts had to appear as natural as possible for the shops under investigation so that the platform would treat them as normal customers. To accomplish this, the profiles of the simulated customers and a method for these bots to interact with the shops in a representative manner had to be developed. The research using this study design revealed the following issues: First, it is impossible to guess all the input variables used due to a lack of fundamental knowledge about the structure of the pricing algorithms used by shop operators and the sometimes massive differences between online shops. Pricing can be based on a large number of inputs, which makes the creation of profiles incredibly time-consuming and bot-based verification nearly impossible. In addition, the unclear input variables for a pricing algorithm caused design issues with the investigated profiles. It was unclear which user behaviors should be simulated when training bot accounts on the website. Consequently, a crowdsourced approach would likely be required. In this particular instance, the difficulty in interpreting the results was due to the fact that a black-box analysis cannot differentiate between personalized and dynamic pricing: Any personalization discovered, even if it indicates unequal treatment based on a protected characteristic, can be easily rationalized away by the retailer by introducing an unknown variable or context to the investigating authority. Alternatively, black-box analysis cannot definitively refute the claim that there is no personalization at all, but only a response to dynamic stock changes. Without more knowledge of the algorithm's inner workings, this is hardly feasible. The issue is triggered among other things, by products with limited supply, such as airplane seats. When a plane is nearly full, the remaining seats can be offered at a premium price in order to avoid having empty seats. In addition, as part of this research, we investigated the behavior of news article distribution on Facebook (T. D. Krafft et al., 2020) and, in collaboration with an undergraduate student, YouTube's automated suggestion algorithm (Schütte, 2019).

The results of these two projects were incorporated into the development of the process model, but they are not specifically addressed here.

Section 8 provides readers with a comprehensive overview of a process model for black-box analyses of ADM systems. The section outlines the necessary steps that must be taken in order to successfully conceive, plan, and execute such an analysis. By following the process model presented in this section, researchers and practitioners can better understand the inner workings of ADM systems and work to uncover any potential biases or errors that may be present. In addition to providing guiding questions for conducting a black-box analysis, section 8 also provides readers with a summary of the experiences we gained from applying such analyses in various contexts of our research. By sharing these experiences, we help readers understand the potential benefits and challenges of conducting black-box analyses, and how they can be effectively utilized to improve the reliability and fairness of ADM systems. The scope of these analyses is defined in greater detail, and typical errors and problem types are identified based on our experience (T. D. Krafft et al., 2019, 2020, 2021, 2023; Reber et al., 2020).

The final section 9 provides a summary of the dissertation's key findings and contributions, as well as limitations of the process model developed for black-box analyses of ADM systems. The section discusses the significance of black-box analyses in uncovering allegations against ADM systems and their operators, involving stakeholders such as consumer advocates, charities/NGOs, regulators, social science and biomedical research centers, and patients.

The implications for policy and practice are also discussed, emphasizing the need for enhanced verifiability, transparency, and accountability of algorithmic systems. The process model and the recommendations derived from the research can guide policymakers and stakeholders in addressing these goals. Recommendations include strengthening data subjects' rights and their enforcement, establishing suitable interfaces for analysis, allowing conditional use of bots for monitoring, ensuring legal certainty for black-box analyses, and conducting large-scale audits of ADM systems.

Overall, this dissertation highlights the importance of black-box analyses in addressing the challenges of ADM systems and emphasizes the need for interdisciplinary collaboration, regulations, and research to ensure accountability and protect consumer rights.

This section presents definitions essential to comprehending the key terms and concepts used throughout this work. One of the significant focal areas is algorithmic decision-making, which involves the use of computer algorithms to analyze data and make decisions or predictions. The resulting decisions can have a profound impact on society and can affect individuals and communities in various ways.

Additionally, this section explores the interdisciplinary field of socioinformatics, which examines the relationship between information technology and social systems. Understanding this relationship is crucial, as technology can significantly shape social interactions and relationships.

The term “black box” is also discussed in detail in this section. It is used to describe an algorithm or system that lacks transparency and whose inner workings are not fully understood by the user or others outside the system. This can make it challenging to identify and address potential biases or errors in the decision-making process, leading to negative consequences.

Finally, the concept of audit is examined, which refers to a systematic review or examination of a process or system to ensure that it is functioning effectively and efficiently. In the context of algorithmic decision-making, an audit can help identify potential biases or errors in the system and provide recommendations for improving its performance or addressing any negative impacts.

## 2.1. Algorithmic decision-making systems

The concept of deriving rules for future decisions with the aid of algorithms emerged with the earliest computers. In the context of decision-making systems, the impact of increasing digitalization on society was demonstrated as early as 1981 (Bonczek et al., 1981). Since the early 1980s, efforts have been made to process collected data in order to make the best decisions possible (Mertens et al., 1988, p. 17). It was predicted that in the information age, computers would be able to assist with a variety of decision-making processes. The problems caused by the volume, variety, and frequency of data

## 2. Definitions and related work

---

are discussed under the term “Big Data”<sup>1</sup>. The hope was to be able to use and evaluate the data generated at an ever-increasing rate for decision-making, so in addition to storage and sorting solutions, procedures were also developed to aid people in such endeavors. Computers should not only be capable of storing data, but also of processing and displaying it. This task is carried out by so-called algorithmic decision-making systems, the definitions of which vary. This dissertation employs the following definition:

### Definition 2.1: Algorithmic decision-making system (ADM System)

Algorithm decision-making systems (ADM systems) contain an algorithmic component that makes a decision based on the input. If the algorithm was created by experts, the system is referred to as an expert system. There are also those that independently derive the rule system from data using machine learning (T. D. Krafft & Zweig, 2018)<sup>a</sup>

<sup>a</sup>Translation by the author; original German text: “Algorithmische Entscheidungssysteme (Algorithm Decision Making Systems, kurz ADM-Systeme) enthalten eine algorithmische Komponente, die basierend auf der Eingabe eine Entscheidung trifft. Wurde der Algorithmus von Experten erarbeitet, spricht man von einem Expertensystem. Daneben gibt es solche, die das Regelsystem mit Hilfe von maschinellem Lernen aus Daten selbstständig ableiten.”

So-called expert systems (Karst, 1992; Lucas & van der Gaag, 1991; Puppe, 1988) were an early approach to ADM systems. These are computer programs that attempt to simulate the specialized knowledge and deductive reasoning of qualified professionals in particular fields. They are used to perform tasks in a specific area.

### Definition 2.2: Expert system

Expert systems are computer programs intended to replicate the specialized knowledge and deductive reasoning ability of qualified experts in narrowly defined application areas (Puppe, 1988, p. 2)<sup>a</sup>

<sup>a</sup>Translation by the author; original German text: “Expertensysteme sind Programme, mit denen das Spezialwissen und die Schlussfolgerungsfähigkeit qualifizierter Fachleute auf eng begrenzten Aufgabengebieten nachgebildet werden sollen.”

As early as 1979, an expert system supported the Pacific Medical Center in San Francisco with the evaluation of lung function tests (Aikins et al., 1982) and achieved 75% (Dreyfus et al., 1986, p. 117) to 96% agreement with the physician’s evaluation (Aikins et al., 1982, Chapter 7). But the use of expert systems has also been

---

<sup>1</sup>Big Data is a term used to describe the issues that arise when processing data that is too large, complex, or growing too quickly to be processed using conventional data processing techniques. It includes both structured and unstructured data and can be found in a variety of fields, including business, science, government, and social media.

successful in non-medical fields, such as industrial contexts, for instance in the diagnosis of faulty engines during the serial production of automobiles (for an overview, see Lucas & van der Gaag, 1991, p. 3 et seqq.). In these systems, the knowledge of experts is transformed through various processes into knowledge bases and decision rules (Puppe, 1988, p. 2). To create a knowledge-based system, it is therefore necessary to have in-depth application domain knowledge.

When one examines the four characteristics of suitable problem domains for the application of expert systems, as identified by Puppe, 1988, p. 148, the limitations of expert systems become apparent:

1. The problem area is manageable and can be delimited using common knowledge.
2. The issue can be effectively resolved by experts.
3. The collection of data is simple, but visual and acoustic data pose particular difficulties.
4. The problem is relatively static.

According to Puppe, items one and three reflect the limited absorption and processing capacity of human experts. In addition, Puppe assumes that, since expert systems are derived from actual experts, they do not possess the same level of predictive accuracy in these situations. In item two, the phrasing of an issue being “effectively solvable by experts” clarifies this property. Therefore, according to Puppe, expert systems could only be used in areas where human decisions are already very good, so that adequate quality could be achieved despite the reduced problem-solving quality of an expert system. Moreover, the last item demonstrates how specialized the rule-based systems of the time were, as deterministic rules cannot effectively address dynamic problems. Therefore, the primary limitation of such systems is their applicability only to domains in which people are able to comprehend the decision-making process well enough to make excellent decisions. In recent decades, the method by which previous decisions are used to develop subsequent decision rules has changed. In contrast to the aforementioned early approaches to ADM systems, in which processed data was presented to human decision-makers in order to derive decision rules (expert systems), applications developed with the aid of more complex methods in the field of Artificial Intelligence use algorithms to derive future decision rules from historical data (Buchanan & Shortliffe, 1984)<sup>2</sup>. Using these techniques, significant progress could be made on the mentioned deficiencies. Machine learning algorithms and heuristics are based on the premise that insights and, eventually, automated decision rules can be derived from historical decision data (Watt

---

<sup>2</sup>The purpose of such an algorithmic decision-making system is either to assign the object or subject being evaluated to a class, which is known as classification, or to assign a numeric value, which is known as scoring. Typically, this score represents the “probability” that an object will exhibit a particular property.

## 2. Definitions and related work

---

et al., 2020). For this purpose, the historical data must be accessible in digital format and meet specific requirements. Each data point contains multiple features, or pieces of information. For instance, CVs can be described by the collection of features they include, such as the applicant’s name, previous jobs, date of birth, and/or additional qualifications. If the expression of a feature is to be predicted using new data, this is referred to as supervised learning<sup>3</sup> (see Definition 2.3).

### Definition 2.3: Supervised learning

Supervised learning is a learning strategy in which the correctness of acquired knowledge is tested through feedback from an external knowledge source. (ISO/IEC 2382: 2015)

In the case of CVs, the attribute to be learned and then predicted could be whether or not the applicant was hired. In conclusion, it requires data, also known as ground truth, which consists of the property vector of the object/subject to be evaluated and the property to be predicted. In the field of machine learning, the corresponding algorithms and heuristics use statistical methods to determine which properties of the data subject the feature most accurately predicts. A statistical model contains decision rules that are derived and stored. This model can then “evaluate” each new data point to make a determination regarding the new data point (Zweig, Wenzelburger, & Krafft, 2018). Despite the fact that this type of ADM system solves a few of the expert system issues outlined previously, the increasing complexity of the decision structure raises new issues.

In contrast to expert systems, which are typically easy to comprehend (Karst, 1992; Lucas & van der Gaag, 1991; Puppe, 1988; Spreckelsen & Spitzer, 2009), machine-learned systems exhibit significantly more issues with regard to transparency and, by extension, explainability (Gunning et al., 2019; Xu et al., 2019). This is in part due to the fact that in expert systems, the experts can be asked to justify the developed decisions, i.e., they can respond to questions. This meta-level is not available in machine-learned systems; creating transparency and conclusions regarding the explainability of made decisions can only be attempted on a technical level. This issue is exacerbated by the fact that both expert systems and machine-learned systems always generate probabilistic decision rules based on statistical models.

As a result, ADM systems do not produce truths; rather, all statements, including binary classifications, are given a probability indicating how likely it is that this classification is correct. Moreover, as the planning, development, training, and use of an algorithmic decision-making system or ADM system becomes increasingly complex, a va-

---

<sup>3</sup>There is also “unsupervised learning” in which there is no specific target feature. Here, general knowledge about the structure of the data is sought: Are there groups of elements that have very similar features? Which element is most similar to a specific element? However, because these systems are less common in current use, and their verification is also quite similar, the focus of this work is on supervised learning.

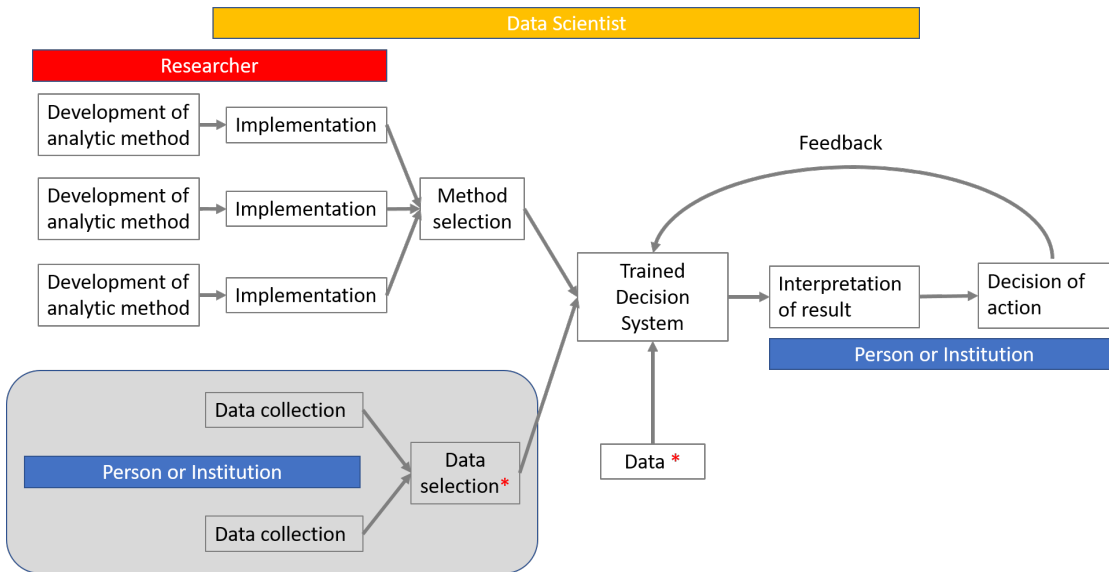


Figure 2.1.: Development of an algorithmic decision-making system: The design of an algorithmic decision-making (ADM) system requires the interaction of various persons and institutions. In a long chain of responsibilities, various decisions are made that all have an influence on the final quality of the resulting system. CC-BY Algorithm Accountability Lab [K. A. Zweig].

riety of problems arise because this process is characterized by an abundance of decisions and assumptions. In a working paper (Zweig, Fischer, & Lischka, 2018), Katharina Zweig placed multiple decisions into a long chain of responsibilities (see Figure 2.1), which we described in detail together with Geörg Wenzelburger (Zweig, Wenzelburger, & Krafft, 2018). It is shown that an ADM system is the result of a complex design process beginning with the selection of the right data and the most suitable data science method and ending with the visualization of the results (M. Haeri et al., 2020; T. D. Krafft & Zweig, 2020; Zweig, Wenzelburger, & Krafft, 2018; Zweig et al., 2021).

Society aims to avoid biased decisions by identifying certain attributes that should be treated fairly. These attributes, known as protected or sensitive attributes, define the characteristics that should not lead to discrimination. Discrimination refers to the unfair treatment of individuals based on these protected attributes in fixed areas (Romei & Ruggieri, 2014). The specific attributes considered protected in a society vary based on cultural norms and laws. The determination of whether a judgment is considered unjustified depends on the specific circumstances and context. Errors, bias, or unintended side effects can easily occur in algorithmic evaluations, which, when applied to humans, can easily result in discriminatory situations. These errors may occur at various stages

## 2. Definitions and related work

---

during the development of an ADM system. As depicted in Figure 2.1, the development process is comprised of numerous process steps. For instance, the historical data used to train the ADM system may already contain instances of unequal treatment, resulting in an ADM system that maintains the status quo but is not discrimination-free. This occurred with the introduction of a predictive system to determine the recidivism rate of offenders in the United States (Angwin et al., 2016). Another source of error is the method and attribute selection used to make decisions. Even if they are harmless on their own, they can indirectly, or in conjunction with one another, result in unequal treatment, as was the case with Amazon’s test introduction of same-day delivery (Ingold & Soper, 2016). In the selection phase of a classification method, the error may be even more subtle. All of these methods have specific requirements for proper operation, but they are also usable without them and produce results that appear reasonable at first glance. Similarly, these methods optimize themselves almost always based on certain quality metrics used to determine when an ADM system is adequate. These two factors have numerous potential pitfalls that are very easy to overlook (Zweig, Wenzelburger, & Krafft, 2018). Even if an automated decision-making system is implemented correctly, there can still be issues with evaluations if the integration of the system’s outcomes into the social process is not done correctly by the individuals or institutions operating it. This can occur, for example, when there is misinterpretation of the results. However, there is also a concern about the presence of inherent bias within decision-making systems. In cases where these systems reproduce value judgments, such as in the context of recidivism, it may not be a matter of misinterpretation, but rather a failure to acknowledge the underlying biases embedded in these systems and originally established by their human developers.

Given that ADM systems have permeated nearly all sectors of society, and given their potential to give rise to complications at different stages, it becomes imperative to establish an effective procedure for the systematic review and assessment of these diverse systems. As will be demonstrated in the following, socioinformatics offers an evaluation strategy that is both practical and effective. Understanding and evaluating whether a technical system makes appropriate decisions requires considering its development process as embedded in a complex social process and examining it as part of a particular socioinformatic system. To facilitate understanding of the socioinformatic approach presented in the following sections, a brief introduction to the research field of socioinformatics and the use of basic terminology is provided next.



## 2.2. Socioinformatics, emergence, and socioinformatics systems

Socioinformatics is a relatively young field of computer science research that focuses on the consequences of interactions between people and software and is gaining importance as the digitalization of society advances rapidly. Together with Katharina Anna Zweig, Anita Klingel, and Enno Park, the author of this dissertation has written a textbook entitled “Sozioinformatik – Ein neuer Blick auf Informatik und Gesellschaft” (Zweig et al., 2021) for this new field of study, where the central terminology is elaborated and a new method for studying socioinformatic systems is introduced. The following remarks regarding ‘emergence’ and ‘model’ are based on the definitions developed there. Socioinformatic phenomena are defined as emergent phenomena between technical and social systems, with a socioinformatic system being a model of both systems in which a socioinformatic phenomenon is detected.

An important task of socioinformatics is technology assessment in the introduction of socially relevant software to ensure that the social goals pursued are achieved in the best possible way. Socioinformatics places particular emphasis on the issue of so-called “emergent phenomena”, which are revealed through the interaction of social actors with software (see Definition 2.4).

### Definition 2.4: Emergent phenomenon

Emergent phenomena are reproducibly observable phenomena (properties or behavior) of systems that are caused by the interaction between elements of this system and cannot be explained without this interaction. In particular, they are not mere aggregations of the individual behaviors or properties of the individual parts, but rather, in the Aristotelian sense, they are greater than the sum of their parts. This “greater than the sum of their parts” is expressed by the fact that something qualitatively new arises as a result of the interaction or that a known behavior (property) changes measurably in causal dependence of the interaction. (Zweig et al., 2021, p. 63)<sup>a</sup>

<sup>a</sup>Translation by the author; original German text: “Als emergente Phänomene bezeichnen wir reproduzierbar beobachtbare Phänomene (Eigenschaften oder Verhalten) von Systemen, die durch die Interaktion zwischen Elementen dieses Systems verursacht werden und ohne diese Interaktion nicht zu erklären sind. Sie sind insbesondere nicht einfach nur Aggregationen der einzelnen Verhaltensweisen oder Eigenschaften der einzelnen Teile, sondern im Aristotelischen Sinne mehr als die Summe ihrer Teile. Dieses „mehr als die Summe ihrer Teile“ drückt sich darin aus, dass durch die Interaktion etwas qualitativ Neues entsteht oder darin, dass ein bekanntes Verhalten (Eigenschaft) sich in kausaler Abhängigkeit von der Interaktion messbar verändert.”

## 2. Definitions and related work

---

Special emphasis is placed on emergent phenomena that involve both technical and social components, such that a socioinformatic phenomenon (see Definition 2.5) emerges from the interaction between social actors and software.

### Definition 2.5: Socioinformatic phenomenon

A socioinformatic phenomenon is an emergent phenomenon that can only be explained by the interaction of social actors (individuals, groups, and institutions) with software. The phenomenon can therefore neither be reduced to the behavior of people alone nor to the behavior of the machine. The existence of a socioinformatic phenomenon therefore implies the existence of a complex socioinformatic system. (Zweig et al., 2021, p. 87)<sup>a</sup>

<sup>a</sup>Translation by the author; original German text: “Ein sozioinformatisches Phänomen ist ein emergentes Phänomen, das sich nur aus dem Zusammenwirken von sozialen Akteuren (Personen, Gruppen und Institutionen) mit Software erklären lässt. Das Phänomen lässt sich also weder auf das Verhalten der Menschen allein noch auf das Verhalten der Maschine reduzieren. Die Existenz eines sozioinformatischen Phänomens impliziert daher das Vorhandensein eines komplexen sozioinformatischen Systems.”

The field of socioinformatics seeks to arrive at a differentiated understanding of how individuals, organizations, or society as a whole interact with software; therefore, it is necessary to first precisely define the system on which the hardware or software acts and vice versa. Zweig et al., 2021 decided to construe the focus of analysis as a representation and the act of capturing the system as modeling. As a result, a socioinformatic system is a model (see Definition 2.6 of System) that consists of a social component and a central hardware and/or software system (Zweig et al., 2021, p. 70). The purpose of such a model is to investigate and explain as precisely as possible an observed emergent phenomenon or to explore possible emergent technological consequences in the use of software by social actors or society as a whole (Zweig et al., 2021, p. 73 et seqq.). Thus, such models consist of at least one component shaped by humans and at least one informatic system, with both subcomponents influencing each other reciprocally, given that humans designed and developed the computer system and only its effects are considered here. This has some advantages because users of the system or affected actors can not only be embedded in different social systems, but also define additional communication rules for the use of the software, which do not have to be written down or be coherent in any way. Furthermore, depending on the desired temporal extension of the technology assessment with only slowly changing social process designs, social actors such as legislators can be defined as either part of the system or part of its environment.

**Definition 2.6: System**

A system is a model that divides a given reality in two halves. On the one hand, it defines a set of entities and their interrelationships that are deemed essential for the context: the system. On the other hand, it also defines the system's environment. These are the parameters that affect system-internal components but do not directly belong to the system because it cannot alter them.

The purpose of the model necessitates this distinction. This purpose usually entails explaining a previously observed phenomenon or predicting a phenomenon that may occur. Therefore, it serves as the starting point for defining the system's boundaries, which must be selected such that all elements pertinent to the respective purpose are included. It should be noted that it is likely that only during the analysis will it become apparent which elements and relationships are essential for comprehending or predicting a phenomenon – and which are not.

In this sense, the system is essentially always modeling and therefore never represents reality. (Zweig et al., 2021, p.54)<sup>a</sup>

<sup>a</sup>Translation by the author; original German text: “Ein System ist ein Modell, das die in die Betrachtung einbezogene Realität in zwei Hälften teilt: Es bestimmt zum einen eine Menge von Entitäten und ihre Beziehungen zueinander, die für einen bestimmten Kontext als wesentlich erachtet werden: das System. Zum anderen wird damit auch die Umwelt des Systems mit definiert. Das sind jene Parameter, die zwar einen Einfluss auf systeminterne Komponenten haben, aber selbst nicht direkt zum System gehören, da sie von ihm nicht verändert werden können.

Relevant für diese Unterscheidung ist dabei der Zweck des Modells. Dieser Zweck ist meistens die Erklärung eines schon beobachteten Phänomens oder die Vorhersage möglicherweise auftretender Phänomene. Er ist damit der Ausgangspunkt für die Festlegung der Systemgrenzen, die so gewählt werden müssen, dass alle für diesen Zweck ausschlaggebenden Elemente enthalten sind. Zu beachten ist dabei, dass wahrscheinlich erst im Rahmen der Analyse klar wird, welche Elemente und Beziehungen wichtig für das Verständnis oder die Vorhersage eines Phänomens sind – und welche nicht.

Grundsätzlich gilt für das System, dass es in diesem Sinne also immer modelliert und daher niemals die Realität stellt. ”

## 2.3. The black-box analysis

There are several theories as to where the term “black box” originates from. Christian Vater reports in his research on the use of the term in the 1930s for the invention of a “feedback box for automated self-regulation of message signals” in telegraphs by Harold Stephen Black. Because of its great utility, this invention had become widely used and was incorporated by many technicians, often without an understanding of its actual functionality. In this regard, it was also known simply as “Black's box”, after the inventor's surname (Vater, 2020, p. 340). Its use in military telecommunications

## 2. Definitions and related work

technology in the 1940s is somewhat closer to the term's later metaphorical meaning. During World War II, a "black box" was understood to mean technology captured from the enemy in a sealed case that could only be examined from the outside and was possibly secured against unauthorized opening (Geitz et al., 2020, p. 5).

Nowadays, the term "Black Box" and the associated idea of viewing a system or system component as a component that cannot be inspected are used in a variety of disciplines, including sociology (Latour, 1994), science and technology studies (H. Weber, 2017), and gender studies (Heßler, 2015).

Besides these definitions from the humanities, there is also a more technical one, which has been mentioned (Geitz et al., 2020) in the "Enzyklopädie Philosophie und Wissenschaftstheorie"<sup>4</sup> (Gabriel et al., 1995) since 1985. The black-box approach is presented here as a method for examining a delimitable component of a system whose functioning is unknown, by attempting to derive insights from the connection between input and output data. The procedure's goal can be both the provision of a mathematical function describing the input-output relationship and the discovery of explanatory hypotheses about the black box's inner structure. What emerges is a subsystem model delimited by the black box (Gabriel et al., 1995, p. 319).

### Definition 2.7: Black box

A definable part of a system, the functioning of which is unknown, is regarded as a black box and examined for the connection between input and output information. The goal of the black-box procedure can be to specify a mathematical function that captures the input-output relationship, and furthermore to find explanatory constructive planning hypotheses about the internal structure of the black box, which (terminologically misleading with respect to model theory) is sometimes referred to as a model of the subsystem delimited as a black box. (Gabriel et al., 1995, p. 319) <sup>a</sup>

<sup>a</sup>Translation by the author; original German text: "Ein abgrenzbarer, seiner Funktionsweise nach unbekannter Teil eines Systems wird als black box betrachtet und auf den Zusammenhang von Eingangs- (input) und Ausgangsinformationen (output) untersucht. Ziel des black box-Verfahrens kann es sein, eine mathematische Funktion anzugeben, die den input-output-Zusammenhang erfaßt, darüber hinaus auch die Auffindung erklärender konstruktiv planender Hypothesen über den inneren Aufbau der black box, der (bezogen auf Modelltheorie terminologisch irreführend) gelegentlich als Modell des als black box abgegrenzten Teilsystems bezeichnet wird.)"

A well-known application of the term can be found in the field of cybernetics. In 1948, Norbert Wiener published a book entitled "Cybernetics - or Control and Communication in the animal and the machine" (Wiener, 1948) which is regarded as the foundation of this area of research. Within the framework of cybernetics, an attempt is made to

<sup>4</sup>Encyclopedia Philosophy and Philosophy of Science.

transfer the behavioral patterns of living organisms to machines and/or social structures via analogies, in order to understand, regulate, or even control them. Wiener defines cybernetics as follows: “[Cybernetics is] the science of control and communication, in the animal and the machine” (Ashby, 1956, p. 1). In addition to biologically motivated areas, cybernetics attempts to predict machine behavior as accurately as possible while being less concerned with explaining the structure or general design (Ashby, 1956, p. 3 (1/5)). The goal of cybernetics is not to predict individual behaviors, but to provide as general a picture of the behavior as possible (Ashby, 1956, p. 3 (1/5)). To achieve the desired generality of the statement, one abstracts from the internal mechanics and works primarily with machine outputs. Cyberneticists therefore make use of theoretical assumptions and scientific hypothesis formation for input-output analysis. In this case, a schematic view of the system as an opaque box is used, which does not reveal anything about the internal mechanics. This abstraction is known as a black box in cybernetics (see Figure 2.2).

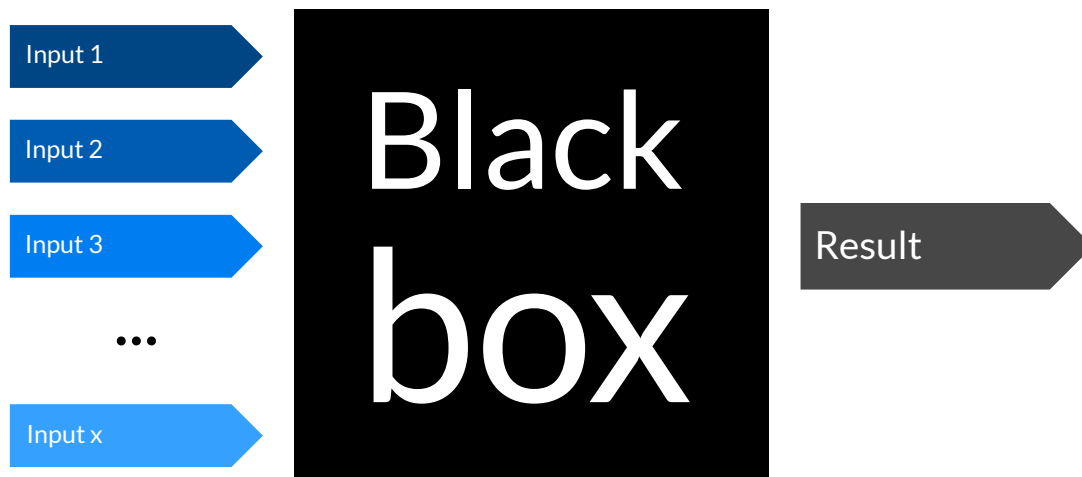


Figure 2.2.: Sketch of a black box with the inputs 1 to x and a result.

All processes and procedures contained within this black box are unknown or unutilized. This type of abstraction is also used, for instance, in abstract schematic drawings in electrical engineering. Individual components, some of which are already quite complex, are represented with symbols in circuit and equipment drawings. In this manner, the overall system can be sketched and further investigated with firm assumptions about the behavior of the individual components - but also without exact knowledge or control over them (Ashby, 1956, p. 86). The concept of a black box is used in this case to reduce overall complexity. Fundamentally, however, the black-box approach to investi-

## 2. Definitions and related work

---

gation addresses the question of which properties can be understood by analyzing the input and output relationship.

When it comes to ADM systems, researchers face a similar situation, whether due to technical constraints in method selection, legal requirements such as commercial confidentiality, or other factors. Society frequently lacks detailed access to ensure important features of ADM systems and, as a result, has little control over their use. This issue is exacerbated by the speed with which ADM systems are now entering or have already entered all facets of society, and society is at risk of being overrun by future technological advances in the absence of suitable, efficient control options.

By analyzing ADM systems as black boxes, the decision-making of an ADM system can be examined and problematic patterns in it can be uncovered. This approach represents a first and necessary, but not sufficient, step in holding the providers of an algorithmic system accountable. Accountability in general can be defined as “a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences” (M. Bovens, 2007, p. 450). Following Bovens’ definition, Wieringa states: Instead of explaining and justifying their own behavior, algorithmic accountability now focuses on the behavior of the algorithm or algorithmic system in question, which must be justified and explained by the person or company using it. Accordingly, this framework requires an actor (individual, collective, or organizational) to explain the algorithm’s behavior and a forum to challenge that explanation. The relationship between the two is shaped by the disclosure and discussion of the explanation and its criteria, and finally the consequences imposed by the forum (Wieringa, 2020).

To put it in simpler terms, if an actor is responsible for the outcomes of proprietary algorithms, those algorithms are usually kept secret to prevent others from exploiting them (Granka, 2010). This makes it difficult to hold the actor accountable since the forum lacks the means to challenge their actions without any knowledge of the algorithm. Therefore, any demand for algorithmic accountability without sufficient insight into the system is likely to fail. Essentially, the actor cannot be held accountable if there is no way to challenge their actions. To date, only a few successful attempts have been made to challenge the performance of such platforms through black-box analyses, e.g., (Andreou et al., 2018; Datta et al., 2015; T. D. Krafft et al., 2019). Usually, these analyses were sparked by concrete evidence or suspicion that determined the further analysis process. The question arises as to why there are not more black-box analyses being done on this important topic, considering that they provide the necessary basis for public discourse.

From an informatics standpoint, black-box analyses are very similar to software testing, as both involve evaluating the results of systems and subsystems (see Definition 2.8).

**Definition 2.8: Software testing (ISO/IEC/IEEE 24765: 2017)**

“An [analytical quality assurance] activity in which systems, subsystems, or components are executed under specified conditions, the results are observed or recorded, and an evaluation is made of some aspect of the system or component” (ISO/IEC/IEEE 24765: 2017).

There are two commonly employed strategies in the field of software testing (Myers et al., 2011, p. 9)(Nidhra & Dondeti, 2012). First, there is white-box testing, where the focus of testing is the internal structure of a program or system. Both what is tested and the test data are determined based on knowledge of the program’s or system’s inner workings (Myers et al., 2011, p. 11). Black-box testing, on the other hand, views the program as a black box, analogous to the usage of the term black box in this thesis. External specifications must be examined without regard to the internal structure of the program or the system. In software development, an external specification is a “precise description of the program’s behavior from the point of view of the end user” (Myers et al., 2011, p. 224). At this point, the similarity between a black-box analysis and black-box testing becomes evident, as research questions that can be investigated with black-box analyses can be abstractly mapped to the question of whether the ADM system complies with a particular external specification. In traditional testing terminology, testing activities that investigate whether external specifications are met are referred to as function testing (Myers et al., 2011, p. 129, 224). Functional testing is a broad field with numerous implementations, typically dependent on isolating the investigated component. In a specification, the criterion ‘fulfilled’ is evaluated using a series of test cases (Myers et al., 2011, p. 224). Therefore, the transmission of knowledge and analytical methods from this domain is purposeful.

Based on the concept of an ADM system as a black box, the goal of this dissertation was to create a procedure that is as efficient as it is simple to use, and which can be used by non-computer scientists to identify potential dangers in the use of ADM systems in advance and then counteract them with appropriate measures. The process model developed and presented in this work is meant to assist researchers in investigating their own questions using this method, but it can also be used as a methodological foundation for investigating the possibilities and limitations.

## 2.4. Audit

The term “audit” is used for a variety of activities, including inspections of compliance by auditing firms and legal reviews by tax authorities. The term is also used in other disciplines, such as public management (Reichborn-Kjennerud & Vabo, 2017) and anthropology (Strathern, 2000). Audit originates from the Latin verb “audire” which means ‘to hear,’ and is a 3rd-person singular noun in the present tense (he, she, it listens

[to]). The term derives from the Latin “auditus” originally used for an official audit of books, which was done orally in the beginning. Insofar as audits describe instruments for checking complex processes, the concept of auditing has survived to this day. An audit is a systematic and thorough examination performed in a variety of areas, including finance, internal audit, and quality assurance, performed by businesses, external audit firms, and regulatory authorities. An audit’s goal is to ensure the integrity and reliability of financial and other information, as well as to increase public trust in a company or organization. An audit, according to (ISO 19011: 2018), is defined as follows:

**Definition 2.9: Audit (ISO 19011: 2018)**

An audit is a “ systematic, independent and documented process for obtaining objective evidence and evaluating it objectively to determine the extent to which the audit criteria are fulfilled” (ISO 19011: 2018, p. 11)

According to what something is audited - which is also called the audit criteria - can be defined at various levels, such as laws and regulations, industry standards, internal company policies, or other standards and requirements. They are typically defined by an independent party, such as an external audit firm or a regulatory authority, and are an important part of any audit because they help to ensure the review’s integrity and reliability.

Audits are processes for systematically checking properties, such as determining whether the object of investigation complies with a company policy, industry standards, or regulations.

We will use the term in various disciplines in the following, beginning with social science, then software development, and finally in the field of algorithmic audit and algorithm audit according to Sandvig et al. (Sandvig et al., 2014).

### 2.4.1. Original use of the term ‘audit’ in social science

Audits have been used as a methodological tool by social science researchers since the 1960s to study difficult-to-prove behaviors such as racial or gender discrimination and decision-making in real-life situations (for a comprehensive list of what is being studied, see Gaddis, 2018, p. 5). Audit studies are a type of field experiment in which audits are used. Field experiments in the social sciences are adapted from the concept of controlled experiments in the natural sciences (Gaddis, 2018, p. 5), in which an attempt is made to implement a randomized research design in the field, i.e., in the real world and environment of the object of study, away from laboratory or survey environments (Gaddis, 2018, p.5).

In general, there are two types of audits: in-person audits and correspondence audits. During the review of the Race Relations Act in England in the 1960s, it was questioned whether interviews could actually detect discriminatory behavior, so an on-site audit was



chosen, where test persons of various origins were assigned to actively try to buy, rent, or ask for a house or to introduce themselves on site in an application process (Daniel, 1968). This type of audit is known as an “in-person audit” because the person performing the testing conducts the audit on-site. Direct interaction is used in these audits to try to get a picture of the organization’s business processes and practices in order to uncover undesirable behavior of people or organizations. However, in addition to scaling issues, Heckman (Heckman, 1998) points out in his study of the weaknesses and limitations of social randomized experiments how complex and difficult it is to form comparable groups that are alike in all dimensions except the characteristic to be studied (Bertrand & Duflo, 2017, p. 318).

Furthermore, it is argued that field auditors are a variable that cannot be assessed because they may have been treated differently due to characteristics or behavior that were not noted. Another limitation of in-person audits is that they cannot be conducted in a double-blind fashion because the people sent out are aware of their status as “test subjects” (Bertrand & Duflo, 2017, p. 318). According to Bertrand and Duflo, there is a possibility that examinees will make an extra effort (see Rosenthal-Effect or Experimenter Expectancy Effect) or unconsciously influence the examination.

Therefore, efforts were made, among other things, to maintain as many characteristics of the inquiries as possible by no longer conducting them on-site and instead transmitting them via telephone, letters, and, later, emails. For example, job postings were sent to recruiters in order to assess the type and number of responses (Bertrand & Mullainathan, 2004; Darolia et al., 2015; Levinson, 1975). As a result, the “correspondence audit” evolved, as the actual process on site was examined, but the interaction occurred via correspondence (Gaddis, 2018). A correspondence audit is thus an audit in which the auditor reviews the audit from a different location. In this case, the auditor receives the organization’s documents and records via post, email, or other electronic means, allowing the review to be conducted from a distance.

In 1969, the first study on correspondence audits was conducted in the United Kingdom. The non-profit institute “Social and Community Planning Research” conducted this study to investigate racial discrimination among employers looking to hire new employees (Jowell & Prescott-Clarke, 1970 as cited in Gaddis, 2018). The authors conducted the review via mail, comparing the responses of British-born whites with four different immigrant groups in order to investigate possible racial discrimination in 128 job advertisements. Even with this type of research, however, certain challenges arise because it is a complex undertaking to ensure that the feature under study is communicated in such a way that it is captured while the flow of communication is not perceived as unnatural (Vecchione et al., 2021).

In conclusion, audit studies in social science refer to a specific type of field experiment in which a researcher randomizes one or more (real or hypothetical) characteristics of individuals and sends them into the field to test the impact of these characteristics on a particular outcome (Gaddis, 2018, p. 5).

Both types of auditing have evolved to the present day<sup>5</sup>, with correspondence audits becoming more prevalent (Vecchione et al., 2021). These enable researchers to make clear causal statements and investigate questions that are frequently difficult or impossible to answer using observational data.<sup>6</sup> It should be noted that this has provided an important foundation for the study of automated decision-making processes. In the next Section, we will examine the growing use of the term “audit” in industrial products, processes, and services that focus on the development and evaluation of software.

### 2.4.2. Use of the term ‘audit’ in the evaluation of software

According to the IEEE Std 1028, 2008 for Software Reviews and Audits, a software audit is defined as follows:

**Definition 2.10: Software audit (IEEE Std 1028, 2008)**

The purpose of a software audit is to provide an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures. (IEEE Std 1028, 2008)

The definition is very similar to the way the term ‘audit’ is used in the verification of products, processes, and services in general. The goal of an audit in this context is to review and assess the effectiveness and efficiency of processes and activities to ensure that they meet the needs of the organization and are carried out correctly. Audits are frequently conducted by specially trained individuals who are responsible for ensuring compliance with and monitoring of company standards and policies. Audits are typically performed at regular intervals, and the results are usually recorded in a report that includes recommendations for improvements that the company can consider. International standards organizations, specifically (ISO/IEC 17065: 2012), distinguish the following basic audit types:

1. A first-party audit is a type of internal audit in which a company’s specific processes or activities are reviewed.
2. A second-party audit is an external audit performed by a customer or business partner to ensure that the company meets the necessary standards and requirements.
3. A third-party audit is performed by an independent, external party that is not affiliated with the company. This type of audit is frequently performed by independent auditing firms to ensure an independent review. Government regulators, industry

---

<sup>5</sup>For a summary, see Bertrand & Duflo, 2017 and Gaddis, 2018, p. 5 et seqq.

<sup>6</sup>See the work of Vecchione et al. for a detailed examination of the various findings of this line of research (Vecchione et al., 2021).

associations, and other external organizations may also conduct third-party audits to establish and/or monitor specific standards or requirements.

Because software is both a product and a service depending on how it is provided and used, the abstract audit descriptions for first- and third-party audits will be transferred to the application context of auditing an ADM system in the following.

A first-party audit or self-assessment describes the systematic recording of a system's (or subsystem's) properties and behavior using measurable test criteria, as well as the evaluation of these properties by the organizational unit responsible for the system's design and (further) development. The self-assessment is performed by the organizational unit that also develops the ADM system (for example, the development department), which is why this type of audit is also known as an internal audit. In recent years, there has been relevant research on internal auditing of ADM systems. Raji et al. present a framework for auditing ADM systems that can be applied throughout the developing organization's development cycle (Raji et al., 2020). The proposed auditing framework seeks to close the accountability gap in the development and deployment of large-scale ADM systems by recommending internal audits as an effective risk-control measure.

Third-party audits are conducted by an external body. If that external body is somehow legitimized, e.g., by the company providing the system to be audited itself in the context of a certification program or a lawyer in the context of a legal dispute, it receives all accesses and information necessary to examine the system. However, if the external body is not legitimized, such as in the case of representatives of an NGO or people affected, options for thorough examinations are limited. For such cases, there is a lack of appropriate, user-friendly verification techniques for ADM systems. A first barrier here is often the external auditor's access to internal information and mechanisms, as society frequently has only limited and also not very detailed access available, making monitoring of such systems extremely difficult, as has been found, for example, in the case of suspect evaluation systems (Angwin et al., 2016) or welfare recipients (Braithwaite, 2020). However, the user interface can also be a technical limitation in auditing. For example, Facebook does not provide an interface that allows for the automatic extraction of information without the involvement of a user or a user interface (T. D. Krafft et al., 2020). In other cases, researchers can only study an ADM system as a black box due to legal requirements such as trade secrecy (Binns, 2018; Brauneis & Goodman, 2018; Mittelstadt et al., 2016; Pasquale, 2015), so there is a significant need for techniques tailored to such scenarios. The same is true for third-party audits, which lack the necessary insights to conduct traditional software or process audits. In such cases, the black-box analysis presented in this dissertation can still be used in the context of a third-party audit. Because this is an investigation of opaque ADM systems, it is a sub-area of the currently evolving 'algorithmic audit,' which is why this subject area will be presented briefly below.

### 2.4.3. Algorithmic audit

In general, algorithmic audits or algorithm audits represent a transfer of the above-mentioned social science audits to the study of algorithms or ADM systems. However, 'algorithmic audits' are investigations into a wide range of analysis situations and research questions. The following definition, provided by Inioluwa Deborah Raji and Joy Buolamwini, clearly shows the range of investigations that fall under this term:

#### Definition 2.11: Algorithmic audit

An algorithmic audit involves the collection and analysis of outcomes from a fixed algorithm or defined model within a system. Through the stimulation of a mock user population, these audits can uncover problematic patterns in models of interest (Raji & Buolamwini, 2019).

Vecchione et al. list the different areas covered by the term 'algorithmic audit' (Vecchione et al., 2021). They range from checks on whether an algorithm considers and responds to legal factors such as non-discrimination (Ali et al., 2019; Hannák et al., 2017; Imana et al., 2021; Wilson et al., 2021) to checks on general performance such as quality claims (Gunawardana & Shani, 2009) or fairness aspects when considering individual quality for specific subgroups (Angwin et al., 2016; Barocas et al., 2021; Buolamwini & Gebru, 2018). Investigations into various approaches, such as attempts to uncover opaque systems and how they work – in which case the term 'audit' is interpreted more in the direction of reverse engineering (Ada Lovelace Institute, 2020; Adler et al., 2018) – or investigations into the balance of political views in content curated by algorithms (R. E. Robertson et al., 2018), also fall under this term. Unlike social science audits, which are by definition experiments (in which the researcher directly manipulates the variable under study and then measures the results), algorithm audits can include both purely observational studies of algorithmic outcomes (in which the researcher measures the outcomes without manipulation) and direct examination of the algorithms themselves.

Christian Sandvig et al. delineated different forms of algorithm audits, which are briefly presented below (Sandvig et al., 2014).

#### Algorithmic audit according to Sandvig

Sandvig et al. present various experimental setups that enable an examination of algorithms on Internet platforms under the name 'algorithmic audit' (Sandvig et al., 2014). An opaque ADM system can be analyzed in a variety of ways, which Sandvig et al. systematized. The forms that apply to an opaque system are detailed below.

A '**noninvasive user audit**' collects data produced by real users (their input and the corresponding system response) in order to perform analyses on it (see Figure 2.3

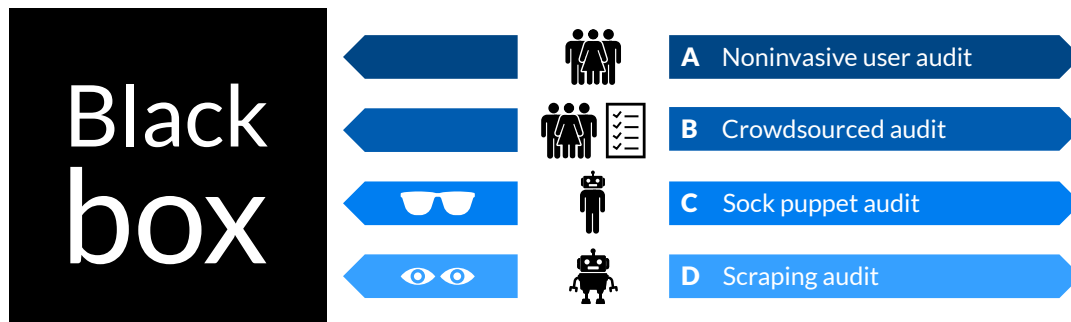


Figure 2.3.: The four audit formats presented by Sandvig et al (Sandvig et al., 2014) that can be applied to opaque ADM systems.

A). Because the auditing institution has no control over the inputs, this type of audit does not allow for any changes to the experimental setup. The system’s results can be obtained using a traditional social science survey format, in which users are asked to report perceived outcomes. However, this method of data collection has the disadvantage of being heavily biased due to unreliable human memories and cognitive biases (Sandvig et al., 2014). The NYU Ad Observatory<sup>7</sup> of the NYU Online Political Transparency Project is an example of a noninvasive user audit investigation. As part of this project, in September 2020, users were asked to install a browser extension to collect and submit political ads they saw while using the social network Facebook.

In a ‘**crowdsourced audit**’, real users provide the system response for their inputs, but the inputs they send to the ADM system are instructed by the auditing institution to allow targeted experimentation with the system (see Figure 2.3 B). These inputs can also be automatically triggered by the user’s equipment. If the ADM system is accessible via a browser, for example, scripts and browser extensions can be created to synchronize as many inputs as possible (type, content and time of request, ...). Several preparations must be made before performing this type of check. It must be determined which types of users must be represented, what their primary characteristics are, and whether specific groups of people must participate (e.g., affected groups or associations). To ensure that the most important groups are included, this definition must be developed based on the research question. The process for developing the definition and potentially selecting a participant group can be roughly derived from the field of medical studies, which has a long history of conducting studies with real people. However, the definition specifies not only what requirements the participants must meet, but also how the study must be advertised, what constitutes a representative user sample, and which partners may

<sup>7</sup><https://2020.adobservatory.org>, last accessed on 28.05.2023

## 2. Definitions and related work

---

be useful in reaching these groups. The simplest way to collect this information, as well as other socio-demographic data required for the study, is through a questionnaire that participants must complete when they register for the study. This includes a request for the participant’s informed consent as well as consent to the study’s processing of their personal data (this is discussed separately in section 8). Following that, a sufficient number of participants must be recruited. This can be difficult and even result in the study being terminated prematurely, especially if the study relies on volunteers rather than paid participants (T. D. Krafft et al., 2021). Opportunities can be enhanced by collaborating with news outlets (T. D. Krafft et al., 2019), which have large audiences and are frequently interested in contributing to a public cause. Other potential partners include thematically related institutions and support groups that have connections to committed people who are also likely to listen to researchers’ advice and be willing to help (Reber et al., 2020). Regardless of how many participants are gathered, the process may introduce an unfavorable bias towards this group if news agencies or interest groups, for example, have specific interests. Another important factor to consider is that the system under investigation may react differently depending on the hardware used by the participants (T. D. Krafft et al., 2021).

A **‘sock puppet audit’** involves no actual users; instead, artificially created accounts controlled by a program (called “bots”) are used to request input and collect system responses (see Figure 2.3 C). The most difficult challenge in this type of audit is to process bots that are not recognized as such by the ADM system to be analyzed (T. D. Krafft et al., 2020). To accomplish this, tools must be developed that interact with the ADM system largely automatically, so that the bots are treated as normal users. The second challenge is to describe the personas under investigation, i.e., which properties all bots “have”, which should only be displayed by a subset, and how the interaction with the ADM system must be designed so that it adequately perceives the bots’ properties. Mikians et al., who studied personalized pricing in online retail, highlighted these difficulties in detail (Mikians et al., 2012). Hannak et al., who used voluntarily submitted browser profiles and algorithmically generated browsing behavior to analyze suspicions of personalized pricing on e-commerce websites (Hannák et al., 2013), as well as Hannak et al. (Hannák et al., 2013) and Datta et al. (Datta et al., 2015), who used bots to investigate transparency and equality in Google search results and ads, provide additional examples of sock puppet audits. The study of algorithmic profiling turns out to be a challenge that can be addressed in very specific areas, but there is no one-size-fits-all solution. In addition to user behavior, information about the user’s location can be included in the ADM system’s result calculation. This approach is known as regionalization, and it is used by Google, for example, to provide users with results for queries that correspond to their location (T. D. Krafft et al., 2019), or by the mobility-as-a-service provider UBER to adjust prices based on the user’s location (L. Chen et al., 2015). Because the information provided by ADM system operators is sometimes imprecise in terms of the exact impact of location data, the two studies cited show examples of how

this was investigated using black-box analyses with a sock puppet audit.

In a '**scraping audit**', no bots are used to simulate human behavior. In this type of audit, input is sent to the system completely automatically, and the results are also retrieved automatically (see Figure 2.3 D). This procedure is aided by an application programming interface (API). If this interface is not available, the reviewing institutions' accesses must be checked manually in order to create the possibility of making fully automatic requests to the system and retrieving the results. While this type of "verification" allows for very precise control over the collection process and the collection of large amounts of data, individual characteristics, such as profiles or system interaction behavior, can be difficult or impossible to simulate. Furthermore, the API may not yield the same results as the user interface.

The auditing institution is frequently constrained by the access an ADM system provides and the resources it can invest in the type of audit: A social network platform may be actively fighting bots, an e-commerce platform might not offer an API, and so on. Furthermore, automating data queries necessitates considerable effort. In the case of a crowdsourced audit of a search engine, for example, a browser plugin that sends identical queries at the same time must first be developed (T. D. Krafft et al., 2019; Reber et al., 2020).

Finding out which type of trial works best in a specific case may require some trial and error, as the best option or possible access is not always obvious. Such studies frequently require the customization and adaptation of software.

## Societal demands for transparency and verifiability of ADM systems

Nowadays, ADM systems are used in a wide range of industries, including online advertising, medical diagnosis, lending, recruitment, and risk assessment in the criminal justice system (Fry, 2018; Saurwein et al., 2015). ADM systems perform complex tasks in each of these areas of application. They supplement or even replace human decision-making, which can have far-reaching consequences for individuals or society as a whole. In her book “Weapons of Math Destruction”, O’Neil, 2016 criticizes the pervasive use of algorithmic decision-making systems in society. She discusses their opacity and unaccountability, their widespread and often unnoticed influence on multiple societal sectors, and their potential to amplify existing biases, leading to discrimination and unfairness. She also warns about their ability to undermine democracy by manipulating perceptions and beliefs, and criticizes their pseudo-scientific basis and lack of validation. Finally, she emphasizes the extensive harm these systems can cause, impacting the lives of numerous individuals. O’Neil’s criticisms serve as a potent reminder of the urgent need for transparency, accountability, and fairness in algorithmic decision-making. Her work underscores the importance of vigilance in ensuring that these powerful tools are used to enhance, rather than undermine, societal values.

ADM systems in the U.S. correctional system, for example, exhibit racist decision-making patterns (Angwin et al., 2016). Search engine results were found to promote racism (Noble, 2018) or include ads with misleading medical advice distributed to users with a serious illness (Reber et al., 2020). Multiple journalistic and academic studies on problems with ADM systems demonstrate that existing control mechanisms are insufficient to prevent negative outcomes (see, for example, Angwin et al., 2016; Braithwaite, 2020). Due to the potentially significant impact on individual well-being and the possibility of interfering with social relations (Brauneis & Goodman, 2018; N. Just & Latzer, 2017; Ulbricht & Yeung, 2020; Yeung, 2017b), a large and rapidly growing scientific community from various disciplines has been addressing the challenges of algorithmic accountability and attempting to solve the problems from the perspective of various disciplines. There are contributions from the field of ethics (Ananny & Crawford, 2018; Binns, 2018; Lanzing, 2019; Mittelstadt et al., 2016) and from the fields of law and politi-



---

cal science (Brauneis & Goodman, 2018; Hildebrandt, 2016; Koops, 2013; Yeung, 2017a) that aim to transfer core ethical principles and develop procedures to ensure that ADM system decisions do not harm or violate the rights of the people involved. Furthermore, technical works demonstrate various ways in which ADM systems can be made more transparent and accountable (Bryson & Theodorou, 2019; Guidotti et al., 2018; Hauer et al., 2021; Kroll et al., 2017; Lepri et al., 2018; Sokol & Flach, 2020; Wieringa, 2020).

The European Union’s General Data Protection Regulation (GDPR) is an important reference point in this context, as it governs the handling of (personal) data at the European level and thus, indirectly – and in some cases directly – the use of ADM systems. The GDPR places high demands on how information is prepared and presented to data subjects, which is why a recital entitled “theme of transparency” (Recital 58 of the European Parliament, 2016) is embedded in the preamble. In the legal field, recitals are used to demonstrate the considerations that led to the adoption of a legal act. In this instance, the recital indicates how future demands for transparency are to be interpreted under the GDPR:

*“The principle of transparency requires that any information addressed to the public or to the data subject be concise, easily accessible and easy to understand, and that clear and plain language and, additionally, where appropriate, visualisation be used. (...) Given that children merit specific protection, any information and communication, where processing is addressed to a child, should be in such a clear and plain language that the child can easily understand.”* (European Parliament, 2016, Recital 58)

Although the GDPR provides a broad framework for data governance, it does not fully define how tools can be used to ensure transparency and accountability for various ADM system applications (Brkan, 2019; Bygrave, 2019). Hence, the GDPR is “not likely to be sufficient” (Koene et al., 2019, p. 1) to adequately enforce accountability of ADM systems, according to a report by the European Parliamentary Research Service. Dreyer and Schulz argue that the GDPR lacks starting points for group and societal goals such as non-discrimination and participation (Dreyer & Schulz, 2018). According to them, the understanding of transparency enshrined in GDPR Articles 12, 13, 14, 15, and 22 deviates both in scope and depth from what is required to review group- or society-related risks (European Parliament, 2016). A review of the systems for discriminatory decisions, for example, in the case of different groups of people, falls outside the scope of both the GDPR and the previous transparency requirements of consumer protection, so the type and concrete implementation of regulation to ensure these requirements remains unclear. Current efforts, however, are aimed at risk-based, differentiated regulation that attempts to cover these cases as well. These efforts will be described in the next section.

### 3.1. Differentiated regulatory efforts of ADM systems

With regard to potentially serious consequences in the case of incorrect decisions of ADM systems or unintended side effects, it is crucial that systems with major societal implications be constructed in a manner that is understandable to society, from the design process to the interpretation of the results. As a result of the high level of transparency frequently demanded, full disclosure of the code may be required.

Viktor Mayer-Schönberger and Kenneth Cukier proposed the so-called TÜV<sup>1</sup> for algorithms in their book “Big Data: Die Revolution, die unser Leben verändern wird” (Mayer-Schönberger & Cukier, 2013), which was then taken up in the media by various actors until 2018 (see, for example, Heise, 2017). It was required in this case that the technical component of an ADM system be demonstrated to and certified by a test center. Several arguments have been raised in opposition to this approach. As some publications on specific areas of application of ADM systems show (Saurwein et al., 2015; van Drunen et al., 2019), there can be no uniform ‘silver bullet solution’ for the regulation of all dangers of algorithmic systems.

ADM systems are always embedded in a social process, and depending on the concrete application and the social setting, the system can cause a wide range of problems and dangers. Each of the technical components is part of a socioinformatic system that must be investigated. A socioinformatic system, as defined in the previous section, is a model that consists of a social component and a central hardware and/or software system (Zweig et al., 2021, p. 87). The influence of a social process on a technical component in a socioinformatic system, and thus on the necessary depth of regulation, is illustrated below through explicit changes to an ADM system’s application scenario.

From a computer science standpoint, different application areas of ADM systems are frequently based on the same ADM system. The widespread use of so-called (product) recommender systems is an example of this (Lü et al., 2012). The underlying technology is the same, regardless of whether the system is used to deliver individualized advertising for goods in online shopping based on a customer’s preferences or to deliver personalized medical advertising. In each scenario, the computer receives a set of objects (i.e., goods,

---

<sup>1</sup>TÜV is the abbreviation for “Technischer Überwachungsverein” in German, which can be translated as “Technical Inspection Association” in English. It refers to a network of independent organizations that conduct technical inspections and issue certificates for various types of equipment, machinery, vehicles, and facilities in Germany. The TÜV organizations were originally established in the late 1800s as private associations to provide technical inspections and certifications for steam boilers and other pressure vessels. Over time, they have expanded their scope to cover a wide range of products and services, including automotive inspections, environmental testing, quality management systems, and more.

In Germany, TÜV inspections are often mandatory for certain types of equipment, such as vehicles and elevators, to ensure that they meet safety and environmental standards. TÜV certifications are also recognized internationally and are often required for exporting German products to other countries.

advertisements for treatments, etc.) that it must independently sort and select for each new user based on its knowledge of the user’s past behavior and the characteristics of the objects.

Figure 3.1 depicts how the use of the same technical component results in different risk assessments based on different data sets because the interpretation of the results varies depending on how it is embedded in a specific socioinformatic system. If it is used for advertising personalized clothing, for example, the risk assessment will be fundamentally different from that of an online search engine because the two are embedded in completely different socioinformatic systems. Looking at the various application scenarios and the associated differences in terms of their damage potential results in very different transparency requirements, despite the fact that the learning method remains a fundamental component of the ADM system. While there is little risk of harm in the case of the above-mentioned advertising for clothing, there are some situations in which even the advertising itself may need to be questioned. For example, personalized medical advertising may have a high risk of causing individual harm if unconventional or unproven methods are recommended, particularly among consumers from a more “unstable” demographic (see section 6).

The same technical component can be at the core of completely different socioinformatic systems in various application areas, each of which places entirely distinct demands on the social process and the technical component. In addition to this explicit context change (similar technical system being used in new application area), there may be an implicit context change that has a significant impact on required reviews and regulations, even when the same ADM system is used for the same data. This was the case in 2018, when the ADM system used by the video portal YouTube<sup>2</sup> to generate recommendations for the autoplay function was the subject of a contentious media debate. While journalistic accusations tended to be limited to the selection of videos from a diversity perspective (see, for example, the article published in the *Süddeutsche Zeitung* on February 4, 2018 by Moorstedt, 2018), James Bridle observed the following: When he viewed videos for young children using the autoplay function, the titles of the suggested videos grew increasingly obscure, and their contents often appeared to have been haphazardly edited. In a Medium blog post (Bridle, 2017) and a TED Talk<sup>3</sup>, he describes how millions of views were accumulated by videos that lacked a meaningful title or compelling content.

Following a journalistic investigation, he hypothesized that a group of YouTube accounts is producing autoplay-optimized films aimed at young children whose parents have autoplay enabled on their browsers or televisions. These videos do not need to be of high quality because the intended audience cannot actively reject the suggestions; the children in this case are too young to operate a computer or television. It appears that

---

<sup>2</sup><https://www.youtube.com>

<sup>3</sup>TED Talk by James Bridle: The nightmare videos of childrens’ YouTube — and what’s wrong with the internet today, 2017, [https://www.ted.com/talks/james\\_bridle\\_the\\_nightmare\\_videos\\_of\\_childrens\\_youtube\\_and\\_what\\_s\\_wrong\\_with\\_the\\_internet\\_today](https://www.ted.com/talks/james_bridle_the_nightmare_videos_of_childrens_youtube_and_what_s_wrong_with_the_internet_today), last accessed on 28.05.2023

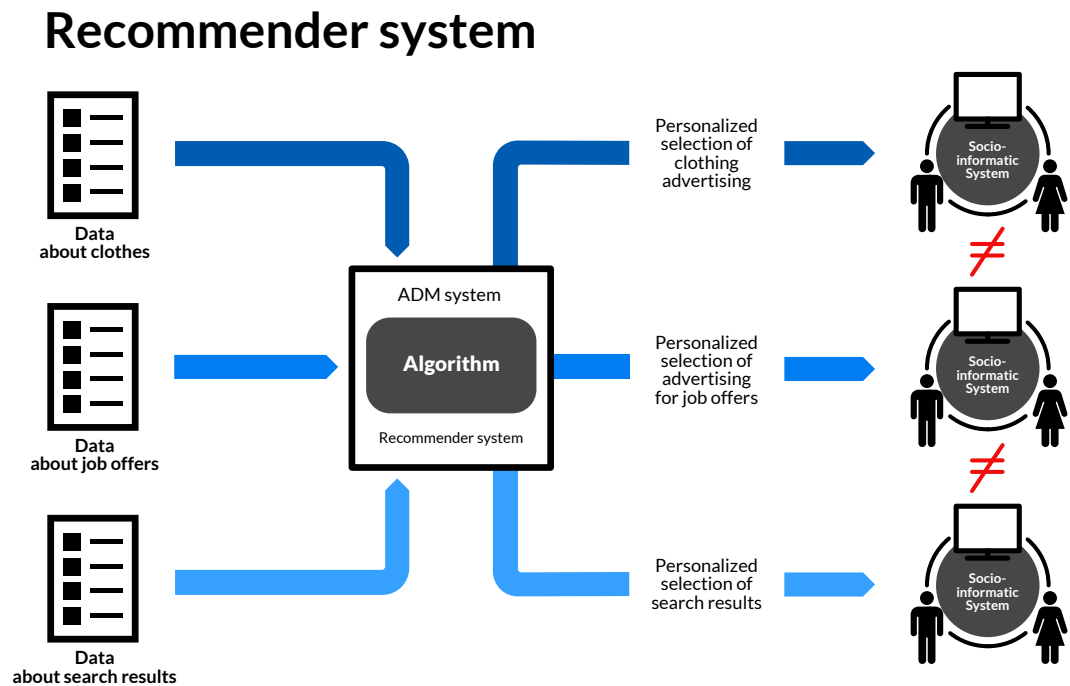


Figure 3.1.: The usage of a recommendation system in various fields of application demonstrates that the resulting socioinformatic systems place varying demands on the technical component (Figure according to T. D. Krafft & Zweig, 2019).

the account owners produce these videos by automatically compiling successful videos in order to attract the attention of the ADM system that generates the autoplay suggestions, so that they are regularly delivered. Although the technological aspect of the ADM system remains mostly undisclosed to protect trade secrets, Google staff have succinctly explained the methodology employed in formulating these recommendations in publications dated 2010 (Davidson et al., 2010) and 2016 (Covington et al., 2016). The ADM system is integrated into a highly intricate socioinformatic system.

While the majority of YouTube users, namely teenagers and adults, have the option to actively interact with the website and reject irrelevant autoplay suggestions, the potential harm from any inappropriate or even disturbing video suggestions increases significantly for (young) children who do not have this option. Thus, while the data used, the overall ADM system, and the actual outcomes remain unchanged, the new audience fundamentally changes the socioinformatic system (see Figure 3.2). This shift

### 3.1. Differentiated regulatory efforts of ADM systems

is accompanied by an increase in the risk of making bad decisions, necessitating significantly higher transparency standards. As a result of the implicit context shift, the two socioinformatic systems necessitate entirely different verification and, most likely, regulatory requirements. Not all uses are equally problematic, so in principle a differentiated regulatory approach to ADM systems is crucial (Saurwein et al., 2015; van Drunen et al., 2019). Systems can be integrated into a wide range of situations, and their objectives, the impact of decisions, and the risks associated with those decisions vary greatly. As a result, regulatory rules must account for the numerous applications of ADM systems. Otherwise, regulation may be too lax in some areas while being too strict in others, discouraging initiative and innovation. Based on the assumption of Jensen & Meckling, 1976 that effective control and regulation at the lowest possible cost to society is generally desirable, the degree of regulation, and thus the approach and the tools for enforcing accountability in the use of ADM systems, must be tailored to the various applications.

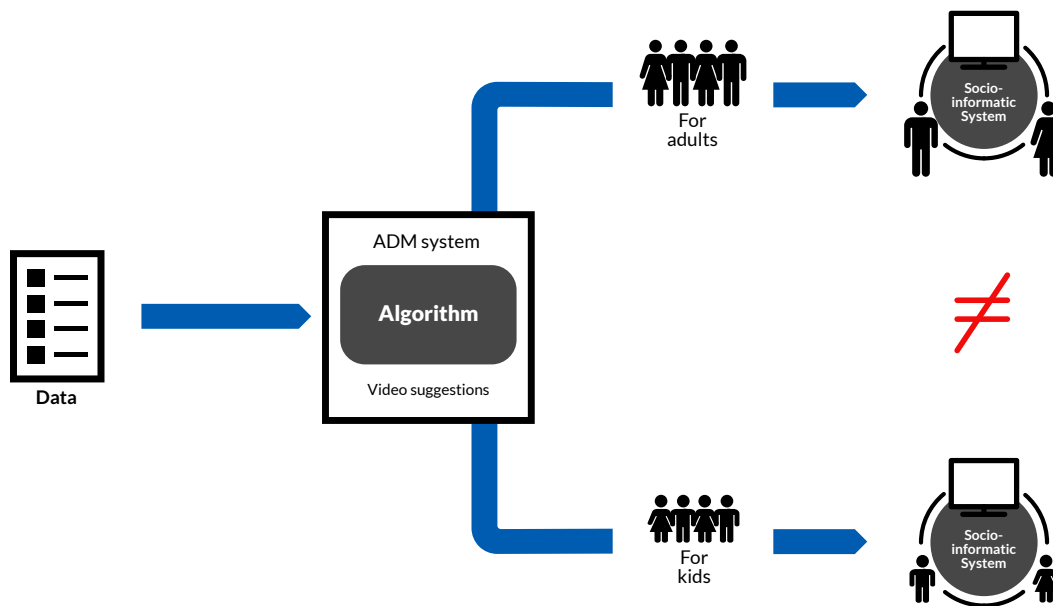


Figure 3.2.: Change of the socioinformatic system as a result of a new target audience.

### 3.2. Risk-based approach to assessing the damage potential of an ADM system in a socioinformatic system

With this in mind, Katharina Zweig and this author published a risk-based regulation proposal for ADM systems in early 2019 (T. D. Krafft & Zweig, 2019). In this proposal, the total damage potential of an ADM system is determined based on the technical component's integration into the socioinformatic system. This takes into account both the potential consequences of a mistake and the extent to which the involved party is affected in terms of the decision's contestability. Following additional research and consultation with various experts (see, for example, DIN & DKE, 2020; Hallensleben et al., 2020; T. D. Krafft et al., 2022), we arrived at the following specification of the two dimensions: The consequences of an incorrect decision are measured by the "extent of possible violations of legal rights and human lives", which includes not only individual consequences but also those for fundamental rights, equality, or social justice, implying that potential super-linear overall societal damage potentials are also included. The following three aspects characterize the contestability of outcomes as "an individual's freedom of action":

- **Control:** Decisions and actions of an ADM system that are additionally controlled or evaluated by human interaction (e.g., purchase of recommended products from an online retailer) have a lower need for regulation than machines that act without human intervention (e.g., emergency shutdown of a nuclear power plant).
- **Choice:** The ability to consult an alternative decision-making system, such as a human decision-maker or an existing variety of providers, means that the overall potential for damage is significantly lower than if only one body is responsible for the decision and this body uses an ADM system. This is a common category for software deployed by the government.
- **Correction:** The ability to challenge or correct an automatically generated decision, as well as the time required to adequately follow up on the corresponding request, have a significant impact on an ADM system's overall damage potential.

We propose increasing the transparency and auditability requirements that the technical component should meet in this area of application based on the level of the overall damage potential determined in this manner. For this purpose, we recommend classification into five regulatory classes (T. D. Krafft & Zweig, 2019):

0. No transparency obligations are required in class 0, and no control processes are permanently installed. In cases of suspicion, a post-hoc analysis is performed, and the risk assessment may need to be repeated.

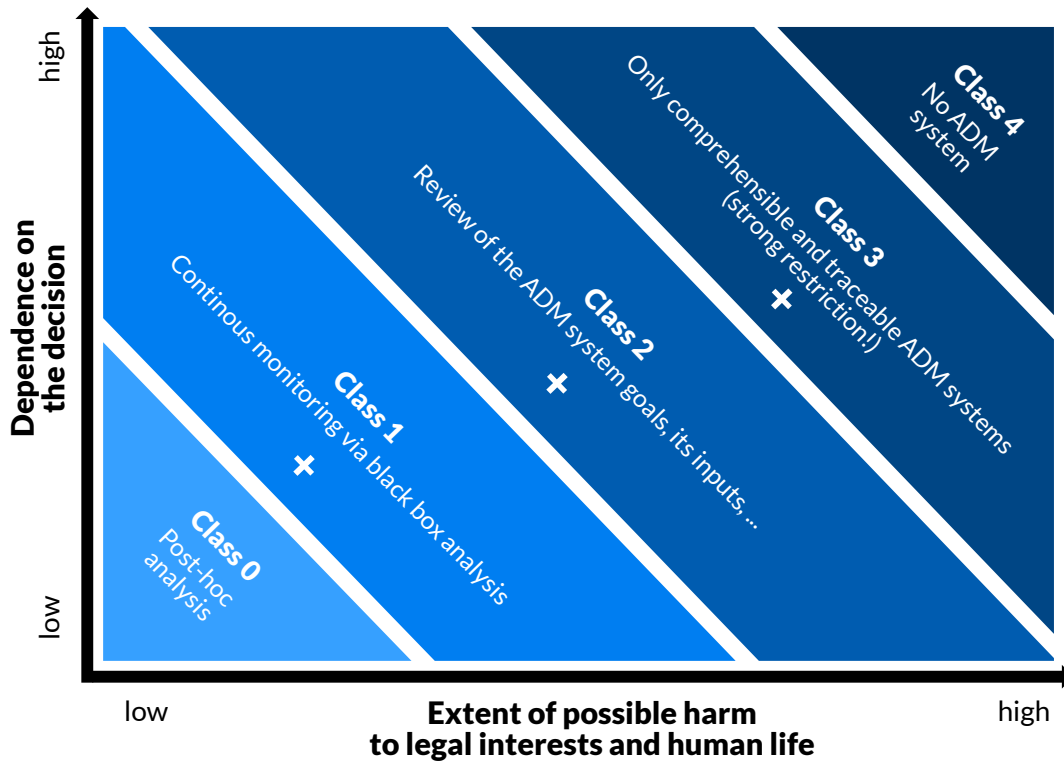


Figure 3.3.: Risk matrix with 5 classes of application areas with risk potentials ranging from ‘post-hoc analysis is sufficient’ in class 0 to the prohibition of AI systems in class 4.

1. Initial transparency obligations are required in class 1. An interface for analyzing the system as a black box must be provided, as well as a description of the ADM system’s integration into the social decision-making process.
2. In class 2, the input data must be completely described (to an audience to be determined) and the information on the quality of the decision-making system must be verifiable.
3. In class 3, all information must be understandable and verifiable within a reasonable time frame for at least a panel of experts. This necessitates different interfaces to the machine’s input data and output.
4. Learning ADM systems with a class 4 evaluation should not be used at all because

### 3. Societal demands for transparency and verifiability of ADM systems

---

the risk is too high, or only when there is a demonstrably sufficient high overall benefit to the contrary.

This proposal was taken up by policymakers both on the national level by the German Datenethikkommission (Data Ethics Commission, 2019, p. 173), the German Parliament’s enquete commission “Künstliche Intelligenz - Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale” (Deutscher Bundestag, 2020, p. 66), and the German Standardization Roadmap AI of the DIN/DKE (DIN & DKE, 2020, p. 102), and on the international level in the AI White Paper of the EU (European Commission, 2020) and the draft of the EU AI Act (European Parliament & European Council, 2021). In a paper for the Special Issue “Algorithmic Regulation” of the journal *Regulation & Governance*, we continued work on the risk-based regulatory proposal with Katharina Zweig and political scientist Pascal König, presenting a framework that differentiates the legal requirements for a variety of ADM systems (T. D. Krafft et al., 2022). In this paper, we developed strategies for adapting the legal requirements when using ADM systems to a variety of circumstances. We used agency theory as a theoretical tool to capture accountability issues and determine how to address them. We focused on solutions that provide transparency and algorithmic control in the context of various accountability mechanisms. Using a risk matrix, we then demonstrated how these tools can be tailored to different ADM applications. The end result is a comprehensive framework outlining the fundamental concepts and standards for regulating various ADM systems.

This dissertation places specific emphasis on automated decision-making (ADM) systems classified as Class 1. These systems require a greater level of transparency and understanding compared to Class 0 systems, which are subject to review only subsequent to an incident. Concurrently, the target verifiability and prerequisites for Class 1 are designed to be less intrusive than those for Class 2 systems, which necessitate stringent, invasive validation protocols to ensure transparency and mitigate potential harm. In Class 1, it is posited that there exist requirements for a non-invasive evaluation framework for ADM systems. In our viewpoint, a considerable subset of these systems will belong to Class 1. Although surveillance is warranted within this class, it should be achieved without internal system access. One viable assessment strategy for such systems is the implementation of black-box analysis.

Nevertheless, to devise and carry out such an investigation, a robust and compelling rationale must be established that an ADM system might indeed inflict the alleged harm. In this pursuit, our technological assessment research has not unearthed any suitable methodology for exploring socioinformatic phenomena in this light. Consequently, the next Section presents the phenomenon-induced socioinformatic analysis formulated by Zweig et al., 2021, which scrutinizes the interconnections in the emergence of socioinformatic phenomena and could also be leveraged for the required justification in this setting.



# Phenomenon-induced socioinformatic analysis

In the preceding section, procedures for regulating and monitoring ADM systems were discussed. Specifically, there was a lack of a procedure to investigate suspected cases in a targeted manner with respect to the allegations made against the involved ADM system. It was often difficult to identify reliable suspicions in order to determine whether a particular technical ADM system was actually a trigger, and what scenarios were associated with it that justified further investigation. Without such suspicions, it was difficult to justify further investigation. As a result, it was often impossible in the previous discourse to determine whether a technical ADM system was responsible for a particular problem or not. This led to a lack of transparency and accountability in the regulation of ADM systems. To address this problem, it is necessary to develop and implement effective procedures for investigating suspected cases in order to clarify the responsibility of ADM systems for specific problems and to ensure transparent and responsible regulation.

During this time, the author of this dissertation co-created the “phenomenon-induced socioinformatic analysis” (Zweig et al., 2021) with Katharina Anna Zweig, Anita Klingel, and Enno Park as a qualitative method for investigating socioinformatic phenomena. This method can also be used to determine which parts of a technical component have a causal effect on an investigated phenomenon, resulting in the confirmation of a previously established suspicious moment. The method is presented and explained in detail in our textbook (Zweig et al., 2021, Chapter 5), which is why this type of socioinformatic analysis is only briefly presented here in the form of an abbreviated and annotated description. The method of phenomenon-induced socioinformatic analysis investigates whether an observable phenomenon is a socioinformatic phenomenon and which factors from the technical component and the behavior of social actors influence the phenomenon. The potential measurable influencing factors are represented as system variables, and their causal relationships are shown in an “effect structure”. Some causal relationships are controlled by software, while others are influenced by social actors’ motivations, and still others cannot be influenced because they are determined by natural laws, for example. The method is designed to identify a phenomenon through as few parameters

and causal relationships as possible, which can then be validated through appropriate experiments and studies.

The main purpose of a phenomenon-induced socioinformatic analysis is to pursue the question of whether an observed phenomenon is indeed a socioinformatic phenomenon and, if so, what elements of the technical component and what actions of the social actors are believed to be causal to the phenomenon. This can also be used as a basis for developing an argument that identifies a particular technical component as the cause of an observed phenomenon. A socioinformatic phenomenon is an emergent phenomenon that results from the interaction between social actors (individuals, groups, and institutions) and software and cannot be attributed solely to human behavior or to the behavior of machines (see Definition 2.5). It indicates the existence of a complex socioinformatics system.

To perform a phenomenon-induced socioinformatic analysis, it is first necessary to detail the phenomenon under investigation. Based on this, it should be verified whether the phenomenon exists at a scale where analysis seems reasonable.

### 4.1. Structure of a phenomenon-induced socioinformatic analysis

After determining the phenomenon to be studied, the phenomenon-induced socioinformatic analysis includes the following five steps:

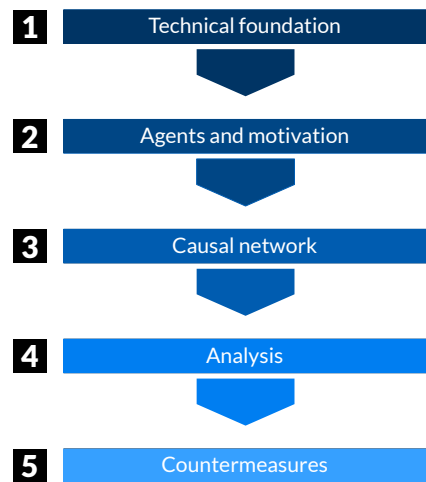


Figure 4.1.: Procedure of a phenomenon-induced socioinformatic analysis (Zweig et al., 2021, p.91)

#### 4.1.1. Phase 1: Description of the technical foundation

The goal of the first phase is to create a description of the technical foundation of the technical component(s) (Zweig et al., 2021, p. 92 et seqq.). To understand who uses the system and why, the description of the technical foundation must be based on a comprehensive analysis of the socio-technical system, i.e., the ADM system and its specific application. In many cases, it becomes clear that involving domain experts can be beneficial because it is often only with their knowledge that it is possible to determine exactly how an ADM system is integrated into a (social) process. While this is still quite simple in the case of an online search engine such as Google, the process becomes far more complicated when a system is used by another entity rather than the people concerned, such as a system used by the state to detect benefit abuse (Braithwaite, 2020). It is essential to identify the software or algorithms used in the various phases of the investigated phenomenon and to determine their influence on the phenomenon's major measured variables in as much detail as possible. These measured variables of the phenomenon, i.e., how to determine whether the behavior described in the phenomenon is actually occurring, are typically derived from the phenomenon's precise description. However, it should be noted that the significance of additional technical components may become apparent only later in some cases. In principle, it is advantageous at this stage to comprehend the technical components to the extent that this is possible in order to discern the relevant incentive structures they provide. However, motivations for specific behaviors are often hidden within the design of the user interface or optimization features, making these internal, less visible functions a substantial obstacle in the examination of socioinformatic phenomena.

#### 4.1.2. Phase 2: Identification of the relevant social agents and their motivations

As a socioinformatic analysis investigates the dynamics of people's behavior when interacting with a technical system, it is necessary to identify all actors involved in the context of the socioinformatic phenomenon under investigation (Zweig et al., 2021, p. 95 et seqq.). This is done within the framework of modeling, where sometimes no clear boundary can be drawn as to which actors (individuals, interest groups, companies, etc.) remain part of the overall socioinformatic system and which do not. As a result, it has proven useful to refer to actors who are unable to change their behavior or actively intervene during the analysis period, known as the analysis horizon, as the "environment" (of the system). Additionally, the method strives to create a model that is straightforward and easy to understand, with the assumption that the model's environment is fixed and unalterable. According to Occam's Razor, the goal of a scientific or philosophical method is to seek the simplest and most parsimonious explanation for a phenomenon. Therefore, the use of Occam's Razor supports the decision to assume a fixed and un-

changeable environment in this case, as it helps to create a simpler and more effective model. Pedro Domingos has written an article on the use of Occam’s Razor for developing simple models in the field of knowledge discovery in databases, which provides an overview of this topic (Domingos, 1999). Existing laws provide a good illustration of the assumption that the environment remains constant throughout the analysis period: While the judiciary can theoretically modify them, they remain unchanged during the investigation of the socioinformatic phenomenon of potentially discriminatory decisions in human resource management and thus belong to the environment.

Because different actors have different incentive structures, i.e., they only act or refrain from acting because certain circumstances motivate them to do so, such structures must also be recorded. In accordance with the terminology used in our basic workbook (Zweig et al., 2021), we use the term “motivation” to describe such incentive structures.

While economic interests predominate in the case of corporations, in the case of people, incentive structures are far more complex. To model such incentives, we believe that the traditional homo economicus (Smith, 1776), which is oriented toward utility maximization, is too one-sided and ineffective. This model assumes that people have complete information about all of their action options and all of the possible outcomes of their actions so that they can always calculate which action will bring them the most benefit. The aforementioned presumption seems unsuitable for numerous everyday situations where socioinformatic phenomena occur (see Definition 2.5). As a result, we suggest employing Maslow’s hierarchy of needs (Maslow, 1943), which offers a more appropriate framework for consideration, given its outline of fundamental human motivations (Zweig et al., 2021, p. 19 et seqq.). Following the completion of the second phase, a preliminary list of involved actors and their respective motivational structures is available and the system’s environment, including any inactive actors, can be determined.

#### 4.1.3. Phase 3: Creating the effect structure

The next step is to employ a so-called ‘effect structure’ to visually process the socioinformatic phenomenon and, by extension, the suspicion under investigation (Zweig et al., 2021, p. 98 et seqq.). An ‘effect structure’ is a simple visual representation of the interdependence of various system variables (variables for short). In the course of our own work and research on socioinformatic phenomena, we have developed the type of visual processing we propose for phenomenon-induced socioinformatic analysis. Relevant system properties of the phenomenon are modeled here as variables, so that the phenomenon can be understood more clearly with the help of the various causal effect relationships between them, and the effect structure already provides explanatory approaches with regard to its creation. The resulting effect structures for reducing complexity make use of rather imprecise causal relationships between measurable system properties or behaviors. The definition of the representable relationships that we have developed requires the system variables to be quantifiable, i.e., measurable. It is therefore important to



Figure 4.2.: Agonistic relationship (Zweig et al., 2021, p. 99).

understand what it means when a variable increases or decreases in order to later evaluate the effect structure. A bad example of a variable is “staff confidence in the ADM system”, whereas “number of staff using the ADM system” is measurable by clearly identifiable identifiers. The measurable properties or system variables are represented graphically as nodes in a graph. A graph is a mathematical structure consisting of two parts: a set of nodes and a subset of all possible pairs of nodes, referred to as edges.

Nodes are represented by an enclosed area in the visualization, and the edges between them are represented by connections of these areas. Edges can be directed or undirected. If the relationship is directed, the arrows on the connections indicate the direction of the relationship. This arrowhead is not present on undirected edges. There are two types of causal relationships between normal variables in the effect structure used here, each of which represents a directed edge representing the effect of a change in the value of system variable A on system variable B. We distinguish between two kinds of relationships:

1. An agonistic relationship is defined by the fact that a change in the variable where the effect originates (variable A) causes a reaction in the variable that is acted upon in the same direction (variable B). A rise in variable A raises variable B, while a decrease in variable A lowers variable B.
2. An antagonistic relationship causes the target variable (variable B) to change in the opposite direction. If variable A increases, variable B declines. If variable A decreases, variable B rises.

As this is a directional relationship, a change in variable B has no effect on variable A. Of course, an inverse relationship can be discovered and plotted, but there is no guarantee that a direct inference in the opposite direction can be drawn.

It has proven useful for the creation of an effect structure to begin with the observed phenomenon and represent it in one or more variables, to which further system variables are added one after the other. In a subsequent step, the (rough) causal relationships are introduced.



Figure 4.3.: Antagonistic relationship (Zweig et al., 2021, p. 99).

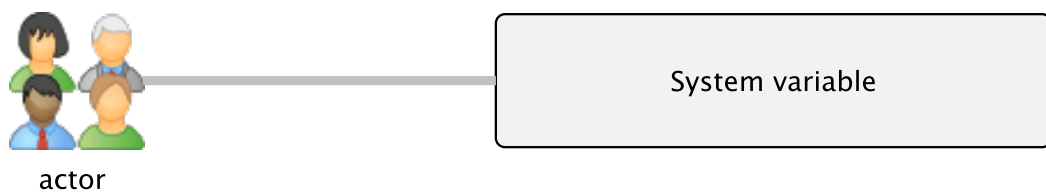


Figure 4.4.: Here, a motivation edge is drawn between an actor and a system variable. It is solid and not directed. At this point, the actor strives to increase the system variable.

To add actors and their associated incentives, another type of node is used for the actor in question. It is connected to the previous system variables via so-called “motivation edges” (see Figure 4.4). They indicate that an actor wishes to alter a system variable. If this variable is to be increased, the edge is solid; if the variable is to be decreased, the edge is dashed. Motivation edges are drawn without an arrowhead as undirected edges.

#### 4.1.4. Phase 4: Analysis of the effect structure

In the context of a phenomenon-induced analysis, the phenomenon under investigation must be explained as thoroughly and plausibly as possible using the effect structure in order to answer the question of whether a socioinformatic phenomenon exists here (see Zweig et al., 2021, p. 108 et seq.). The purpose of this investigation is to determine whether an emergent phenomenon resulting from the interaction of social actors and technical components is present. All components of the socioinformatic phenomenon being investigated must be recorded and explained in a logically coherent manner. To determine the social component, it is best to begin with the actors’ incentives in phase 2 and examine their direct and indirect influence on system variables. The actors act

## 4.2. Phenomenon-induced socioinformatic analysis of the filter bubble theory on Google

in accordance with the motivations displayed, so it is now necessary to determine what they influence and which possible causal relationships they can use to achieve their own goals. If an actor aims to increase a variable, such as personal income, without having direct control over it, but possesses influence over a related factor, like the company's customer count, which is indirectly connected to their salary, the individual will endeavor to increase the number of customers. The role of the technical component must be investigated once the involvement of one or more social actors has been clarified and proven. The technical system involved is captured by system variables, which is followed by an explanation of how the phenomenon arises from the interaction of the two components in the final step. It is critical to understand that the effects cannot be derived from a single component. Only when both components are involved in some way does this prove the existence of a (complex) socioinformatic system that causes or could cause the socioinformatic phenomenon. The resulting argument that the ADM system, as a technical component, at least contributes to the emergence of the socioinformatic phenomenon, i.e., the suspicion, can then be used to justify the targeted investigation of this system. Furthermore, the established socioinformatic phenomenon as a whole can be used to argue for the ADM system's potential for harm.

### **4.1.5. Phase 5: Identification of possible countermeasures**

The effect structure serves as a basis for identifying various control options that can be applied to the socioinformatic system. These control options encompass interventions in the incentive structures of the actors involved and potential modifications to the technical components of the system. However, as this particular aspect is not pertinent to the use of the method within the context of this dissertation, further information and guidance can be found in section 5.1.5 of (Zweig et al., 2021, p. 110 et seq.).

## **4.2. Phenomenon-induced socioinformatic analysis of the filter bubble theory on Google**

The process of phenomenon-induced socioinformatic analysis is illustrated in this Section using the filter bubble theory on the Google search engine as an example. This will demonstrate both the specific procedure and the potential of this method to address contentious issues related to the use of ADM systems. The so-called filter bubble was a much-discussed phenomenon in 2015/2016. The topic of the discourse was the possibility that political opinion-forming processes could be distorted and influenced by the personalized dissemination of content, particularly news on the Google platform. Our textbook (Zweig et al., 2021, Chapter 5) explores this phenomenon by means of a phenomenon-induced socio-informatic analysis. The descriptions provided in the textbook serve as the foundation for the subsequent analysis. First, the phenomenon under

investigation had to be described in greater detail: The days of searching for information in a media environment that was manageable and simply structured, with a limited number of directly accessible sources of information, are long gone. The centuries-old method of disseminating information has been fundamentally transformed as a result of digitalization. Whereas the population used to always read, hear, or see the same news through information sources such as newspapers, television, and radio, these sources are now being replaced in the course of the digital transformation by increasingly diverse and personalized media offerings. ADM systems enable this level of personalization: An algorithm, not a human, determines what content might be of interest to users, and only that content is made available on various news platforms and social networks. The same holds true for - similarly personalized - search engines such as Google, Yahoo, and Bing. With over 3,200 billion search queries on these platforms back in 2016 (Statista, 2023a) already, sorting by humans is impossible.

According to the 'MedienVielfalts-Monitor'<sup>1</sup> published by the Bavarian Regulatory Authority for Commercial Broadcasting (BLM)<sup>2</sup>, more than a third of the population in Germany aged 14 and older uses search engines at least once a day - led by Google Search - to learn about the current situation.

Already in 2016, it was demonstrated that information relevant to opinion formation is no longer limited to traditional media (Siegfried et al., 2016, p. 29). This trend is also supported by the distribution of opinion-forming weight, which shows that the Internet already holds the top spot with a share of 46.0% among the lower age group of 14- to 29-year-olds (Siegfried et al., 2016, p.7). Parallel to this transformation, in the age of digitalization, new information intermediaries are emerging between the information source and the information seeker. Intermediaries collect, arrange, and structure news and organize access to information content for users in their role as information mediators, so they play an increasingly important role in the process of opinion formation for both providers and users. According to Ecke, 2016, 57.3% of all online users in Germany used at least one intermediary for information purposes every day in 2016 (Ecke, 2016, p.13)<sup>3</sup>. Due to its ever-increasing importance as a point of access for the development of information, this new form of information transfer fundamentally raises the question of potential risks and dangers for the formation of public opinion and how these could be legally mitigated. The fear that information intermediaries such as Google or Facebook use their capabilities to manipulate public opinion is growing (Kramer et al., 2014). In this context, the accusation of algorithmically generated filter bubbles has emerged, which is explained in more detail below based on Eli Pariser's filter bubble theory.

---

<sup>1</sup>Media Diversity Monitor

<sup>2</sup>The 'MedienVielfalts-Monitor' consistently releases the evaluation of media opinions, which is determined by Kantar TNS Infratest, a market and opinion research institute.

<sup>3</sup>The presentation by Oliver Ecke is being referenced due to the study's unavailability to the general public.



## 4.2. Phenomenon-induced socioinformatic analysis of the filter bubble theory on Google

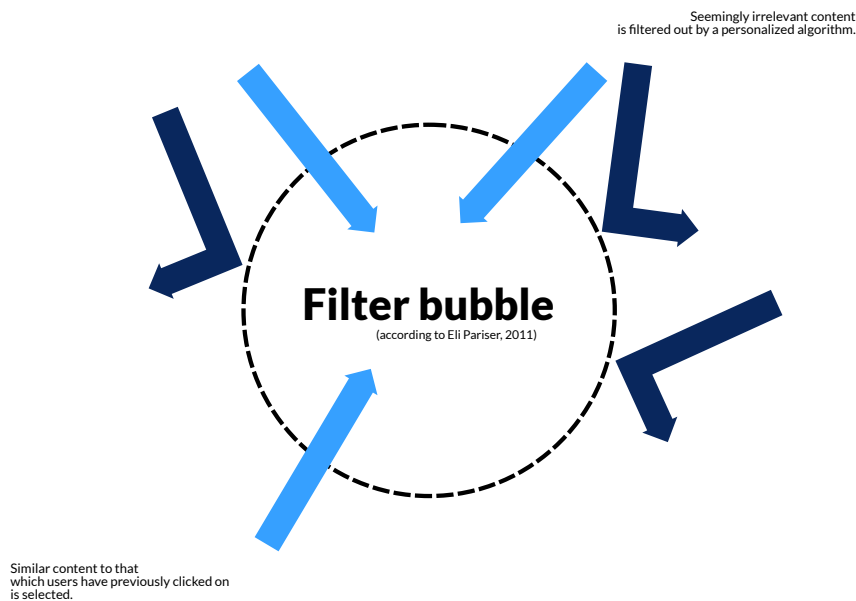


Figure 4.5.: The filter bubble refers to the result of compiling content solely according to perceived preferences (Zweig et al., 2021, p. 167).

### Filter bubble theory according to Eli Pariser

In the context of search engines, the term 'filter bubble' represents one aspect of Eli Pariser's filter bubble theory, which he presented in his 2011 book "The Filter Bubble: What the Internet Is Hiding from You" (Pariser, 2011). His theory was sparked by significantly different search results on the online search platform Google that two friends received in 2010 immediately following the oil disaster on the Deepwater Horizon oil rig in the Gulf of Mexico when they searched for "BP". Pariser then developed the theory that social media facilitate the formation of distinct information spheres, each containing completely different content or opinions, through the use of personalization algorithms. Individuals in these so-called 'filter bubbles' are primarily fed news and information from their preferred opinion space, according to Pariser, 2011 (see Figure 4.5). Pariser hypothesized that personalization algorithms discover a pattern in previous usage behavior that reveals this preference, but fail to recognize that less frequently clicked content may still be relevant for the user (see Figure 4.5). The number of clicks is associated with relevance, while non-clicking is associated with a lack of relevance.

Pariser's 'filter bubble theory' highlights the risk of a restricted perspective and the subsequent constriction of one's worldview due to information filtration by personal-

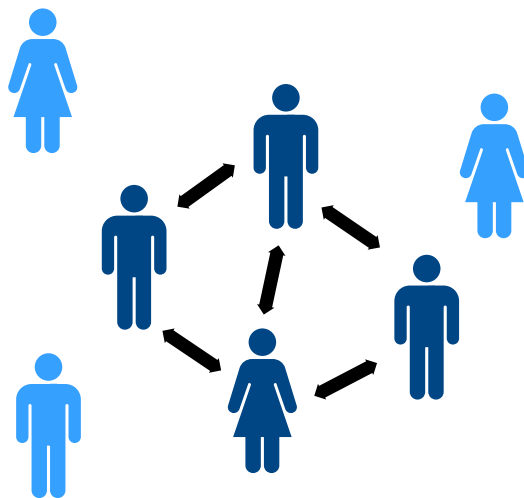


Figure 4.6.: Echo chamber in which people (e.g., friends, club members, etc.) support and confirm each other's opinions. The colors light blue and dark blue represent different attitudes here. Figure according to (Zweig et al., 2017).

ization algorithms. Meanwhile, the 'Echo Chamber Effect,' a term originating from communication science, denotes a similar hazard stemming from typical human behavioral patterns. Originating from acoustics, the term was initially employed in 2001 by the American legal scholars Sunstein & Chambers, 2001. In communication science, the 'Echo Chamber Effect' describes the occurrence wherein an individual's viewpoint is fortified by their social surroundings, mirroring their own perspective back to them, akin to an echo. This can be attributed to the human inclination towards social homophily, characterized as the preference for interacting with individuals who share similarities with oneself. Various factors can contribute to this, including age (Burt, 1991) or shared perspectives (Lazarsfeld, Merton, et al., 1954; Verbrugge, 1977). Consequently, in both analog and digital environments, individuals tend to engage with others who hold similar beliefs, share information that aligns with their own views, and participate in groups centered around a collective narrative (McPherson et al., 2001).

Figure 4.6 depicts the tendency of avoiding opposing voices and seeking communication in a group whose members confirm each other's world view or opinions. This reinforces the individuals's own convictions until eventually, only the group's limited perspective is perceived (see, for example, Garrett, 2009; Sunstein & Chambers, 2001). According to cognitive psychology, such phenomena are rooted in confirmation bias, a key area of selective perception. The human brain tends to store only information that corresponds to one's own opinion on a subject (Baer, 2019, pp. 59-68) (Zweig et al., 2021, p. 24).

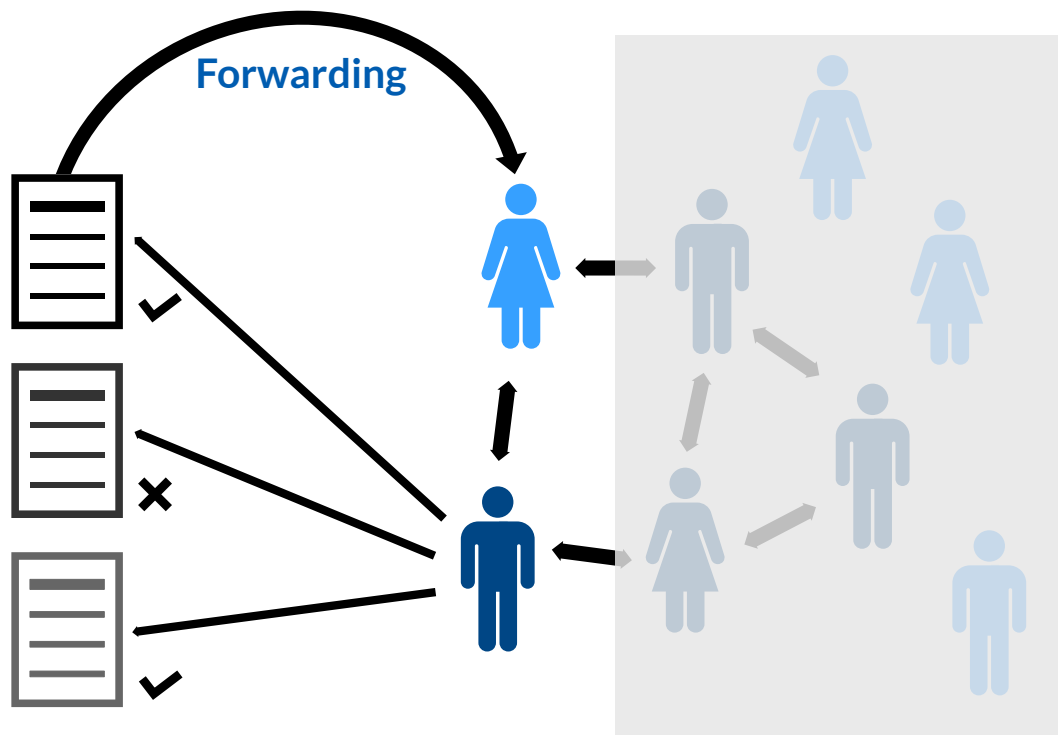


Figure 4.7.: A recommender system generates suggestions by analyzing user behavior when interacting with websites and with other users. This system uses this information to learn about a user’s preferences for certain websites and then suggests them. Figure according to (Zweig et al., 2017).

Contradictory experiences or opposing points of view, on the other hand, are ignored or faded out, so that only information that fits the corresponding presuppositions and reinforces people in their existing opinion enters consciousness through self-selection of the environment.

Eli Pariser applies these theories to online knowledge acquisition and opinion formation. According to him, personalized information flow filtering can lead to groups or individuals being informed about different facts, i.e., facts that “create a unique universe of information for each of us” (Pariser, 2011, p. 10). The interaction behavior of users and their contacts with websites is how algorithmic recommender systems learn which websites a person may like; see Figure 4.7 (Zweig et al., 2017).

The ‘filter bubble’ is the result of a one-sided selection of content based solely on the individual preferences of the users, according to Pariser. The filter bubble denotes

the stream of news that reaches a person and thus distinguishes it from the stream of news that does not. This is especially problematic when the content is politically extreme, and the one-sided viewpoint impairs citizens' ability to converse. According to Pariser, the filter bubble, described as the selection of news that corresponds to one's own perspectives, has the potential to lead to the consolidation and strengthening of one's own political position. This circumstance described by Pariser can be verified using phenomenon-induced socioinformatic analysis because the network between the social actors Google and the user, as well as Google's search engine as a technical component, is a socioinformatic phenomenon.

In the following, we will describe a phenomenon-induced socioinformatic analysis of the filter bubble effect on Google's search engine. This will provide a line of reasoning that suggests the possibility of the ADM system contributing to the potential existence of a filter bubble effect.

### **Phase 1: Description of the technical foundation**

According to the filter bubble theory, search engines like Google and social media platforms like Facebook and Twitter provide users with personalized results based on their previous interactions and preferences. The filter bubble theory is technically based on the use of algorithms and machine learning to predict what information is most relevant to the user.

In the case of Google, the search engine evaluates the relevance of websites for a specific search query using various algorithms and ranking factors (Google, 2023f). These factors include the use of keywords on the webpage, the quality and relevance of backlinks, the website's user-friendliness, and the content's topicality.

When a user enters a search query, Google analyzes their search history, location, demographics, and other information to provide personalized results based on the user's specific needs and interests. This can result in certain information and perspectives being amplified while others are concealed, leading to the formation of a filter bubble.

### **Phase 2: Identification of the relevant social agents and their motivations**

Two participants in the socioinformatic phenomenon can be identified: search engine operators and users.

The incentive structure of search engine users is primarily based on the expectation of finding relevant information that meets their needs and interests quickly and efficiently. As a result, most users prefer to use search engines that can provide relevant and useful results that match their search queries (see motivation edge VIII in Fig 4.9).

A verifiable system is another important concern for search engine users. They expect search engine operators to be open and communicate clearly which factors are considered when determining the relevance of websites (see motivation edge IX in Figure 4.9).

## 4.2. Phenomenon-induced socioinformatic analysis of the filter bubble theory on Google

The search engine operators are primarily obligated to their users in terms of the search engine algorithm in order for them to continue using the service, which is most easily measured by the benefit to the search engine users, i.e., the quality of the personalized prediction (see motivation edge X in Figure 4.9).

### **Phase 3: Creating the effect structure**

Figure 4.9 depicts an effect structure containing the most significant causal relationships underlying a personalization of content recommendations based on the evaluation of human behavior. In addition to the previously mentioned actors, the following variables are listed here:

1. The amount of behavior observed: In addition to previous searches, search engines like Google collect a wide range of characteristics that they use to build profiles. According to Pariser, the Google search engine uses more than 50 user signals (Pariser, 2011, p. 6). The user's language, geolocation, search query history, and social connections on Google+ are all explicitly mentioned by Google (Singhal, 2011). Today, there appear to be more than 200<sup>4</sup>.
2. User profile granularity: The profiles that a search engine creates for its users can be arbitrarily granular: from simple tagging of known characteristics to the calculation of more specific key features or group affiliations. A quick glance at this author's own Google advertising profile<sup>5</sup> as a user reveals that interests were derived from personal search history and are used to granularize his user profile (see Figure 4.8). Even if this profile is rewritten for advertising purposes, it is reasonable to assume that similar key figures also contribute to the personalization of the search.
3. Granularization of users into subgroups: In his book "The Granular Society: How the Digital Dissolves Our Reality" (Kucklick, 2014), Christoph Kucklick refers to the process of individualization and ever finer characterization as granularization (see Definition 3). Users can be treated more and more individually based on increasingly granular user profiles (variable 2).

---

<sup>4</sup>John Mueller, a Google Webmaster Trends Analyst, explaining in a video from Google on 22 March 2019 how Google search works: <https://www.youtube.com/watch?v=ykfJUyD7y0A&t=170s>, last accessed on 28.05.2023.

<sup>5</sup><https://myadcenter.google.com/customize>, last accessed on 28.05.2023.

#### 4. Phenomenon-induced socioinformatic analysis

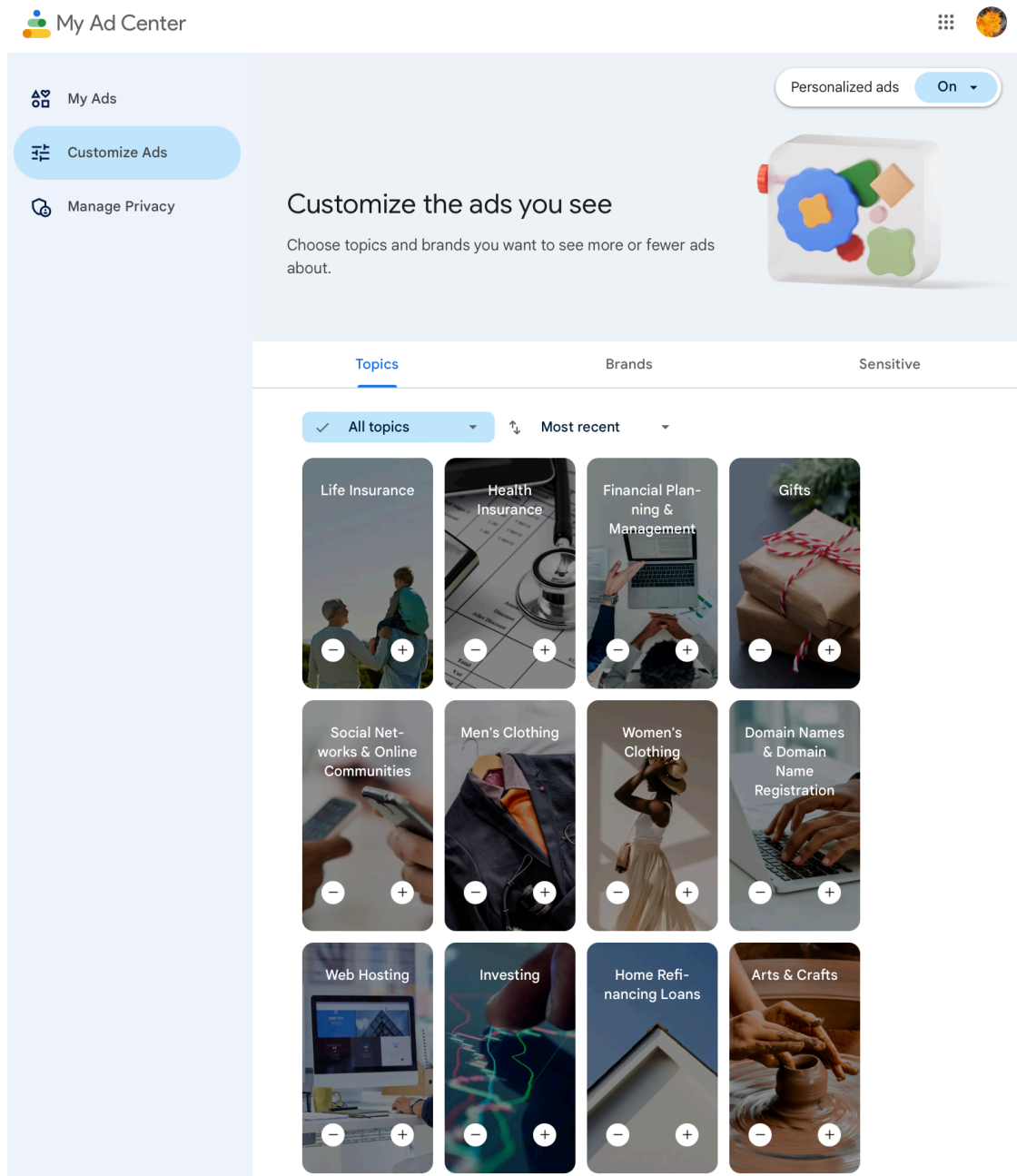


Figure 4.8.: View of the author's advertising profil on 19 Nov. 2022.

**Definition 4.1: Granularization**

Granularization refers to the ever greater individualization of digital services, based on ever more detailed information about the user. In extreme cases, each individual user receives a service that is personally adapted to them. Kucklick calls this Singularization (Kucklick, 2014). (Zweig et al., 2021, p. 162).<sup>a</sup>

<sup>a</sup>Translation by the author; original German text: “Als Granularisierung [...] bezeichnet man die immer stärkere Individualisierung von digitalen Diensten, basierend auf immer detaillierteren Informationen über die Nutzer:innen. Im Extremfall bekommt jeder einzelne Nutzer und jede einzelne Nutzerin den auf ihn oder sie persönlich angepassten Dienst. Kucklick spricht hier von Vereinzlung bzw. Singularisierung”

4. Different system treatment: This variable measures how different the responses to the same search queries are.
5. System verifiability: The more dissimilar the rolled out search result lists are, the more difficult it is to verify the system.
6. Correctness of user profile entries: Many of the entries in user profiles are not verified, but are derived from search behavior, for example, and are thus always subject to error. This variable indicates the accuracy of the information in user profiles.
7. Personalization accuracy: This variable indicates the quality of personalized content prediction.

Relevant cause-effect relationships must be included in order to model and explain the phenomenon under investigation using these system variables:

- I. The more information is available about a search engine user’s behavior, the more precise and granular the user profiles that can be created. There are a number of approaches to this. Attempts are made to extract information from the current search session (White et al., 2009), long-term usage behavior (Matthijs & Radlinski, 2011), or a combination of the two (Bennett et al., 2012). Other approaches rely on using ontologies (Sieg et al., 2007) or topic modeling (Harvey et al., 2013). Information from similar people is sometimes used to supplement a person’s profile (Teevan et al., 2009). All of these approaches have one thing in common: they provide more precise user profiles as more information about the users becomes available.
- II. The more granular the user profiles, the more individualized and personalized the search results.

- III. The more granular the subgroups, the more likely they will be treated differently in order to best address their respective user profiles.
- IV. As the subgroups that are treated in the same way become smaller and smaller, it becomes more difficult to examine and evaluate the overall behavior.
- V. The more data collected about user behavior, the more accurate the user profiles become (see references in I).
- VI. As the user profile becomes more precise, the personalization can be implemented more effectively.
- VII. The more granular the user profiles, the better prepared, selected, and sorted the content can be. Personalization cannot be as precise as with a subdivision by 50 or more characteristics if the user profiles are very coarse: for example, people over 30 and people under 30.

Eli Pariser is concerned that the “different system treatment” modeled in variable 4 causes or promotes the filter bubble effect he describes. Another, smaller effect structure is depicted in Figure 4.10. It shows when algorithmically generated filter bubbles are dangerous.

#### **Phase 4: Analysis of the effect structure**

Eli Pariser’s filter bubble theory, with its troubling implications for society, is based on four basic mechanisms (Pariser, 2011):

1. Personalization (XI in Fig 4.10): A customized selection of content that achieves unprecedented granularity and scalability.
2. Minor overlap in the new/different results (XI in Fig 4.10): Minor or non-existent filter bubble overlap, i.e., news and information from one group remain unknown in another.
3. The explosive nature of the content (XII in Fig 4.10): The nature of the content, which becomes problematic only in the case of politically explosive topics and vastly different perspectives.
4. Isolation from other sources of information (X in Fig 4.10): People in groups with homogeneous, politically charged, and one-sided news situations rarely use other sources of information, or only those that place them in extremely similar filter bubbles.



4.2. Phenomenon-induced socioinformatic analysis of the filter bubble theory on Google

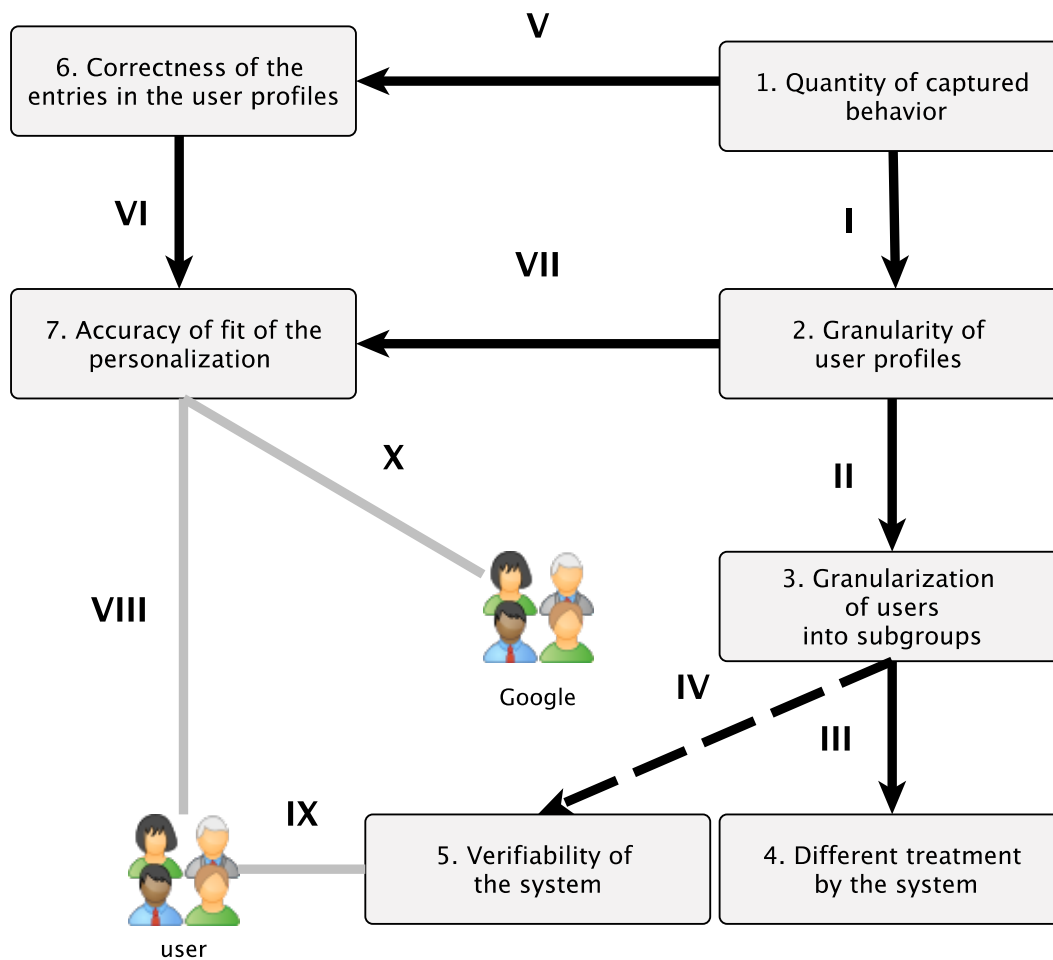


Figure 4.9.: Effect structure of a personalization of content recommendations based on the evaluation of human behavior (translated Figure from Zweig et al., 2021, p. 165).

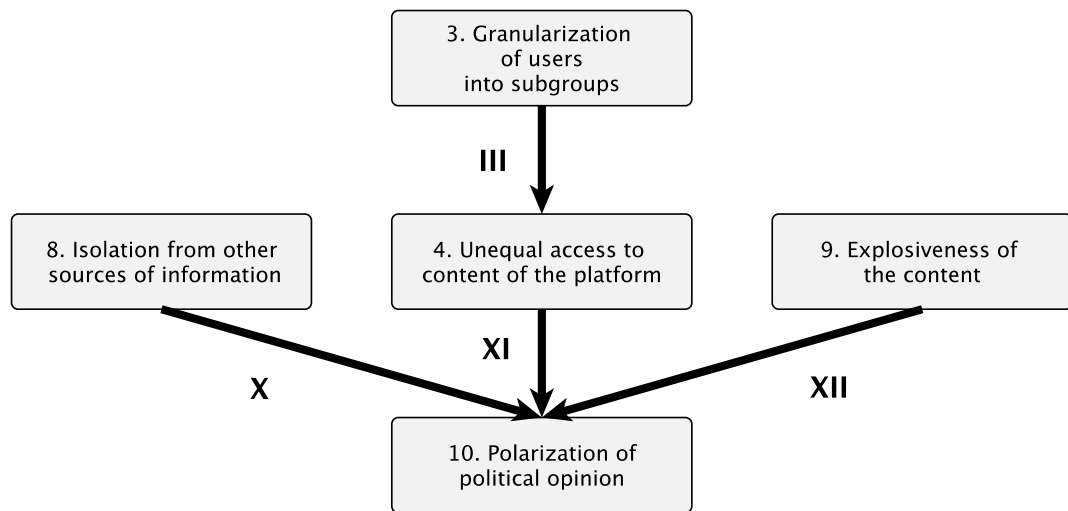


Figure 4.10.: Effect structure of the implicit causal structure of the filter bubble theory as a basis for political polarization. It represents an extension of the effect structure in Figure 4.9. (Translated Figure from Zweig et al., 2021, p. 168)

The more pronounced these four mechanisms are, the stronger the filter bubble effect, including its potentially harmful consequences for society. The degree of personalization of news streams is therefore critical, because politically relevant filter bubbles do not occur when an algorithm responsible for selecting news does not highly personalize this selection. A high level of personalization and verified filter bubbles, on the other hand, do not necessarily have political implications if their content is not political or users use other sources of information. For instance, information delivered to users of different languages is, by definition, non-overlapping if the results are displayed in those languages; regardless, these users are not in separate filter bubbles in terms of content.

At this juncture, multiple approaches could have been employed to further examine whether the filter bubble effect on a search engine like Google did, as Pariser had feared, influence political opinion-shaping processes prior to the 2017 German federal elections. In collaboration with communications and political scientists, it would be possible to examine whether the content disseminated in each instance had the potential to significantly impact the formation of political opinions. This would be an investigation into Pariser's third pillar, content. However, it can be assumed that such an investigation would take a great deal of time and involve numerous value-based decisions, as the task would involve assessing the extent of influence that specific posts generate among various parties and types of information seekers.

#### 4.2. Phenomenon-induced socioinformatic analysis of the filter bubble theory on Google

Furthermore, the isolation of search engine users from other media in the formation of political opinion could be investigated. This analysis would imply a wide-ranging investigation into the media's opinion-forming weight, similar to the 'MedienVielfalts-Monitor' (Siegfried et al., 2016), specifically during the period preceding the elections.

However, because only the interaction of all four pillars of Eli Pariser's filter bubble theory results in a potentially problematic effect for society, an investigation into how much the search results rolled out in each case actually differ can be revealing. The next section will therefore present a black-box analysis that examines personalization on the Google search engine.

# Black-box analysis – Filter bubble on Google

This section's black-box analysis of the Google filter bubble is based on published research results (T. D. Krafft et al., 2019) as well as on the interim (T. D. Krafft et al., 2017, 2018c) and final project reports (T. D. Krafft et al., 2018a, 2018b) that I co-authored. Katharina Zweig, who was instrumental in the design of the experiment and the validation of the results, as well as Michael Gamer, whose extensive analytic skills made the data analysis significantly more robust and efficient, deserve my gratitude.

In the previous section's phenomenon-induced socioinformatics analysis, it was determined that Google's search engine could potentially trigger the filter bubble effect predicted by Eli Pariser. In collaboration with Katharina Zweig, I developed this black-box analysis of the search engine prior to the 2017 German federal elections.

Since search engines on the Internet are able to control the flow of information directed at users, they can be assigned a role similar to that of a gatekeeper in traditional journalism (see Moe & Syvertsen, 2007). Therefore, it is necessary to investigate the actual extent of potential biases exhibited by various intermediaries and search engines. Prior to the 2017 German federal elections, however, there were few credible studies on this topic. Since 2000, research in the field of information retrieval has demonstrated the benefits of personalized search results, such as increased effectiveness (Fan et al., 2000; Liu et al., 2004; Sieg et al., 2007). However, research on the actual roll-out of personalization by major search engine operators and a systematic analysis of the potential risks has been lacking.

The number of websites associated with a search term exceeds the number that can be displayed meaningfully on a web page for (almost) all search queries, therefore users pay the most attention to the websites that appear at the top of the page (Keane et al., 2008; Web searches: Granka et al., 2004; Google: Pan et al., 2007). It is therefore essential that search engines filter the potential search results via selection and sorting. Undoubtedly, the language of the user is one of the most important filters, but topicality, popularity, and embedding in the entire WWW (as measured by the PageRank algorithm Brin & Page, 1998) also play a role. In this context, in 2004 Google experimented with a test version of a personalized search engine (see this CNN article by Matt Hines, 2004), which

---

selects and sorts search results based on the individual traits of the user. On a technical level, this sorting and selection is carried out by so-called “recommender systems”. For a detailed look at recommender systems and how they function, see Lü et al., 2012, which compare the interests of a currently searching user with those of other people who have exhibited similar (click) behavior in the past and determine which articles may be of interest to the searching user. Such systems could also create a profile for each individual based on their individual click behavior and known categorizations of the clicked content, stating, for example: This person prefers travel and celebrity news and enjoys reading short texts and news that are less than one day old (for a detailed overview of Google’s implementation, see Google’s patent for “Personalized Search,” Weare, 2006).

In November 2005, this personalized search was progressively transferred from the test environment to daily operation (Google, 2005), and in 2009, Google spoke of “personalized search for everyone” (Google, 2009). Google’s 2018 privacy statement continued to address personalized search: “We use the information we collect to customize our services for you, including providing recommendations, personalized content, and customized search results” (Google, 2018)<sup>1</sup>. Meanwhile, the number of user characteristics used for personalization grew steadily; in 2011, Eli Pariser wrote that over 50 signals were employed (Pariser, 2011, p. 6), and Google itself mentioned the user’s language, geolocation, search query history, and social connections on its former social media platform Google+ (Singhal, 2011). Currently, there appear to be more than 200 signals<sup>2</sup>. Due to the large number of users and the complexity of the tasks, according to Zweig et al., 2017, this can only be accomplished algorithmically through the use of various machine learning and statistical models. Since in its May 2012 privacy policy, Google published that all Google services can share information about users (Whitten, 2012), it was expected that users who are logged in to their Google accounts would receive more personalized search results than users who are not logged in, given that user information and behavior can be analyzed from different services. However, it remained unclear to what extent Google personalized the returned results; moreover, recent findings indicate that Google search does not employ filter bubbles (Haim et al., 2018). What existed at the time were either qualitative studies, such as the study by Jacob Weisberg, who had only five people search for topics in the context of a Slate article and found that the results were very similar (Weisberg, 2011), or small-scale quantitative studies, such as the attempt by Hannak et al. to quantify the degree of personalization in online search engines. Hannak et al. conducted a study to quantify personalization in search engines by paying 200 Amazon Mechanical Turk users to submit identical search queries

---

<sup>1</sup>Can only be accessed via the Internet Archive “Wayback Machine”: <https://web.archive.org/web/20180703113624/https://policies.google.com/privacy?hl=en>, last accessed on 28.05.2023.

<sup>2</sup>In a video from Google dated 22 March 2019, John Mueller, a Google Webmaster Trends Analyst, explains how Google search works: “Wie funktioniert die Google-Suche? | ‘Frag doch Google’ #20” by “Google Deutschland”: <https://www.youtube.com/watch?v=ykfJUyD7y0A&t=170s>, last accessed on 28.05.2023.

to Google and Bing in order to compare the results (Hannák et al., 2013). Prior to 2017, no more extensive quantitative assessment of the degree of personalization for a larger user base with “real” users was conducted, despite the fact that important questions regarding the degree of personalization and the overlap of individual news streams can only be answered with a large user base.

Conducting such an investigation seemed especially important in light of the debate over the influence of filter bubbles in social networks sparked by Donald Trump’s presidential election victory in 2016 (T. D. Krafft & Zweig, 2017). During the 2015/2016 U.S. presidential election, the question arose as to whether Google used its reach and near-monopoly position on the search engine market to actively shape political opinion. Daniel Trielli, Nicholas Diakopoulos, and Sean Mussenden attempted to demonstrate Google’s tendency to deliver news in this regard in 2015 and reported on their findings in a Slate article (Trielli et al., 2015). In light of the upcoming German federal elections in 2017, the debate was also transferred to Germany, where it was discussed whether and to what extent citizens would encounter personalized, distinct information in their online research, and thus whether the political opinion-forming process would be threatened by possible filter bubble effects.

As described in the previous section, Eli Pariser’s filter bubble effect relies on the interaction of four mechanisms. To investigate this effect on the Google search engine, we employed a black-box analysis to examine the degree of actual personalization rolled out and thus the first two mechanisms (1. Personalization / 2. Minor overlap).

The research project<sup>3</sup> was conducted in collaboration with the NGO AlgorithmWatch<sup>4</sup> and funded by media authorities of six German federal states<sup>5</sup>. The objective was to monitor Google’s search engine in the weeks leading up to the 2017 German federal elections to determine the extent to which politically relevant search queries yielded personalized search results.

The following sections will first present the black-box analysis in detail and then summarize the experiences derived from it.

### 5.1. Conception of the black-box analysis

Domain-specific variations exist in the usage of the term personalization. For the study presented below, the definition of Zuiderveen Borgesius et al., 2016 was used. Based on this understanding, personalization involves selecting content tailored to users that

---

<sup>3</sup><https://datenspende.algorithmwatch.org/en/index.html> This page can currently only be reached via the Internet Archive “Wayback Machine”: <https://web.archive.org/web/20181122010223/https://datenspende.algorithmwatch.org/en/index.html>, last accessed on 28.05.2023.

<sup>4</sup><https://algorithmwatch.org/en/>

<sup>5</sup>Media Authority of Bavaria (BLM), Media Authority of Berlin and Brandenburg (mabb), Media Authority of Hesse (LPR Hessen), Media Authority of Rhineland-Palatinate (LMK), Media Authority of the Saarland (LMS), Media Authority of Saxony (SLM).

they have not yet engaged with, but which may interest them due to its relevance to individuals with similar preferences. Before designing a black-box analysis, it should be determined whether there is truly no alternative to conducting such a complicated analysis. It may also be possible to investigate the question by gaining insights into the technical or ADM system and its corresponding process. These so-called white-box approaches, such as code reviews, produce results with significantly higher validity than time-intensive black-box analyses that rely on numerous assumptions. In this instance, however, Google as the search engine operator does not provide adequate transparency for this approach. Complete transparency regarding the algorithmic implementation of Google's sorting and selection process is also not recommended, as this has historically led to manipulation attempts by a number of website owners. In 1999, Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd published the PageRank algorithm (Page et al., 1998), the algorithmic foundation of their search engine Google. With this knowledge, website operators were able to identify algorithm flaws and develop methodologies to be falsely identified as relevant by the Pagerank algorithm and thus placed in a high position by it (for a detailed overview, see Zweig et al., 2021, Chapter 7.1). Since the publication of this article, a fierce competition has emerged between search engines and website owners for the top rankings (see, e.g., Grimmelmann, 2008). Google has remained silent on the design of its algorithms since then in order to prevent further opportunities for manipulation. Due to the lack of knowledge about the internal mechanisms of Google search, the only option is to examine the system as a black box. Here, the next step is to examine in detail the input and output situation, which will be referred to as the "black-box scenario" in the subsequent process model (see Section 8.2.1).

Google did provide a general overview of the parameters used to personalize and regionalize search results on their blog (Singhal, 2011), but additional information is missing. All that is known is that Google incorporates over 200 signals regarding user behavior, including the stored behavioral profiles of other Google services, into their evaluations<sup>6</sup>. In this regard, it is unclear whether the search query with all of its metadata, such as query time, IP of the query, and so on, uses other information unknown to us to calculate the results page, in addition to the parameters we can control.

As depicted in Figure 5.1, it can be assumed that Google uses some meta-information about the query for the calculation, such as which Google profile was logged in at the time of the search or when the search query was submitted. Although it is possible to examine the system-generated list of results, it is not possible to determine or modify the system's additional input parameters used for possible personalization. However, this uncertainty was not an impediment to the planned investigation, since it is not

---

<sup>6</sup>In a video from Google dated 22 March 2019, John Mueller, a Google Webmaster Trends Analyst, explains how Google search works: <https://www.youtube.com/watch?v=ykfJUyD7y0A&t=170s>, last accessed on 28.05.2023.

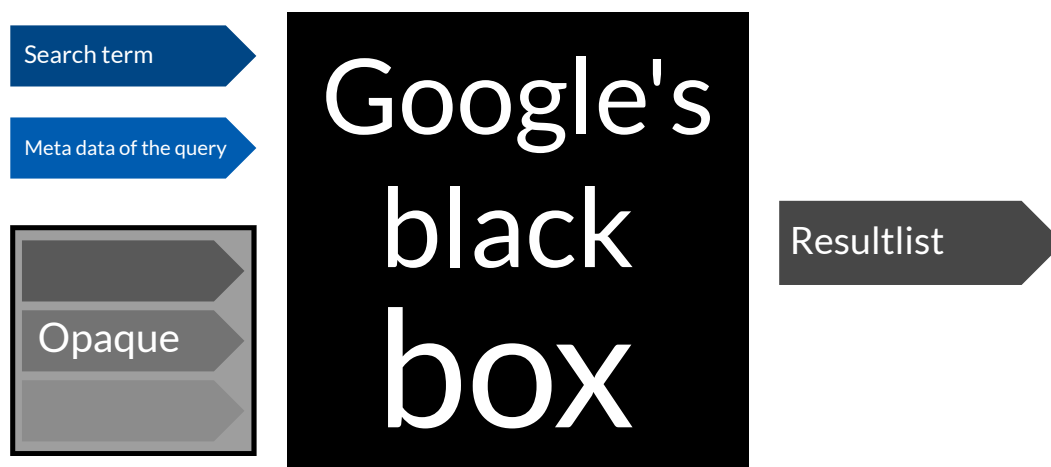


Figure 5.1.: The investigation of the black box of Google’s search engine algorithm represents a black-box scenario of the second category (for details, see Section 8.2.1) There is only knowledge about the outputs of the system and part of the input. However, there are also indications of further inputs, e.g., an individual personalization vector.

the causes of personalization that were to be investigated, but rather its existence and manifestation.

A sensitivity analysis allows for an abstract examination of the personalization of search results. This type of analysis can be used, among other things, to determine the impact of individual input parameters on a classification result. (for a more detailed definition, see Section 8.2.2). The study design involves conducting a scientific experiment using a black-box analysis to test a hypothesis regarding the influence of individual input parameters on the system’s output. Figure 5.2 illustrates the abstract concept underlying our application of this concept to a black-box study. By having real users a and b send identical queries (search term  $x$ ) to Google with as much identical metadata as possible, such as the same time (time of the search  $y$ ), the effect the property “comes from another user” on search results can be determined. A deviation would indicate that search results have been personalized.

Here, all search query properties should be identical (as much as possible), with the exception of the supposedly opaque personalization vector of the user submitting the query. Since a largely deterministic behavior is assumed, it would be expected that the result lists that two users receive for the same query if no personalization is applied would be identical. To determine the presence or absence of personalization, one can



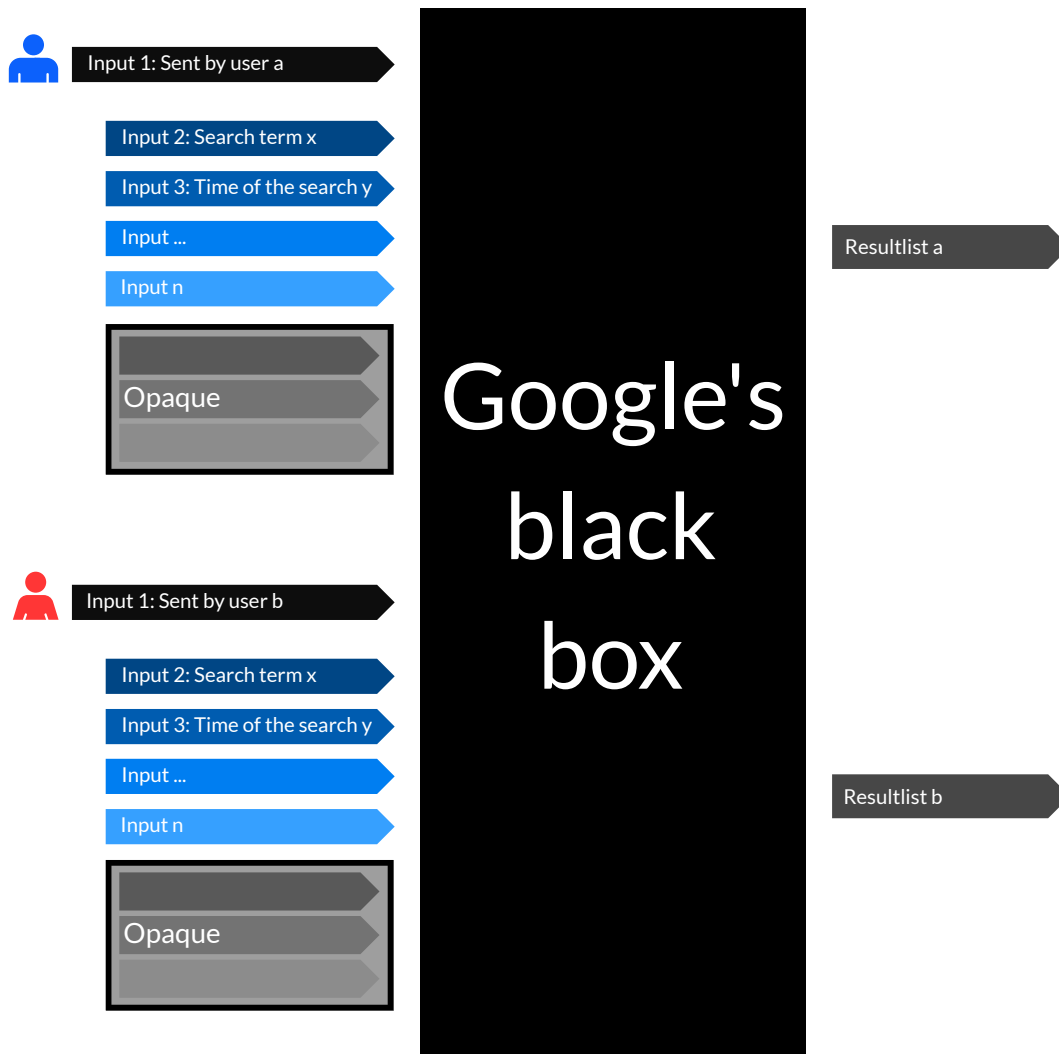


Figure 5.2.: An outline of a sensitivity analysis of Google's search engine. While the Inputs 2...n remain unchanged, the independent Input 1 "sent by user" is modified such that the search query is sent to the black box twice, once by user a and once by user b, to determine whether the answers Result a and Result b differ. A difference would indicate that the search results have been personalized.

assess the degree of variation in the results obtained using different metrics. If the delivered results show significant differences when evaluated based on various metrics, this suggests a high likelihood of personalization being implemented. Conversely, if the results are very similar across different metrics, this indicates a lack of substantial personalization. To analyze this, a sensitivity analysis can be conducted, as described in Section 2.3. Sensitivity analyses are commonly used to examine how modifications in input parameters influence the outcome of a model or system. In this context, sensitivity analyses help determine the extent to which a system responds to changes in input variables and how such changes impact the output variables. Depending on the design of the unknown influencing variable of a user profile, one is dealing with either a one-dimensional sensitivity analysis (if the user profile is entered as a single value) or a multi-dimensional sensitivity analysis, and the results must be evaluated in relation to these unknowns. Section 8.2.2 will elaborate on these two types of sensitivity analysis.

The next step is to select the actual audit form, i.e., determining exactly how the queries are sent to the system and how the respective responses are collected (see Section 2.4.3).

This black-box analysis, as described above, focuses on the impact of the user profile on the search results delivered. Since search queries should be as similar as possible, it is impossible to conduct a “non-invasive user audit”. Here, both the search terms and the respective query time would be different, making it impossible to conduct a sensitivity analysis on this data due to the presence of too many variables. For both a “sock puppet audit” and a “scraping audit”, appropriate user profiles must be created beforehand. However, because too little is known about the actual influencing factors of these profiles, a substantial number of these factors remain undetermined for both of these audit types. Even though these forms of investigation would be a possible next step in the event personalization were detected, it is more effective for the current research question to rely on the user profiles of real people and to request that they send the corresponding queries synchronously from their respective computers to Google as part of a “crowdsourced audit”. To implement this type of audit, a client-server infrastructure consisting of Internet browser plugins that can be installed on participants’ computers and a central server that receives and stores the submitted requests had to be established. Due to a very short-term funding commitment shortly before the 2017 German federal elections, an external service provider, the company ‘Lokaler’<sup>7</sup>, developed the browser plugin according to our specifications.<sup>8</sup> The plugin was developed for the two most popular browsers (Google Chrome and Firefox, see Statista, 2023b). With the browser open and the Internet connection active, the plugin searched for 16 predefined search terms on the German version of the Google search engine ([www.google.de](http://www.google.de)) and on

---

<sup>7</sup>Lokaler UG (HRB 134923) was a software development company owned by Lorenz Matzat, which was liquidated by him in 2019.

<sup>8</sup>The plugin can be downloaded at <https://github.com/algorithmwatch/datenspende>.

Google’s news portal Google News ([news.google.com](https://news.google.com)) every four hours (12 am, 4 am, ..., 8 pm) between 21 August 2017 and 01 October 2017. As search terms, the seven major political parties in Germany and their respective candidates for chancellor were chosen (see list in Table 5.1). The search terms were listed explicitly in the plugin’s source code and were therefore immutable during execution. With these search queries, we wanted to clearly emphasize the process of political opinion formation.

If the browser was turned off during one or more consecutive searches, the next time it was turned on, a new round of searches would begin immediately. Consequently, there are search result lists with varying timestamps.

Search term: party	Search term: person
AFD	Alice Weidel
Bündnis90/Die Grünen	Dietmar Bartsch
CDU	Alexander Gauland
CSU	Katrin Göring-Eckardt
Die Linke	Christian Lindner
FDP	Angela Merkel
SPD	Cem Özdemir
	Martin Schulz
	Sahra Wagenknecht

Table 5.1.: All 16 search terms that were searched for in the data donation for the 2017 German federal elections.

When the plugin performed a search, the entire Google results page was sent to the server, where it was processed and stored for later evaluation. The data available for the investigation was organized as follows: First, it was determined whether the search results were directly on Google or on the search engine operator’s news page. The entirety of the available data was separated into these two categories. When searching on Google’s regular search engine, in addition to typically ten hits, the user is also shown up to three headlines, which we refer to as top stories, or short news articles (see Figure 5.3). These were also submitted and included.

While Google searches typically lead to the websites, social media accounts, and aggregator topic pages of parties and individuals, Google News Search only displays news from registered partners. The lifespan of news is relatively short there, as the majority of news is only available for a limited time in sequential searches, whereas the websites and social media accounts of parties and politicians are displayed almost always. In this regard, it makes sense to evaluate the results of the news search and the Google search separately.

For the standard Google search results, an additional distinction was made between




## 5. Black-box analysis – Filter bubble on Google

Angela Merkel

Alle News Bilder Videos Maps Mehr Einstellungen Tools

Ungefähr 53.000.000 Ergebnisse (0,45 Sekunden)

### Schlagzeilen

 <p>Ex-Ministerpräsident Beckstein: "Ich mache Angela Merkel keinen Vorwurf" T-Online vor 1 Tag</p>	 <p>Bekommt Poroschenko Hilfe von Merkel? Russischer Politiker erklärt Spiegel vor 1 Tag</p>	 <p>Streit in der Union: Richtlinienresistenz Spiegel Online vor 15 Stunden</p>
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

→ Mehr zu Angela Merkel

### Angela Merkel

<https://www.angela-merkel.de/> ▼  
Die persönliche Internetseite der Vorsitzenden der CDU Deutschlands, **Angela Merkel**.

### Angela Merkel – Wikipedia

[https://de.wikipedia.org/wiki/Angela\\_Merkel](https://de.wikipedia.org/wiki/Angela_Merkel) ▼  
Angela Dorothea Merkel (\* 17. Juli 1954 in Hamburg als **Angela Dorothea Kasner**) ist eine deutsche Politikerin (CDU) und seit dem 22. November 2005 amtierende Bundeskanzlerin der Bundesrepublik Deutschland. Am 14. März 2018 wurde **Merkel** vom Bundestag zum vierten Mal zur Bundeskanzlerin gewählt. **Merkel** ...  
[Angela Merkel](#) · [Merkel-Raute](#) · [Joachim Sauer](#) · [Am Kupfergraben](#)

Figure 5.3.: Google search for Angela Merkel with three additional news headlines (top stories).

the top stories (not always delivered) and the “organic” search results, i.e., the eight to ten results in the lower-left sub-field (Figure 5.3 shows two of the organic search results for the search term “Angela Merkel”). Google search results pages may also include information in the lower right-hand corner, such as advertisements or information boxes about people and events. These were not transmitted to our servers; only potential top stories and search results were. From the submitted result lists, the following properties were saved in addition to the type of search (“Google Search”, “Google News Search”):

- The search term of the search
- The timestamp of the search
- The approximate location based on the transmitted IP address of the user
- The login status of users on their Google account: A user can be “logged in” or “not logged in”.
- The language set in the browser (not the search language, which one sets in the Google account settings).
- A unique identifier generated by the plugin that does not reveal any information about the user, but remains the same for all data donations collected from this device as long as the plugin is not reinstalled.
- The URL of the search result
- Where available, a descriptive text of the links was added (available for most organic search results, but not for all).
- If it was a top story, the corresponding time shorthand (e.g., “54 minutes ago”, “3 hours ago”) was also stored.
- In the case of top stories and results from Google News Search, the medium (the news source) and a title were added, if available (e.g., “Dresdner News” with “Das sagen unsere Leser zum Auftritt von Cem Özdemir”).

Since our project only received its funding very shortly before the elections and the funding amount was not sufficient, it was not possible to recruit a representative sample of the German population, so participation was voluntary. It would have been prohibitively expensive to compensate a representative sample of 1,000 participants for 10-15 minutes of computing time per day on a private computer. On the project’s landing page, the volunteer participants were informed in detail about our project and the data donation procedure. The interaction of the plugin with the browser was to be transparent to the user, which is why a pop-up was displayed to inform the user that the data collection was about to begin (see Figure 5.4).

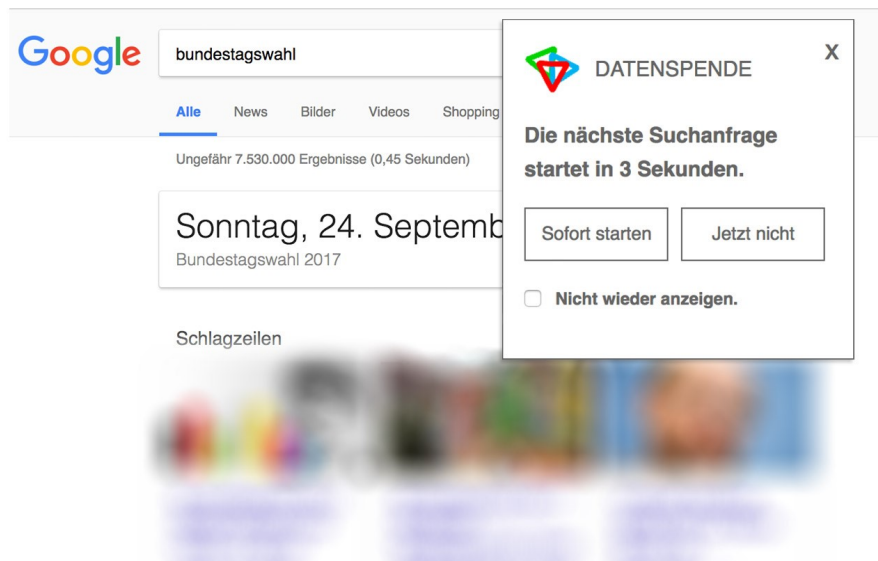


Figure 5.4.: Browser plugin just before a search was triggered on a user’s device, with the first search results page then being forwarded to the provided server structure and thus being “donated”.

In addition, we decided to publish the source code<sup>9</sup> of the browser plugin in order to establish a high level of acceptance and trust in the study and eliminate any possible doubts. It was understood that such an approach would increase the vulnerability of the investigating body to attacks. On the one hand, if the data submission process is made public in this manner, it is possible that malicious manipulation attempts, such as code injection, will be made to counteract the study results or obtain sensitive data. We took precautions to avoid this risk by securing the server, and as far as we are aware, there were no significant attacks on the server infrastructure. The second risk is posed by the investigated ADM system’s potential responses or attempts at adaptation. Given the media attention surrounding the call for participation, it is possible that Google was aware of the investigation prior to the completion of the data donations. This does, for instance, imply the risk that personalization might have been reduced or altered during the investigation period. However, if such an action were to be discovered by Google or published by whistleblowers, for instance, the public response would be enormous, which would undoubtedly have a negative impact on the company’s reputation and finances. Even if this possibility cannot be ruled out, Google’s willingness to take such a risk remains doubtful.

<sup>9</sup><https://github.com/algorithmwatch/datenspende>

## 5.2. Operationalization of personalization

In order to examine the extent of possible personalization, meaningful metrics must be defined to compare the search results of individual users. There are numerous metrics in the field of information retrieval that compare ordered lists based on various characteristics. A first metric could involve a comparison of the actual overlap (later referred to as “commons”), which provides an overview of how many search results (i.e., URLs) are uniquely attributed to a user. This metric places greater emphasis on the selection of pages made by the search engine provider and less emphasis on the order in which the results are presented. In reality, according to a 2007 study by Edward Cutrell and Zhi-Hui Guan, it is the order in which search results are presented that plays a crucial role. Cutrell and Guan had participants perform simple search tasks using a search engine while eye trackers recorded their eye movements (Cutrell & Guan, 2007).

The subjects tended to devote the majority of their time to the first suggestion, significantly less time to the second suggestion, and no more than half of their time to the third or subsequent suggestions. The two researchers also observed that, on average, those who ultimately clicked on the first link had only viewed the first four results, while those who clicked on one of the first four links had only viewed the first five links. For this type of search query and with a limited number of participants, the study was able to demonstrate that the majority of people rely on pre-sorting by the search engine. In 2004, Granka et al., 2004 conducted an eye-tracking study with 36 students regarding the interface design of search engines and the question of how users interact with a displayed search results list. It revealed that as the position decreased, both the amount of time spent on each result and the number of clicks decreased dramatically. While linearly decreasing reading times for results and descriptions on the page were observed for the top five results, the number of clicks decreased from approximately 150 of 397 searches at the first rank to less than 40 at the second rank. In a similar study, Jansen & Spink, 2006 confirmed these findings. Regarding the evaluation of the relevance of search results on Google, Pan et al., 2007 attempted to determine whether the decision of what to click on when evaluating the relevance of an abstract is based on the content, the ranking or a combination of both. They, too, were able to demonstrate that the subjects were strongly affected by the order in which the results were presented, whereas the relevance of the respective content had a negligible effect. Users appear to have such a high level of confidence in Google that they click on abstracts in higher positions even if they are less pertinent to the search query posed (Pan et al., 2007). The study was repeated by Schultheiß et al., 2018, replicating the results. Keane et al., 2008 demonstrate in a study that the proportion of clicks on each position diverged even further. In their experiment, thirty students were tasked with locating sixteen computer-related pieces of information using a search engine, e.g., the creator of the Java programming language. The authors examined the first click made in each instance to answer a question. 70% of these clicks went to the link located at the top of the page, while only 10% went to the

second link. Less than five percent of the participants clicked on any of the remaining positions. This raises the question of whether the links were only clicked due to their positioning.

In their research, Keane, O’Brien, and Smyth approached the issue from two perspectives. First, they solicited feedback from other participants regarding the search results’ relevance to the question posed. Human raters and the search engine largely concurred that there were frequently significant differences in the quality of the results based on the ratings. This was interpreted as evidence that searchers should trust the top results. Next, a second experiment was conducted in which half of the participants were shown the initial ten search results in the exact reverse order in which they had been displayed by the search engine (while the interface looked exactly the same). In this instance, 40% of the time, the participants clicked on the first link in each case. Actually, the last two links were clicked on first in roughly 10% of the instances each, despite the fact that they would have been the best links according to the standard sorting. This suggests that users anticipate finding the best solution at first glance, but are also capable of discovering a superior link in a less prominent location.

Following these explanations, it can be concluded that the positioning of search results has a significant impact on the likelihood that a user will click on a search result, and that this factor should be considered when evaluating any possible personalization.

The actual deviation of the rank (position 1 to 10 in the results list) is the first indication of different sorting, as Hannák et al., 2013 demonstrate in a previously presented study with participants recruited via Amazon Mechanical Turk. Here, the search results lists must be compared in pairs for each rank or position in order to determine how many URLs in each rank are identical so that the study period’s mean value can be calculated. The percentage of results that differ at each rank can then be calculated; this will be referred to as “deviation per rank” in the following. This metric provides a preliminary overview of various classifications, which must be examined in greater detail.

As a third metric, we used the longest common subsequence (LCS) to determine whether the results lists contained identical sub-lists. This measure of similarity is utilized in a variety of applications, including text analysis (Akinwale & Niewiadomski, 2015), the study of clusters of genome sequences (Namiki et al., 2013), and the detection of trajectories for mobile devices (Niedermayer et al., 2013). In addition, it is employed in the management of high-frequency financial data (Guo et al., 2022).

As a final measure, Kendall’s Tau was applied to the results lists. Kendall’s Tau is a correlation measure used to quantify the degree of agreement or disagreement between two ordered lists or series. A value of 1 represents a perfect match, whereas a value of -1 represents a perfect discrepancy. This measure is discussed in greater mathematical detail in Stepanov, 2015. Kendall’s Tau is especially helpful when comparing agreement between two ordered lists with varying lengths or no numerical order of the elements. The measure is also resistant to outliers, making it suitable for analyzing data that may contain incorrect or missing values. Kendall’s Tau was utilized by Hannák et al., 2013



to analyze the personalization of web searches, and by Fagin et al., 2003 to compare the results of various search engines. The latter utilized a modified version of Kendall's Tau with a range of only 0 to 1. This metric was also applied to search engine results by Webber et al., 2010.

Despite the decision to use the above-mentioned metrics to investigate whether personalization is present, the level of these metrics at which one can speak of the existence or absence of personalization remains unknown, as no statement regarding this aspect had been made in the scientific literature prior to this study.

The relationship between the location and the relevance of search results is an important factor to consider when investigating personalization. This concept is referred to as "regionalization" or "geographic similarity" (Andrade & Silva, n.d.) and is the final central aspect we investigated regarding possible personalization on Google. Regionalization in online searches refers to the selection of websites for a group of individuals who conduct a search from a specific region or who are known to be from a specific region. For instance, the current location can be roughly deduced from the IP address of the searching device, from the smartphone's location information, or from the profile known to the search engine (Cambazoglu & Altingovde, 2012; Teevan et al., 2011). Kliman-Silver et al., 2015 demonstrate that location does affect Google search results, and that the differences increase with spatial distance<sup>10</sup>. Considering this, an evaluation of personalization in a search engine should not solely identify the degree of personalization but also investigate the extent to which variations in search engine outcomes can be attributed to regionalization. Eli Pariser's filter bubble effect only influences the political opinion formation of citizens if the selection (commons) and the deviation in sorting (rank, LCS, and Kendall's Tau) of the search results rolled out to citizens by Google's algorithms are sufficiently large and this effect is not due to regionalization.

### 5.3. Data collection

Information about the project and the associated call for data donation was disseminated through the communication channels of our project partners and via our media partner Spiegel Online (Horchert, 2017). During the course of the project, the plugin was downloaded and installed 4,384 times for the Chrome or Firefox browser; between 300 and 600 of these devices were active at the search times of 12 noon, 4 pm, and 8 pm and provided us with the results of the first search results page and the Google News page calculated by Google for them in each instance. Almost all of the 5,991,500 donated search results are freely accessible to the public for analysis<sup>11</sup>. Because there was a risk of de-anonymizing participants by aggregating their data, the UserID was removed. For

---

<sup>10</sup>Here, the searches were conducted on 30 days from 59 locations in the USA in a fully automated manner as part of a scraping audit. A total of 3,600 searches were evaluated.

<sup>11</sup>The data can be accessed via the DOI <https://doi.org/10.26204/DATA/1>.

research purposes, the complete data can be requested from Katharina Zweig.

Since the users were not part of a representative sample, but instead self-selected to participate based on factors such as our media partner's reports, it can be assumed that the user group is not representative. It might, for instance, be more homogeneous in terms of age or level of education than the entirety of Internet users in Germany. The following results are therefore not conclusive, but their form is so unambiguous that we are confident that they would be largely replicated in a representative sample.

### 5.4. Examination of the results

Before the collected results could be evaluated effectively, they first had to be processed appropriately. Before we examine the occurrence of personalization in the processed data sets, the subsequent Section describes and justifies the data cleansing and preprocessing steps in greater detail.

#### 5.4.1. Data cleaning and preprocessing

It was determined through data collection that the initial Firefox plugin assigned the same ID to all users. This corrupt data could have been handled differently at this point. However, we decided to completely exclude the data from the study, despite the fact that this error affected 34% of all donated URLs in Google Search and 41% of all donated URLs in Google News Search. When the study was designed, it was unclear how many participants would download and use the tool, so each time a browser was launched, a search for the search terms was conducted in addition to the coordinated searches conducted at uniform intervals. Therefore, even if the number of installations had been reduced, it would have been possible to increase the number of submissions and, consequently, the evaluation's quality. As the number of participants during the study period was sufficient, however, data submissions that were more than 30 minutes outside the search windows were eliminated. Only with the unique user ID could duplicate submissions from the same person be removed during this time period. An important finding from this project is the need for the ability to correct errors in the study software without the active participation of the study participants.

Following this initial step, the length of each submission was evaluated. Due to the fact that Google can display up to 200 results on the initial results page based on the user's preferences, the length of the submitted results lists varied. Here, we chose the default length of ten results per page employed by Google and shortened longer results lists. As there were isolated users who were shown a tape containing videos, which the data donation tool recorded and processed as a news article, condensing of the top stories was also required. Since these were not actual search results, they were cleaned up so that after this step, a search results list would contain no more than ten search results and three top stories. Additionally, two types of incorrect results were eliminated at this

point. On the one hand, a programming error in an early version of the Firefox plugin caused the initial URL to be submitted ten times instead of the first URLs. On the other hand, there were instances in which a search result did not contain URLs to web pages, but rather so-called redirect links, which route the page request through a Google service. Initially, the structure of these forwarding links appeared to be comprehensible; however, because they contain additional individual information in addition to the URL, it was not possible to extract the actual URL with precision. However, this error only affected 0.5% of the results, so these results lists were also eliminated in their entirety. Due to these two cleaning procedures, the Google search data set was reduced by 19.1%<sup>12</sup>. To determine the distribution of our participants across Germany, their IP addresses were used to locate data donors on the map of Germany shown in Figure 5.5. This map demonstrates that we were able to recruit data donors from all across Germany. The red points on the map represent data donations that were excluded by the aforementioned cleaning procedures; as expected, no pattern emerges.

Subsequently, an overview of the registration behavior during the project period was obtained. Figure 5.6 demonstrates that there were significantly fewer submissions on weekends compared to weekdays. Therefore, significantly fewer users opened their browsers and contributed data to us on weekends, most likely as a result of reduced computer usage on weekends. Consequently, only weekday search results were considered for the study between 21 August 2017 and 24 September 2017 (Note: the only weekend considered was the election weekend of 23-24 September 2017). On a daily basis, an average of 506.9 users contributed search results.

Noon, 4 pm, and 8 pm were chosen as search times because they yielded the most results. As expected, evening submissions were low. Often, there were fewer than 50 records during the night. For each of the three specified time points, all searches submitted within a maximum time difference of 30 minutes before or after were considered.

For the subsequent investigations, four data sets were created (see Table 5.2): The first data set contains all data donations according to the steps outlined above and is referred to as “all” in the following sections. Since some of the results lists contained search results in a different language, we attempted to compile a data set containing only German search results. Evidently, the language settings of the respective searchers led to English or French results, for instance. Unfortunately, the language settings in the browser or Google profile of the searcher are not queried, so we determined the presumed language of the searcher based on the “published” field of the top story, since those searching in German were provided with a time in German (“vor 4 Stunden” instead of “4 hours ago”). As soon as one of the results was written in German, we flagged the entire list as German and created a “German” data set. In this regard, our second database contains all results lists for users who searched for German results; this data set is referred to as “German” in the following.

---

<sup>12</sup>From 4,416,585 to 3,707,302.

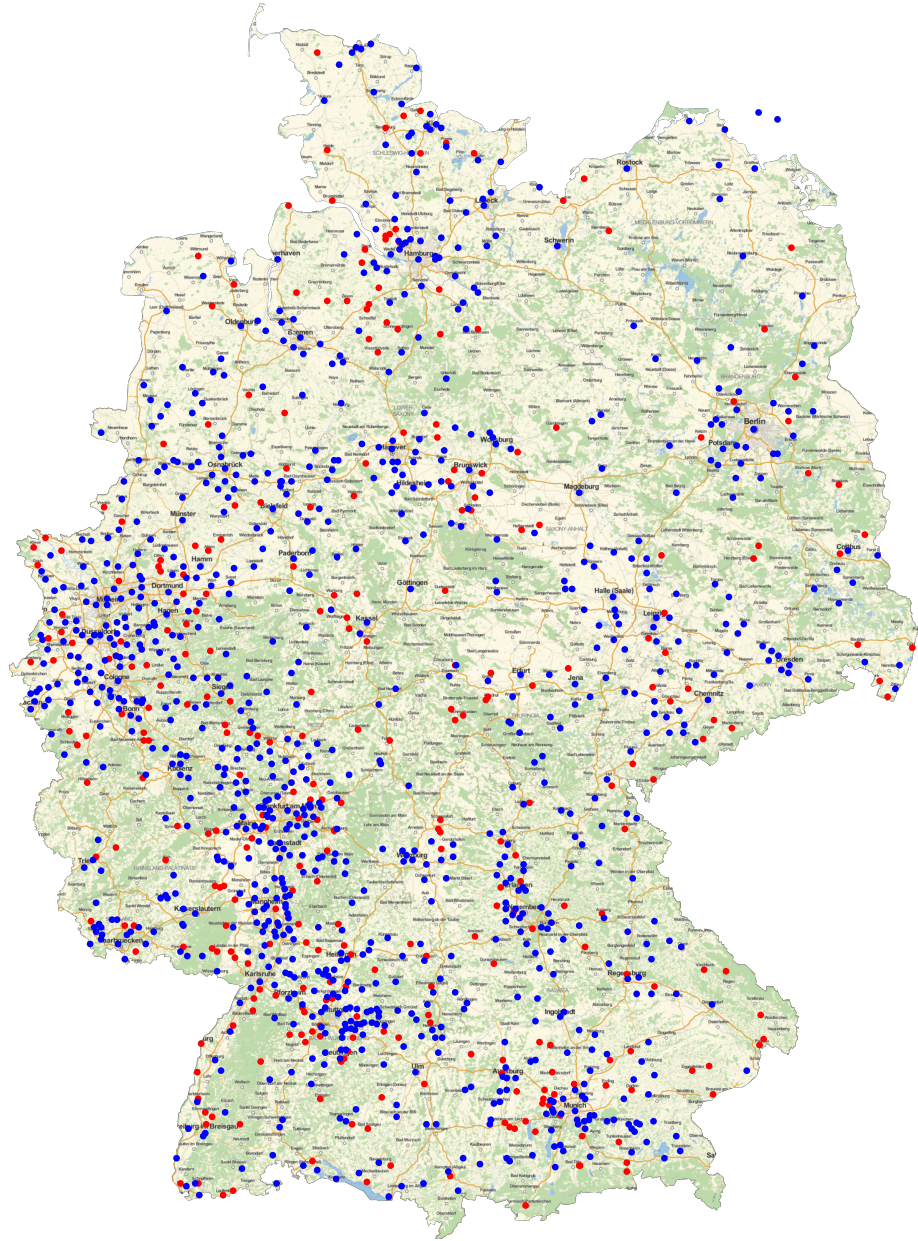


Figure 5.5.: Distribution of study participants in Germany based on their IP address. Locations dropped from the dataset after data cleaning are marked with a red dot; blue dots indicate locations that remained after data cleaning (T. D. Krafft et al., 2019).

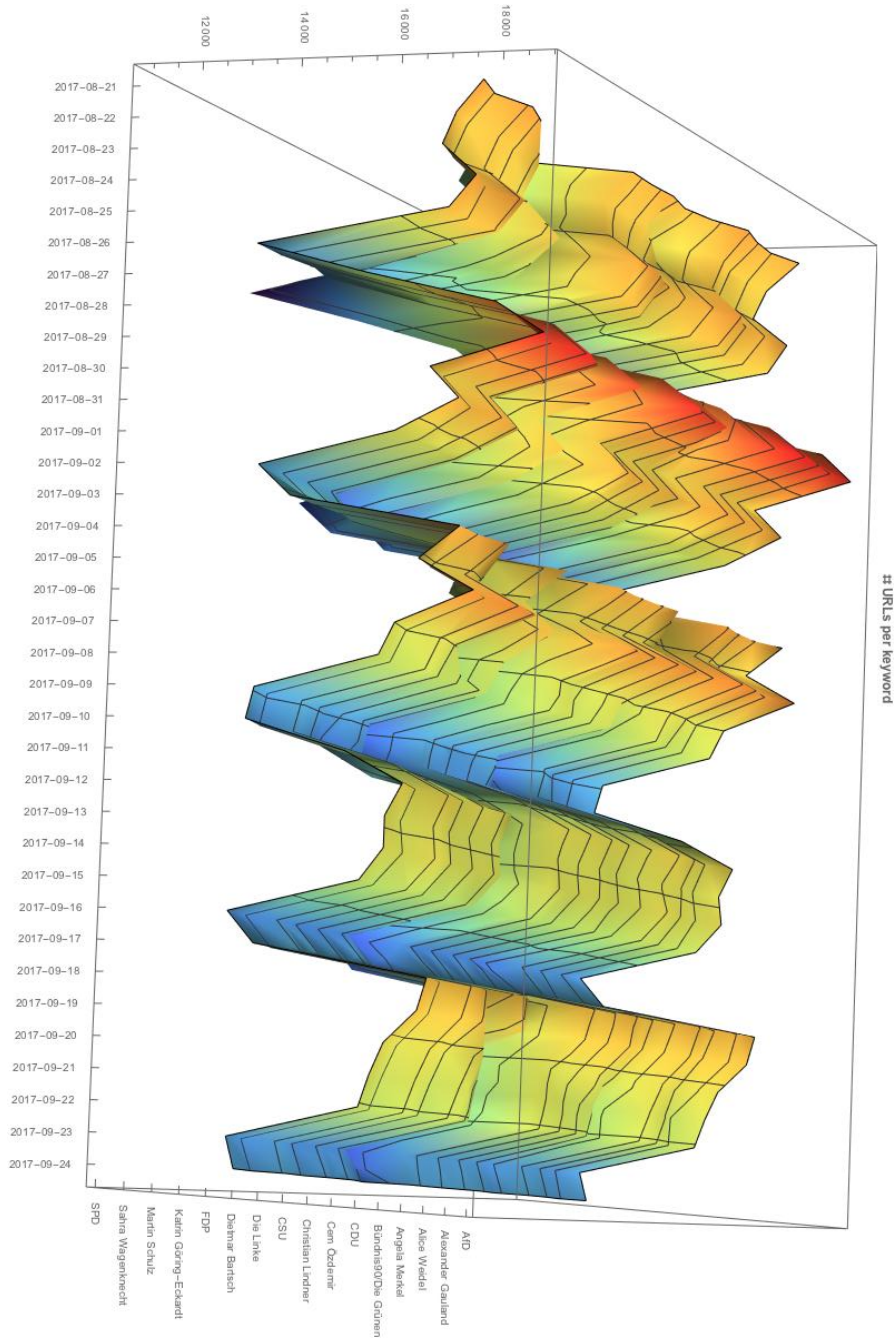


Figure 5.6.: Number of URLs per keyword and day. The x-axis shows the days for which the data was collected, the y-axis shows all search terms. The z-axis shows the total number of URLs for the specific day and search term (T. D. Krafft et al., 2019).

Based on the IP address allocation, two data sets on Berlin were compiled in preparation for a planned investigation into the possibility of regionalizing the delivered results lists. Even though regional assignment of IP addresses is no longer required, it was the only predictor of the data contributors’ location. Therefore, we initially compiled a dataset containing only German-language search results whose IP addresses were located in Berlin. This data set is referred to in the following as “Berlin”.

Since we discovered that Google does not regionalize based solely on IP address and that the IP address assignment is not precise enough, each URL in the Berlin data set was manually tagged for regionality. This allowed the Berlin data set to be reduced by eliminating URLs with no regional significance (Berlin regional). Table 5.2 provides a summary of the number of submitted search results and the number of submitters for the four data sets.

	All	German	Berlin	Berlin regional
Data records	3,707,302	3,287,401	249,928	220,863
Users	1,759	1,597	177	177
Results lists	315,197	276,276	20,990	20,906

Table 5.2.: Overview of the size of the data sets used.

#### 5.4.2. Data evaluation

This Section describes the data evaluation, beginning with the investigation of “commons”. The focus is on the extent to which searches conducted simultaneously with the same search term differ in the results displayed.

##### Commons

Calculated as the cardinality of the intersection of two search results lists, the commons are defined as an operator for precisely two results lists.

Let  $l_1$  and  $l_2$  be two lists of search results, then:

$$commons(l_1, l_2) := |l_1 \cap l_2|$$

The space for personalization in the search results, i.e., the number of potentially unique URLs for the user, is determined by calculating the commons for all tuples of a data set, separated by search terms and time points. To create comparability between the data sets, the average length of the results lists in each case was determined, and the difference between the number of common results (commons) and the average length of the results list in each data set was computed. Figure 5.7 displays the results for searches by person, while Figure 5.8 displays the results for searches by party.

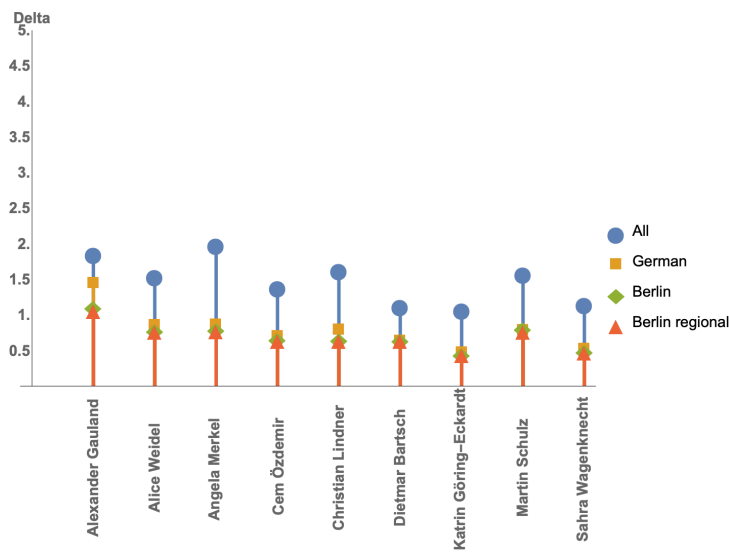


Figure 5.7.: Similarity measure commons for individuals: The space available for personalization is the difference (delta) between the average length of the results lists and the average number of commons per tuple for the persons (T. D. Krafft et al., 2019).

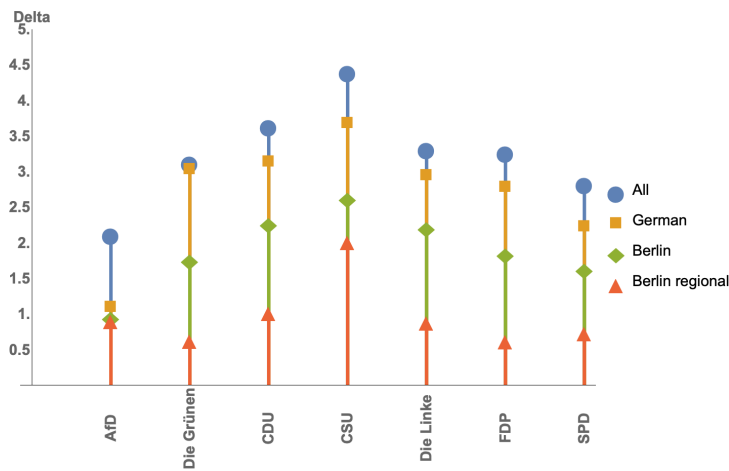


Figure 5.8.: Similarity measure commons for parties: The space available for personalization is the difference between the average length of the results lists and the average number of commons per tuple for the party in the respective data set (T. D. Krafft et al., 2019).

For party-related searches, it can be observed that the space for personalization decreases as search results are restricted to increasingly localized regions. The data set containing all data (all) has the highest values, while results from the Berlin region (Berlin regional) have the lowest; the results for the data set “German” are in the middle. Here, the “CSU” party, which is only active (and electable) in Bavaria, has the greatest room for personalization.

There is significantly less room for personalization in the case of searches for individuals. The results lists that each user receives at each timestamp differ by only one or two URLs. In searches for individuals, the drastic reduction in the personalization space caused by the restriction to the German data set are remarkable. When searching for parties, however, all three reductions of the data set size have an obvious impact. The maximum personalization value for parties in the entire data set is approximately 4.5 (CSU, see Figure 5.8); for individuals, it is approximately 2 (Angela Merkel and Alexander Gauland; see Figure 5.7).

Thus, the more a data set is restricted to local information, the less space is available for personalization.

Despite the limited space for personalization, as previously explained, different sorting of the displayed content may influence perception and, as a result, the formation of political opinion. Consequently, the metrics that incorporate the order in their evaluation were evaluated as follows.

### Deviation per rank

The deviation per rank is a discrete metric calculated on a set and was implemented as follows: For a set  $M$ :

$$\delta: M \times M \rightarrow \mathbb{R}, (x, y) \mapsto \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$$

Let  $L := \{l_1, l_2, \dots, l_n\}$  be a results list for a given timestamp and search query, then every entry in the results list consists of a list of (up to 10) URLs,

$$l_\nu = \{e_{\nu,1}, \dots, e_{\nu,k}\}, \quad k \leq 10$$

Then the deviation  $\Delta_k$  of rank  $k$  for  $1 \leq k \leq 10$ , is

$$\Delta_k = 1 - \frac{\sum_{1 \leq i, j \leq n} \delta(e_{i,k}, e_{j,k})}{n(n-1)}$$

The deviation per rank was calculated for the data sets All, German, and Berlin because the manual removal of “regional” links from the data set Berlin regional would have rendered this metric incomparable.



	1. Position	2. Position	3. Position	4. Position	5. Position	6. Position	7. Position	8. Position	9. Position	10. Position
<b>All</b>	22,78	16,67	40,71	52,37	57,15	61,11	65,62	66,48	65,99	75,06
<b>German</b>	6,66	11,89	29,30	43,55	50,00	54,55	59,98	61,34	60,76	65,39
<b>Berlin</b>	8,76	13,71	28,23	43,04	47,76	51,53	56,19	58,41	57,61	66,36

Figure 5.9.: Average pairwise deviation (in %) at each rank when searching for a person in the data sets “All”, “German”, and “Berlin” (T. D. Krafft et al., 2019).

	1. Position	2. Position	3. Position	4. Position	5. Position	6. Position	7. Position	8. Position	9. Position	10. Position
<b>All</b>	10,91	34,33	56,18	72,05	79,92	84,76	87,09	87,99	88,16	88,92
<b>German</b>	7,34	25,12	50,76	68,01	77,27	82,42	85,05	85,92	86,98	87,41
<b>Berlin</b>	3,17	25,21	44,95	61,17	68,20	72,19	75,35	75,87	76,39	75,24

Figure 5.10.: Average pairwise deviation (in %) at each rank when searching for a party in the data sets “All”, “German”, and “Berlin” (T. D. Krafft et al., 2019).

The manual review of the individual deviations at rank level revealed that they are very similar for searches for parties and persons; consequently, the calculated deviations for the individual criteria ‘search term’ and ‘search time’ average at rank level. Table 5.9 displays the deviations at rank level for person searches in the three data sets All, German, and Berlin; Table 5.10 shows those for parties searches.

A record in this table indicates the percentage of searches at a given time for a particular search term that differ at which position in the results list. It can be observed that, for all positions and search terms, the proportion of divergent results per rank decreases as local containment increases (from All to German to Berlin). While the mean values for searches for parties increase monotonically with the position in the results list, the mean values for searches for individuals in the data set ‘All’ vary in the top positions. The first place has a larger deviation than the second place. Moreover, the proportion of deviations increases to 75.06% for tenth-place individuals, whereas this value is already exceeded in searches for fifth-place parties and rises to a maximum of nearly 90% for tenth-place. In terms of individual positions, and thus in terms of sorting, the parties display a significantly greater degree of diversity. Even though both tables exhibit a clear increase in pairwise deviations, their slopes and maxima differ significantly.

### Longest Common Subsequence

As an additional metric for investigating personalization, the so-called “longest common subsequence” (LCS) was utilized. The length of the longest common substring between two strings is one of the most fundamental metrics for measuring string similarity. Given two (ordered) lists  $l_1$  and  $l_2$ , a sequence  $s := s_1, \dots, s_p$  is the longest common subsequence if  $s$  is a subsequence of both  $l_1$  and  $l_2$  and  $p$  is maximal. This measure of similarity was utilized to rank the search results for each search term and timestamp. This provides a more accurate view of sorting search results lists, as the order of the results is also considered. Figures 5.11 and 5.12 depict the outcomes of searches for individuals and parties within the data sets All, German, and Berlin.

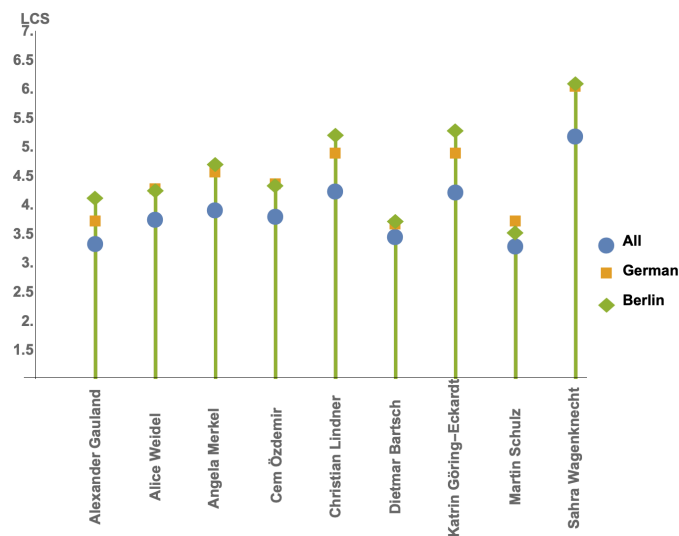


Figure 5.11.: Mean value of LCS for searches for persons in the data sets “All”, “German”, and “Berlin” (T. D. Krafft et al., 2019).

It can be seen that the average length of the LCS for individuals is significantly longer than for parties. This is consistent with the previous findings regarding the common similarity measure. The average length of the LCS tends to increase when more local data is considered, regardless of whether we are examining parties or individuals. Among all parties, the CSU, which is only active in Bavaria, has the shortest average length of LCS at less than two.

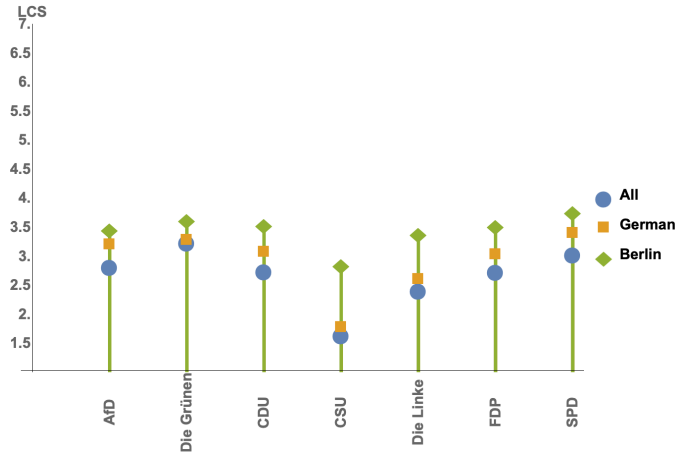


Figure 5.12.: Mean value of LCS for searches for parties in the data sets “All”, “German”, and “Berlin” (T. D. Krafft et al., 2019).

### Kendall’s Tau

Kendall’s Tau can be used to compare and measure the correlation between two ordered lists of values. Before calculating Kendall’s Tau, all possible pairs of values in the two lists must be considered. The pairs are then separated into concordant and discordant pairs. The elements (A, B) in the lists [A, B, C] and [A, D, B] are concordant because they are in the same order in both lists. A pair of elements in the lists is referred to as discordant if they appear in different orders. For instance, the elements (A, C) in the lists [A, B, C] and [C, B, A] are discordant because A comes before C in the first list and C comes before A in the second.

Kendall’s Tau  $\tau$  is determined as follows:

Given two lists  $x$  and  $y$  of length  $n$ , Kendall’s Tau is computed as follows. Let  $P$  be the set of all tuples  $(i, j)$  and  $1 \leq i, j \leq n$  then  $c$ , the number of *concordant* pairs, is defined as

$$c := |\{(i, j) \mid x_i < x_j \text{ and } y_i < y_j\}|$$

the number of *discordant* pairs is defined as

$$d := |\{(i, j) \mid x_i < x_j \text{ and } y_i > y_j\}|$$

In cases where elements of one list are not present in the other list, so-called “ties” are defined as follows:

$$n_x := |\{(i, j) \mid x_i = x_j \text{ and } y_i \neq y_j\}|$$

and

$$n_y := |\{(i, j) \mid x_i \neq x_j \text{ and } y_i = y_j\}|$$

then

$$\tau = \frac{c - d}{\sqrt{(c + d + n_x)(c + d + n_y)}}$$

Notably, Kendall’s Tau only considers the relative order of the values in the two lists and not their actual values.

In each instance, Kendall’s Tau is applied to all pairs of results lists with matching search time and search term from the data sets All, German, and Berlin; the individual results can be found in the appendix of T. D. Krafft et al., 2019. Due to the high degree of similarity between all search times, the mean value is calculated and plotted in Figures 5.13 and 5.14. As mentioned previously, the commons similarity measure has a limited scope for personalization, so users typically receive similar search results. The Kendall’s Tau measure, which evaluates the matching order of two distinct search results lists, demonstrates that not only do the results lists contain similar links, as determined by the commons similarity measure, but they also match to a high degree in their order. The majority of Kendall’s Tau values are close to 0.9. As shown in Figure 5.13 for individuals, this is particularly evident for the German and Berlin results. The search term Angela Merkel yielded a slightly different result, as she was the only candidate for the German federal elections who had greater international influence than the others. Figure 5.14 depicts the results for parties with Kendall’s Tau value. Similar to the evaluation of the LCS results, the Bavarian CSU party received different (lower) values than the other parties. The evaluation reveals that the mean Kendall’s Tau values for both parties and individuals are positive, indicating that the order of the results has a similar orientation.

In the next step, Kendall’s Tau was computed for all results lists and binning was performed to examine the dispersion of the Kendall’s Tau values. Figure 5.15 depicts the deviation of the Kendall’s Tau coefficients as a histogram. The x-axis displays the proportion of Kendall’s Tau values that fall within the respective interval  $[y, y + 1]$  on the y-axis. The three distinct Kendall’s Tau results are displayed for all results lists in the data set, grouped by results lists associated with Berlin users, German users, and all users. It is also evident that there are few differences between regional outcomes.

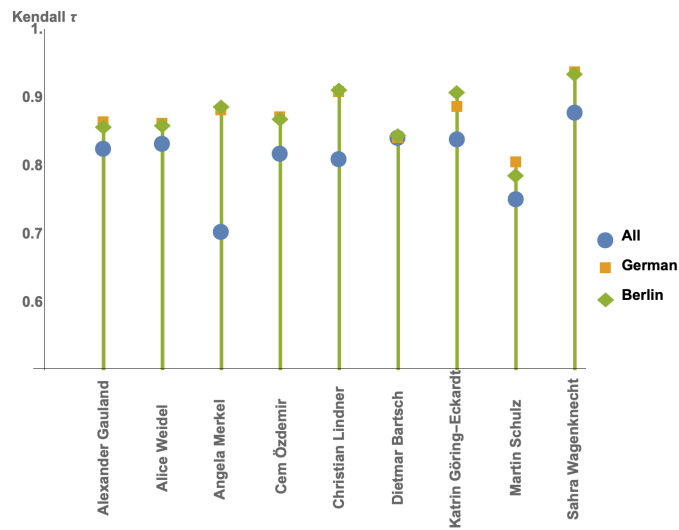


Figure 5.13.: The mean values of Kendall's Tau for all results lists in the data sets All, German, and Berlin for the search for persons are shown (T. D. Krafft et al., 2019).

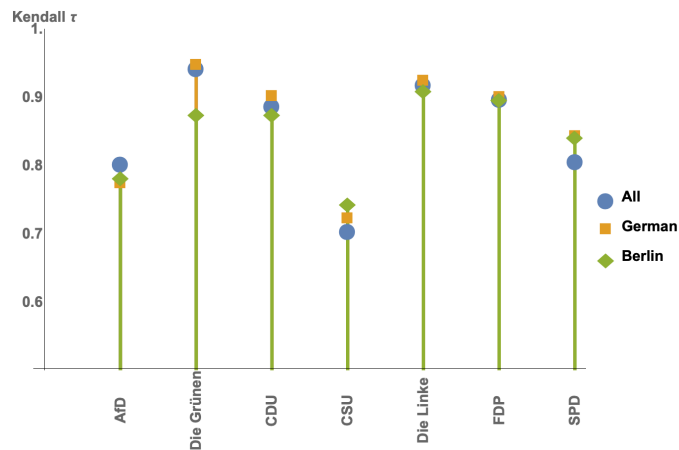


Figure 5.14.: The Figure shows the mean values of Kendall's Tau for all results lists in the data sets All, German, and Berlin when searching for parties (T. D. Krafft et al., 2019).

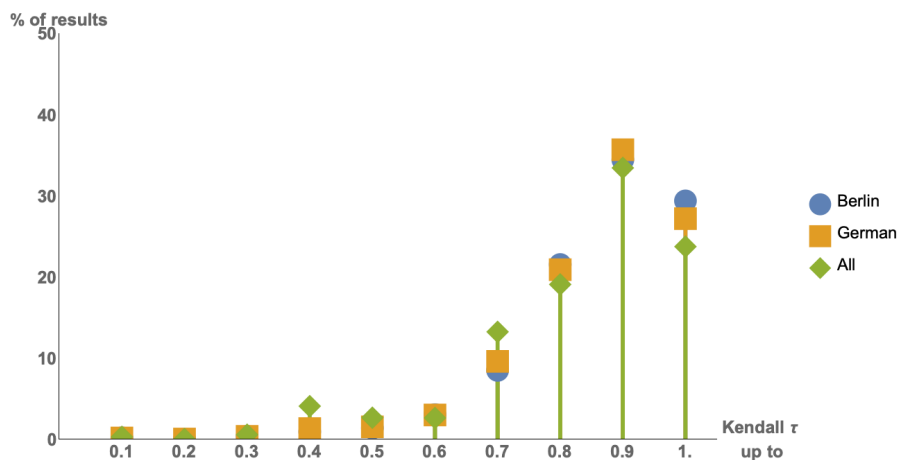


Figure 5.15.: Histogram of the occurrence of the respective Kendall’s Tau values in the preceding evaluation (T. D. Krafft et al., 2019).

### 5.4.3. Threats to the validity of the results

Drost, 2011, p. 114 et seqq. provides a schematic guide for the evaluation of social science studies to assess the validity of study results and the limits of their interpretation. This guide identifies numerous threats to validity that can be investigated to improve the interpretability of empirical results. Drost describes four types of validity that can affect a study’s reliability (for details, see Section 8.3.2).

#### Statistical conclusion validity

The statistical analysis employed in the study is referred to as statistical significance. Numerous variables can influence the statistical significance of research results. These include inadequate sample size, violation of assumptions that can render measurements unreliable, the possibility of random confounding factors in the experimental environment, and random differences between respondents (if humans were involved) that can bias the results (Drost, 2011, p. 115). In our study, the number of participants and thus the sample size was not controlled due to the self-selection of the study participants. Even though I personally believe that more than 4,000 participants constitute a large enough sample to investigate the effects, I am unable to cite any scientific evidence to determine whether this number of participants is sufficient. Similar to the Dieselgate scandal (L. Bovens, 2016), Google’s system could have identified the data donation as a “test” and treated the participants differently from the rest of the population. In theory, there is also the possibility that individuals could have examined the program code we published for the user plugin and submitted fake search results lists that appeared

authentic in order to falsify the study results. There were no indications of this, however.

### **Internal validity**

Internal validity refers to the extent to which a study's results can be genuinely attributed to the independent variable tested and not to other external factors. For this purpose, the temporal progression of the study must first be discussed. Throughout the duration of the study, millions of citizens initiated searches, so their queries may have been factored into Google's ranking of website relevance. However, because this effect is not a sign of personalization, the factor could have affected the results. In contrast, an example-based and manual evaluation conducted by me revealed that the results were very similar when comparing the beginning and the end of the data collection.

### **Construct validity**

Construct validity is the degree to which a measurement or instrument accurately reflects the theoretical concept or construct it is intended to measure. First, it is possible that the appropriate measures for determining selection and sorting were not selected when comparing the submitted search results lists. While a review of similar studies in this setting and academic research comparing ordered lists in information retrieval have been conducted, it remains possible that there may be better or more meaningful measures. We rely on our subjective perspective when assessing the results and determining whether they fall within the "high" or "low" range because we do not have a basis for evaluation through identical studies.

### **External validity**

External validity refers to the question of generalizability, or the extent to which the observed effect can be transferred to other areas, demographics, or situations (Drost, 2011, pp. 120(Campbell & Stanley, 1963, p. 5)). The self-selected group of participants may have resulted in "random" heterogeneity, so that the collected data might not permit representative statements about the German population. However, there are no indications of this, either.

## **5.5. Interpretation of the results**

The primary objective of the data donation project was to examine the level of personalization of Google search results. To this end, lists of results for each timestamp and search term were collected and analyzed using a variety of similarity measures in order to compare the results for each user and determine whether and to what extent search results are personalized for each user. First, the selection of content to be rolled out

was evaluated by determining the average number of common links displayed to users at each timestamp. This allowed calculating the average number of potentially personalized links presented to users. As mentioned in earlier sections, this number was small. For individuals, there were typically fewer than two potentially personalized URLs (see Figure 5.7), whereas for party searches, the personalization margin ranged from 0.5 to 4.5 (see Figure 5.8). Thus, the search results for people offered significantly less space for customization than those for parties. For the analysis with the commons similarity measure, the data was further divided into four data sets of decreasing size, demonstrating that the space for personalization decreases when the data set is limited to more local data (see Figures 5.7 and 5.8).

As stated in the introduction, the sorting of search results has a significant impact on the interaction behavior of users (see Cutrell & Guan, 2007; Granka et al., 2004; Jansen & Spink, 2006; Keane et al., 2008; Pan et al., 2007; Schultheiß et al., 2018); therefore, three key figures are examined in this Section. In conclusion, the three outcomes are always the same:

- The sorting is very similar.
- Searches for individuals are typically more similar than party searches.
- The more regional the data sets, the more similar the sortings.

Regarding the study’s underlying research question, it is possible to conclude from the data analysis that Google’s search results for the 2017 German federal elections had limited potential for personalization. It was discovered that the level of personalization decreases when data is restricted to a regional scale. This, in turn, suggests that Google’s search function does not support one of the central tenets of the “filter bubble” theory.

In addition, the presented methodology enables comprehensive investigation of online search processes without the need for algorithmic specifics. The study design demonstrates that it is theoretically possible for society to continuously monitor the level of personalization of search engines for any search terms. In order to establish similar trustworthiness, the general design may also be transferred to other intermediaries if suitable APIs restrict selective access to the study-relevant content. On Facebook, for example, this would include selective access to media news in a timeline or political election ads restricting access to friends’ private messages. A small study of our own on the rollout of posts from a Facebook page to followers failed because of this lack of insight (T. D. Krafft et al., 2020). The data donation project demonstrates that society can examine a significant algorithm for relevant phenomena that generate publicity and shape public opinion without understanding the algorithm’s underlying code.



## 5.6. Lessons Learned

During the analysis presented, additional findings emerged, which are briefly outlined below.

### 5.6.1. General Learnings

The large number of participants, the overwhelmingly positive responses on their own social media accounts when they reported on the project, and the responses of readers to reports from the project indicate that citizens are generally highly motivated to donate data to investigate opaque algorithms. However, the difficulties associated with the short-term and financially constrained approval of the project indicate that an exhaustive study with a representative sample must be prepared. Recruiting real individuals for research purposes also involves dealing with various challenges such as different devices, operating systems, browsers (that have different versions), and other software that may interfere with data collection. During our black-box analysis, we encountered several participants who were unable to install the plugin or had issues with sending data. In some cases, the plugin hindered the normal usage of their browsers by consuming excessive computing power. Unfortunately, we could not determine whether this was caused by their ad-blocking software. Additionally, we faced minor problems due to differences in the settings of the participants' Google user accounts, such as the preferred language or the number of search results displayed on one page.

### 5.6.2. Technical Learnings

On a technical level, the study revealed the significance of extensive software testing, the existence of adaptable software, and monitoring. Due to the short-term approval, we had very little time for the software development process, which resulted in inadequate preparation and execution of the software and integration tests. Therefore, it was not until after the first versions were released that we realized that all Firefox plugins had been submitting same ID; an error that could have been avoided by conducting a brief trial study under real-world conditions. Unfortunately, we were also unable to directly contact the study participants and request that they install a new, revised version of the plugin. As a result, we received numerous results that we were unable to use. A client-side, active search for updates, followed by a user request to install the new updates, would have been a simple solution at this point. In general, as much as possible must be modifiable centrally, and the client side must be flexible.

Google changed the structure and layout of their page shortly before the start of the project, necessitating adjustments in the area of parsing the search results page. Likewise, a functional update process would have been advantageous. According to Google and the market observers moz, Google releases multiple updates per day in an effort to

enhance their service and adapt to shifting search behavior (Illyes, 2017; Moz, 2023). Observers noted that the updates typically aim to optimize the search engine for user-focused, high-quality content, prevent malicious search engine optimization attempts, comprehend a searching user’s context and intentions, and increase the number of editable queries (Vinoth, 2017).

Another consideration is the active monitoring of search results pages. Since the software development process was outsourced and there was no direct monitoring access to the server, this also slowed down the error detection and correction procedure.

The collection of data presents a common issue due to the constantly changing nature of the content featured in ads or search results. Often, we encountered links from ads that were no longer valid at the time of analysis. To avoid this, it would have been more beneficial to crawl these links during the data collection process and save the relevant pages for later analysis. However, considering the prevalence of A/B testing (Kohavi & Longbotham, 2017), where different users are presented with different versions of a website, it would be necessary to follow the link from within the plugin. This means that the participant’s browser would open not only the Google web page but also any other web page advertised or displayed on the results page. This raises concerns about safety and data privacy that are difficult to address and may even be in violation of Google’s search engine service terms and conditions.

### 5.7. Further use of the collected data

During the course of the project, we conducted additional research. Some of the findings are introduced below. One of the main discoveries Katharina Zweig made is that, in the Google News data sets, there are consistently small clusters with significantly different results lists (T. D. Krafft et al., 2017, Section 4). These clusters only share an average of two or three results with other search lists, indicating that the algorithms behind the search engine can produce variable results for similar search queries. However, despite their relatively homogeneous nature, manual examination of these clusters did not reveal any explicitly political content. This lack of political material raises questions about the origins of these small clusters, which remain unknown at this time. Nevertheless, we believe that social and communication scientists could perform content-based analyses on either these clusters or the entire Google and Google News data set.

Numerous URLs, such as the respective Wikipedia entry or personal web pages, are permanently represented in the search results of the standard Google search engine. On Google’s news portal, the situation is significantly more dynamic: A URL is rolled out to users for an average of less than two days. However, isolated URLs achieve significantly longer display periods (T. D. Krafft et al., 2018c). Even on Google’s news portal, there is not much room for personalization; only 4-5 of the top 20 results differ on average (commons) (T. D. Krafft et al., 2018c).

The majority of first-page Google search results for parties led to websites whose content is managed by the parties themselves. 34 percent of the hits were websites of a party, party members, or local chapters. An additional 17% of the search results were social media profiles of the parties, whose content can also be controlled directly. 11% of the results contained links to Wikipedia entries that can be at least partially controlled. In contrast, media offers accounted for 26% of Google search results for parties (T. D. Krafft et al., 2018a, Section 3.4), indicating that traditional news sources are still relevant in the online political landscape.

The parties' ability to place self-controlled websites, such as their homepages or social media accounts, on the first page of search engine results varied. While more than 80 percent of the results for the search term "Bündnis90/Die Grünen" led to the party's own websites<sup>13</sup>, social media profiles, or Wikipedia entry, only 27 percent of the results for the search term "AfD" did so. The proportion of self-moderated results for "Die Linke" was approximately 82%, for "FDP" 75%, for "CDU" 63%, for "SPD" 52%, and for "CSU" approximately 52% (T. D. Krafft et al., 2018a, Section 3.4).

Therefore, it is crucial for future research to focus on the impact of search engines on political information dissemination and the formation of public opinion. Such research should examine how search engines influence the visibility and ranking of political content, and how political parties and interest groups utilize search engine optimization techniques to manipulate search results in their favor.

Overall, the insights gained from this research have significant implications for political communication and media studies, as they contribute to our understanding of the role of search engines in shaping the public's perception of politics and political actors. It is important to continue exploring these topics to ensure that the democratic process will not be undermined by biased or manipulated search engine results.

---

<sup>13</sup>For non-German readers, it may be helpful to note a possible explanation as to why the search results for "Bündnis90/Die Grünen" predominantly lead to party-affiliated websites: few sources outside the party itself use this long and formal name, much like with "Die Linke". In media reporting, for example, the party is often referred to simply as "Die Grünen". This could result in a high number of references to party-owned sites when searching for the official full name.

## Black-box analysis – Stem cell advertising ban on Google

Verifying the functionality of ADM systems for personalization is challenging because these systems rely on vast quantities of data and intricate algorithms to make decisions and predictions. Individual errors can be difficult to identify due to a lack of oversight caused by the increasingly granular personalization capabilities. In the area of ad distribution on search engines, the lack of transparency regarding who sees which ads can become a serious problem, as demonstrated by the following black-box analysis examination of ads for unproven stem cell therapy on Google Search.

The idea of the research project originated from Anna Couturier’s doctoral research at the University of Edinburgh and her role as Project Manager of the EU-funded public engagement initiative EuroStemCell, a biomedical researcher-led collaboration to improve public knowledge, patient decision-making, and researcher engagement around stem cell research. Through her work at EuroStemCell creating collaborative resources for patients and researchers regarding stem cell research and the movement of therapies from lab to clinic, larger questions arose around the dissemination of medical treatment information through digital channels, including the ubiquitous use and growing importance of algorithmically mediated platforms. Specifically, she explored secondary economies that have developed around emergent stem cell therapies, including the direct-to-consumer marketing of stem cell treatments on platforms like Google Search. This work was first and foremost motivated by feedback provided by high-risk stakeholders within the community including patients. Through her work in patient engagement, she noted that patient advocates from the Parkinson’s Disease and Multiple Sclerosis communities in the United Kingdom (UK) flagged personal anecdotes around the frequency of encounters with ads from private clinics offering unverified treatments while searching for information on their conditions. Some of these impressions were gathered at a workshop called “Patienthood and Participation in the Digital Era: findings and future directions”, which was funded by the Wellcome Trust Seed project and hosted by the Usher Institute at the University of Edinburgh in August 2018 (Erikainen et al., 2019). As a result, an initial investigation was conducted into the promotion of unverified stem cell treatments via advertisements in the United Kingdom, as documented in (Erikainen

et al., 2020). However, this inquiry did not encompass the most significant player in the dissemination of information, which is Google Search. This omission prompted Anna Couturier to collaborate with us to perform a black-box analysis to determine the frequency of these ads and whether they were specifically targeted at patients, as opposed to a control group of healthy individuals.

## 6.1. On the problem of stem cell therapies

This research was made possible through collaboration with the EuroStemCell project, but more specifically through the research and work of Anna Couturier, Project Manager at EuroStemCell, and carried out as part of her doctoral research in Science, Technology, and Innovation Studies at the University of Edinburgh. The EuroStemCell project is a network of over 400 stem cell researchers from the European Union that aims to bring stem cell researchers and the general public together through science communication, the development of educational resources, and training. The stem cell information website<sup>1</sup>, which is visited by a million people each year, is central to the project. Every year, thousands of patient inquiries about stem cell therapies are processed here, and the sheer volume of these inquiries demonstrates how many people with serious illnesses are seeking information about stem cell therapy. The extent of suffering in the search for ways to cure or at least alleviate pain can be gauged by reading the request our cooperation partner Anna Couturier received from a patient: “Don’t use lab rats – use me instead.” (Zarieczny et al., 2019, p. 1145).

To understand the intricacies of this case study and why it presents an environment susceptible to online misinformation, it is important to understand what a stem cell is. A stem cell is a unique type of cell that possesses two fundamental abilities (EuroStemCell, 2023a). First, a stem cell has the capacity to self-renew, meaning it can produce identical copies of itself. Second, it can undergo differentiation, which involves specializing into more specialized cell types. This process can be likened to the branching of tree roots. Just as a root can either continue as the main stem or branch out into specific directions, a stem cell can either generate additional stem cells or differentiate into cells with specific functions. To illustrate, consider bone marrow stem cells. These cells have the ability to replicate themselves or transform into blood cells, which play a critical role in the immune system. This is why bone marrow transplants from healthy donors are employed as a treatment for blood disorders. The main issues regarding stem cells revolve around the potential applications and, notably, the groundbreaking possibilities unlocked by the discovery of induced pluripotent stem cells (iPS) by Kazutoshi Takahashi and Shinya Yamanaka in 2006. These iPS cells are derived from adult stem cells and possess the remarkable ability to be reprogrammed to an earlier state and then guided to develop into specific types of cells (Takahashi & Yamanaka, 2006). To give an extreme example,

---

<sup>1</sup><https://www.eurostemcell.org>

envison the transformation of a skin cell into a neural cell.

This breakthrough has brought about two significant changes since 2006. First, it offered an alternative to the more traditionally “controversial” use of fetal or embryonic stem cell tissue, which has sparked more controversy in the United States than in Europe. Second, it opened up avenues for utilizing a patient’s own cells to generate healthy cells, eliminating the need for donor cells.

Therefore, and due to the increased understanding around how cells work, repair, can be genetically repaired or manipulated through intervention, the research into stem cells and their potential to treat various diseases is a rapidly expanding field of study. In the field of regenerative medicine, the utilization of stem cell therapies aims to replace, construct, or regenerate human cells, tissues, or organs in order to restore or establish their normal functions (Biehl & Russell, 2009). However, the problem lies in the fact that the practical application of these treatments (autologous stem cell transplant) for patients is significantly less advanced today than what is being portrayed by both public narratives and unethical private entities (Sipp et al., 2017).

Some regard it as highly modern and efficient, while others are more sceptical (Herberts et al., 2011; Nadig, 2009; Strauer & Kornowski, 2003). According to EuroStemCell, 2023b, there were relatively few stem cell-based treatments in human medicine at the time of the black-box analysis, such as bone marrow or haematopoietic stem cell transplantation for the treatment of leukaemia and other blood disorders (Buchholz & Ganser, 2009; Giralt et al., 2009). Their use in bone marrow transplants as an alternative treatment for aggressive multiple sclerosis, in which the patient’s immune system is reset, has also been investigated (Patani & Chandran, 2012). Since the 1980s, stem cells have also been used for skin transplants to treat common skin diseases (Hirsch et al., 2017). Stem cell therapy for the treatment of corneal injury is the first stem cell-derived therapy that has been shown to be safe and effective for a condition other than blood or skin. Since the European Union’s approval of this treatment in 2014, it has been utilized effectively to treat corneal damage. It has been shown to improve visual acuity, reduce pain and inflammation, and be a safe alternative to conventional corneal injury treatments (Knapton, 2014). Other than these treatments, which are specific to their biomedical use cases, no other stem cell-derived treatments have met the standard of clinical evidence for the treatment of other conditions or diseases (EuroStemCell, 2023b). For a detailed discussion of the issue of unproven stem cell therapies and stem cell information online, see the book “Stem cell tourism and the political economy of hope” by Petersen et al., 2017. Caulfield et al., 2016 report that despite the lack of established treatment protocols, private clinics have significantly increased their direct marketing of stem cell treatments for a wide range of conditions, including Parkinson’s disease, multiple sclerosis, and diabetes. Online marketing of private clinic treatments is feasible in the context of utilizing stem cells, particularly fat stem cells (so-called mesenchymal stem cells) obtained from the patient’s own adipose tissue, without significant genetic or cellular modification. Subsequently, the reintroduction of these cells into the patient’s

body is categorized as a cosmetic procedure rather than a medical intervention that must comply with regulatory guidelines (Berger et al., 2016). This distinction allows such clinics to function within the legal framework, although some establishments may operate in countries like Panama or Thailand, where more extensive cellular manipulation is permissible due to its legal status (Petersen et al., 2017).

This growth has been facilitated in part by the expanding diversity of digital advertising options, which will be discussed in greater detail below.

### **Online advertising for questionable stem cell therapies**

Over the past three decades, online advertising has evolved from simple static advertising spaces to integrated networks that deliver personalized multimedia advertising (Rashtchy et al., 2007). This new form has a number of advantages over conventional advertising. For instance, it is possible to optimize advertising through a variety of media and to reach target groups in a quantifiable manner, with (almost) no geographical limitations. Moreover, by targeting customers based on their characteristics, personalized advertising enables a targeted approach with custom-tailored offers (Rashtchy et al., 2007). As digitalization continues, the healthcare industry is also increasing its use of these new forms of advertising, so that the distribution of medicines derived from stem cells by medical research is complemented by an ever-expanding Internet-based direct market and patients have access to treatment options outside of the established local healthcare infrastructure, which are then, of course, also subject to other regulations. Consequently, marketing strategies have increasingly shifted to the Internet (Petersen et al., 2019; Tanner et al., 2019; Turner, 2017), with search engines, web advertising, and ad exchange (Google, 2023a, 2023c; Muthukrishnan, 2009) playing a significant role for stem cell therapy providers in achieving the highest possible rankings in organic search results (Rashtchy et al., 2007, p. 184). Direct marketing via search engines has emerged as the preferred method and most popular form of communication for businesses (Master et al., 2014; Yuan et al., 2012). It contributes to the “branding” of search terms, so that certain search queries are associated with a particular brand, product, or service (Rashtchy et al., 2007, p. 184). In the healthcare industry, this can result in certain clinics gaining legitimacy by being associated with relevant search terms, such as when their names appear in the top search results for stem cell treatments. While providers of controversial stem cell therapies argue that direct advertising guarantees patients’ freedom of choice and self-determination, critics note that the concept of informed consent is disregarded when patients make decisions based on unreliable and untrustworthy claims or offers (Turner, 2018). The companies in question publish actual data about their procedures and success rates only on occasion; therefore, online communication about stem cell therapies frequently lacks medical information and truthful claims about the specifics and efficacy of a treatment (Connolly et al., 2014). The

assertions of private providers of questionable stem cell therapies that they are registered or certified in some way, as well as their assurances that there are no ethical or health concerns, are not convincing. Despite the fact that many providers cite credible experts, memberships in professional associations, and testimonials and publications to bolster their credibility (Munsie et al., 2017), this is criticized from multiple perspectives. There are allegations that they would minimize risks, disregard warnings, and fail to document informed patient consent (Enserink, 2006; Master et al., 2014; Ryan et al., 2010). In addition, there is a lack of safety or efficacy evidence as well as comprehensive patient information and support (Enserink, 2006; O'Donnell et al., 2016; Ryan et al., 2010) for the majority of the advertised therapies, which are typically not proven at all or only through poorly designed clinical studies. In addition, it is often difficult for patients, who typically lack the requisite knowledge, to distinguish between reputable and questionable providers, considering that both utilize very similar advertising strategies and messages (Sipp et al., 2017). Given the size of the online advertising market, it is crucial to consider the possibility of false or deceptive advertising campaigns for commercial stem cell treatments with limited or no clinical evidence of their safety or efficacy (EuroStemCell, 2023b; ISSCR, 2023; Sipp et al., 2017). Due to the risks to patients and the lack of scientific evidence supporting such treatments, traditional medical institutions are issuing strong warnings following a number of cases in which patients suffered serious injuries or died as a result of these therapies in private clinics (Mendick & Hall, 2011) and are calling for stricter consumer regulation (see, e.g., C. B. Cohen & Cohen, 2010; Lysaght, Lipworth, et al., 2017; Regenberget al., 2009; Sipp et al., 2017).

Another significant concern highlighted by Petersen et al., 2017 is the substantial financial investment made by private clinics in search engine optimization and search engine marketing. These clinics allocate significant funds, often in the range of tens of thousands of dollars per month, towards these auxiliary industries with the explicit objective of targeting patients as consumers. As the world's largest provider of online advertising by a wide margin, Google deserves special consideration in this context, if for no other reason than its monopoly position. Google prohibits advertising in certain sensitive categories in principle (Google, 2023g).

Anna Couturier informed me early in 2019 that despite this general ban, advertising for unproven stem cell therapies was being implemented on online search platforms. However, quantitatively robust results were lacking. Google's above-mentioned statement prohibiting advertising for unproven stem cell therapies had not yet been published at the time, but the company's general stance was that no advertising for such therapies should be placed. Nonetheless, the following advertisements were discovered (see Figure 6.1).

However, there were no structured methods to collect and validate these anecdotal reports.

The majority of sociological research on stem cell therapies has focused on capturing the perspectives and experiences of patients contemplating or undergoing these treat-



## Stem Cell Treatment | Swiss Medica Clinic

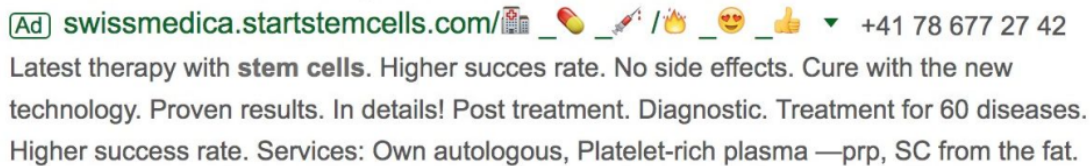
  
Latest therapy with **stem cells**. Higher succes rate. No side effects. Cure with the new technology. Proven results. In details! Post treatment. Diagnostic. Treatment for 60 diseases. Higher success rate. Services: Own autologous, Platelet-rich plasma —prp, SC from the fat.

Figure 6.1.: Advertisement of a questionable provider of stem cell treatments. This screenshot was provided to me by Anna Couturier.

ments. Decision-making, perceptions of risks and benefits, and experiences with the healthcare system have all been examined (Langstrup, 2011; Petersen et al., 2014). Access to stem cell therapies has also been studied, including the phenomenon of “stem cell tourism”, where patients travel to other countries for treatments not yet approved or available in their home country (Petersen et al., 2017). In addition, researchers have investigated the regulatory environment of stem cell therapies by mapping country-specific approaches to monitoring private clinics offering stem cell therapies, including international comparisons (Lysaght, Kerridge, et al., 2017) and country-specific accounts of regulatory oversight of stem cell clinics (Sipp & Turner, 2012).

On the technical side, there have been studies on the personalization of Internet advertising (Barford et al., 2014) and the relationship between the recording of user behavior in the browser (user tracking) and Google advertising (Datta et al., 2015). Latanya Sweeney, 2013 investigated the discriminatory potential of rolling out ads, highlighting the problem that online ads indicating criminal records appear more frequently to people with “black-sounding” names than to people with “white-sounding” names. She was unable to examine the system directly as a white box, so in a combination of crowd-sourced audit (in which she was the only participant) and sock puppet audit, she sent over 2,000 search queries with different names to Google and Reuters to evaluate the advertisements displayed.

What has been missing is research specifically examining the impact of digital search platforms on stem cell treatment knowledge and accessibility. As the availability of information on digital platforms can have a significant impact on how and where patients seek and ultimately use medical treatments, more research is required in this area. While digital search platforms make it easier for patients to locate information about unproven stem cell therapies, they can also make it challenging to distinguish between credible and unreliable sources of information. Therefore, a better understanding of the role of digital platforms in the decision-making process of patients could provide valuable insights.

## 6.2. Socioinformatic analysis

A socioinformatic effect structure is used in the following to demonstrate how the incentive structures of the involved actors continue to trigger the socioinformatic phenomenon of advertising for unproven stem cell therapy offers.

The effects depicted in Figure 6.2 can be deduced from the analysis presented in section 5. The business model envisioned by Alphabet Inc. is depicted on the left. It consists of the combination of the knowledge gained through Google Search and its own advertising network AdSense<sup>2</sup>. As stated in section 5, Alphabet Inc. collects not only previous search queries through its Google search engines, but also a variety of attributes that are used to construct user profiles. Online advertising employs “Information Retrieval, Machine Learning, Data Mining and Analytic, Statistics, Economics, and even Psychology to predict and understand user behavior” (Yuan et al., 2012, p. 1) to predict and comprehend user behavior based on the quantity of captured behavior (variable 1). Thus, as behavior is captured, the granularity of user profiles also increases (variable 2). The AdSense advertising network enables advertisers to rent advertising space on numerous websites with relative ease (Google, 2023e). Here, advertisers can plan campaigns and select extremely specific target groups. The advertisers might select a specific audience based on a number of factors such as demographic characteristics, affinity, purchasing interests, particular behavior, or similarity to another audience. Additionally, it is possible to target individuals based on the topics and content of the websites that appeal to them or the keywords they have entered. Furthermore, users can be targeted in certain contexts of the searching person, for instance, during a specific life event (like marriage), or situational context of the search (time of day, location, mobile device usage) (Google, 2023b, 2023h). In this process, advertisements can be arbitrarily or deliberately placed on specific websites, applications, or media platforms (Google, 2023d).

While a simple advertiser can only limit the advertisement’s target audience by 20 to 30 characteristics, the implications of Google’s expansion of its primary profit mandate around the provision of relevant offers for business customers suggests that there are additional modes of audience targeting within its platform model. As a broker concerned with supply and demand, AdSense calculates matching based on keywords and search terms, website content, and user data (Google, 2023f). Since AdSense acts as an intermediary, specific ad targeting is possible, either based on website information, such as topic or target audience, or on user characteristics (Guha et al., 2010; Mayer & Mitchell, 2012). Consequently, the granularity of the user profiles improves the precision of the advertising spaces (variable 3). The greater the relevance of the ads to the searchers’ interests, the greater the likelihood that searchers will click on an ad (variable 4). Since the operator of the AdSense advertising network is interested in automatically increasing revenues through the sale of advertising space and through high click

---

<sup>2</sup><https://www.google.com/adsense/>

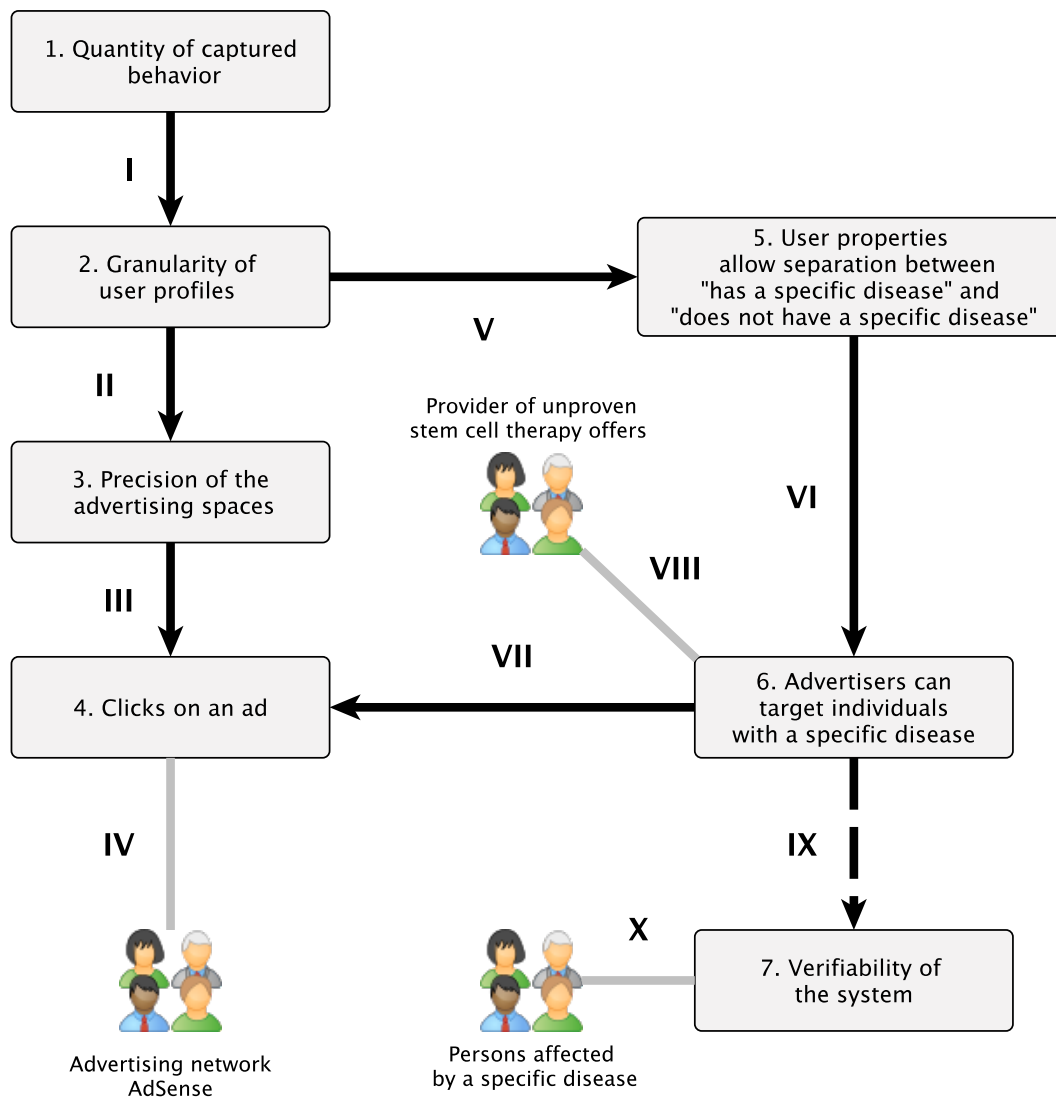


Figure 6.2.: Effect structure of the socioinformatic system of advertising network, providers of unproven stem cell therapy offers, and persons affected by a specific disease.

rates, they strive to continuously improve the granularity of user profiles, as it has been demonstrated that the use of behavioral targeting in advertising significantly increases click-through rates among similar users of a specific target group compared to controls without targeting (Yan et al., 2009). However, there is a risk that disease patterns can be mapped in user profiles (Edge V). This is probably not the case because, as far as I am aware, advertisers do not specifically target people with diabetes, for example. However, through interactions with the rolled-out advertisement for a therapy against disease X, they discover the characteristics of the groups of individuals interacting with this advertisement. Consequently, signals that indicate a disease pattern are implicitly recorded, and advertisers can target individuals with the disease despite the absence of an explicit disease pattern (variable 6). AdSense allows providers of unproven stem cell therapies not only to target their advertising offers based on user profiles, but also to select specific search terms for which the ads are to be displayed. Verifiability (variable 7) suffers as a result of this individual addressing of advertising because those affected cannot currently prove which advertisements they have received and why. Thus, there is a number of ways that advertisements for unproven stem cell therapies can be distributed to patients with relevant medical conditions. There is no direct way for affected users to determine which path led to the display of a particular advertisement. Nonetheless, it is vitally important for the discourse between those affected and Alphabet Inc. to collect credible evidence of the occurrence of particular advertisements. Only then can the risks and effects on the unique socioinformatic system of patients with Parkinson's disease, multiple sclerosis, and diabetes be addressed.

In the following, our black-box analysis of Google advertising will be presented, with an emphasis on unproven stem cell treatments. The focus of this Section is not on the differences between the displayed posts, but rather on the documentation of the advertisements deployed in each instance. Our project's objective was to document the enforcement and efficacy of Google's ban on unproven stem cell therapy (Biddings, 2019; Google, 2023g) in order to provide domain experts with the evidence required to encourage further research on the impact of Google's ADM-based advertising modalities on end-user outcomes, assuming appropriate results. To this end, the study investigated the potential risk posed to vulnerable patient groups when they search Google for health-related information. To accomplish this, the following research questions were formulated:

**Research question 1 (RQ1):** Does this type of advertising continue to appear despite Google's ban on unproven stem cell therapies?

**Research question 2 (RQ2):** Are advertisements for unproven stem cell therapies significantly more prevalent among individuals with Parkinson's disease, diabetes, and multiple sclerosis?

To protect its users from potentially harmful offers, Google announced in September 2019 that, effective October 1, 2019, it would explicitly prohibit advertisements for

stem cell-related experimental medical treatments (Biddings, 2019). The timing of this announcement was shortly before the initiation of our study, so we plan to examine and compare the patterns of rollout behavior both prior to and subsequent to the mentioned date.

### 6.3. Conception of the black-box analysis

Before beginning an elaborate investigation of a technical system as a black box, it is critical to ascertain whether any assertions leveled against Google can be substantiated or refuted by scholarly articles or insights into the deployment of advertising. However, this assessment was hindered by Google's opaque practices concerning the dissemination of advertisements from their proprietary advertising network, AdSense, within their search engine infrastructure. Determining the black-box scenario relies on similar data to that used in the black-box analysis in section 5. Google does provide a rough overview of the parameters that are used to roll out personalized ads (Google, 2023g) by showing users very vague information about why they see a specific ad (Google, 2023i). From a researcher's perspective, however, it is unclear whether, in addition to the controlled parameters, the search query with all of its metadata about query time, people requesting other unknown information, etc. is used to determine the advertisements displayed. As depicted in Figure 6.3, it is reasonable to infer that besides the search term and certain meta-information pertaining to the query, such as the identity of the user conducting the search (e.g., the logged-in Google account) or the search history accessible to Google, additional undisclosed parameters are employed by Google to determine appropriate search results and advertisements. These supplementary parameters remain unknown to external observers. Although it is possible to examine the relationship between the search term/metadata and the advertisements displayed by the system, it is not possible to identify or modify the opaque input parameters used by the system. According to section 8.2.1, Classification of black-box scenarios, this is also a second-category black-box scenario.

As stated previously, the two underlying research questions seek to determine whether advertisements for unproven stem cell therapies continue to be distributed despite Google's ban (RQ1) and, if so, whether these advertisements are displayed specifically to individuals with relevant medical conditions (RQ2).

While the first part of the investigation is a straightforward oracle-based analysis (see Section 6.4), the second part is a sensitivity analysis focusing on the searching user's health. Figure 6.4 depicts the research strategy that combines the two types of analysis. While it is sufficient for an oracle-based analysis to examine RQ1 to determine whether the advertisements displayed contain ads for unproven stem cell therapy offers, RQ2 requires a sensitivity analysis. Similar to the donation of data for the 2017 German federal elections in section 5, it is necessary to establish study groups and compare the

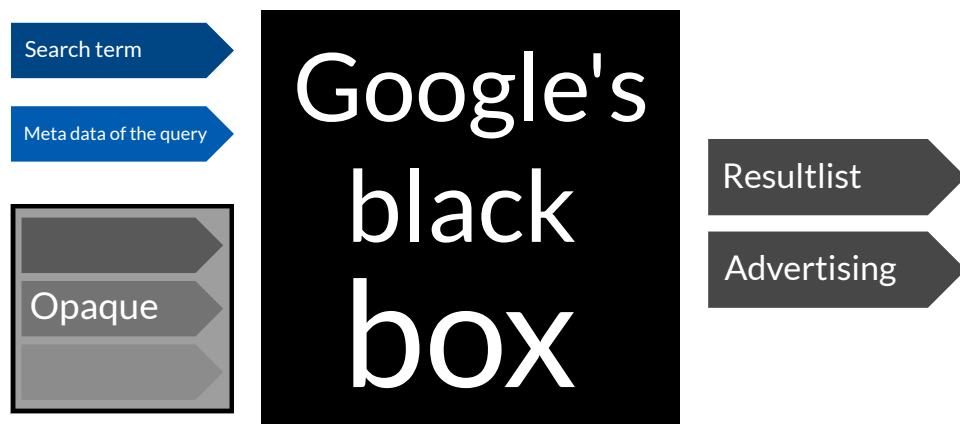


Figure 6.3.: The examination of the black box of advertisements displayed by Google represents a black-box scenario of the second category (for details, see section 8). For systems in this category, there is knowledge about the system’s output and part of the input, and there are also indications of further input, e.g., an individual personalization vector.

results of these two groups. In order to determine whether the characteristic “is ill or a carer”<sup>3</sup> actually increases the proportion of advertisements for unproven stem cell therapy products. In the black-box analysis described below, the displayed advertisements of both subgroups are evaluated and compared to determine whether users with one of the diseases (Parkinson’s disease, multiple sclerosis, or diabetes) are exposed to more critical advertisements than members of a control group (see Figure 6.4).

According to Sandvig et al., 2014 (see Section 2.4.3), the form of an analysis has a direct impact on the audit procedures that can be selected. Since the objective was to send identical inputs to the black box and, thus control as many parameters as possible, including search time and search term, it was impossible to conduct a non-invasive user audit to evaluate “natural” user behavior. Specifically, the second research question (RQ2) could be verified using a sock puppet or crowdsourced audit. However, the profiling required for a sock puppet audit, as in the case of the data donation for the 2017 German federal elections, posed too many difficulties for us. Since it is unknown how exactly Google processes the user behavior of a logged-in user, it is impossible

<sup>3</sup>The so-called “carers” are also included among the ill. “A carer is anyone who looks after a family member, partner, or friend who needs help because of illness, frailty, disability, a mental health problem or addiction and cannot cope without their support. The care they give is unpaid.” (as defined by the National Health Service England: <https://www.england.nhs.uk/commissioning/com-m-carers/carers/>, last accessed on 28.05.2023)

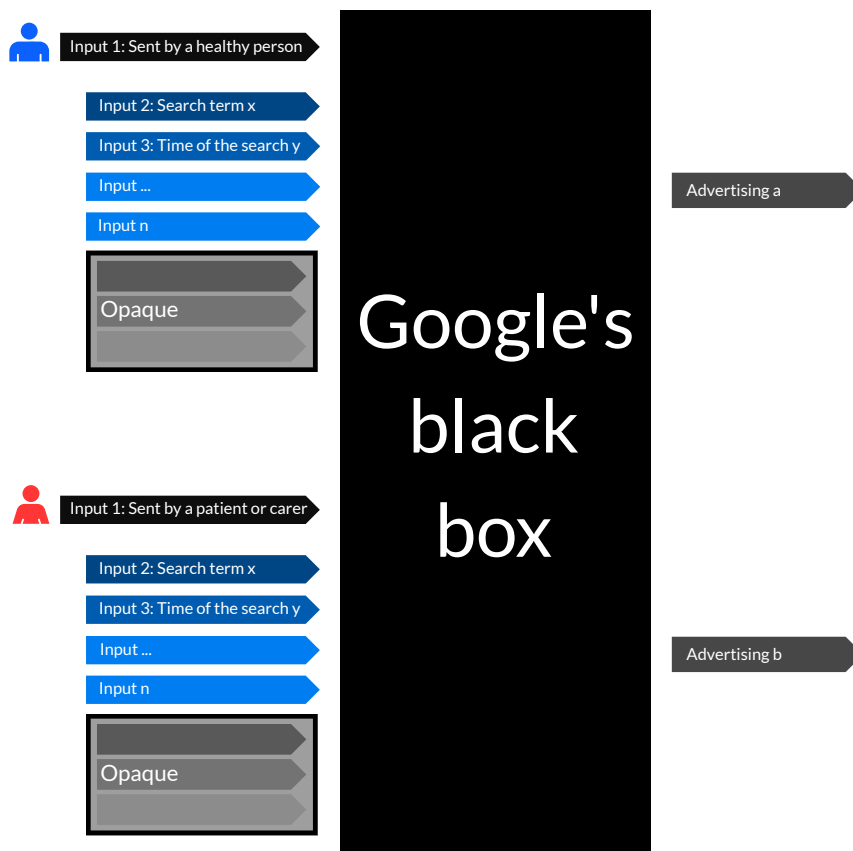


Figure 6.4.: Overview of a sensitivity analysis of Google’s advertising campaigns. While the inputs 2...n remain unchanged, the independent variable “sent by” is altered in each case by sending search queries from healthy users and ill users/carers to the black box to determine whether advertisements differ on the resulting search results pages and, if so, how frequently advertisements for unconfirmed stem cells are included. In addition to the known inputs, however, there may be others that cannot be influenced in the present study design.

to state with certainty which user behavior must be simulated in order to create user profiles on Google’s pages that are sufficiently accurate for such a study.

In the context of a crowdsourced audit, we had to rely on individuals actually affected by these diseases Anna Couturier was particularly helpful in this regard. In her role as Project Manager at EuroStemCell, she recruited actual patients and their families

to participate in the data donation by using her connections and approaching them through the project. Thus, she was able to connect us with patients in Canada, the United Kingdom, Australia, and the United States. Since it was assumed that the ads displayed are influenced by the geolocation of the search due to different national laws as well as the target groups selected by the advertisers, we formed the study groups according to the countries and divided them by the three disease conditions diabetes, multiple sclerosis, and Parkinson's disease. This resulted in the following twelve study groups:

- Canada + Diabetes
- Canada + Multiple Sclerosis
- Canada + Parkinson's
- UK + Diabetes
- UK + Multiple Sclerosis
- UK + Parkinson's
- Australia + Diabetes
- Australia + Multiple Sclerosis
- Australia + Parkinson's
- United States + Diabetes
- United States + Multiple Sclerosis
- United States + Parkinson's

As the objective of RQ2 is to make a direct comparison between healthy people and ill patients/carers in terms of the advertisements rolled out, a suitable control group was required for each study group. Therefore, participants were sorted into these control groups if they reported not having a relevant illness and not providing care for anyone with these conditions. We attempted to have 50 participants in each of the control groups when distributing the participants. We started by "filling" the Parkinson's disease control group, as Anna Couturier's primary research objective was to comprehend the online advertising landscape in relation to Parkinson's disease. The objective was to create a "healthy" control group for each of the twelve study groups.

Based on the experience from the project on the 2017 German federal elections, a way was sought to compensate for irregular submission behavior of real study participants. Therefore, we additionally rented Virtual Private Servers (VPS) in all four countries so



that identical search queries could be sent to Google synchronously. We conducted an additional scraping audit as the machines sent queries to Google without any search history. Three machines were used per country, and one was added in Florida because Anna Couturier observed a high concentration of stem cell therapy practitioners there. Thus, a total of thirteen VPSs regularly submitted search results to us. The servers were monitored and updated frequently. In addition, a server-side logging process was implemented to expedite troubleshooting in the event of anomalies. Due to the fact that Google does not offer an API for search and advertising, we were forced to conduct our investigation through Google’s web interface. For this purpose, the plugin from the 2017 German federal elections data donation was adjusted and extended. A client-server infrastructure consisting of Internet browser plugins that could be installed on the participants’ computers and a central server to receive and store submitted requests was required to conduct the study. Anna Couturier, Roman Krafft, Martin Reber, Katharina Zweig, and I conceived the black-box study in close collaboration, with me overseeing the technical direction and contributing the conceptual foundation. The programming of the server-side infrastructure in the development of the technical components was handled by Roman Krafft, a student assistant at our Chair who was supervised by me; the client-side browser plugins were developed by Martin Reber, who adapted the idea of the plugin from the data donation of the 2017 German federal elections as part of his Master’s thesis, and the data evaluation was validated by me. In the end, the findings were published in a collaborative work (Reber et al., 2020). Martin Reber developed the browser plugins for the latest versions of Google Chrome and Firefox, as these were the most popular Internet browsers in 2019 (Statista, 2023b) and their respective plugin stores provided a professional and trustworthy platform for the distribution and installation of our plugins. The development procedure is described in detail in Martin Reber’s Master’s thesis (Reber, 2020). At this point, only a brief overview of the plugins’ behavior as well as relevant interface design decisions will be provided.

Since the study required recruiting patients with Parkinson’s disease or one of the other diseases mentioned above, we first sought to gain a better understanding of the demographic characteristics of this patient population. This led us to conclude that the average age at Parkinson’s disease diagnosis, which is over 60 years (Pagano et al., 2016), may pose a challenge for potential participants, as they may have limited technological knowledge and low willingness to engage with complicated software. In order to mitigate this problem, our strategy focused on enlisting individuals specifically afflicted by early-onset Parkinson’s Disease (typically between the ages of 30 and 50). This recruitment had been carried out in collaboration with the UK-based entity Spotlight YOPD (Young Onset Parkinson’s Disease)<sup>4</sup>, Parkinson’s UK, and the Edinburgh Parkinson’s Research Interest Group as well as partners within the EuroStemCell network (see Couturier, 2023). Moreover, the specific condition of the participant might also affect their physical

---

<sup>4</sup><https://spotlightyopd.org>

## Privacy Statement

The purpose of this statement is to ensure that you have read and understood the information about the study and are fully aware of your rights should you decide to take part. If you would like to take part, please indicate this by reading the following questions. Consent is required in order to download and install this plug-in.

**Please note: since data is anonymised at the point of contact, we cannot retroactively withdraw any data collected before uninstallation.**

### Declaration of consent:

With the installation of the plugin I confirm that the first 10 search results and ads of Google on Google.com (in "All") including the above mentioned additional data (plugin ID, time, exact query, approximate location) on my browser will be made available to the public under a CC-0 license for analysis. I understand that the plugin regularly searches for all healthcare related queries listed above.

I agree that the following data will be collected, processed and published:

- A general location, derived from the IP address, corresponding in precision to approximately your postal code.
- The plug-in ID
- The exact search query
- The time of the search
- The result of the search (everything on the first page of the search results)
- The language setting of the browser
- Whether you are logged in as a user with Google.

I can disable or de-install the plugin at any time. No further data is sent with the deactivation of the plug-in or with its de-installation.

**By downloading the plugin, I confirm that I have read and understand the privacy statement, [www.eurostemcell.org/datadonation](http://www.eurostemcell.org/datadonation), for the above study. I have had the opportunity to consider the information and ask questions which have been answered satisfactorily**

**I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason.**

**I understand that any personal data collected during the study will be treated with confidence and handled in accordance with the Data Protection Act 1998 and GDPR.**

**I agree to take part in the above study.**

Figure 6.5.: This privacy statement was displayed on the website where we published the plugin. Screenshot from the website of the study taken on 27 January 2023.

ability to operate a computer. Common indicators of Parkinson's disease include shaking, difficulty initiating movements, and muscle stiffness (NHS, 2023). Studies indicate that people with poor health have significantly less Internet experience and are less likely to

use the Internet (Houston & Allison, 2002; Li et al., 2016).

We aimed to make the process of enrolling in the study as simple as possible. A frequently asked questions section<sup>5</sup> was provided on the study page, and a physical onboarding event in collaboration with Parkinson’s UK<sup>6</sup> was held whenever possible to explain our intent. Once they had downloaded the plugin, participants were guided through a straightforward registration procedure, beginning with a request to accept the Data Protection and Privacy Statement, which was also published on the website of the study<sup>7</sup> (see Figure 6.5). Afterwards, a questionnaire was displayed that requested socio-economic information required for study group assignment and subsequent evaluation (see Table 6.1). The plugin was always active whenever the browser was open and the computer was connected to the Internet. Figure 6.6 depicts the processes during a data donation executed by the plugin. At the specified search times (12 am, 4 am,..., 8 pm), the plugin sent the search terms we had previously selected to the Google search engine. In the course of this, the website [https://www.google.\[toplevel\]/search?q=\[term\]](https://www.google.[toplevel]/search?q=[term]) was requested. [term] was substituted with the search query, and [top level] corresponded to the respective top-level domain of the participant’s study group’s region.

Thus, the interaction with the browser window was significantly simplified compared to the first data donation (see Section 5), where the plugin was required to actively fill in the search field on the search engine page and click the “search” button. This change was intended to make the plugin unobtrusive and not have it interfere with normal browsing. This allowed us to collect data in inactive tabs of the current browser window, so users were not “forced” to watch and interrupt their work each time their browser conducted a search. Even though this process was relatively quick (10-20 seconds for 20 searches), participants in our study on data donation during the 2017 German federal elections (see Section 5) found this interruption to be very upsetting. Despite the searches in inactive tabs, an attempt was made to provide transparency by providing a tool that displays the most recent contributions so that users could visualize their contribution to the study.

The selection of the search queries was based on the following criteria and considerations:

- These are frequent search queries associated with the topic we are researching.
- They include stem cell and therapy as well as the names of the respective diseases (Parkinson’s, multiple sclerosis, diabetes).
- We also included “natural-sounding” questions because we assumed that older searchers, in particular, would ask direct questions to search engines if they do not fully comprehend how search engines function. This assumption is based on

---

<sup>5</sup><https://www.eurostemcell.org/datadonation#paragraph-1575>

<sup>6</sup>Parkinson’s UK is a Parkinson’s disease research and support charity in the United Kingdom: <https://www.parkinsons.org.uk>

<sup>7</sup><https://www.eurostemcell.org/datadonation#paragraph-1576>

Table 6.1.: This questionnaire was displayed by the plugin during onboarding.

Number	Question	Possible answers
1	Are you or someone close to you impacted by {condition}?	I am a patient. I am a carer.*
2	Are you a stem cell researcher or medical professional?	Yes No
3	What is your country of residence?	United Kingdom United States Australia Canada
4	Age:	18-29 30-39 40-49 50-59 60-69 69+
5	Gender:	Male Female Other I prefer not to say
6	How often do you use your computer?	Daily (More than 2h a day) Daily (Less than 2h a day) Weekly Monthly
7	How often do you use Google Search?	Daily (More than 2 times a day) Daily (Less than 2 times a day) Weekly Monthly
8	Have you ever paid for or inquired about stem cell treatment?	Yes No

research indicating that both age and search engine experience influence the way people use search engines (I. Weber & Jaimes, 2011).

In total, we selected the following 14 searches, where [disease] was replaced with the disease of the respective study in each case:

- stem cells
- stem cells cost
- stem cells treatment
- stem cells cure
- stem cells therapy
- can stem cells help me?
- can stem cells cure [disease]?
- [disease] cure
- [disease] therapy
- [disease] treatment
- [disease] cells cost
- [disease] stem cells treatment
- [disease] stem cells cure
- [disease] stem cells therapy

Similar to the data donation for the 2017 German federal elections (see Section 5), when the plugin was activated, it immediately began querying and submitting search queries. Consequently, there are search results with varying timestamps. As shown in Figure 6.6, the Google results page displayed in this study was evaluated on the client side, and only relevant HTML components were sent to the server. This procedure differs from the first study in that less information is transmitted to the server. This reduces the possibility of unintentionally processing and storing personal information, such as Google account information. Only the search results, top stories, and advertisements displayed were therefore submitted:

- Advertisements
  - Name

## 6. Black-box analysis – Stem cell advertising ban on Google

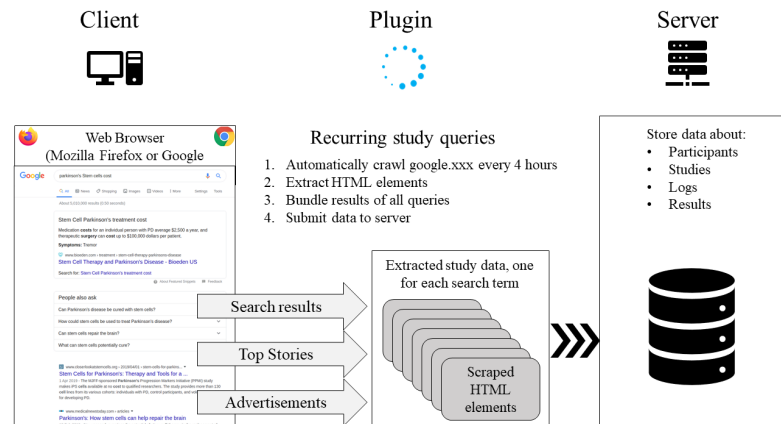


Figure 6.6.: This Figure shows the communication process between the client with the plugin installed and the server in the EuroStemCell data donation. Figure from (Reber, 2020), with permission from Martin Reber.

- Title
- URL
- Content
- Search results
  - Title
  - Content
  - URL
  - Position
- Top stories
  - Title
  - Author
  - URL
  - Position

After all searches were completed, the plugin transmitted the results to the server along with administrative and statistical data (ID, version, time, language) pertaining to the participants and the plugin. To ensure transparency and reproducibility, Martin

Reber uploaded the plugins' source code to a public repository licensed under GNU GPL v3<sup>8</sup>.

This project and the associated methodological plugin were reviewed and approved through the University of Edinburgh School of Social and Political Science (Approved on 29 March 2019 in association with the doctoral research of Anna Couturier "Google Search and the Mediation of Knowledge on Treatments: A Case Study on Unproven Stem Cell Treatments" [UoE Internal ID: 259679]). Additionally, we consulted with Clare Blackburn (School of Biological Sciences, University of Edinburgh, MRC Centre for Regenerative Medicine), Claire Tanner and Professor Megan Munsie from the University of Melbourne, the Anne Rowling Clinic (Alison Irving, Pamela MacDonald, Dawn Lyle, and Judith Newton), and patient advocates from Parkinson's UK (Alison Williams, David Adams, David Melton), Edinburgh Parkinson's Research Initiative (Martin Taylor), Young Onset Parkinson's Disease (YOPD) (Gaynor Edwards), and the many other patient advocacy groups and individuals who participated throughout the process. Every effort was made to ensure that anonymity and informed consent were preserved for the study participants.

## 6.4. Data collection

The duration of the data collection was slightly longer than four months (134 days; 30 September 2019 to 11 February 2020). Despite the high level of attention the topic received and the extensive reach of the EuroStemCell partner network, the installation numbers remained significantly lower than those for the 2017 German federal elections data donation. 138 participants took part in the data donation and installed the plugin, with 102 actively submitting data. In addition, the data set includes the submissions of the 24 participant VPS servers that submitted data automatically. From a health sociology and biomedical patient research perspective, this was a successful number of participants (Couturier, 2023). Here the differences between the disciplines are quite clear. Moreover, the targeted group and topic were extremely specialized, with much interest coming from within the research and health community but outside the data donor profile. Figure 6.7 displays the actual participants by region, color-coded by disease. It is evident that the majority of the participants had Parkinson's disease.

The small number of participants in the other study groups compelled us to evaluate only the Parkinson's disease study data and shift the study's emphasis from quantitative to qualitative analysis. Thus, the second research question became: Are advertisements for unproven stem cell therapy significantly more prevalent in Parkinson's disease [diabetes and multiple sclerosis] patients?<sup>9</sup>

---

<sup>8</sup>[https://git.cs.uni-kl.de/m\\_reber16/EuroStemCell](https://git.cs.uni-kl.de/m_reber16/EuroStemCell) last accessed on 15 Dec. 2022

<sup>9</sup>So the part in the square brackets was excluded.

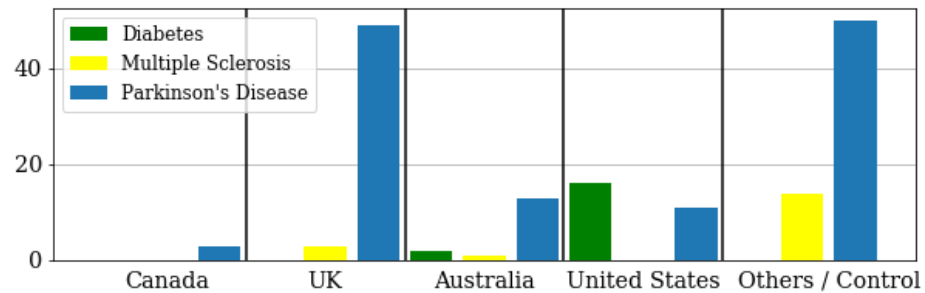


Figure 6.7.: Size of all study groups grouped by region, color-coded by disease. Figure from (Reber, 2020), with permission from Martin Reber.

As part of Martin Reber’s Master’s thesis, he and Anna Couturier conducted a qualitative evaluation of the submitted advertisements by having Anna Couturier rate the “explosiveness” of the advertising providers and using these categories for evaluation.

Between 30 September 2019 and 11 February 2020, a total of 177,756 records were submitted to searches for the Parkinson’s disease studies, of which 63.8% were submitted by VPS donors. Figure 6.8 depicts the time progression of the submissions, with each day’s real and VPS submissions represented separately. In addition to a rise in submissions from actual participants following November 2019 onboarding events, this graph also depicts the activation of the VPS as well as fluctuations in VPS submissions caused by disruptions or DDOS attacks.

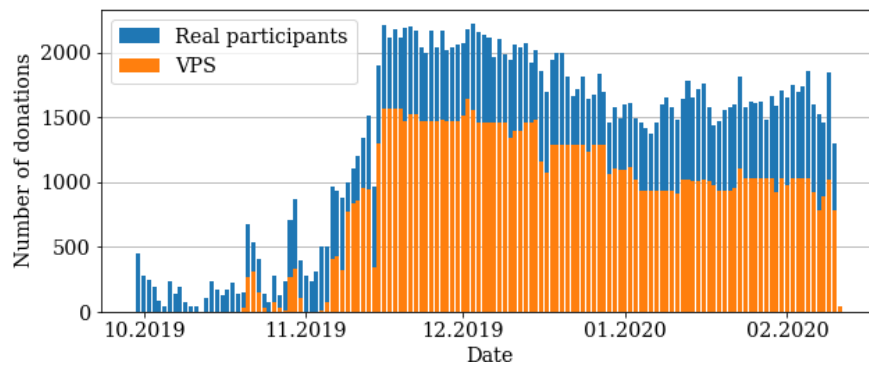


Figure 6.8.: Donations over time show a spike after onboarding events and VPS rollout. Figure from (Reber et al., 2020).



Only 5.7% of the submitted records contained advertisements, according to the analysis. To calculate this number, only those records with values in the “ads” field were chosen. Due to the fact that some of the submitted records contained multiple advertisements, these had to be extracted to ensure an accurate count. This selection and extraction revealed that there were a total of 21,188 advertisements available for evaluation.

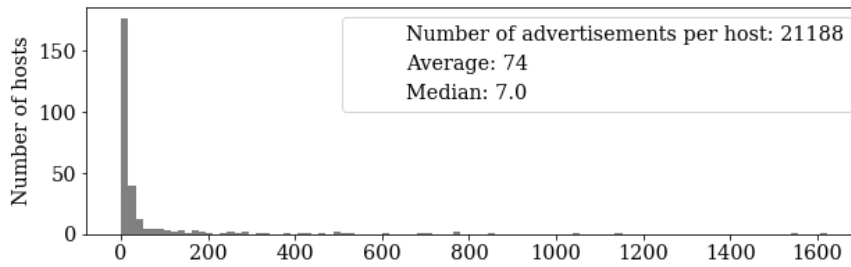


Figure 6.9.: Histogram of the distribution of advertising media by number of advertisements. Figure from (Reber, 2020), with permission from Martin Reber.

An evaluation of the top-level domains of the advertisements’ landing pages revealed that they linked to 285 unique websites. Figure 6.9 is a histogram depicting the distribution of advertising providers by the number of advertisements displayed. The median number of advertisements per host is only seven, while the average is 74. In addition, 80 percent of advertisers appear in the data fewer than 50 times. Assuming that the 285 hosts represent individual advertisers and not a smaller number of advertisers operating multiple websites, this indicates that there are a large number of small advertisers who must compete with a small number of powerful competitors.

## 6.5. Examination of the results

The following text provides a brief presentation of the research findings. First, the evaluation of the submitted advertisements for the first research question is described. The results of the study are then discussed with regard to the second research question, namely, whether advertisements for unproven stem cell therapy offers actively target Parkinson’s disease patients.

### 6.5.1. Captured Ads Analysis

Martin Reber, 2020 conducted an analysis of the collected advertisements in order to better comprehend their nature. By collaborating with Anna Couturier and thus Eu-

Most problematic	Commercial clinic
Quite problematic	Clinical trials - private Clinical trials - commercial Complementary treatment - commercial Blood banking - commercial Health news - commercial
Potentially problematic	Political lobby organization Pharmaceutical company Commercial non-health specific Conference - commercial Biopharma supplies
Neutral	Health news - public Research institute Blood banking - public Clinical trials - public Conference - public Governmental Healthcare provider - institution Non-profit health organization Patient groups Social Crowdfunding Other News
Not to determine	Unknown
Possibly drugs	Needs review

Figure 6.10.: Advertisement host labels and categorization proposed by Anna Couturier.

roStemCell, it was possible to categorize the advertisers as belonging to the biomedical industry, public health, and commercial medical services. To assess the risk potential for users, a “traffic light” analysis was applied. Anna Couturier created the labels in Figure 6.10 and assigned them to the 285 advertisers in order to accomplish this.

The analysis uncovered potentially problematic sources such as private clinics, for-profit websites, pharmaceutical websites, and private biobanks within the advertisements. However, the advertisers also included non-profit organizations and patient advocacy groups, such as Parkinson’s UK and the Michael J. Fox Foundation. Figure 6.11 displays the twenty providers with the highest display volume in the dataset, with each color representing Anna Couturier’s rating.

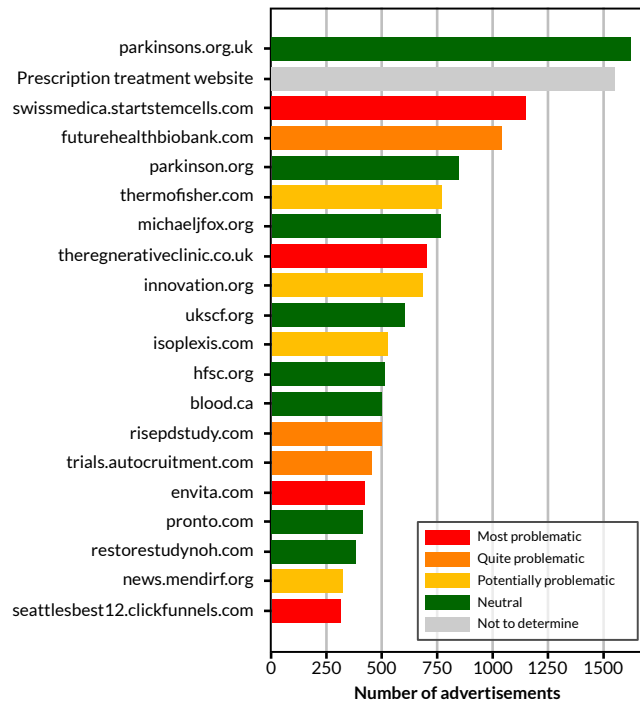


Figure 6.11.: The top 20 advertisers are shown with the number of ads collected. The color-coding is based on an assessment by Anna Couturier. Figure based on (Reber et al., 2020).

Our study shows that Google continued to display advertisements for unproven stem cell therapy offers until the end of the study, despite the self-imposed ban (see Bidings, 2019) on them, and that the advertisements displayed in the process did not only originate from trustworthy, informative providers (Reber, 2020).

### 6.5.2. Roll-out behavior of the web displays

As shown in Figure 6.12, data donors who identified as patients or carers received more advertisements than participants in the control group or VPS data donors. This suggests that Google’s ad targeting includes additional methods that impact users identified as having Parkinson’s disease. Notably, the content of these advertisements did not contain a disproportionate number of highly problematic sources (for a detailed analysis, see Reber, 2020). The increased display of advertisements for patients and carers, however, suggests that Google may have already identified vulnerable groups to target. Due to the limitations of the study and the participants, the reasons for the visibility of these

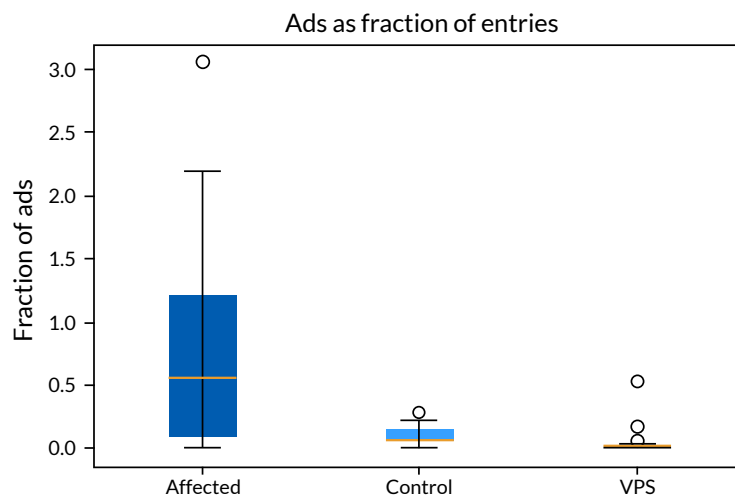


Figure 6.12.: Fraction of ads in total donations per participant. Means: Affected 0.76, Control 0.10, VPS 0.05 Medians: Affected 0.57, Control 0.07, VPS 0.01. Figure based on (Reber et al., 2020).

modalities could not be determined. There is an urgent need for additional research to investigate this trend and its impact on potentially vulnerable users.

## 6.6. Threats to the validity of the results

In social science research, Drost, 2011 provides a schematic guide for evaluating the validity of study results and the limits of their interpretation. The guide provides a list of various validity threats that can be investigated to improve the interpretability of empirical results. It describes the internal validity, external validity, construct validity, and statistical conclusion validity that can influence the validity of a study. These four types of validity are essential for determining the accuracy and dependability of study results, and they are applied to the study in order to interpret the results below (for details, see Section 8.3.2).

### Statistical conclusion validity

Due to the fact that statistical significance refers to the statistical analysis conducted in the study (Drost, 2011), this factor is only relevant to the investigation of research question 2. Due to the small number of participants in this study, we can only speak of an initial indication; a generalization to the entire roll-out behavior of Google AdSense is

not possible. But in digital sociology and digital ethnography, this can absolutely be used to support larger platform behavior statements. For example, the fact that problematic advertisements continued to exist suggests a limitation in terms of intervention and a mismatch between Google's policy and its actual platform practices.

### **Internal validity**

Internal validity is the extent to which a study's results can be attributed with certainty to the independent variable tested and not to other external factors (Drost, 2011). Due to the participation of real study participants in the investigation of the second research question, it cannot be guaranteed that Google or AdSense correctly identified them as persons with or without a disease based on their previous Internet usage behavior. Even if we assume, as described above, that it is not the disease itself that can be selected as a target for unproven stem cell advertising, Google must correctly identify the respective combination of characteristics indicating the disease. Moreover, a very complex software system with highly interdependent, networked algorithmic systems is being studied here, so it cannot be ruled out that other factors that we did not monitor played a role in the study's findings.

Due to the small sample size, it is also not possible to rule out the presence of A/B testing (Kohavi & Longbotham, 2017). Google admitted in 2006 that it uses such techniques to optimize its search engine (Pansari & Mayer, 2006), so it can be assumed that the same techniques are employed when distributing advertisements.

### **Construct validity**

Construct validity is the degree to which a measurement or instrument accurately reflects the theoretical concept or construct it is intended to measure. Since the investigation of the first research question focused solely on the existence of advertisements for unproven stem cell therapy offers, the discovery of such advertisements does not permit drawing a false conclusion. Therefore, there is no room for interpretation. The second research question compared the advertising distributed to individuals with the disease to that distributed to healthy individuals. Again, no measure or metric was used that allowed for interpretation.

### **External validity**

External validity refers to the question of generalizability, or the extent to which the observed effect can be transferred to other areas, populations, or situations (Campbell & Stanley, 1963, p. 5)(Drost, 2011). Due to our reliance on voluntary participation in the study described above, we might not have been able to eliminate "selection bias"<sup>10</sup> in

---

<sup>10</sup>Selection bias is a type of bias that can occur in studies and surveys in which participants or data are chosen unequally. This may result in biased results and incorrect conclusions regarding the underlying

the participant selection (see Ellenberg, 1994). The participants we invited to participate were selected to the best of our knowledge, but there is no assurance that they or their carers were actually affected by the respective diseases. Due to the small number of participants, there is also a high likelihood of “sampling bias”<sup>11</sup>.

## 6.7. Interpretation of the results

The study investigated the behavior of advertisements for unproven stem cell therapy offers on the Google search engine. To collect data that could be used for black-box analysis, a browser plugin was created. The purpose of the study was to determine whether a change in Google’s advertising policy had an effect on problematic health-related advertisements (research question 1) and whether they were more likely to be displayed to people with Parkinson’s disease than to healthy people searching for specific keywords (research question 2). This black-box analysis revealed that despite Google’s policy change (Biddings, 2019), questionable stem cell advertisements persist on its on-line platform. Consequently, individuals suffering from serious illnesses continue to be targeted by providers offering unverified stem cell treatments and by other questionable entities (Reber et al., 2020). This situation presents a societal concern as it discriminates against a vulnerable user group. Therefore, society, particularly advertisers and intermediaries within the online advertising ecosystem, should take into account users’ perception of ads, including potential confusion and apprehension regarding personalization and misuse.

It has become clear that Google’s published ban (Biddings, 2019) is not a guarantee, and there is still a risk of false information being disseminated or deceptive advertisements being served to vulnerable individuals (Reber et al., 2020)<sup>12</sup>.

It is essential to continue monitoring the effects of these systems and to make any necessary adjustments to ensure that they are used ethically and responsibly. It is crucial for regulators and the medical community to collaborate closely to protect patients from potentially harmful and unproven therapies and to ensure that all medical treatments provided are scientifically proven to be safe and effective.

It has also been discovered that there are a number of competitors in the field of stem cell therapy with varying commercial and educational objectives (Reber et al.,

---

trends and ratios. Self-selection bias, in which participants take part of their own free will, and selection bias, in which participants are chosen from a predetermined group, are two examples of the various ways in which selection bias can occur.

<sup>11</sup>Sampling bias refers to an incomplete or inaccurate representation of a population or group of interest within a sample. This may occur as a result of unequal probabilities of selecting certain units in the sample, the selection of a non-representative sample, or other methods that introduce bias. The result is a reduction in the reliability and validity of the research.

<sup>12</sup>It is important to mention that the ban implemented by Google is no longer in effect. This raises the question of whether the ban was lifted because Google faced difficulties in fully intervening without compromising their essential commercial operations.

2020). Between cautious medical associations and questionable actors, there is a constant competition for attention. At the very least, the narrative of stem cell tourism was detectable in the data set used in this study, as multiple actors from various nations promoted unapproved treatments. It was apparent in our findings that the current state of Google Search around the topic of stem cell treatments transcends geographical borders and regulatory boundaries.

To sustain the web search ecosystem, it is essential for users to interact with advertisements in a safe manner. To ensure a safe environment when using the Internet and the online advertising that accompanies it, society, and especially advertisers and intermediaries in the online advertising ecosystem, must actively promote ethical practices. Our study revealed that although monitoring and reviewing ad content, implementing anti-discrimination policies, and maintaining transparency in the targeting process is necessary, these measures are not consistently enforced. In order to promote safe online advertising, it is crucial for all parties to assume their responsibilities and adhere to ethical standards. Because this is about patient safety, oversight, and freedom from biomedical and economic exploitation.

As ADM systems become more prevalent in the medical field, it is crucial to minimize their negative effects and protect users. Despite existing policies to restrict advertisements in sensitive categories and prohibit misleading or unproven medical treatments, our study demonstrates that these policies are not always consistently implemented (Reber et al., 2020). Economic interests of advertisers and intermediaries that conflict with ethical practices appear to be the greatest obstacle. In 1999, Page and Brin, the founders of Google, stated: “We expect that advertising-funded search engines will be inherently biased towards advertisers and away from the needs of consumers” (Brin & Page, 1998, Appendix A. Advertising and mixed motives). Therefore, it is even more crucial that all parties assume responsibility for their actions and promote non-discriminatory online advertising.

This type of research, utilizing black-box analysis, is both possible and crucial in shedding light on important phenomena that would otherwise remain elusive. Black-box analysis enabled us to conduct a comprehensive and data-informed digital ethnography, focusing specifically on a critical topic — stem cell therapy advertisements — and capturing a significant moment of platform policy change. The black-box analysis facilitated exploring the impact of Google’s advertising policy change on problematic health-related advertisements. It uncovered the persistence of questionable stem cell advertisements on the online platform, despite the stated ban. This finding demonstrates the ineffectiveness of the policy change in eliminating misleading or unproven medical treatment advertisements. The study’s findings highlight the importance of employing black-box analysis to investigate and monitor critical topics in the digital landscape. This approach allows researchers to uncover the hidden dimensions of platform policies and their impact on user experiences. It offers a nuanced understanding of how policy changes may fall short of their intended goals and fail to effectively address societal concerns.

Overall, this research demonstrates the power of black-box analysis in capturing significant moments of platform policy change and conducting in-depth digital ethnographies that provide invaluable insights into complex phenomena that would otherwise remain obscured. It emphasizes the need for continued efforts to monitor and regulate online advertising practices to protect vulnerable user groups and promote ethical standards in the digital ecosystem.

## **6.8. Lessons Learned**

The study yielded additional learnings, which are briefly described below.

### **6.8.1. General Learnings**

The benefits of a black-box analysis performed using a crowdsourced audit include the natural interaction with the web service by participants with real profiles and the ability to obtain a wide variety of input configurations from a large number of users. This opens up an entire world of analysis that was not possible in the past. The implications for fields like digital health sociology and health studies are significant. The use of black-box analysis presents several benefits, including the opportunity for participants with genuine profiles to interact naturally with web services and for researchers to gather input configurations from a diverse range of users. This approach enables researchers to tap into a vast array of data that was previously inaccessible, opening up a whole new realm of analysis.

In the field of health studies, the use of black-box analysis enables researchers to uncover previously unseen patterns, trends, and biases within digital health information ecosystems. It allows identifying gaps in knowledge, evaluating the quality and accuracy of health information sources, and understanding user experiences and needs. This approach helps build a more comprehensive understanding of the challenges faced by individuals who rely on online platforms for critical healthcare information. From a digital health sociology perspective, black-box analysis offers a unique opportunity to examine the complex interplay between technology, platforms, and users within the context of healthcare. Researchers can, for example, explore how profit-driven structures shape the visibility, accessibility, and trustworthiness of health information online. They can analyze the power dynamics between platform owners, advertisers, and users, and assess the implications for public health, equity, and informed decision-making.

Nonetheless, the group of stakeholders investigated in this study revealed one of the inherent limitations of black-box analysis: Due to the small size and widespread distribution of the affected population targeted by the study, it was difficult to gain access to affected individuals despite the efforts of a number of leading research institutions. The average age at first diagnosis of Parkinson's disease is over 60 years (Pagano et al., 2016), which presented additional challenges. Consequently, we needed a more elaborate



study design than for previous data contributions. To increase acceptance, additional support measures, such as practical training or instructional videos, could have been implemented. However, such measures must be tailored to the needs and expectations of the intended audience. This was unfortunately not possible within the scope of the study, as it would require a comprehensive analysis of the target audience. It has been demonstrated that the individual characteristics of stakeholder groups have a substantial impact on the design and implementation of a crowdsourced audit within the context of a black-box analysis.

### **6.8.2. Technical Learnings**

The use of virtual private servers (VPS) as benchmarks in a crowdsourcing audit offers numerous benefits, but also presents risks and difficulties. One of these obstacles is the effort required to utilize VPS hosts. Numerous virtual systems reside on VPS hosts, which increases the risk of an IP range being blocked by Internet services. This can lead to complications when a VPS server is used to make requests to a social media platform by an automated process, such as in the current study design. An incident in which requests to Google Web Search were continuously blocked at a U.S. VPS server location in Dallas is an example. Other websites required manual intervention to deactivate captchas, a process that verifies the accuracy of the information and the non-robotic nature of the user. This can result in an increase in impracticality if manual intervention is required to support automated processes.

Large online service providers, such as Google, may also be capable of detecting automated audits. Although there is no evidence of such action in this study, some virtual private servers (VPSs) were blocked due to increased traffic. This at least demonstrates the existence of mechanisms for handling suspicious traffic. Researchers have already suggested the existence of such a mechanism (Datta et al., 2015).

To conduct a study that is truly reliable, one must accurately imitate users and simulate the usage scenario as closely as possible. This problem was already pointed out by (Diakopoulos, 2014a, p. 17), as it turned out that the results of real interactions with the system did not match the results of pure API requests. To achieve a realistic result, it is essential for the emulated user activity and usage scenario to be tailored to the respective target audience of the crowdsourced audit. The simulation should also take into account the demographics of the target group, which can have an effect on Internet and media literacy.

The black-box analysis presented in this section highlights the interrelationships, complex levels and sociotechnical issues involved in dealing with ADM systems. At its core, the purpose of this section is to illustrate the transformative role of black-box analysis in examining our perceptions and approaches to the citizen-biomedical research interface. Indeed, this collaboration represented an innovative interdisciplinary opportunity. Utilizing a black-box analysis in this project was vital as it allowed us to take an in-depth

look at the underlying algorithms that dictate platform behaviour and their potential implications on various facets of life, including health. This work demonstrates the capacity of black-box auditing to enable not just computer science researchers but also stakeholders from non-technical backgrounds to understand the technological dynamics that shape everyday lives. With real-world impacts, this research underscores the importance and the potential of black-box analysis. It serves as a key tool that significantly alters the approach of investigations, revealing insights that might remain concealed without its application. Ultimately, this section underscores the critical role of black-box analysis within biomedical research, offering a fresh, interdisciplinary approach to the investigation of intricate technological systems and their intersection with human life. This highlights the immense potential of such methodologies to address real-world challenges, thereby redefining the ways in which scientist approach and solve problems within the realm of ADMS in biomedical research.

## Black-box analysis – Price differentiation in online retail

The capabilities of online commerce reduce buyer-seller interaction to the presentation of goods and services along with a price tag. Digitalization offers significant benefits to sellers, as it enables the use of faster, more covert, and even completely new pricing options, such as dynamic or even personalized price adjustments (White House, 2015). Both strategies are familiar from offline retail, but their implementation is limited. Throughout the day, for example, gas stations dynamically adjust the price of gasoline for their customers, and when selling cars, car dealers negotiate a personalized price with each customer. In the past, the effects of dynamic price management were highly dependent on the application context and population acceptance. For instance, a dynamic reduction of the price of chilled beverage cans above a certain air temperature in a Spanish amusement park was well received (H. Simon & Fassnacht, 2019, p. 211), whereas in 1999 a press release about dynamic prices at Coca-Cola vending machines regarding a higher can price on hot days resulted in a notable drop in the share price of the Coca-Cola group, despite a profit increase (H. Simon & Fassnacht, 2019, p. 210 et seq; Hays, 1999; Leonhardt, 2005). These are examples of the fact that in physical salesrooms, buyers can exchange information about price offers and perceive and denounce unequal treatment, whereas in online retailing, this capability does not exist or requires significant effort to implement.

In addition to such dynamic pricing, which affects all customers equally, the remote selling of goods significantly expands the scope of personalized pricing. In principle, personalized prices pose a risk of unequal treatment of individuals, particularly if directly or indirectly protected characteristics are included in the evaluation of buying behavior (hereafter referred to as “profiling”) and personalized prices are offered based on this profile. In general, retailers are permitted to customize prices for specific consumers or consumer groups. This may also be accomplished through automated decision-making or consumer behavior profiling<sup>1</sup> (see European Commission & Council of the European Union, 2019, recital 45). However, because such pricing could be viewed as a risk in the purchase decision, this topic has now entered the political discourse. Consequently, the

---

<sup>1</sup>Here, an explicit assessment of the consumer’s purchasing power may also be made.

EU Commission requests that retailers inform consumers when the prices of goods or services have been personalized based on automated decision-making and profiling of consumer behavior (European Commission & Council of the European Union, 2019, recital 45). In its coalition agreement, the German government addresses the issue of transparency and advocates proactive consumer protection. Here, algorithm- and AI-based decisions, services, and products are to be made verifiable so that they can be checked for discrimination, disadvantages, and fraud (CDU, CSU and SPD, 2018, p. 6, 354 et seq.). Unequal treatment does not have to be incorporated intentionally; it can also manifest itself indirectly through other characteristics. For instance, Amazon’s “Same Day Delivery”<sup>2</sup> was initially only available in certain areas of the United States, which were determined based on non-discriminatory factors such as the number of Prime customers and proximity to a shipping center. However, journalists were able to demonstrate that African-American residential areas were largely excluded from the test areas (Ingold & Soper, 2016). In Germany, too, there are growing concerns about the discriminatory effects of digitalization. The German Council of Consumer Experts, for instance, published an expert report on consumer-friendly scoring in 2018 (SVRV, 2018). The report focuses on the appropriate use of assessment procedures, such as in the assessment of creditworthiness. The Monopolies Commission has also addressed algorithm-based price management, but from the perspective of illegal price fixing (Monopolkommission, 2018).

In order to detect the existence of unjustified or even punishable discrimination, one must first determine whether a specific population group is systematically placed in a worse position to a statistically significant extent. Individual consumers cannot compare prices, making it difficult to prove personalized pricing, particularly in online commerce. Such pricing strategies can only be detected by comparing the prices displayed to various customers in real time. In Germany, such price differentiation has only been observed in the tourism industry (Schleusener & Hosell, 2015). Journalists also rightly caution consumers against revealing their precise willingness to pay<sup>3</sup>. The refusal of this personalization-related data processing should not result in access discrimination. Otherwise, important supply platforms could be cut off from certain consumer groups. In a draft bill based on EU Directive (EU) 2019/2161 by the European Commission & Council of the European Union, 2019, the German Federal Ministry of Justice demanded information requirements for price personalization by the end of 2020 (Bundesministerium der Justiz und für Verbraucherschutz, 2020), which demonstrates that politicians are aware of the potential dangers and are contemplating restrictions. To ensure compliance with these obligations, the legislator must be able to determine whether an online retailer actually employs a price algorithm with a personalization component. There is a general suspicion that online retailers employ personalized pricing. In a study conducted by the

---

<sup>2</sup>Offer of the online department store Amazon.com, Inc. to guarantee same-day delivery of orders.

<sup>3</sup>Evidence of this risk has so far only been anecdotal, such as in an article in the *Handelsblatt* by Anja Stehle, 2016 dated 15 Feb. 2016 .

---

Vienna Chamber of Labour, for instance, a number of products from various online retailers were evaluated for this purpose<sup>4</sup>. Although the primary focus was on the tourism industry (airline and travel companies), Amazon products were also examined. There were price fluctuations for all observed products, ranging from 0.1% to 22.9% (0.01 euros to 50 euros) for Amazon products and up to 480% for flight providers on the online booking portal for air travel Opodo<sup>5</sup> (Delapina, 2019). In the absence of a detailed description of the exact query process, it is often impossible for researchers to determine whether the deviations found were personalized pricing or simply dynamic pricing, i.e., only a change in price over time for all customers. Moreover, air travel is a highly volatile market with limited quotas, which makes differential pricing a common and widely accepted practice (Clark & Vincent, 2012). During a study of the tourism industry in the United States, it was discovered that changes in location led to interesting price differences between 24 U.S. hotels and six car rental companies (Schleusener & Hosell, 2015). It was discovered that rental car requests made from a German IP address were, on average, just under 5% more expensive at Los Angeles dealers than local requests; Chicago providers added an average of nearly 25% for German customers (Rose & Rahman, 2015). Even though these results are merely a snapshot of the study at its respective points in time, they demonstrate the options available to online shop providers. This section summarizes my explorations and research in this context. An important basis for it is provided by a study I wrote with Roman Krafft, Marcel Wölki, Michael Rahe, and Katharina A. Zweig for the Ministry for Family, Women, Youth, Integration and Consumer Protection of the German federal state of Rhineland-Palatinate.

Regrettably, the Ministry had not published the pertinent study by the time of submission for this dissertation. This constitutes a considerable loss, as legal scholar Michael Rahe's research added a well-founded legal analysis of the issue of potential discrimination arising from personalized price management in online commerce. Additionally, his work explores the legally mandated requirements for transparency and observability by black-box approaches in this context. Although my background as a computer scientist limits my ability to adequately summarize Rahe's contributions, it is essential to recognize and acknowledge his expertise in this field.

---

<sup>4</sup>The survey period was from 2 to 13 April 2019. The products were queried from different terminals on Tuesday, Thursday, and Saturday in order to identify possible differences. The terminals were distributed all over Austria in order to avoid location-based prices.

<sup>5</sup><https://www.opodo.com/>

## 7.1. Dynamic and personalized prices in online retailing

First, the distinction between dynamic pricing and personalized pricing must be clarified. To that end, the concept of price differentiation will be examined in greater depth first (Section 7.1.1) in order to build on it and distinguish between dynamic and personalized pricing. This will be followed by an overview of consumer reactions to personalized prices (Section 7.1.2) and an examination of whether and to what extent automated pricing can be determined with the aid of customer-specific software. The necessary technical foundations for automated pricing will then be clarified, that is, algorithmic customer profiling (Section 2.3).

### 7.1.1. Price differentiation

In literature and journalism, discussions of personalized pricing are frequently complicated by ambiguous and inconsistent terminology. For instance, depending on the investigation and reporting, it is not always possible to distinguish between identical and similar but not identical products. Occasionally, prices aimed at specific groups are also considered personalized, or the terms “dynamic prices” and “personalized prices” are incorrectly used interchangeably. For this reason, these terms and ideas will first be defined in greater depth and distinguished from one another.

#### Personalization of products and prices

Personalization enables companies to separate themselves from the competition. A basic distinction must be made between the personalization of products & services and the personalization of prices. The former is done by optimizing the fit of the product to the individual consumer while the latter is based on the customer’s willingness to pay (Belleflamme & Peitz, 2010, p. 219 et seq.).

When investigating price discrimination that may be illegal or immoral, it is crucial to determine whether or not customers are offered identical products. Jentzsch identifies a total of four distinct price-product relationships based on product customization (homogenous vs. personalized product) and/or pricing strategy (standard vs. personalized) (see Figure 7.1).

Besides the usual process of offering a standardized product for the same price each time, i.e., a standard price (IV in Figure 7.1), the following options are available: The first combination consists of a personalized price and a personalized product (I in Figure 7.1). The second variant involves the combination of a standard price and a personalized product (II in Figure 7.1): Although all customers pay the same price, the product is individually tailored to each customer’s preferences, as with a coffee cup that is individually printed and sold at a standard price. The third alternative describes the combination of a personalized price and a standard product (III in Figure 7.1). In

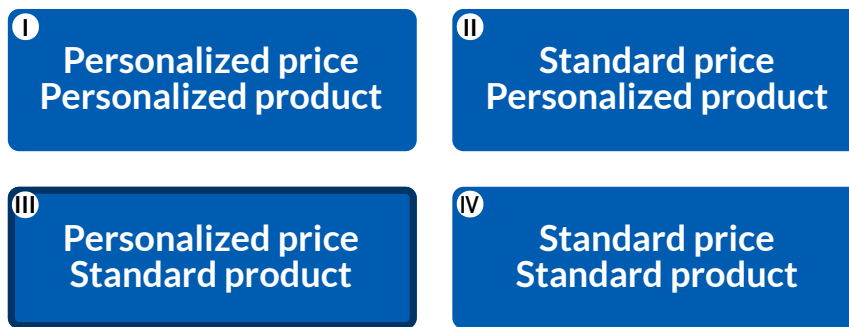


Figure 7.1.: Possible combinations of price and product personalization; modified representation according to Jentzsch, 2017.

this variant, the customer pays a different price than other customers, despite receiving the exact same product. Varied pricing for a standard product is known as price differentiation. It is irrelevant whether this product is sold to the same consumer or to separate consumers at different prices (Pigou, 1920, p. 244 et seq.)(Tirole, 1989, p. 133) (Fassnacht, 2003).

Inequitable treatment based on legally protected individual characteristics can only occur in the third quadrant, which is why this case will be examined in greater detail below. In economics or marketing, a distinction is made between three degrees of price differentiation, which are depicted in Figure 7.2 based on the amount of information required to execute differentiation. Note that the degree of price differentiation was not initially based on the amount of information required, hence the unusual numbering.

*First-degree price differentiation* assumes fully personalized pricing, which is why it is also known as “perfect price differentiation” (Klein & Steinhardt, 2008, p. 43). To achieve this objective, the customer’s presumed willingness to pay would be determined based on their identity using Big Data analyses, for instance, and prices would be adjusted accordingly. However, because it is not yet possible to precisely estimate the willingness to pay of individual potential customers, Stole, 2007 claims that no pricing policy based on first-degree price differentiation has yet been implemented in practice. Fundamentally, it should be noted that price differences resulting from different manufacturing costs do not constitute first-degree price differentiation (Jentzsch, 2017, p. 9). If the same product is produced in China and also in India, and one portion of the customers pay a price based on the manufacturing costs in India while the other pays a price based on the manufacturing costs in China, this is not price personalization. The differentiation is not based on customer characteristics, but on production conditions.

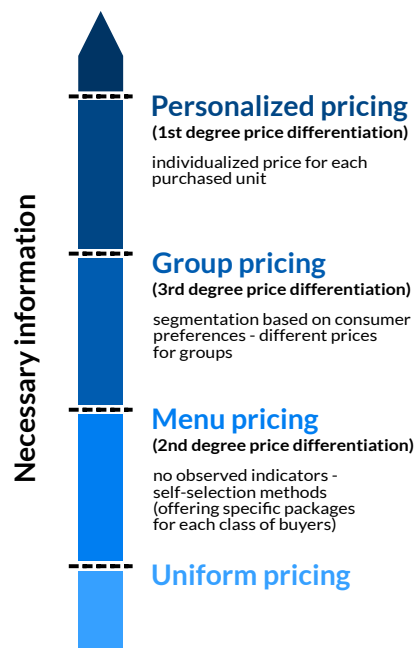


Figure 7.2.: Visualization of the three degrees of price differentiation based on the amount of information required. Figure according to Belleflamme & Peitz, 2010, Part IV Chapter 8: Group Pricing and Personalized Pricing.

*Third-degree price differentiation* is based on clear group affiliations, such as students, retirees, or residents of a particular city (Klein & Steinhardt, 2008, p. 44) (Jentzsch, 2017, p. 9). This requires much less information than individual pricing.

*Second-degree price differentiation* calls for the least amount of information and describes menu differentiation (Klein & Steinhardt, 2008, p. 44). In this instance, the customer is presented with multiple prices for variations of a product or a product package, allowing them to choose the variation of the desired product with the price that best suits them. Thus, when selling a standard product, it is possible to differentiate prices based on the characteristics of the buyer. There is also the possibility that the seller will adjust the price based on the timing of the offer. This is referred to as dynamic pricing, as depicted in Figure 7.3. This Section examines the distinction between dynamic pricing and personalized pricing in greater detail.



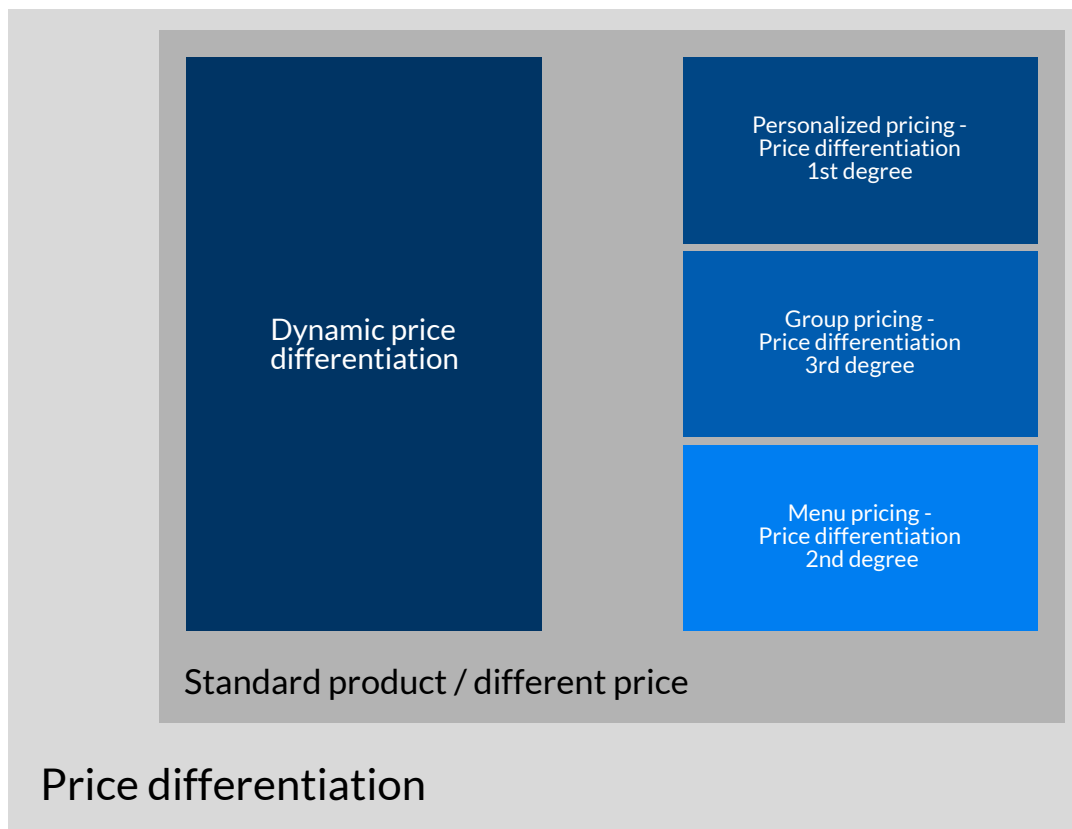


Figure 7.3.: Visualization of dynamic pricing. Price differentiation of first to third degrees as variants of price differentiation for a standard product.

### Distinction between dynamic and personalized prices

As mentioned at the beginning of this section, the terms dynamic and personalized pricing are frequently defined inconsistently in the literature and are sometimes confused, so a conceptual distinction between them is established here.

**Dynamic prices** Dynamic price management refers to the practice of vendors adjusting their prices over time, as is the case with gas prices at gas stations. Such a procedure is regarded as a unique form of price differentiation, as it involves the differentiation of an individual's payment mentality in relation to time, which cannot be classified within the scale of price differentiation (den Boer, 2015; Popescu & Wu, 2007). Instead, dynamic price management refers to the fluctuation of a seller's product prices over time.

Dynamic prices are distinguished from personalized prices by the fact that the price is not (subjectively) dependent on buyer characteristics but rather changes (objectively) in the same manner for all potential buyers (Gallego & Van Ryzin, 1994; Gönsch et al., 2009). Contingent products are frequently subject to dynamic pricing (Klein & Steinhardt, 2008, p. 77 et seq.). For example, airlines may offer the cheapest flight tickets for the first few seats on a plane, then raise the price until just before departure in order to sell the remaining seats at a steep discount as last-minute deals (McAfee & Te Velde, 2006, Belleflamme & Peitz, 2010, p. 60 et seq. and p. 197 et seq.). Since all customers see the same price, based on the number of remaining seats and the date, prices are not personalized. Also to be distinguished from the concept of dynamic pricing are prices that are offered differently in the context of randomized A/B testing, as long as the group formation is also randomized, for example to investigate the effect of a website's design on a customer's acceptance of a price (see Kohavi & Longbotham, 2017).

**Personalized prices** Personalized pricing, as depicted in Figure 7.3, refers to the area of price differentiation where different pricing for the same standard product is based on individual characteristics (1st-degree price differentiation), group characteristics (3rd-degree price differentiation), or menu prices (2nd-degree price differentiation). Personalized price management is thus understood to be the indication of different prices for the same product at the same time for different groups of people or individuals, with the personal characteristics of the customers (as a group or as an individual) factored into the pricing. In algorithmic pricing, properties that are not used in analog pricing and which may be unexpected for customers, such as the device used for shopping or product search (such as a PC or smartphone) or the most recently visited website, can also be considered. Age or gender could be considered by vendors when setting prices, despite the fact that society may not want these characteristics to be considered for pricing purposes.

By issuing vouchers or discount codes, or directly in the online store, a personalized price can be displayed (Schleusener, 2017, p. 74).

### 7.1.2. Consumer reactions to personalized pricing

In this Section, customer responses to personalized pricing for a standard product are discussed. In the literature, it is assumed that consumer responses to personalized prices can be quite variable (Schleusener, 2017, p. 83; Reinartz et al., 2017). On the one hand, customers can derive clear benefits from a discounted offer, which could elicit positive emotions. On the other hand, there is also the possibility that price differentiation is unfavorable to the customer, which can result in angry responses. Amazon's randomization of price differentiation in 2000 to test personalized pricing is an example of an experiment that was received negatively by customers. On selected items, random discounts

between 20% and 40% were offered to customers. It was demonstrated that customers in online retail were able to quickly identify price differences between comparable goods, eliciting emotions such as anger and indignation at the unequal treatment (Iyer et al., 2002, p. 298). The customers' anger reached such proportions that Jeff Bezos, the CEO of Amazon, apologized to them personally and promised compensation. Customers place a high premium on the perception that they are receiving a fair price, and on their privacy. Customers will only continue to accept the business model if this is taken into account (Schleusener, 2017, p. 83). The perspective that personalization measures are unfair ranges from price differentiation based on time or location to differentiation based on individual characteristics or product preferences. According to a 2016 survey by the Consumer Policy Institute ConPolicy, the vast majority of respondents (87% of 879) considered only price differentiation in the form of customer loyalty rewards to be fair (Thorun & Diels, 2016, p. 10). Despite this, the majority of respondents (57% of 856 in total) favored constant prices (Thorun & Diels, 2016, p. 6). This can be explained by the fact that personalized prices appear to instill the fear that the loss of a reference price will necessitate greater search effort (Schleusener, 2017, p. 84). Regarding the personalization of online offers, however, one can observe a habituation effect. Those who use the Internet frequently view increased search effort less negatively (Schleusener, 2017, p. 84). Customers who perceive booking differences (such as time differences) are more likely to agree to different prices than those who do not. To demonstrate this further, consider the tourism industry, where prices vary depending on when they are booked (Schleusener, 2017, p. 84 et seq.). The significance of the privacy factor is reflected by the fact that the protection of personal data about one's own behavior and person is in the customer's best interest, even if it results in a price increase (Schleusener, 2017, p. 86). Possibilities for customers to take action against personalized pricing in the case of a negative attitude are numerous, but require a significant amount of effort. Altering one's own surfing habits to make it more difficult to identify the individual would be one potential countermeasure. Alternately, the use of technical measures to encourage the submission of low-cost bids is possible (Schleusener, 2017, p. 83).

The perception of unjustified price differences results in uncertainty and the associated abandonment of the purchase, dissatisfaction, negative word-of-mouth, and boycott. From the company's perspective, these potential reactions carry a high degree of risk, which is why the majority of suppliers avoid price personalization (Schleusener, 2017, p. 84).

### 7.1.3. Profiling

For a large number of customers to receive personalized pricing, a profile must be created for each customer. The required data and properties can originate from a variety of sources (Hornung & Engemann, 2016; H. Simon & Fassnacht, 2019; Wiedmann et al., 2002):

1. Data can be purchased. There are two types of data: factual data (such as age and location) and aggregated data, such as a credit score or an estimated annual income.
2. Customers may have entered the data themselves and therefore knowingly for the purchase of the product (for example, the number and type of products purchased or the billing address).
3. It can be behavioral data, such as the amount of time a user spent on a product website or the length of time a product remained in the shopping basket before being purchased. It also includes user-entered data (search queries, rating comments, etc.) where it is evident that this information can be used for pricing.

At this point, it is necessary to note that profiling does not likely correspond to the traditional marketing profiles of customers employed in analog marketing. While these profiles are still relatively simple and based on easily collected characteristics such as gender and age, algorithmic profiling systems can divide customers into groups based on a wide range of criteria, such as the length of time it takes to make a purchase decision, whether customers have already purchased the latest bestseller in the product category or are more likely to order classics, or a credit rating according to any third-party company. In addition, depending on the product category, the classification of a single customer may be based on entirely distinct criteria. Individuals do not need to comprehend the groupings, nor do the selection criteria need to always be based on the same individual characteristics of the customers. Last but not least, the criteria utilized by the software may superficially have nothing to do with the legally protected characteristics, but correlate so strongly with them that discrimination is ultimately unlawful. Users are typically unaware of the profile derived from their actions. However, Google and Facebook provide at least some insight into the advertising characteristics (see Figure 4.8 in Section 4)<sup>6</sup>. Identifying the customers is a fundamental requirement for profiling, which is required for personalized pricing. Personal information is analyzed and stored in order to profile individuals. This data includes, for instance, a person's name, location information, or identifying characteristics (European Parliament & Council of the European Union, 2016, article 4). Customer identification enables the dynamic observation of user behavior by the provider using software-based personalization technologies and is also

---

<sup>6</sup>See: <https://adssettings.google.com/>

referred to as monitoring or tracking in the literature (Iyer et al., 2002). Customer identification can be accomplished in either an active or passive manner. Active customer identification involves the customer’s participation in data collection and storage. The creation of a customer account by the consumer is one possible method of active identification. The customer is aware of the data entry (Jentzsch, 2017, p. 12). Passive identification is characterized by the absence of user participation in the identification procedure. Consequently, the individual is typically unaware of the data collection and storage. Passive identification typically occurs through data mining (for a detailed explanation, see Witten et al., 2011) and the recording of user behavior (Yan et al., 2009), for example in the form of cookies, which are stored by websites on the user’s device via the browser in order to have information about them permanently retrievable and to uniquely identify them across current and subsequent page visits. In theory, a cookie can store any information a website has about a user, allowing the user’s behavior to be tracked in any way (Park & Sandhu, 2000). On the other hand, it is also possible to access information about the customer or customers across multiple pages (Google Analytics, 2020).

## **7.2. Is it possible to capture personalized prices with a black-box analysis?**

ADM systems can be utilized by online retailers for automated price management based on profile data. Based on the information about their customers available to them or purchased, retailers can make predictions regarding whether a customer will pay for their order or the likelihood that a certain group of people will purchase a product at a certain price (Dominique-Ferreira et al., 2016; Gao et al., 2020). Hannák et al., 2014 assert that there are numerous factors that can be used for personalized price differentiation. The best way to check pricing software would be for control institutions to actively cooperate with online retailers. The users of the pricing systems could then permit a group of experts to conduct a code review and respond directly to the reviewers’ questions. This method would constitute a “white-box analysis” because it would expose the algorithmic decision-making system’s inner workings. An investigation with direct access to the code would reduce effort and maximize the usability of the results, thereby fostering confidence in the system itself, as quality claims made by the system operator could be verified. Nevertheless, this cooperation for the evaluation of the pricing system is not possible for a number of reasons. Companies often consider their pricing strategies and related information as valuable trade secrets. Revealing these trade secrets to external parties, even for the purpose of evaluation, could negatively affect the company’s competitive advantage and market position. The fear of unauthorized disclosure of sensitive pricing information could lead companies to be reluctant or unwilling to participate in such cooperative evaluations. Consequently, a reviewing body must currently rely on black-

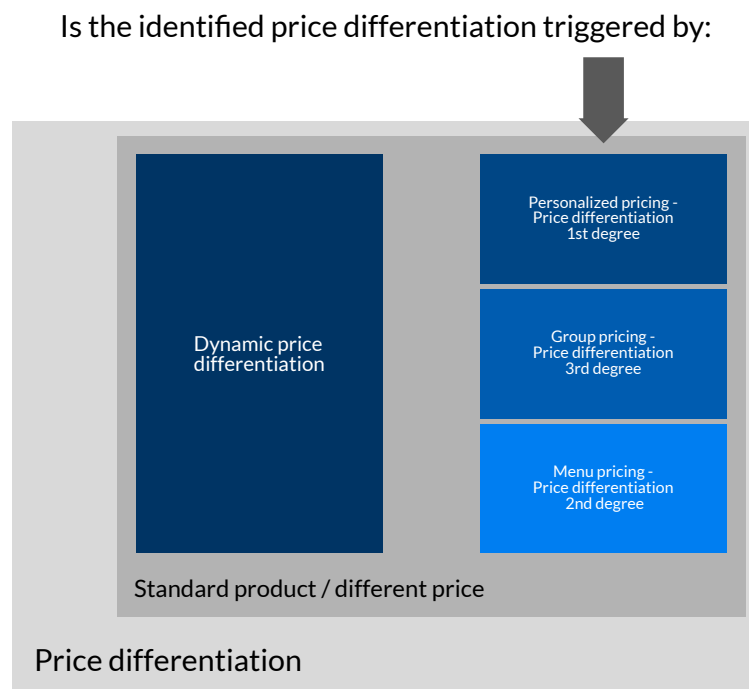


Figure 7.4.: Visualization of the second part of the investigation. Price differentiation found can be traced back to person-related characteristics and thus dynamic price differentiation can be ruled out.

box analyses. It can, for example, systematically query for example the prices and draw conclusions about the pricing mechanics. The next Section describes the projects conducted as part of my doctoral research into the algorithmic pricing of online shops in relation to possible price personalization. In each case, the course of the investigation is divided into two phases:

1. First, a price difference for the same product must be documented.
2. In a second step, the (potentially improper) price difference based on buyer characteristics is examined (see Figure 7.4) in an effort to answer the following questions:
  - Can this price difference be distinguished from potential temporal dynamics (e.g., contingent pricing) in pricing (dynamic price differentiation)?
  - Is there undesirable personalized pricing based on individual characteristics or group affiliations for which unequal treatment is neither desired nor permitted?

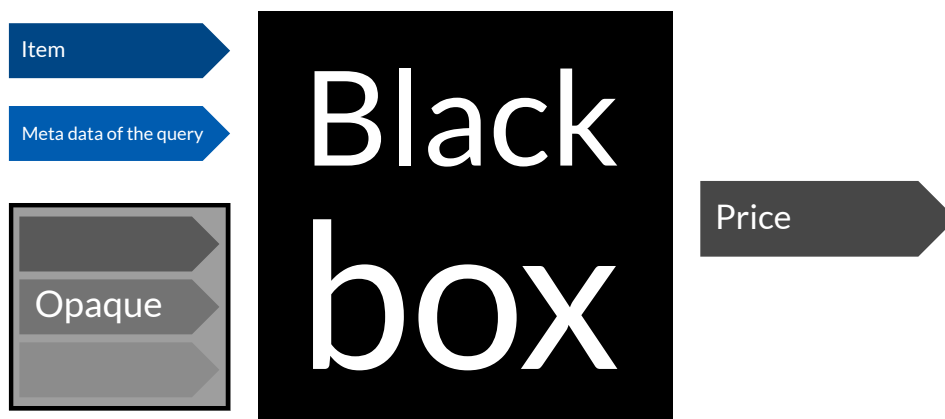


Figure 7.5.: The investigation of pricing in online shops represents a black-box scenario of the second category (for details, see Section 8.2.1). There is knowledge about the outputs of the system as well as a part of the input; there are also indications of further inputs, e.g., an individual personalization vector (see, e.g., Hannák et al., 2014).

The second phase proves to be exceptionally difficult when examined in the context of a black-box analysis.

When determining the black-box scenario, only a subset of the inputs used to calculate prices are known. Metadata of the request, such as time, IP address, etc., can be used in addition to the examined product, but the operators of online shops disclose little about what additional information is included in the pricing (see Figure 7.5). Consequently, the study of personalized price management presents a black-box scenario of the second category (details can be found in Section 8.2.1).

The objective of a black-box analysis is to derive a model of an unknown decision structure. To understand and comprehend the causal relationships between customer characteristics and the respective proposed price, it is necessary to investigate the influence of individual parameters on pricing. In order to analyze the pricing of an online store, it is necessary to attempt to keep all inputs constant that may be used for pricing and allow only the input being examined to vary. If the investigated property is one that is legally protected, this could be considered unjustified price differentiation. This could be the case, for instance, if two groups of people with identical behavior and characteristics, except for their gender, are offered different prices for the same product. Within this context, unjustifiably treating individuals unequally can be understood as an act of discrimination (Romei & Ruggieri, 2014).

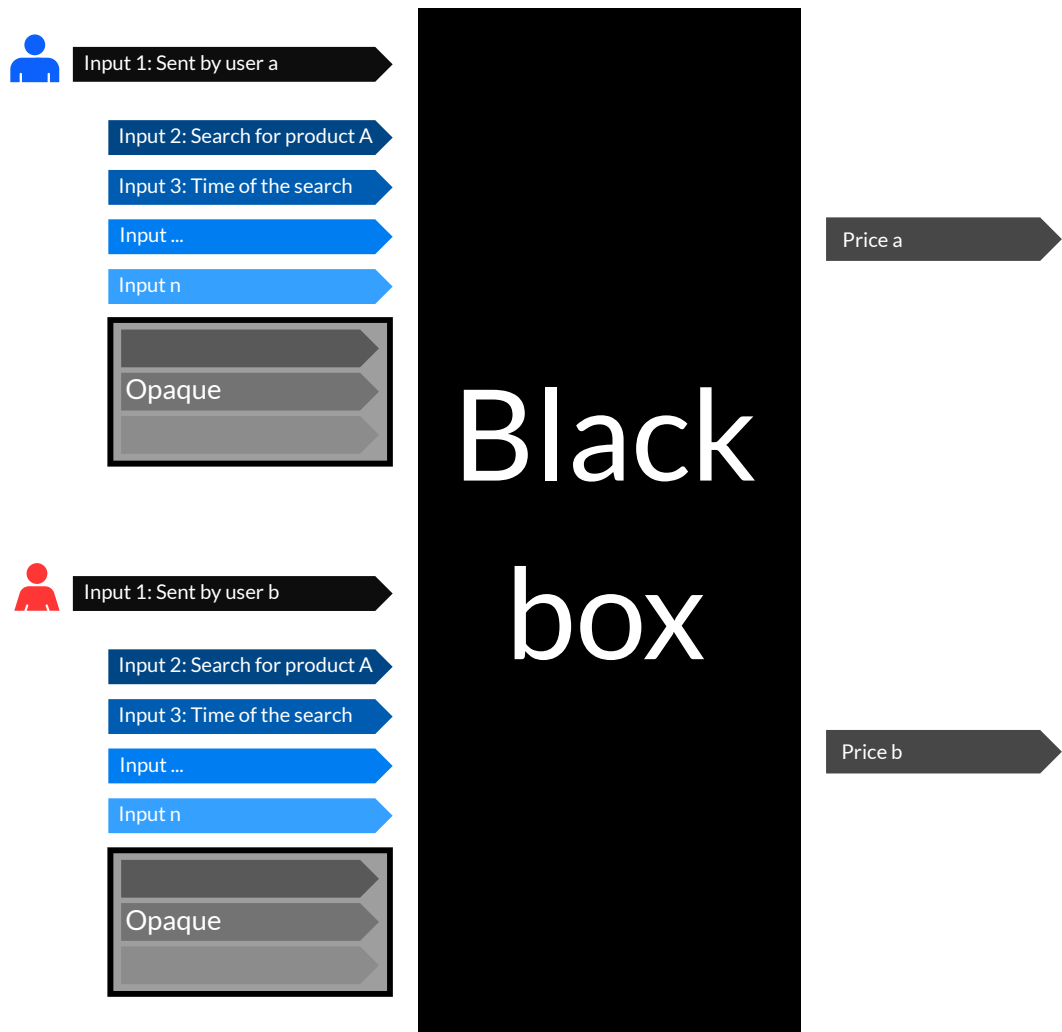


Figure 7.6.: Pictorial representation of the investigation of an algorithmic decision system of an online shop for pricing with a sensitivity analysis within the framework of a black-box analysis (see Section 2.3). Here, all known inputs (in the picture: input 2 to input n) are kept constant except for the querying person, so that the delivered price can then be compared.



This type of investigation represents a sensitivity analysis (see Figure 7.6). Within the framework of a black-box analysis, there exist a variety of options for examining the pricing of online stores, which are presented and discussed below. Such an analysis might be conducted, for instance, in the context of a data donation (see Section 5 and 6). It would be necessary to solicit customers' participation in the data donation and submission of their naturally created and managed profiles. The issue, however, is that exact knowledge of the person behind the profile is required; otherwise, only price differences, and not the reasons for them, can be identified. For a differentiated analysis of the results, it is necessary to have insight into numerous, even personal, aspects of the characteristics of the study's participants; however, even with this knowledge, there is a high risk that these will be overlooked in the study or intentionally concealed during the survey. Moreover, real-world user studies can be difficult to scale due to distrust and accessibility (see Section 6). In addition, it is unlikely that two sufficiently large user groups can be identified whose behavior is identical with the exception of a certain trait to be investigated. Due to the limited number of available data sets, however, this would reduce the interpretability of the results or make it impossible to examine any characteristics. While such issues can be resolved by offering monetary incentives, convincing a representative group of members of the public to install the analysis-capable software would require a substantial amount of money. In addition, these individuals would have to be willing to open their customer account to the extent that price inquiries can be made, which would require an exceptionally high level of trust in the regulatory body. Due to these factors, conducting a black-box analysis in the context of a data donation would be extremely expensive and methodologically challenging. A scraping audit would not be possible due to the massive number of requests, as this virtually impossible behavior of a human user would trigger security measures that the site operators have installed as protection against bot attacks (T. D. Krafft et al., 2020).

Therefore, the only practicable study method remaining is the automated sock puppet approach, which employs carefully designed user profiles and will be analyzed here in detail in terms of its potential applications. First, two groups of user accounts must be generated systematically in such a way that, with the exception of a predetermined property, they appear identical to the pricing machines. If these two user groups, which differ in only one legally protected characteristic, were to pay different prices for the same product, this would be a strong indication of disparate treatment. If the prices for one group were statistically significantly higher than those for the other, then this method could be used to prove unlawful unequal treatment. Due to the business-sensitive nature of the data used to create a user profile at an online store, the shop operators do not publish the exact processes and information regarding which data is used to create a user profile. The generation of suitable profiles is therefore a difficult problem that can only be solved with great effort. Without concrete insight into the exact processes of the individual online shops, it cannot be determined with certainty which, if any, internal

factors are defined and how they are extracted from the available data. In addition, as profiling is typically performed indirectly through behavioral observation, it is difficult to determine what the pricing engine uses to determine prices. A user account that is intended to appear “female” or “financially strong”, for example, must also behave accordingly, because an online store will only adjust prices based on its model if the characteristics it may filter for are actually identified. If there are no direct indicators of a user’s financial situation, the online store will not adjust prices based on this factor. While it is possible that the programmers of the pricing engine specifically sought out behaviors that target characteristics such as gender or purchasing power, profiling may have been left entirely to the software. In this instance, a machine learning procedure would search for digitally observable behaviors that are associated with higher prices and those that suggest the opposite when prices are lower (Wiedmann et al., 2002). The actual characteristic, however, can only be assumed. It could be, for instance, that the average number of days a product remains in a user’s shopping basket or the number of visits to the same product’s website by a user before the product is purchased are the determining factors for the price. Additionally, these may or may not correlate with gender and/or purchasing power. Typically, such data is tracked through cookies left by visited websites, or the user has a profile in which the store maintains their past purchases and other data of interest to them (Eirinaki & Vazirgiannis, 2003). Without knowledge of the source code, it is typically impossible to reliably manipulate individual data points in order to comprehend how they influence pricing. Essentially, a precise definition and description of the profiles used for the studies is required to ensure that the websites capture the characteristics to be studied optimally and that the user groups can be distinguished optimally. In order to imitate genuine interest, real user behavior on the websites must be simulated as naturally as possible. Random opening and closing within a few seconds, which would be simple to implement, would likely be filtered out and not used for profiling by the website operators. All of this demonstrates that the creation of such realistic user profiles is extremely complex, which explains why it has been implemented so infrequently (Badmaeva & Hüllmann, 2019).

In the past, there have been attempts to create such profiles in order to quantify the degree of personalization of prices. On the premise that lower-income users are offered lower-quality goods, general profiles have been developed to provide indications of user creditworthiness. There is a possibility that they will be shown different products or prices than people who are assumed to be wealthy and therefore have a supposed greater willingness to pay (Mikians et al., 2012; Schleusener & Hosell, 2015; Vissers et al., 2014). Such profiles are based on the premise that online retailers view as more affluent users who frequently visit luxury goods websites but rarely or never use comparison portals (Mikians et al., 2012; Schleusener & Hosell, 2015; Vissers et al., 2014). Although this assumption sounds plausible in theory, it is not sufficient to draw conclusions about the actual financial situation by merely observing a luxury item. It would be equally possible that a user enjoys browsing luxury goods and frequently visits related websites,

but is unable to make a purchase due to their limited finances. Therefore, they should be included in the low-income category, as they apparently cannot afford the desired items.

The user's financial situation can be deduced with greater precision based on the user's hardware and operating system. These can be read via the browser and, due to the significant price differences between the individual systems (Schleusener & Hosell, 2015; Vissers et al., 2014), enable a more accurate determination of which of the two income groups the user would fall under. A smartphone with the Android operating system, for instance, can be purchased relatively inexpensively, whereas even older models of Apple's laptops remain expensive. A person's geographical location is yet another criterion that could be used to investigate pricing. Despite the fact that there may be substantial price differences due to taxes and customs surcharge at national borders, prices within a country should be stable and independent of the current electronically recorded location. This information is relatively simple to simulate, as the location is disclosed by the end device itself when accessing the Internet and can therefore be actively masked or redirected to a predetermined location (Schleusener & Hosell, 2015; Vissers et al., 2014). Since the simplicity of such redirection is also familiar to online store developers, it is questionable whether they rely on this information. In the event that an automated pricing system uses behavioral characteristics that are not directly protected by law, but strongly correlate with it, it would not be appropriate to use automatically generated accounts that clearly designate gender (as is possible with Google and Facebook accounts, for instance) but do not display the actual relevant activity that the pricing agent is looking for. Here, price differentiation "by gender" would not be observed - not because it does not exist, but because the price-setter's decision-making process is unknown. Amazon's attempt to assess job applications using an ADM system is an illustration of this. According to journalistic accounts, the system discovered the word "Women" as in "Women's Chess Club" or a degree from a women's university as a direct dependency (confounding variable), even though the gender of the individuals was purposefully withheld. Therefore, without knowing the gender explicitly, the system made decisions indirectly biased regarding gender (Dastin, 2018).

On the basis of the preceding explanations, it can be assumed that the profiling of online retailers does not or need not adhere to the traditional dimensions of age, gender, purchasing power, etc. On the other hand, it is possible that neither an explicitly stated characteristic nor any grouping that is comprehensible to the operator is used, since the ADM system records and uses characteristics for each product category to determine the price. These can correlate strongly with legally protected properties without the above method being able to detect it if the artificial profiling does not include the characteristics used by the ADM system.

In order to "train" user accounts in this scenario, it is possible to execute the following procedure, which is based on multiple assumptions: At first, various online actions could be carried out by controlled computers with accounts that have been created. To make

these automatically created user profiles as authentic as possible, the use of freshly initialized computers is recommended. Then these devices should simulate specific user behavior in the browser for a sufficient length of time so that the profiling possibly used by the shop operator will treat them as genuine customers.

### 7.2.1. Measuring price differentiation and personalization

To determine price differentiation in general, all study participants, whether humans or pre-trained bots, must submit their requests to an online shop at the same time (see Figure 7.6). Each request that results in the same price quote can be ignored. Different prices for the same product, on the other hand, must be analyzed in greater detail, as there may be multiple causes for this. Even requests made at precisely the same time do not need to arrive at the respective server (computer of the online retailer) simultaneously. Even if it were possible to control this, it is possible that they would be processed sequentially. This is comparable to the behavior of individuals who arrive almost simultaneously at a single checkout and then agree on an order in the line in front of it. Different prices may therefore occur, but they are dynamic in nature and do not indicate personalization. The delayed delivery of a price change due to so-called caching is an important issue when measuring price differentiation and personalization. Due to the size and complexity of the Internet, it is imperative that web pages are cached in multiple locations. This is the only way to ensure timely delivery of web content in accordance with user or customer expectations. These restrictions are extremely stringent. For instance, Google has discovered that more than half of mobile users will abandon a page if it takes longer than three seconds to load (Osmani & Grigorik, 2018). Therefore, a local copy of web pages should be stored if possible so as not to exceed this time limit. This process is referred to as caching, and it can result in changes to web pages not being immediately visible to users if they retrieve an outdated copy instead of the original from the origin server (Kaul et al., 2012). Such differences are typically imperceptible to users and have no significant impact. However, this is not the case for websites that frequently adjust their prices. It is possible, for instance, for two people to access the same product under the same address (URL) but be shown two different prices, not because the operator intends to do so, but because the two requests have received two different copies, one of which was loaded from an outdated cache and the other came directly from the operator or an already updated cache (Wessels, 2001). For the study presented here, it follows that prices must be queried as quickly, frequently, and from as many geographical locations as possible in order to identify any outliers and avoid falsely denouncing them, which could diminish the credibility of genuine discoveries.

Due to the varying update frequency of these caches, different prices may exist in the system at the same time; given the locally organized nature of Internet connections, it may even appear that a particular region has been given a price advantage or disadvantage. Examining the response server for the respective price offer can provide additional

information, but this would only be possible in cooperation with the respective hardware operators. Particularly with contingent products such as travel, hotel rooms, train or concert tickets, it is fundamentally difficult to differentiate personalized prices from dynamic ones (Klein & Steinhardt, 2008, p. 177 et seq.). Here, it is always possible to argue that the drawn sequence in the processing of more or less synchronously arriving inquiries has resulted, for instance, in the later inquiries being offered higher prices. To demonstrate personalization, particularly personalization based on legally protected characteristics, it is necessary to initialize automatically created accounts so that they correspond to profiles that differ only by a single characteristic.

### **7.2.2. Challenges and state of the art in monitoring prices in online retailing**

If one wants to monitor prices for online retail, a number of technical obstacles must be considered. Mikians et al., 2012 describe three fundamental issues: There is no standard website layout for online shops. This makes it more challenging to query prices on the various pages, as the script for querying prices may need to be adapted for each layout. The second issue relates to potential fluctuations, which may occur as a result of A/B testing (Kohavi & Longbotham, 2017) or time differences between the original and comparison query. Factors such as physical locations, system problems, caching, and browser history constitute the third issue. Hannák et al., 2014 describe additional difficulties: They state that all queries that are to be compared with one another must occur within a very brief time interval; otherwise, measurement distortions may occur. The authors elaborate on additional biases in distributed infrastructures. It is possible for different data centers to charge different prices for identical cached search queries (Wessels, 2001). Despite these obstacles, various research methodologies have already been employed to investigate pricing in online shops. Mikians et al., 2013, for instance, conducted a study in the form of a crowdsourced audit using a small sample size. For this purpose, the researchers developed a Chrome and Firefox browser extension. This software provided the researchers with a means to compare direct requests for the individual prices of various products. To accomplish this, they initially marked a price on a website. After confirming that the price had been correctly identified, the request was sent to 14 observers, and the prices were subsequently queried. These observers were geographically dispersed, and the results of the various queries were saved in a database for subsequent analysis. Over the course of five months, data from a total of 1,500 queries was collected and analyzed. During this period, more than 188,000 price inquiries were made. The authors discovered a dynamic price distinction. They also discovered a price difference between logged-in and non-logged-in users on Amazon. This could, however, only be determined by three observers. Mikians et al., 2012 developed their own framework for measuring personalized prices in a subsequent study. It includes a browser, a measurement server, and a proxy server. The proxy server has three distinct functions in this context: All traffic from participating users is

routed through the server and stored there as the first function. The second function is the technical adaptation of the HTTP response header for the browser. The proxy also permits the addition of additional privacy features, such as “Do Not Track” options, to the HTTP header. In this configuration, the measurement server acted as a controller for the individual browsers, which also stored the results. In turn, the browsers were executed on distinct, separate, and local machines located at different sites and running on different operating systems. In addition, it was ensured that the browsers always operated in a “clean” environment, i.e., that they represented a brand-new, unidentified system from the shop owner’s perspective. In the study, the websites were queried with a total of eight distinct server-browser combinations, which were initially trained as personas for seven days. Using these personas, 20,000 unique measurement points were saved over the course of four days. According to the authors’ interpretation, there was no evidence of personalized price differentiation. However, different prices were discovered at different locations. Hannák et al., 2014 published another study in 2014, including two experiments. The objective of the first experiment was to determine which prices are displayed to individual users using Amazon Mechanical Turk. The users then automatically queried previously selected websites through a proxy server operated by the researchers. This proxy logged traffic and saved the entire HTML code of requested websites for analysis. The second experiment investigated what factors can result in personalized pricing. The measurement tool PhantomJS was used for this purpose. This tool is a complete implementation of a Webkit browser that allows JavaScript execution, cookie management, and much more. Compared to a standard browser, it is significantly more efficient to work with a large number of machines in an automated manner. In addition, user-defined scripts can be utilized there. Due to the total control, various attributes (cookies, browsers, and devices) can be set as desired, which could lead to differentiation and thus be used as inputs for black-box analyses. This enabled the analysis to determine that significant disparities exist, particularly in the travel industry (hotels and car rentals). In the car rental industry, for instance, Amazon Mechanical Turk users exhibited variances of up to 3.6%. (Hannák et al., 2014) However, the nature of the industry suggests that this is dynamic price management, as Amazon Mechanical Turk cannot coordinate tasks in sufficient time to recognize personalized pricing. The following characteristics were identified as factors in personalization: operating system/browser, account ownership on the website, and history of clicked or purchased products. In addition, A/B testing was identified on specific websites in an effort to entice users to book more expensive hotels.

On the basis of the results of the previously described approaches and studies on the measurement of personalized price differentiation, special care was taken in the design of a black-box study on this topic to ensure that the measurements to be compared would fall within a very short time interval; otherwise, as described above, confusion with dynamic price differentiation might occur. Therefore, queries at precisely the same time would be ideal. However, this is not possible, as even queries sent simultaneously

do not arrive at the server at the same time. In the following, two black-box studies are presented that I supervised, which aimed to examine personalized pricing. The initial analysis addressed how a black-box study can detect price differentiation. In a second step, it was determined whether potentially legally protected customer characteristics form the basis of personalization, thereby raising an initial suspicion of illegal price differentiation. The study design of the first step was therefore expanded to include an initial bot training process for this purpose.

### **7.3. Findings of our black-box analysis of price differentiation in online retailing**

The following section describes the first step of the presented study design, where price differences are recorded without the subject of profiling. My primary emphasis regarding this investigation was on the conceptual level. First, this research aimed to extrapolate and apply insights gleaned from prior black-box analyses to this particular domain, and to evaluate the effectiveness of various approaches. Second, it was apparent from the outset that this area of investigation presented significant obstacles to the methodology of black-box analysis. Efforts were made to surmount these challenges throughout the course of this study. However, the actual implementation and evaluation processes were delegated as components of subsequent student theses. The EU-Preis Plugin, which was designed and implemented as part of two theses at the Algorithm Accountability Lab of TU Kaiserslautern, is presented for the modular study design of the investigation of price differentiation in online retailing. In a preliminary study, the instrument was evaluated to identify any changes in the structure of the shops and to adapt the instrument to these changes. The initial concept was a crowdsourced study design, similar to the data donation for the 2017 German federal elections (Section 5), but it quickly became apparent that automated bots could also be operated with the tool with a few modifications. EU-Preis is a browser extension that attempts to identify price differences in online retail. It was developed during Roman Krafft (R. Krafft, 2018) and Marcel Wölki's (Wölki, 2018) respective Bachelor's theses, which I supervised. The development of the software component is described in detail in their works, while a brief summary is provided below.

In order to implement the strategy, the plugin was divided into the Creator Tool and the User Plugin (see Figure 7.7). The Creator Tool is responsible for the creation of studies and the collection of submitted data, whereas the User Plugin retrieves prices automatically and transmits them to the server infrastructure. Studies are initially created using the Creator Tool. During this phase, the researcher can establish fundamental study parameters. These include the study's title, a detailed description, and a notation of the beginning and end of data collection. In addition, the query interval that the user plugin uses to query the price is set here. The creator can then visit the web-

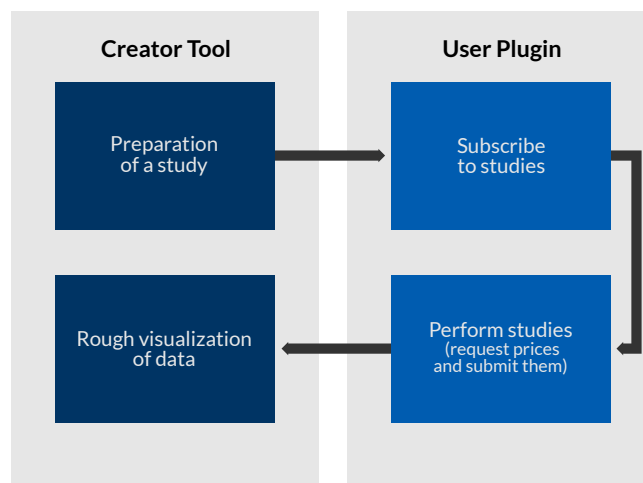


Figure 7.7.: Allocation of the individual tasks to the different plugins.

site and select the price to be analyzed to determine which value will be fetched and submitted by the user plugin (see Figure 7.8). After the studies have been created, the researcher can upload them to a self-hosted server and launch them there, at which point the users or machines configured for the study can log in (see Figure 7.9). Both a sock puppet audit and a crowdsourced approach with the assistance of citizens are possible with this tool. The principal component of the tool is the user’s browser plugin, which is compatible with Google Chrome and Mozilla Firefox, the two most frequently used browsers (Statista, 2023b). These plugins can request the websites specified in the study parameters at the given intervals, collect the prices displayed, and send this data to a central server.

After the user selected one or more studies to participate in, the plugin assigns a randomly generated, anonymous user ID for each study which will be submitted along with any price data collected. From this point on, the data collection is performed at predetermined times for all participating user plugins until the set conclusion date. After the query, the user plugin sends the data to a central server. All submitted data can then be viewed and downloaded as a comma-separated values (CSV) file using the Creator Tool.



### 7.3. Findings of our black-box analysis of price differentiation in online retailing

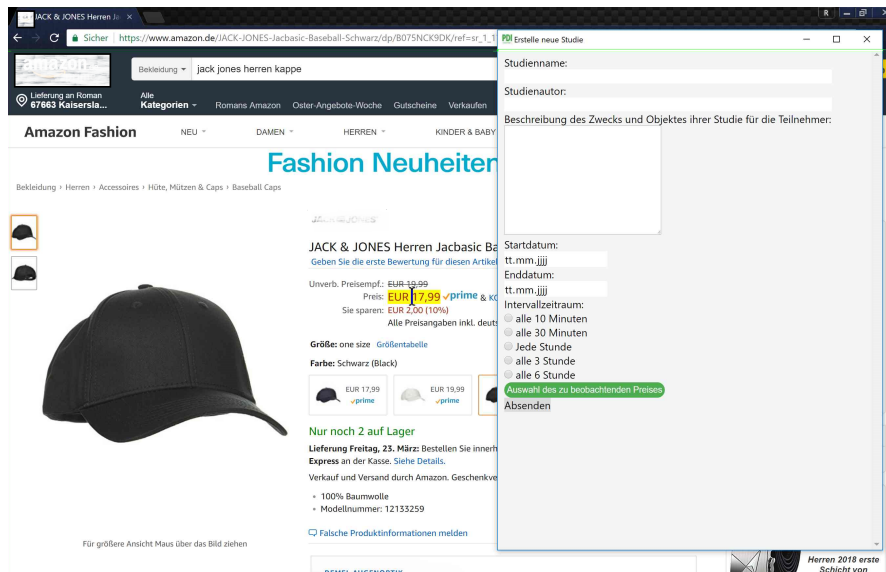


Figure 7.8.: Creator Tool when selecting a price.

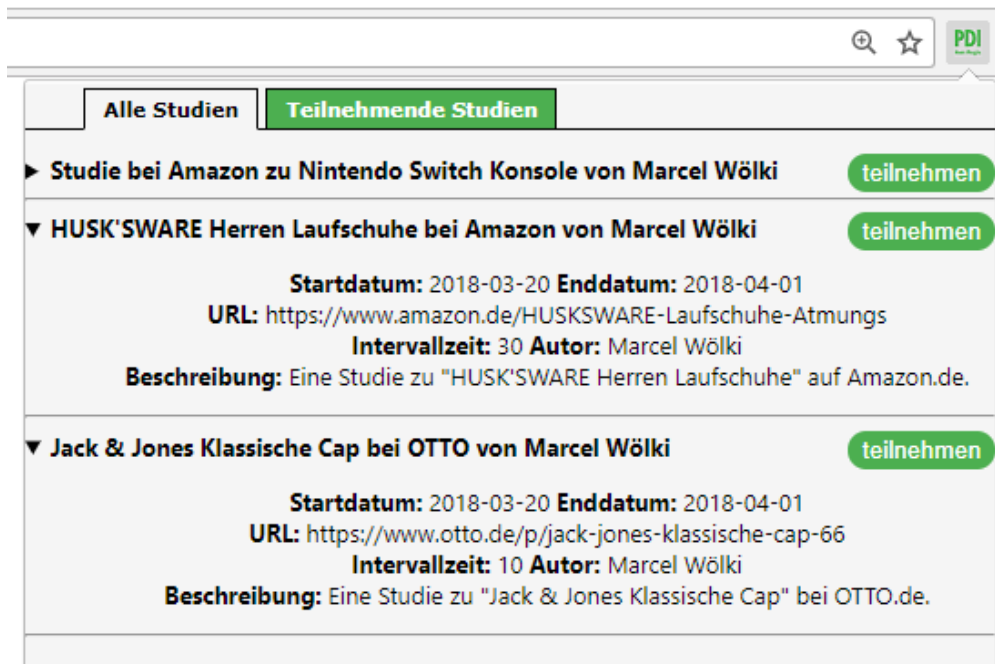


Figure 7.9.: Browser plugin where the user can participate in predefined studies.

### 7.3.1. Preliminary study with EU-Preis

As part of Roman Krafft and Marcel Wölki’s Bachelor’s theses, a preliminary study (09 Feb. 2018 - 23 Feb. 2018) of the eight largest online shops by revenue in 2016 (R. Krafft, 2018) was conducted. The study focused on investigating both long-term viability and compatibility with the highest-volume online retailers. To achieve the highest degree of comparability possible in the studies, the same products were observed in as many of the selected stores as possible.

Table 7.1.: Overview of the preliminary study from 09 Feb. 2018 to 23 Feb. 2018.

Online store	Query interval	User plugins
Amazon	10 minutes	13
Zalando	30 minutes	11
Cyberport	60 minutes	6
Otto	60 minutes	13
Notebooksbilliger	180 minutes	13
Tchibo	60 minutes	12
Bon Prix	360 minutes	14
Conrad	360 minutes	15

However, it was immediately apparent that the presentation of prices at various retailers varies greatly. For example, the price of a product at the online retailer MediaMarkt is not displayed as a string of characters, but rather as a series of pictures of individual numbers. As a defensive measure against “price comparison sites”, online shop providers could employ this and other methods to make the automated reading of current prices more difficult.

Table 7.1 provides an overview of the studies, including the investigated online shop, the query interval, and the number of self-configured computers using the browser plugin to collect and submit data.

During this preliminary investigation, no direct evidence of personalized price differentiation was discovered. Some products experienced dynamic price fluctuations, but there was never any significant disparity between participants. Minor lags were detectable, but they nearly always disappeared in subsequent measurements. However, during the course of the study, the presentation of prices on the websites varied on occasion. This was first observed when Amazon’s website displayed a “price reduction” in red letters, but other retailers also displayed prices differently. Since the Creator Tool stores a fixed process for extracting the price from the HTML code of the web page for each study, this observation prompted us to actively request a new extraction process from the server when starting the client browser so that we could respond to a changed price display. This feature was not implemented, however, due to its complexity.

### **7.3.2. Study with EU-Preis**

Following the preliminary study, a more comprehensive study with a total of 41 studies was conducted between 20 February and 6 March 2018 at four of the largest online shops. To reduce product-related price fluctuations, the decision was made to favor products sold by as many retailers as possible. Twenty identical computers were set up for each study. Since these instances submitted identical requests to the online retailers, the outcomes should have been identical. However, a comprehensive comparison of all stores is impossible because some of them, such as Zalando and MediaMarkt, have dissimilar product selections.

### **7.3.3. Interpretation of the results**

Throughout the duration of the study, the prices of the examined products remained remarkably consistent and did not fluctuate significantly. Out of the 41 studies conducted, only thirteen of them experienced any changes in prices. Additionally, among those thirteen studies, only four of them revealed variations in prices during the same submission time. A closer examination of the data from these studies revealed that prices only differed at a single point in time, a variation that can be explained by the argument of caching (for a more in-depth discussion of the study results, see (R. Krafft, 2018; Wölki, 2018)). Excluding the price differences that were only measured at a single point in time and can be accredited to the aforementioned necessity to cache web pages, only the MediaMarkt study showed a price difference in a 10- to 29-minute time frame. Nonetheless, this difference can be explained by the fact that it was measured at two distinct times, 10 minutes apart. The actual duration may therefore fall within this range.

The results indicate that without comprehensive profiling, identifying potentially problematic pricing practices using black-box analysis is exceedingly difficult. The analysis requires a significant amount of data and requires working under several assumptions to provide meaningful results. In legal discourse, however, such assumptions may be deemed too uncertain, for instance in the context of an evidentiary hearing.

### **7.3.4. Pre-training profiles with bots**

In Arthur A. Just, 2019's Bachelor's thesis, the EU-Preis plugin was expanded to include the "training" of bot accounts to investigate whether a detected price difference is caused by temporal effects (such as contingent prices) or buyer characteristics. In order to accomplish this, a browser plugin was created that could be given instructions for automated browser control and then perform "natural" web browsing according to the specifications (A. Just, 2019). The plugin could load web pages, follow links based on predefined criteria or at random, and include suitable, partially random waiting times.

During research and implementation, it became apparent that it was impossible to validate whether the bot “acquired” the desired properties.

If the study design wanted to determine whether young and old people were shown different prices, it was not possible to determine whether this property was accurately recorded by the online shop after, say, two weeks of simulated browser use. A small study conducted as part of the Bachelor’s thesis revealed that careful modeling of website usage for the categories “gender” and “wealth” as well as a two-week “training” of the bot accounts based on this had no real effect on the prices displayed on the websites we analyzed.

However, the following issue arises with profiling by online retailers: It is impossible to determine whether all the variables by which the website differentiates profiles are also reflected in the profile, as these variables are unknown. How an online retailer can generate a specific user profile is therefore the subject of current research. The current status of previous research on this topic is summarized in Section 7.1.3. When the involvement of pre-trained bots is necessary or planned, the following points should be considered:

- The development of a scientifically robust description for the targeted profiling of ‘natural’ online behavior and the subsequent creation of automated browser control for the unobtrusive implementation of this online behavior is so complex that it can be considered a separate field of study.
- For each characteristic under investigation, a profile imitating this behavior needs to be designed.
- Enough accounts must be configured for each specific behavior. For instance, one group of accounts may visit a number of publications per day, click on some of the links to articles they provide, and remain on the pages for an arbitrary amount of time, whereas another group of accounts may simply browse social media.

Only with the assistance of these automatically generated accounts can a variety of price measurements be performed then.

### 7.3.5. Lessons Learned

Finally, the general and technical insights obtained from the studies in this section will be presented in order to investigate the possibility of personalized pricing in online retail. In general, it must be stated that it is impossible to determine all of the input variables due to the lack of information about the structure of the systems used by shop owners for pricing and the sometimes vast disparities between online retailers. Pricing can be based on a large number of inputs, making profile creation disproportionately time-consuming and bot-assisted verification nearly impossible. In addition, the unclear input variables for a pricing algorithm lead to problems in the design of the profiles to be examined, so

for the training of bot accounts, it cannot be determined precisely which user behaviors should be simulated when using the website. In this particular instance, interpreting the results is complicated by the fact that a black-box analysis cannot differentiate between personalized and dynamic prices. In this regard, any personalization found, even if it indicates unequal treatment based on a protected characteristic, could be rationalized relatively easily by the retailer through an unknown variable or context. Black-box analysis cannot definitively refute the claim that there is no personalization or that pricing is merely a response to dynamic stock changes, as is the case with contingent pricing, in the absence of further insights into the pricing algorithms used.

#### **7.3.6. General Learnings**

Even though the collection of prices was a success and price differentiations could be identified using this method, I find the interpretability of the black-box approach in this context to be challenging. The following obstacles argue against investigating pricing on the Internet using black-box analyses:

1. Due to a lack of basic knowledge about the structure of the pricing algorithms used by shop operators and the differences between online shops, it is impossible to determine all of the input variables. Nonetheless, as depicted in Figure 7.6, this information is necessary to ensure that a difference in output is caused by a change in the controlled input and not, for instance, by one or more hidden variables.
2. Setting up bot accounts as part of a sock puppet audit is incredibly challenging; more on this in the next section on technical learnings.
3. Crowdsourced auditing necessitates massive citizen participation: When real user profiles are used to cover all eventualities, the amount of data required would continually increase, making anonymization equally challenging.
4. A black-box analysis cannot conclusively determine whether the observed price difference was caused by dynamic pricing rather than personalized pricing.

#### **7.3.7. Technical Learnings**

Setting up bot accounts as part of a sock puppet audit introduces significant challenges that complicate both the initial setup and ongoing management. One critical factor is the complexity of modern web structures. Websites increasingly use dynamic content, such as JavaScript-driven elements and complex layouts, which make it difficult for bots to reliably extract relevant data. This complexity demands advanced parsing techniques and often requires bots to emulate browser behaviors precisely to avoid detection.

Risk of detection is another major concern. Many sites deploy sophisticated anti-bot measures, including CAPTCHAs, device fingerprinting, and behavioral analysis, which

can identify non-human activity patterns. As these detection mechanisms grow more advanced, the likelihood of bots being flagged or blocked increases, which can disrupt the data collection process and lead to incomplete or biased results.

Moreover, mimicking genuine human behavior is a resource-intensive task. Bots must perform actions with human-like timing, navigation patterns, and even personalized responses to avoid detection. This requires creating diverse behavior profiles and fine-tuning interaction sequences, adding considerable complexity to the bot setup and maintenance. Additionally, each bot instance must account for randomized or context-specific interactions that resemble typical user behavior to reduce the risk of detection.

Uncertain input variables for pricing algorithms significantly complicate the design of the profiles to be studied. It is unclear which behaviors must be simulated during bot account training. This lack of validation capability raises questions about the reliability and interpretability of the collected data in such black-box investigations.

The conducted preliminary study was able to reveal incorrect assumptions and programming flaws. Due to this, it should be regarded as an essential component of a black-box analysis. The preliminary study also revealed that in order to react appropriately to any changes in the examined site, the capability to influence the precise behavior of client-side software is of utmost importance. As a potential solution, it has become apparent that the client should routinely check for server-side configuration updates.

The experience of the research and the projects discussed in this work serves as the foundation for the process model for black-box analysis that is offered here.

As stated in the introductory section, one of the goals of developing the process model was to create an instrument based on interdisciplinary knowledge and provide non-computer scientists with a straightforward and efficient method for developing black-box analyses. For this reason, each individual process stage is accompanied by associated guiding questions that aid in the planning and execution of a black-box analysis.

A black-box analysis of an ADM system is frequently prompted by a strong suspicion that an ADM system induces undesirable side effects. Even if it is assumed that ADM providers operate in good faith and to the best of their knowledge, the numerous critical investigations of ADM systems suggest that a review of such systems must be possible whatever the circumstances.

The purpose of the process model developed in this Section is to ensure that false system behaviors can be detected and documented, regardless of whether they are the result of system mismanagement, unintended false positives, or deliberate fraud.

Throughout my investigation, the process model introduced in this dissertation has been developed and adapted: An initial version can be found in my paper “Why Do We Need to Be Bots? What Prevents Society from Detecting Biases in Recommendation Systems ” from 2020 (T. D. Krafft et al., 2020). Furthermore, Roman Krafft utilized a prototype of my process model in his Master’s thesis to aid in devising a comprehensive research design for examining personalized political advertising on Facebook. The current process model has been updated to reflect the experience gathered during my research process.

The process of a black-box analysis is always highly context-specific; however, as seen in Figure 8.10, there are abstract process steps that must be considered while developing a scientific research setting.

In the discipline of cybernetics, William Ross Ashby refers to three key questions that researchers must consider in a black-box analysis (Ashby, 1956, p. 87):

*“How should an experimenter proceed when faced with a black box ?”*

*“What properties of the Box’s contents are discoverable and what are fundamentally not discoverable ?”*

*“What methods should be used if the Box is to be investigated efficiently ?”*

In the context of black-box analysis of ADM systems, these can be transferred as follows:

1. What is the nature of the analysis, i.e., what specific research question is to be examined?
2. Which aspects of the investigated ADM system are opaque?
3. What is the precise method for investigating this question?

The process model of a black-box analysis of an ADM system depicted in Figure 8.10 is based on these three questions. In addition to findings from social science audits (see Section 2.4.1) and software testing, the process model is based on my own experience conducting black-box analyses (see Section 5, 6, 7) and the review of a large number of studies conducted in recent years (Bandy, 2021; Vecchione et al., 2021, see, among others).

Referring to the fact that the Royal Society, the world’s first scientific organization, adopted the motto “nullis in verba”<sup>1</sup> (Hunter et al., 1989, p. 17) in 1660, scientific knowledge in the sense of “trust but verify” should only be acknowledged if independently repeatable experiments demonstrate its truthfulness. Consequently, the presented process model intends to design the experimental environment as optimally as possible, following the course of a traditional scientific experiment. For this reason, the conception of the process model’s individual steps is based on social science field experiments (see Section 2) that effectively implement this procedure.

The following individual phases have been shaped by experiences from external sources as well as my own research. There are guiding questions at the start of each Section that allow for proper planning and implementation of the appropriate phase.

---

<sup>1</sup>Latin for “on the word of no one” or “take nobody’s word for it”



## 8.1. From suspicion to falsifiable statement



Figure 8.1.: Process step 1: From suspicion to falsifiable statement

As indicated previously, a black-box analysis is conducted on the basis of a strong suspicion that an (opaque) ADM system might produce unwanted side effects <sup>2</sup>. Therefore, the first stage is to find and verify a suspicion. Following that, a falsifiable, i.e., verifiable, statement must be established, because only a falsifiable assertion can be scientifically investigated (see Figure 8.1).

### 8.1.1. Find a suspicion

The first guiding question to be addressed is:

#### ?? Guiding Question 1.1 What (undesirable) behavior occurs and under which conditions?

To determine precisely what is undesirable about a system’s behavior, Jack Bandy’s three “problematic behaviors” of an ADM system can be used as a starting point. In his comprehensive literature review, Bandy, 2021 systematizes the myriad of concerns surrounding the use of ADM systems, ranging from individual misconceptions to potential societal dangers, by reducing the objections to three problematic behaviors of ADM systems examined using black-box analysis:

- Discrimination: The ADM system treats individuals unequally based on, for example, age, gender, location, socioeconomic status, and/or intersectional identity (Bandy, 2021, p. 7).
- Distortion: The ADM system presents a selection of reality that is distorted or obscured (Bandy, 2021, p. 7).
- Exploitation: The ADM system misuses content from other sources and/or sensitive personal information of individuals. This can occur, for instance, through the aggregation of data from which conclusions about individuals or their individual characteristics can be drawn (Bandy, 2021, p. 7).

<sup>2</sup>Possible applications not based on such a suspicion are discussed in greater detail in 9

While Bandy’s description of discrimination focuses heavily on the legal concept of discrimination (“treats or disparately impacts people on the basis of their race, age, gender, location, socioeconomic status, and/or intersectional identity” Bandy, 2021, p. 7), this field can be expanded to include the unregulated area of unequal treatment. The legal fact of discrimination is only met when individuals are treated differently based on fixed characteristics in particular settings (Hoffmann et al., 2022). In Germany, for instance, the General Equal Treatment Act prohibits discrimination based exclusively on the personal characteristic ‘age’ within the material scope (domain) of trade union membership. Depending on the focus, these two restrictions (fixed characteristic; specific domain) can be lifted, and black-box analyses can provide valuable insights in cases of unequal treatment based on other properties or in a (previously unprotected) domain. This is illustrated by the personalized pricing study described in the previous section. Personalized pricing in online retail was not prohibited at the time of the study, despite ongoing international efforts to do so (see European Commission & Council of the European Union, 2019).

### 8.1.2. Verify the suspicion

Since a black-box analysis is always associated with costs and efforts that must be justified, it is essential to carefully substantiate the identified suspicion that the ADM system significantly contributed to or was the trigger for the observed undesired behavior. The following guiding questions can assist in formulating and substantiating a suspicion:

**?? Guiding Question 1.2 Which groups of people are affected?**

**?? Guiding Question 1.3 Which ADM system is involved?**

**?? Guiding Question 1.4 How can the ADM system contribute to the suspicion?**

Depending on the nature and form of the suspicion, a variety of approaches may be used to substantiate it; however, the primary objective here is to provide solid evidence and convincing justifications for why and how the ADM system is involved in the formation of this specific suspicion. Utilizing the phenomenon-induced socioinformatic analysis developed by Zweig et al., 2021 is an effective method for achieving this, as its structured methodology encompasses a broad range of applications. As previously described in section 4, the method is suitable not only for elaborating which actors and technical components of the ADM system are involved in the socioinformatic phenomenon, but also for corroborating whether and to what extent an ADM system in its socioinformatic environment can support the problematic behavior described above. Specifically, it allows for determining which components of the socioinformatic system could be the trigger of the undesired behavior.

In principle, it should be determined in advance, if possible, whether the technical

component is even capable of exhibiting the “undesired” behavior. This concept is exemplified by the following hypothetical scenario:

If a human resources manager uses an ADM system in the course of applicant selection to evaluate the spelling of the applications received, but then makes discriminatory decisions based on the pictures in the CV, for example, by inviting fewer or no women to an interview, this is an undesirable behavior, but it is not due to the technical component involved. If, on the other hand, the ADM system is intended to make genuine recommendations for applicant selection, the same result - fewer women in the interview - could be due to the system’s recommendation.

In conclusion, it can be stated that the grounds for suspicion can be complex for various reasons. It is advised determining whether the suspicion can be attributed to a socioinformatic phenomenon (see 2.5). If this is the case, then by definition at least one social actor is involved in the emergence of the suspicion, and this influence must be examined in addition to the technical system. When the suspicion has been identified and specified in an argumentatively robust manner, the phase has been successfully completed.

### 8.1.3. Formulate a testable hypothesis

Given that a black-box analysis is a scientific investigation that applies statistical methods to empirical data, the previously established suspicion must be translated into one or more hypotheses that can be tested. The next step of the process therefore concerns the question:

#### ?? Guiding Question 1.5 Which hypothesis should be tested?

Initially, a hypothesis is an assumption whose truthfulness has not yet been demonstrated. It should be formulated objectively and precisely. Often, the assumption chosen as the so-called null hypothesis ( $H_0$ ) is not the assumption that is truly of interest, but rather the one that one wishes to disprove. This indirect argumentation is necessary due to the fact that empirical experiments can never prove a statement, only disprove it, which is the idea of the paradigm of science that is dominant today. In other words, a statistical test can never prove a hypothesis; it can only reject or fail to reject it with a degree of certainty. When formulating the null hypothesis ( $H_0$ ), it is crucial to ensure that it can be refuted by subsequent experiments; this is known as the falsifiability of a hypothesis (Popper, 2005, p. 57 et seqq.). As a counter-hypothesis, also referred to as an alternative hypothesis ( $H_1$ ), the inverse of  $H_0$  is always assumed, so that in the real world, either  $H_0$  or  $H_1$  is always true.

There is an abundance of literature on hypothesis generation and testing as a scientific method that discusses best practices and common pitfalls (e.g., Gauch, 2003, p. 11, Gigerenzer et al., 2004). An important step here is the written documentation of the hypothesis to be tested. My own research has shown that in the later evaluation of the

data, one might want to modify the actual hypothesis in order to simplify the evaluation or because, for instance, the data for the intended investigation is insufficient or contains errors. Referring to what has been documented discourages casual modifications. Clearly, other hypotheses can be investigated or evaluated at any time using the available data. Nonetheless, it is always necessary to ensure that all framework conditions for the investigation of the new hypothesis are met.

After this phase of hypothesizing, one has reasonable assumptions as to which technical component might trigger the problematic behavior, as well as a testable hypothesis that can be used to investigate this connection.

## 8.2. Design Decisions



Figure 8.2.: Process step 2: Design decisions

Once a testable hypothesis has been formulated, the next step is to determine which investigation methods are appropriate for the ADM system (see Figure 8.2). To be able to choose an appropriate form of analysis, it is necessary to first determine what information and insights about the ADM system are available. These questions will serve as guidelines here:

### 8.2.1. Identify black-box scenario

**?? Guiding Question 2.1** In which black-box scenario does the planned black-box analysis take place?

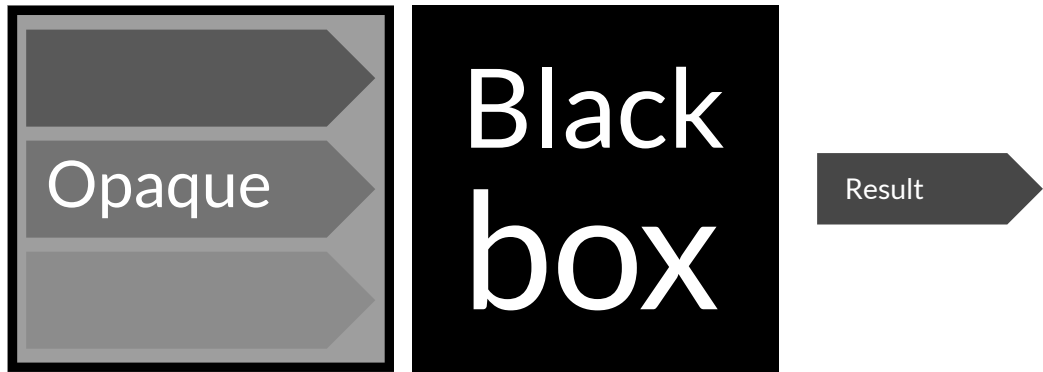
**?? Guiding Question 2.2** Can the hypothesis under investigation be examined in this scenario?

If the operator of the ADM system does not permit official audits or if there is no access to the internal processes of the ADM system for other reasons, it must be determined in each case which forms of analysis are feasible for the opaque system.

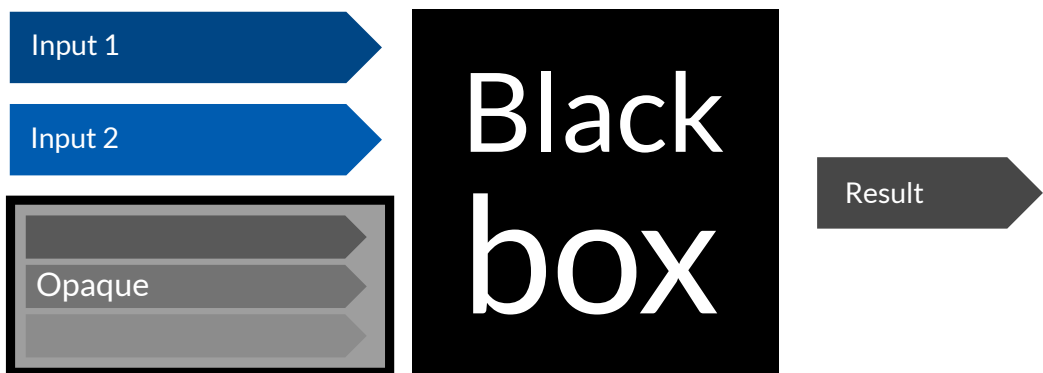
The initial phase in the design of a black-box analysis is to answer the second question derived from Ashby, 1956, p. 87: “Which aspects of the investigated ADM system are opaque?”. Depending on the system, not only the internal processes but also parts of the system’s inputs or outputs can be opaque, so three distinct black-box scenarios can be distinguished based on the information about the system’s inputs and outputs gathered during the investigation (see Figure 8.3), allowing for three distinct cases:

1. Black-box systems in which only the output is observable while the input remains concealed (Diakopoulos, 2014b, p. 10).
2. Black-box systems in which only some inputs can be observed and manipulated, such as during testing, while other inputs are unknown or cannot be manipulated.
3. Black-box systems in which both the input and output are observable (Diakopoulos, 2014b, p. 10).

Scenario 1



Scenario 2



Scenario 3

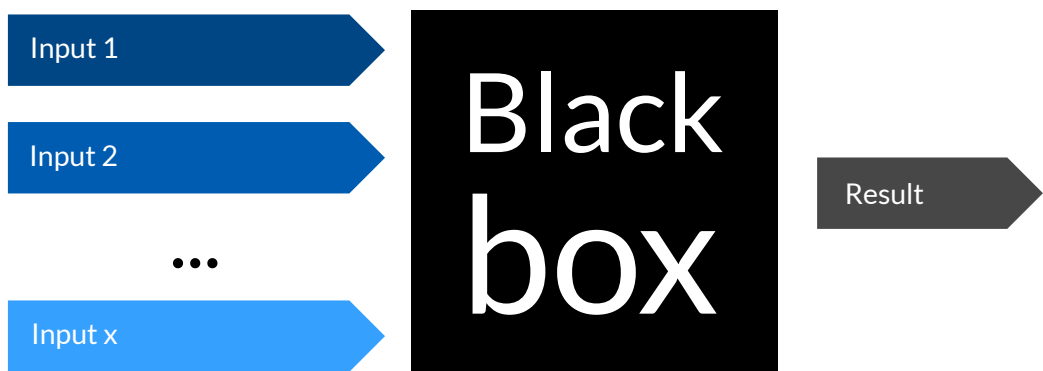


Figure 8.3.: Three black-box scenarios that differ in what knowledge is available about the inputs of the black box.

Analyses in the first category are extremely challenging, since no information about the type or number of inputs is known. The system's outputs can thus only be evaluated to a very limited extent. If, for instance, an unusual system behavior is observed, it cannot be ruled out that it is caused by inputs that are extreme but expected. When analyzing the properties of such a system, only a handful of properties can be examined, since statements about the inner mechanics are largely excluded. These include requirements for the output format and the overall utility of the system, such as performance or similar aspects, in addition to guaranteeing that no wholly unexpected or permitted data is output. It is therefore only feasible to determine whether and how quickly an output is produced, but not whether or not it is correct or adequate. Here, Nicholas Diakopoulos notes that data journalists investigating these types of ADM systems frequently begin their investigations in such an environment (Diakopoulos, 2014b, p. 10). If, for instance, the investigation concerns the algorithmic curation of social networks, it is possible that only the ADM systems' output can be accessed.

During a study conducted with Rhein-Neckar Fernsehen, a regional, private television broadcaster in Germany, to investigate the principles of objectivity and impartiality of reporting, diversity of opinion and the balance of their offerings (RStV, 1991, §11/2) on the Facebook platform (T. D. Krafft et al., 2020), one of the issues that became apparent was the difficulty of investigating an ADM system with opaque inputs over which we as investigating researchers have no control. For results of an ADM system to be comparable, as many inputs as possible should remain unchanged. However, such investigations are only possible to a very limited extent due to Facebook's lack of trustworthy information regarding the information used to determine the individual timelines<sup>3</sup> of its users.

In black-box scenarios of the second category (see Figure 8.3), information is available about at least a portion of an ADM system's inputs so that it can be determined whether these inputs were chosen appropriately for the ADM system's intended purpose. Nonetheless, the analysis has the same fundamental flaw as the scenarios in the first category. Although correlations between inputs and outputs can be observed, it is difficult to assess the influence of unknown inputs. Since unknown influences cannot be investigated, it is difficult to make definitive statements about systems in this scenario. Lastly, the algorithm may take into account inputs that are not observable and therefore not measurable (Pedreschi et al., 2018).

Throughout the investigation conducted during the 2017 German federal elections, our aim was to examine the potential impact of Eli Pariser's filter bubble effect on the shaping of political opinions (see Section 5). However, it was not possible to ascertain until the conclusion of the study whether Google incorporates undisclosed factors, in addition to the known criteria for determining search results such as search terms, past search

---

<sup>3</sup>Facebook users have individual timelines, which include their posts as well as the posts of friends and followed Facebook pages on the platform.

history, location, and similar parameters, that may influence the selection and ordering of search results. The presence of a unique personalization factor or information about the end-user device used for searching is entirely conceivable. Therefore, investigations in a second-category black-box scenario must always account for the influence of uncertain inputs on the evaluation.

In the third type of black-box scenario (see Figure 8.3), all inputs and outputs of the ADM system are provided in detail. Therefore, there are no constraints imposed by unknown inputs, and the relationship between inputs and outputs can be readily investigated. While operators of an ADM system can easily set up a third-category black-box scenario if they do not wish to examine or test the system directly with knowledge of its internal properties, i.e., as a white box, it turns out that for outsiders, clear information about the inputs to an ADM system is rarely available.

Following the previous outline, it should be noted that when analyzing an ADM system, it is possible that any number of unknown parameters will influence the decision-making process of a black-box system. In this circumstance, it can only be assumed that the scenario falls into black-box scenario two (some inputs are known) or three (all inputs are known). This assumption should be documented and taken into consideration when publishing the results of the analysis, as it has a significant impact on the study's validity.

After determining the black-box scenario, it is possible to consider whether the hypothesis and, by extension, the suspicion can be investigated. It has to be considered that in a scenario of the second category, in case of an accusation, the operator can identify the unknown other inputs as the cause of the behavior of the ADM system. In such a circumstance, one would have reached the limits of black-box analysis, necessitating the use of more intrusive investigation techniques, such as white-box investigations.

### 8.2.2. Design of the analysis method

If it is determined that an investigation of the hypothesis in the current black-box scenario is theoretically possible, the next significant design decision is the selection and adaptation of an appropriate analysis method. In light of this, the following Section clarifies which characteristics merit special consideration. The fundamental guiding question is as follows:

#### ?? Guiding Question 2.3 Which method of analysis should be used to test the hypothesis?

In theory, there are numerous forms of analysis, but two distinct analysis techniques can be distinguished based on the respective investigation objective:

1. **Oracle-based analyses:** When the objective is to match a collection of input-output instances with expected results, a so-called oracle, and then evaluate them statistically, this is known as an oracle-based analysis.



2. **Sensitivity analysis:** A sensitivity analysis is required when examining individual input-output instances in respect to one another.

Both objectives can be pursued concurrently, so they are not mutually exclusive; however, combining both analyses increases the complexity of the study design (see Section 6). The two types of analysis are described in detail below.

### Oracle-based analysis

The concept of oracle-based analysis can be transferred from the software testing domain. In the discipline of functional testing, also known as black-box testing, there exist a variety of test methods (Nidhra & Dondeti, 2012) that can be applied to the investigation of black-box systems. As stated previously in section 2, functional testing depends on the isolation of an investigated component in order to verify compliance with a specification through a succession of test cases (Myers et al., 2011, p. 224). Similarities to black-box analysis become apparent when the entire ADM system is considered to be the isolated component.

In the field of functional testing, there are numerous approaches, with *Equivalence Class Partitioning* and *Boundary Value Analysis* being notably applicable for comparison with black-box analysis:

- Equivalence class partitioning implies that the input space can be partitioned and that the system output is identical for all outcomes within an equivalence class (Bhat & Quadri, 2015; Murnane et al., 2007).
- Boundary Value Analysis reduces the output requirement for Equivalence Class Partitioning. Here, it is not determined whether inputs from an input space activate the same output, but rather any output from an output space. The term is derived from the fact that input and output boundaries are formulated (Bhat & Quadri, 2015; Murnane et al., 2007).

These two approaches share the characteristic that, for fixed inputs, the system output is matched with expected outputs. In the community of classical testing, this “expected output” is known as an oracle. This explains the titular name of this category of analysis, as a test result must be validated with the aid of a so-called test oracle as a benchmark to determine whether the system has produced the correct output (Howden, 1978).

A prevalent type of test oracle is the so-called ground truth, where the correct output for a restricted set of inputs is known. A ground truth is utilized in the development and quality assessment of supervised learning systems, for instance. In this case, it consists of the property vector of the object/subject to be evaluated and the “correct”/“expected” system output. Using this data set, it is possible to compare whether a decision system’s output matches the expected outcome. The precise evaluation is conducted using a set of metrics known as quality measures (M. A. Haeri et al., 2022). In addition to directly

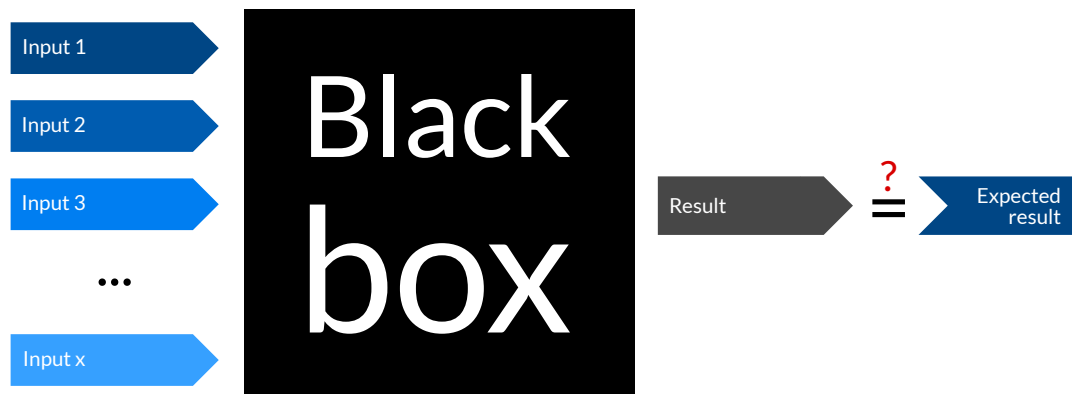


Figure 8.4.: Sketch of an oracle-based analysis of a black box: Here, for a chosen input vector (Input 1; Input 2; Input 3; ...; Input x), it is checked whether the results provided by the black-box system match the expected result, i.e., the oracle.

observing how good the respective decisions of an ADM system are, fairness aspects can also be analyzed by comparing them to the expected results of the test oracle, for instance by examining whether the ADM system treats subgroups differently, i.e., unfairly (Barocas et al., 2019, p. 18 et seq.). The approach is similar to the equivalence class partitioning discussed above. Such measurements of the quality of an ADM system can also be used to compare different ADM systems (Hauer et al., 2020; König & Krafft, 2021).

In addition to an adequate oracle, a comparison metric is required if a black-box analysis is based on matching the output of the ADM system (for a given input vector) to an expected output value or oracle to assess its validity and/or suitability (see Figure 8.4). To determine both, experiences from the testing literature can be used and parallels can be drawn from the following questions: “How do I find the appropriate test oracle?”; “How do I evaluate the information about what can be considered a correct output for the majority of possible inputs, such as a very large one?”; and “How do I use it?”

While there are probabilistic test oracles for which a certain error rate is acceptable, a ground truth typically assumes values to be absolutely accurate.

### Sensitivity analysis

When making decisions regarding people or human behavior, there is typically no ‘perfect’ oracle, as there is frequently no clearly specified right or wrong; instead, previous decisions made by people in this or comparable situations are relied upon. Such decisions

pose a significant challenge for many ADM systems, not only in terms of the accuracy of the prediction (in the sense of being correct or incorrect), but also in terms of whether people with various protected characteristics are treated equally. Partly to resolve these issues, testing methods that do not require an oracle have been developed. They test for a particular ratio or distribution in the results, regardless of whether the results are correct or not. Depending on the definition, this desired distribution may also be regarded as an oracle, although the literature is ambiguous on this classification (Barr et al., 2014).

In order to test for a particular distribution, it is necessary to collect data on these (protected) attributes; otherwise, it is impossible to determine their impact on a decision (Žliobaitė & Custers, 2016). In the absence of pertinent data, it is presumed that an algorithm cannot take it into account. This is true at first glance, but the bias caused by these protected variables can also hide behind other, ostensibly acceptable parameters, so-called proxy variables, making it imperative to conduct appropriate tests in any case (Žliobaitė & Custers, 2016). So-called sensitivity analysis is one possible form of analysis without an oracle that can detect the influence of given features on the distribution of outputs.

Different domains interpret the term sensitivity analysis differently (Saltelli, 2002; Saltelli et al., 2004, p. 42; Iooss & Saltelli, 2017). In general, sensitivity analyses are employed to determine how alterations in input parameters affect the outcome of a model or system. In this regard, sensitivity analyses are conducted to determine how sensitively a system responds to changes in the input variables and how this affects the output variables. Typically, these mathematical models are used to investigate the impact of specific inputs on real-world systems. They make it possible to evaluate the robustness and dependability of models and simulations, and are thus an essential method for validating results. In numerous fields, such as economics, environmental modeling (Hamby, 1994), risk assessment (Frey & Patil, 2002), and chemistry (Saltelli et al., 2005), sensitivity analyses are utilized.

In addition to these applications, sensitivity analyses are utilized to examine opaque ADM systems (Cortez & Embrechts, 2011; Kewley et al., 2000). Here, the systematic variation of individual inputs, also known as perturbation, is used to determine which input parameters have the most impact on a classification result (Cortez & Embrechts, 2011). With the input vector (input  $y$ ; input 2;...;input  $x$ ), an effort is made to develop a (computational or mathematical) model  $f$  of an algorithmic decision system. Using a black-box analysis, a scientific experiment is conducted to test a hypothesis regarding the influence of input  $y$  on the system's output. This method attempts to develop a (computational or mathematical) model  $f$  of an algorithmic decision system with the input vector (input  $y$ ; input 2;...;input  $x$ ). In a one-dimensional sensitivity analysis, only a single input value is altered at a time; however, there are methods that attempt to alter two or more input values (Petitti, 2000, p. 234), which are summarized as multidimensional sensitivity analyses in the following sections.

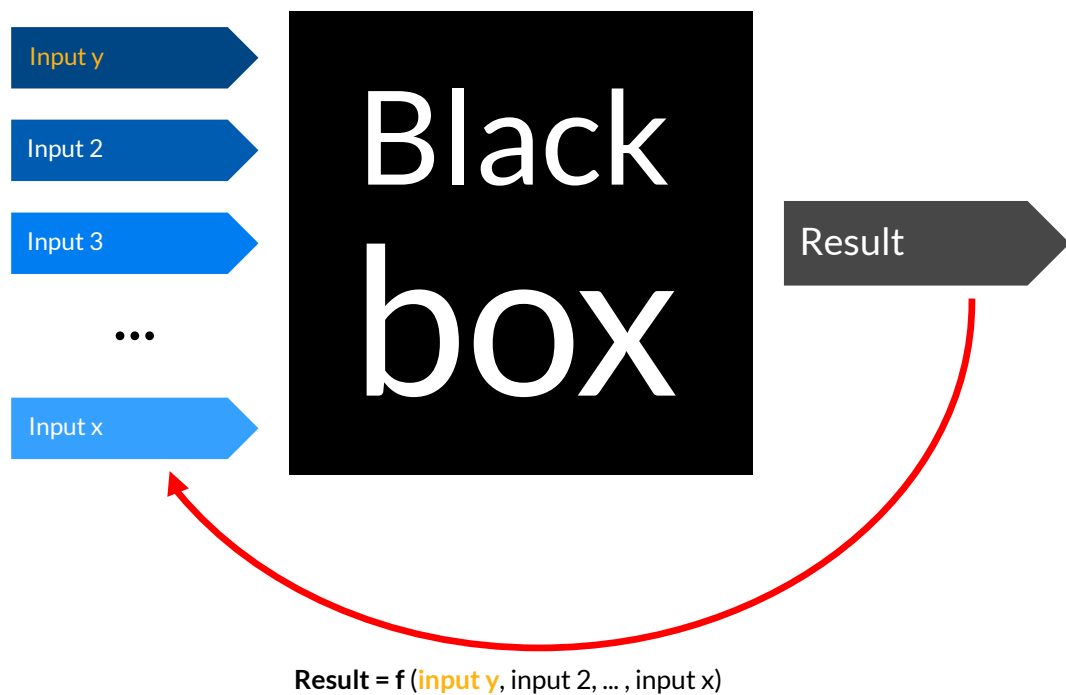


Figure 8.5.: Sketch of a black-box analysis that includes a one-dimensional sensitivity analysis. In the context of a scientific experiment, the impact of  $y$  on the result of the black box is evaluated.

In addition, there is a distinction between local and global sensitivity analysis. While a local sensitivity analysis examines the impact of small “local” changes on the inputs, a global sensitivity analysis considers the full range of the input variables’ definitions (Pettiti, 2000, p. 234). Both properties of sensitivity analyses (one-dimensional vs. multidimensional and local vs. global) reflect fundamentally distinct characteristics of the context and the hypothesis, so they can be used to investigate distinct types of assertions in each instance.

The situation required for a one-dimensional sensitivity analysis is depicted in Figure 8.5.

Such a procedure is comparable to “paired testing” or “matched tests” in social science audits. Here, the audited individual or organization is questioned by two or more auditors who have largely similar characteristics, but differ in one characteristic to be tested (Gaddis, 2018, p. 7). Allegations of discrimination are common issues that can be investigated using a sensitivity analysis. This analysis may provide evidence that a

system's output varies based on a particular characteristic. For instance, gender discrimination investigations have been conducted on CV search engines that enable recruiters to conduct proactive searches for applicants using keywords and filters (L. Chen et al., 2018). When unequal treatment is based on a protected trait, such as gender (Buolamwini & Gebru, 2018), this can be considered discrimination. It should be noted, however, that the legal concept of discrimination is only met in cases of unequal treatment based on certain characteristics in clearly defined application contexts (Hauer et al., 2021).

In principle, the presented analysis methods (oracle-based analysis and sensitivity analysis) allow investigating various system properties. Therefore, their application enables the analysis of various types of statements. If a quality statement is to be examined through a direct comparison of the ADM system's results with "correct" results, the possibility of an oracle-based analysis should be considered. When testing a hypothesis with this type of analysis, it is necessary to not only use the metric, but also to determine at what threshold the investigated hypothesis is accepted or rejected. Frequently, the determination of the threshold value is a question derived from the hypothesis being tested. For instance, the quality of decisions made by humans prior to the implementation of the ADM system can be used as a benchmark. If, however, for a given hypothesis it is examined whether, for instance, subgroups of women and men are treated equally by the ADM system used in an application procedure, a gender-focused sensitivity analysis must be conducted. The question is what effect the "gender" parameter has on the output of the ADM system.

In conclusion, it can be stated that with oracle-based analysis and sensitivity analysis, two distinct types of analysis exist, each requiring distinct framework conditions. In light of the gathered data, it is now necessary to determine which of the two should be chosen for continuing the investigation of the hypothesis.

To assist in the development of the study design at this stage, the following guiding questions have been formulated for each analysis type. These questions aim to provide valuable insights and offer relevant comments and documentation notes. Addressing these guiding questions for each analysis type makes it possible to establish a robust study design and make an informed decision regarding the preferred method for investigating the hypothesis. It is crucial to carefully consider the unique requirements and implications associated with each analysis approach to ensure accurate and reliable results.

#### **Oracle-based analysis:**

##### **?? Guiding Question 2.4 Which metric is used to compare the results?**

In an oracle-based analysis, various comparison metrics can be used depending on the research question. For instance, if the ADM system under investigation is a binary

classifier, i.e., a system that assigns inputs to one of exactly two classes, there are already over 20 different quality measures that can be used to evaluate decision quality (M. A. Haeri et al., 2022; Tharwat, 2021). My research has shown that it is sometimes necessary to conduct extensive inquiries in multiple domains to find comparable studies and thus appropriate metrics. In addition to an interdisciplinary team, it was helpful to reduce the problem to its underlying mathematical problem in order to investigate which domains have already conducted studies with a similar objective.

### ?? Guiding Question 2.5 To which “expected” results are the results of the ADM system compared and where do they come from?

The source of the expected results, i.e., the oracle, has a significant impact on the validity of the subsequent investigation. Therefore, the sources from which this data is obtained should be thoroughly documented. If the result is numeric and will be compared to a range of values, this must be documented explicitly for each test case. Likewise, it is essential to explain why the respective value range was selected.

### ?? Guiding Question 2.6 What is the quality of the “expected” results?

The purpose of this question is to determine whether the comparative data is error-free and whether human decision-makers, who can also make mistakes, are used for comparison. If, for instance, the data was generated by human decision-makers, cognitive bias like the so-called observation bias (also called the John Henry effect in Saretsky, 1972, p. 27) must be considered, for example. Observation bias refers to the fact that people who are aware that they are being observed behave differently, e.g., they put more effort into the task at hand than usually (see “Social Desirability Effect” by Fisher, 1993). Even though the effect is scientifically debated, it may be prudent to avoid any possible occurrence in principle by keeping the observed persons in the dark about this condition or by extracting the data from the normal decision-making process.

### ?? Guiding Question 2.7 At what threshold is the hypothesis accepted or rejected and how is this threshold determined?

There is no general answer to the question of what threshold should be used for accepting or rejecting a hypothesis. In general, when determining threshold values, the primary literature on the metric used should serve as the foundation for employing the evaluation standards used by the developers. Attempts should be made to locate comparable studies that can serve as benchmarks if the employed metric is not conventional. However, the rationale for the choice of thresholds should be the more detailed the less robust the external evidence is that is used to determine the threshold. Drawing inspiration from the principles of the test-first approach (Shore & Warden, 2021), there is strong emphasis on the significance of documenting thresholds before implementing a

solution. This particular practice holds paramount importance as it serves to mitigate any potential bias during the evaluation phase. By documenting predetermined thresholds prior to implementation, developers establish a framework that ensures objectivity and impartiality in assessing the effectiveness of the implemented solution. Moreover, this approach fosters transparency and precision in evaluating the overall quality of the developed black-box analysis. For both types of analysis, guiding questions are provided below:

### **Sensitivity analysis**

#### **?? Guiding Question 2.8 Will there be a one-dimensional or multidimensional sensitivity analysis?**

If more than one input value is observed, the study design becomes significantly more complicated. While a rather simple study design was feasible for the data donation associated with the 2017 German federal elections (Section 5), the multidimensionality of the data received from the participants of the study on advertising for unproven stem cell therapies (Section 6) in terms of countries and diseases required the formation and coordination of significantly more study groups.

#### **?? Guiding Question 2.9 Will a local or global sensitivity analysis be conducted?**

If the dependent variable encompasses only a few values, the study design is simplified because the number of study groups needed can be kept low. If there are several values or if the variable is continuous, the number of study groups increases substantially.

#### **?? Guiding Question 2.10 Are the dependent variable(s) categorical or discrete?**

When considering the dependent variable(s) in a study, it is essential to determine whether they are categorical or discrete in nature. If the dependent variable(s) fall under the category of discrete variables, it becomes crucial to document and justify the precise subdivision of the study into domains. It is necessary to document the specific domains or subcategories into which one or more discrete variables are divided. This documentation should clearly outline the rationale behind the chosen divisions and provide justification for their relevance to the research question at hand. By precisely documenting and justifying the subdivision of the study into categories, researchers can maintain transparency and facilitate effective interpretation of the results. In the case of categorical variables, it is important to provide a documented explanation alongside the categories they represent. Furthermore, this documentation serves to enhance the reproducibility and replicability of the study, enabling other researchers to understand and potentially build upon the findings. Additionally, explicitly justifying the chosen

domains assists in minimizing potential biases and ensuring that the subdivisions align with the research objectives.

### 8.2.3. Determine access to the system

The next step is to investigate what access to the black box is available, i.e., how data is exchanged with the black box. This leads to the following guiding question:

#### ?? Guiding Question 2.11 Which access to the ADM system is used for the black-box analysis?

As described in section 2.4.3, several types of audits can be considered for this purpose (Sandvig et al., 2014):

- Non-invasive user audit: Actual users are requested to submit black-box outputs for review. No specifications are made regarding the input of the users; they are supposed to collect the results during their “natural” interaction with the system.
- Crowdsourced audit: Here, too, actual users are tasked with submitting results of the black box for examination, though there are input requirements. Nevertheless, it is typically unknown which user-related data is processed, such as the user’s past interactions with the system.
- Sock puppet audit: The use of an opaque system is simulated to appear as realistic as possible, but all interactions are fully automated and can therefore be precisely designed. However, there is the possibility of being identified as a bot and consequently blocked or treated differently (T. D. Krafft et al., 2020).
- Scraping audit: This form is a fully automated result query, primarily via application programming interfaces (API).

Each of these auditing methods has its own benefits and drawbacks. If the hypothesis includes the normal behavior of the ADM system’s users, a non-invasive audit of the users should be pursued. This has the benefit of producing the most realistic results that can be collected. If the hypothesis can be examined with this type of audit, the validity risk is minimal. It must be kept in mind, however, that there is no control over the user’s input, and thus only very specific hypotheses can be investigated using this method. Furthermore, this type of audit requires the most effort. In addition to a primarily monetary incentive structure for the actual participants, trust-building measures and corresponding data protection requirements must be implemented.

A non-invasive user audit is not suitable if it is necessary to manipulate the inputs, as this form of audit does not allow control of these parameters.



In the context of a non-invasive user audit or crowdsourced audit, when working with real users, it is also important to focus on their characteristics and behavior so as not to undermine the validity of the results collected. Metaxa et al., 2021, p. 312 discuss the problem under the topic “Avoiding Personalisation”. A scraping audit would be preferable in this instance, as it permits control over as many input parameters as possible within the ADM system. However, absolute statements regarding the influence of inputs are only possible if all inputs that enter the ADM system and influence the output are known, i.e., if a third-category black-box scenario exists. If, on the other hand, a black-box scenario of the first or second type is present, this must be taken into account during the evaluation of the results, since the reliability of the results is then reduced. The results should then be interpreted in light of the potential limitations caused by the unpredictability of the input parameters.

When conducting a sock puppet audit, where multiple user accounts are simulated for evaluation purposes, or a scraping audit, which involves automated data extraction, proactive steps are taken to ensure the integrity and accuracy of the input. These audits can provide valuable insights and help identify any discrepancies or anomalies in the data.

Nevertheless, it is important to recognize that employing such methods may raise suspicion within the ADM system. Being flagged as an unnatural user can result in different treatment or restrictions, such as increased monitoring, limitations on access, or even potential penalties.

To minimize the risk of detection and unfavorable consequences, it is advisable to carefully consider the terms and conditions of the system or platform being audited. It is essential to understand their policies regarding data collection and ensure compliance with any applicable regulations or guidelines.

As the selection of the black-box audit type has a significant impact on the study’s design, further guiding questions are presented below, with their respective audit types depicted in Table 8.1.

### **?? Guiding Question 2.12 Which queries should be sent to the ADM system?**

When conducting an audit, it is crucial to carefully plan and articulate the content of the system requests. The specific information that needs to be included in these requests should align with the research hypothesis and the objectives of the investigation. By ensuring the requests are well-prepared and tailored to the study’s focus, researchers can gather relevant data that directly addresses the research question at hand.

Moreover, it is important to be mindful of the meta information that the system can extract from incoming messages. In addition to the actual content of the messages, various metadata, such as timestamps, sender information, or message characteristics, can provide valuable contextual information. These additional details play a significant role in analyzing and interpreting the data accurately.

Table 8.1.: Assignment of further guiding questions to the four types of audits presented.

#	Short title	Non-invasive user audit	Crowdsourced audit	Sock puppet audit	Scraping audit
2.12	What queries	✗	✓	✓	✓
2.13	Personal data	✓	✓	(✓/✗)	✗
2.14	Participant acquisition	✓	✓	✗	✗
2.15	Informed consent	✓	✓	✗	✗
2.16	Submission process	✓	✓	✗	✗
2.17	Software for participants	✓	✓	✗	✗
2.18	Bot training	✗	✗	✓	✗
2.19	Treated like humans	✗	✗	✓	✓

**?? Guiding Question 2.13 Is personal data collected, and if so, how is it recorded, processed, and stored?**

Depending on the design of the study, it may be necessary to collect personal information, for instance if the effect of the gender parameter on the output is being actively investigated. But it may also be necessary to collect personal data to ensure a balanced or even representative sample of participants. In accordance with Sec 9 and Sec 24 of the Declaration of Helsinki (WMA, 2022), one of the responsibilities of researchers is to protect the participants' dignity, integrity, right to self-determination, privacy, and confidentiality of personal information. Moreover, the collection and processing of personal data is subject to stringent legal requirements (see European Parliament & Council of the European Union, 2016). A professional data protection report confirming that all necessary precautions have been taken to protect the data of the participants or data subjects should be prepared.

**?? Guiding Question 2.14 How are participants recruited?**

Since real users are recruited for non-invasive user audits and crowdsourced audits, it is essential to document the recruitment process in great detail. In addition, when recruiting, specific biases must be considered. Most notably, sampling variance and sampling

bias (Bautista, 2012, p. 40). Sampling variance pertains to the variation in results observed when drawing multiple samples. With each sample, there is a possibility of obtaining slightly different outcomes. However, determining the precise variance is often challenging due to practical limitations in conducting multiple surveys under identical conditions. Conversely, sampling bias arises when specific individuals in the population are either excluded entirely or have unequal probabilities of being selected. This imbalance in selection probabilities can result in a biased sample. Ensuring that all individuals in the population have an equal opportunity for selection minimizes this bias.

Openly inviting citizens to participate, for instance, in a study can result in so-called sampling bias. In this case, it occurs because the participants choose to be a part of the sample, which is why it is also known as self-selection bias (Heckman, 2010; Gideon, 2012, p. 72 et seq.). For our study investigating the filter bubble phenomenon, prospective participants were contacted via a news portal as a means of recruitment (T. D. Krafft et al., 2019). Due to this pre-selected reach, no population-representative sample was collected, so this decision impacted the subsequent representativeness of the study. Moreover, it must be ensured that the people in a study group are not only in the correct group in terms of the dependent variables, such as gender, but that their characteristics are completely identical or as diverse as possible. In social science, this is discussed under the term “balance between experimental groups”

If, for instance, a system that sends job interview invitations based on CVs is examined and all the women have IT skills but none of the men do, the results of this study would be falsified.

The technical aspect of participant registration should be made as simple as possible for a study involving elderly and already impaired patients. Such preliminary considerations were particularly important to the investigation of advertising for unproven stem cell therapy services (Section 6), as the circumstances faced by our target groups could make enrollment considerably more challenging. At the time of initial diagnosis, patients with Parkinson’s disease were, on average, over 60 years old (Pagano et al., 2016). This could make it difficult for them to participate in the study, as that generation’s familiarity with using the necessary technology cannot be assumed.

In some cases, it is very difficult to achieve a sufficiently large turnout: For example, in the investigation of unwarranted advertising of unproven stem cell therapy offers, it was possible to recruit approximately 100 patients (Reber et al., 2020), which is a relatively large sample size for a social-anthropological medical study but insufficient for a comprehensive statistical analysis of the results.

Various hardware and software environments present an additional obstacle when attempting to interact with real people. Data collection can be affected by a variety of devices, operating systems, browser versions, and other software applications on the devices. In our analysis regarding the 2017 German federal elections (see Section 5), a number of participants were unable to install the plugin. In other cases, it failed to transfer data or restricted browser functionality due to increased CPU load. It was unclear

whether this was caused, for instance, by ad-blocking software. The different settings of the participants' Google user accounts, such as preferred language and preferred number of search results per page, also caused a minor issue. Although the crowdsourcing approach has the advantage of collecting user-generated data, it is nearly impossible to fully manipulate the search engine's input to better comprehend the system's actual behavior. The input to a search engine with a personalized account, for instance, depends on a number of factors, such as the keywords used, the time of day, the computer's IP address, the user's search history, general web usage, and the system's assumed human user characteristics, such as age, income, and gender. These variables are not easily modifiable in order to consistently check results for a specific user profile.

### **?? Guiding Question 2.15 How are participants informed about participation in the study, and how is informed consent obtained and recorded?**

In accordance with the World Medical Association's Declaration on Ethical Principles for Medical Research Involving Human Subjects, all participants in a study involving human subjects are required to provide written informed consent. The "Declaration of Helsinki" is regarded as the standard of medical ethics, and Sec 26 states:

In medical research involving human subjects capable of giving informed consent, each potential subject must be adequately informed of the aims, methods, sources of funding, any possible conflicts of interest, institutional affiliations of the researcher, the anticipated benefits and potential risks of the study and the discomfort it may entail, post-study provisions and any other relevant aspects of the study. The potential subject must be informed of the right to refuse to participate in the study or to withdraw consent to participate at any time without reprisal. Special attention should be given to the specific information needs of individual potential subjects as well as to the methods used to deliver the information.

After ensuring that the potential subject has understood the information, the physician or another appropriately qualified individual must then seek the potential subject's freely given informed consent, preferably in writing. If the consent cannot be expressed in writing, the non-written consent must be formally documented and witnessed.

All medical research subjects should be given the option of being informed about the general outcome and results of the study. (WMA, 2022)

Depending on the scope of the study and the nature of participant involvement, it may be necessary to submit the study design to an ethics committee or an Institutional Review Board (IRB). The purpose of the IRB is to review the appropriateness of research

studies in order to protect the rights and well-being of the subjects (Amdur & Bankert, 2010, p. 5), for instance by conducting a risk-benefit analysis (Amdur & Bankert, 2010, p. 147 ff.). It is also advisable to have an external body review this report to ensure its quality and impartiality.

**?? Guiding Question 2.16 How is the submission process designed for participants?**

This relates to both the design of the submission process and the corresponding expectations, such as whether continuous submissions are expected or data is sent to the researchers only at the conclusion of the study. For some studies, it is also worthwhile considering whether participants could/should be allowed to make modifications to the submitted data sets, such as removing unpleasant information from data sets that participants do not wish to share. Despite the fact that this could affect the validity of the study, it could also be a way of increasing study participation. When advertisements displayed to users are to be collected for the study of election advertisements, it may prevent people from participating in the study because they are unable to screen the advertisements before submitting them and, for instance, remove advertisements for medical products that are mistakenly collected as election advertisements or those containing personal data such as personal salutations.

Participants would ideally grant the scientific study teams limited access to their account on the relevant platform (T. D. Krafft et al., 2020) by, for instance, registering for the study with their Google account. Still, search results and advertisements would be automatically collected and sent to the person leading the study. However, there is currently no way to retrieve specific information from social media accounts, even with the user's permission. This data cannot be accessed in a targeted manner on platforms such as Google's various services, Facebook, or Twitter. Although Facebook used to provide the Facebook Graph API, which allowed targeted access to users' accounts if they consented, access was severely restricted after the Cambridge Analytica scandal in 2017 (Kirchgaessner, 2017), so black-box analyses focusing on certain aspects, such as advertising distribution or certain messages in the timeline, are no longer possible from outside of Facebook (T. D. Krafft et al., 2020).

**?? Guiding Question 2.17 What software components are required for the participants?**

Depending on the type of audit chosen, different automation tools must be prepared. In the case of a crowdsourced audit of a search engine such as Google, for example, a browser plugin (see, for example, the tools developed for our own studies (T. D. Krafft et al., 2019; Reber et al., 2020)) is required to send as many identical queries to the ADM system simultaneously as possible. For a sock puppet audit the corresponding bot personas and their behaviors must be developed and implemented. The necessary software for such research must also be designed and adapted on a case-by-case basis.

In the majority of cases, the software used to conduct the study must be developed individually, for instance to store and submit only the data required for the study on the participants' devices. Although the necessary software components must be developed for nearly all black-box analyses, separate requirements should be formulated for the software components utilized by study participants. In addition to the data protection issues already mentioned, usability issues are also relevant, as participants may be unwilling to continue supporting the study in the event of errors (see Section 5) or may have limited knowledge of how to use their devices (see Section 6), so the installation and use process must be made as simple as possible and must be explained in a clear and understandable manner. A variety of devices, operating systems, and browsers running different versions and software can interfere with data collection (T. D. Krafft et al., 2021, p. 149). Several participants in our black-box analysis project for the 2017 German federal elections had trouble installing the plugin, submitting data, or using their browsers normally due to excessive resource consumption. It was impossible to determine whether ad-blocking software was responsible. Variations in the settings of the participants' Google accounts, such as preferred language or number of search results displayed per page, caused a second minor issue.

**?? Guiding Question 2.18 What type of behavior should bots simulate? Is there a training phase prior to the study, and if so, when and how are the bots trained?**

The fact that bot-based approaches are routinely detected and blocked by the majority of online platforms is a fundamental disadvantage. While these measures are necessary to detect malicious bot attacks, they also impede scientific research supported by public interest (T. D. Krafft et al., 2020). In addition, the creation of fake accounts presents complications. We were fortunate in our study, as we were able to analyze search engine results a) without logging in and b) with relative ease using an HTML scraping technique. A bot-based strategy is nearly impossible if it requires the creation of fake accounts and/or the use of a software provider's application. In the paper titled "Why do we need to be bots?" (T. D. Krafft et al., 2020), we elaborate on some of the difficulties we encountered while creating fake Facebook accounts.

To eliminate room for interpretation during implementation, the behavior of bots in the context of a sock puppet audit with regard to interactions with the ADM system should be described as precisely as possible. In addition to the sequence of interactions, this includes the schedules indicating when each interaction is planned and which actions are to be logged, i.e., documented.

In some instances, ADM systems may process past interactions with the ADM system in addition to direct input. This is the case when the system customizes the delivered results based on the user's past actions. If there is a suspicion that such personalization is being used or even investigated, it may be necessary to provide the bots with historical

behavior. It is important to understand that this information goes beyond the metadata of a query, which is explained in guiding question 2.12. Hannak et al.'s browser profiles can be used for this purpose (Hannák et al., 2014), or algorithmically generated browsing behavior can be implemented. As described in Section 7, this has been shown to be difficult.

**?? Guiding Question 2.19 Is it monitored that the ADM system treats bots and APIs similarly to actual humans?**

One of the greatest challenges of a sock puppet/scraping audit is ensuring that the system treats automated queries as a typical user interaction. Therefore, it may be useful to conduct random or periodic checks. To accomplish this, the results gathered by the automated tool can be compared to the results a real user would receive for the same input. As some changes in system behavior, such as the pricing examined in section 7, have a temporal element, it is important that both queries are executed simultaneously, if possible. In this instance, if the timestamp deviates, the availability of the checked product may be the determining factor for a price change.

In this step, it is also possible to determine that the investigated suspicion or the derived hypothesis cannot be tested using the available investigation techniques. If this is the case, one can either attempt to alter the status quo, i.e., acquire additional or alternative access points, or modify the hypothesis to a statement that can be verified using the available resources.

### 8.3. Concept of the Black-Box Analysis

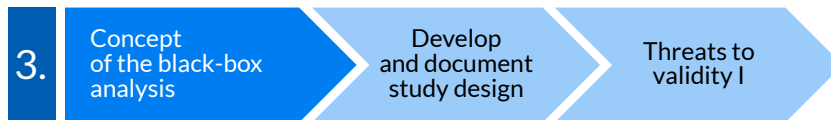


Figure 8.6.: Process step 3: Concept of the black-box analysis

After identifying the black-box scenario and determining the precise type of analysis and method of access to the ADM system, the black-box analysis can be designed (see Figure 8.6). For this purpose, the study design that will be used to test the formulated hypothesis must be described in detail. In addition, a good design addresses potential threats to the validity of the planned study. Even if portions of this evaluation are only possible after the data has been collected and analyzed, it makes sense to consider and document these issues beforehand.

#### 8.3.1. Develop and document the study design

Additionally, it is essential to precisely describe the expected system behavior. Due to the adaptability of ADM systems in particular, it is possible for the system to evolve over time without the users' knowledge. Google, for instance, releases numerous new versions of its search engine each year and tests multiple versions of its algorithm simultaneously to determine which performs best (Pansari & Mayer, 2006). Such a process is commonly referred to as A/B testing (Kohavi & Longbotham, 2017); therefore, when investigating ADM systems, this possibility must always be considered, necessitating extremely close monitoring (Bucher, 2016; Gillespie, 2014; Introna, 2015). Consequently, it becomes essential to provide a precise depiction of the anticipated behavior of the system, which can then serve as a benchmark for comparing the observed behavior. This subject was addressed in greater detail in section 5.

To be able to respond to a potential change in system behavior, it is also advisable to design the study to allow for the rapid modification of a large number of study parameters when it is conducted later on. The development of the study design is a crucial step in the planning of a black-box analysis, during which all considerations and conceptual planning for the study must be systematically carried out and documented. The specification of all individual steps in the implementation of a black-box analysis presented in this section can be used as a planning guide. It should be noted that documentation of the design decisions is an essential component of the planning, which includes a precise description of the planned procedure as well as the design and detailed description of the statistical test to be performed as part of the study. An essential



requirement for a valid and meaningful study is thus the meticulous development of the study's design. The general steps of a statistical test will be outlined first in the following Section. Then additional guiding questions will be provided, which are a useful addition to the study design development.

When designing a statistical test, it may be useful to refer to the procedure described in Henry R. Neave's 1976 paper (Neave, 1976).

In his book "100 Statistical Tests" (Kanji, 2006), Gopal Kishore Kanji presents and discusses one hundred statistical tests with accompanying examples. It is recommended comparing the question to be investigated with this collection in order to draw not only on a proposal for implementation but also on any experiences related to the tests outlined in the collection. On this basis, potential methodological implications can be examined in greater depth. When designing a black-box analysis, it is recommended working with an expert on statistics or at least having the results validated by one in order to be made aware of potential errors and to avoid them. In addition to the technical design of the test, the later analysis of the data, i.e., the values that are incorporated into the "statistics" mentioned above, should be developed at this stage to ensure that the investigated hypothesis can be tested with the data to be gathered. For instance, if the objective is to determine whether Google personalizes its search results, it is necessary to consider how this "personalization" can be measured by comparing the search result lists (see Section 5.2 and T. D. Krafft et al., 2019). The precise evaluation should also be documented as far in advance as possible. Such a procedure is also referred to as a "pre-analysis plan" for which Coffman & Niederle, 2015; Lin & Green, 2016; Olken, 2015 have provided comprehensive overviews and outlined key advantages and disadvantages. Although these analysis plans aid in both the preparation and execution of data exploration, it should be noted that they may lead to questioning the data and conducting additional data explorations for secondary analyses in conjunction with or after the primary analysis has been completed (Lahey & Beasley, 2018, p. 93).

Due to the highly context-specific nature of black-box analyses, additional questions that can aid in the development of the study design are provided below.

#### **?? Guiding Question 3.1 What study groups are necessary?**

Various study groups are derived depending on the hypothesis. Here, it must be specified in which characteristics these differ and which characteristics all members of a group must share.

#### **?? Guiding Question 3.2 How large does the sample have to be?**

The term statistical power, also known as statistical significance, is important in the field of statistics. It is defined as the probability that an effect will be detected if it exists (J. Cohen, 1988, p. 2). This probability is directly linked to determining the minimum sample size required to detect the desired effect size. The effect size is used to determine the

practical applicability of statistical test results by describing the relationship of empirically measured variables. A large effect size therefore indicates a high degree of practical applicability. Various effect measures are utilized to quantify the effect size. A “suitable” effect size can, in the best-case scenario, be derived from previously completed studies or related studies, or in some cases directly from the suspicion to be investigated with the black-box analysis. In the absence of this kind of information, reference should be made to the work of American psychologist Jacob Cohen (J. Cohen, 1988), who, in his book “Statistical Power Analysis for the Behavioral Sciences”, established norms for small-, medium-, and large-effect sizes as well as a straightforward method for estimating the required statistical power for planned studies. In the field of statistics, the determination of the sample size is considered a standard procedure. Consequently, it is necessary to refer to relevant literature in order to obtain guidance and recommendations regarding this matter (Krejcie & Morgan, 1970).

### ?? Guiding Question 3.3 Can regionalization effects influence the study results?

As described above, ADM systems can use the location of the inquirer to customize the output, i.e., regionalize it. This possibility should be accounted for in the study’s design, as the results may be skewed if all requests originate from the same IP range (which in most cases still provides clues to the location, but in any case provides a clue to the same origin of the requests). In contrast, regionalization effects can also account for divergent results. In our data donation for the 2017 German federal elections (see Section 5), we examined the submitted search results for signs of personalization. When regionalization effects were taken into account, the first indications of strong personalization were eliminated.

### ?? Guiding Question 3.4 Who develops the necessary software components for the study?

Depending on the team’s technical resources and/or time constraints, the decision can be made to outsource the software components to external developers. In such a case, the following points should be included when drafting the application’s specifications:

1. For which monitoring tasks are developers responsible and for which do they provide interfaces?
2. During software implementation, the capacity for possible short-term software modifications is required. As described in the study in section 5, short-term adjustments to the ADM system by the operators can impede or even prevent data collection, necessitating a prompt response.
3. In what format does the software collect data, and how is it transferred to researchers?

4. Will the source code be provided to the researchers?

**?? Guiding Question 3.5 Will the code of the study software be published?**

Even if publication of the study software enables operators to gain insight into the study design and use it (similar to the diesel emissions scandal, where car manufacturers were able to install a software that recognizes whether a car has been in a test situation L. Bovens, 2016), publication likely increases the acceptance and willingness of potential study participants.

**?? Guiding Question 3.6 What is the project management structure and the exact schedule with milestones?**

Black-box analyses can span weeks, months, or years and require the active participation of a large number of stakeholders. Such a scope may necessitate expert project management (e.g., in accordance with DIN 69901: 2009 or ISO 21500: 2012. In such project management, time management is typically structured via so-called milestones. A study design entails establishing appropriate project management and dividing the duration of the project into appropriate, attainable milestones.

### 8.3.2. Threats to validity I

The evaluation of social science studies by Ellen A Drost, 2011 provides a schematic guide for assessing the validity of study results and the limitations of their interpretation. This guide identifies a variety of threats to validity that can be evaluated to improve the interpretability of empirical results. Four areas identified by Drost as having an impact on the validity of a study are briefly described and discussed in the context of black-box analyses.

#### Statistical conclusion validity

Statistical conclusion validity refers to the extent to which the data and results of a statistical analysis support the conclusions of a study (Cook & Campbell, 1979). It is essential because it ensures that conclusions are based on solid evidence and not on chance or error. A study with high statistical conclusion validity is more likely to reflect the actual relationship between the studied variables. According to Drost, 2011, several factors can compromise this. If the statistical tests used to test the hypothesis make assumptions about the data, then a violation of these assumptions can lead to incorrect conclusions about the cause-effect relationship; therefore, a violation of assumptions threatens statistical conclusion validity. Errors in measuring results or potential distractions during execution have similar consequences (so-called random irrelevancies). Random heterogeneity of the study participants can also be problematic, as it can lead

to the identification of erroneous relationships or an increase in the variance of the results. This is especially important for studies involving actual people (non-invasive user audit and crowdsourced audit). The following are the important threats to statistical significance:

**Sampling errors:** As discussed in guiding question 2.14 sampling error consists of two components: sampling bias and sampling variance (Bautista, 2012, p. 40). Sampling variance refers to the variability in outcomes when taking multiple samples. Each sample can yield slightly different results, but determining the exact variance is difficult due to practical constraints in conducting numerous surveys under the same conditions. On the other hand, sampling bias occurs when certain individuals in the population are excluded or have unequal chances of being selected, leading to a biased sample. To reduce bias, it is important to provide equal opportunities for selection to all individuals in the population.

**Confounding variables:** In an experiment or observational study, researchers are typically interested in understanding the relationship between an independent variable (the variable that is manipulated or changes naturally) and a dependent variable (the outcome that is measured). However, there may be other variables that also influence the dependent variable. These are called confounding variables. They are variables that can cause or prevent the outcome of interest, are not intermediate variables, and are associated with the factor under investigation. They can interfere with the experimental outcome and lead to inaccurate results if not properly controlled or accounted for. Particularly when investigating opaque systems for which - depending on the black-box scenario - neither information nor knowledge about all inputs is available, confounding variables can be a threat to internal validity. In this instance, a change in the dependent variable can be attributed to changes in a related confounding variable (Brewer & Crano, 2014, p. 50). As previously explained in the description of the three black-box scenarios, such an effect cannot be ruled out in the absence of complete knowledge of the ADM system's input. However, even with a black-box scenario of the third category, such initially unknown correlations should be tested for in order to exclude these effects to the best possible extent by selecting appropriate study groups or mitigating the threats to internal validity posed by confounding variables.

**Lack of statistical power:** This may be the case if the sample size is too small to accurately detect differences between groups. See guiding question 3.2.

**Errors in data collection:** These can be measurement errors or errors in the way data is collected, leading to inaccurate results.

**Errors in data analysis:** This may involve the incorrect application of statistical procedures or the incorrect interpretation of results.

Basically, it must be stated that with each effect that compromises the internal validity of a research study, there is the possibility of falsifying the results, thereby calling the statistical significance into question. This internal validity will therefore be discussed in greater depth in the next Section.

### **Internal validity**

Internal validity refers to the extent to which the results of a study can be confidently attributed to the independent variable being examined and not to other external factors. Establishing internal validity is crucial for determining the causal relationship between an experiment's independent and dependent variables. It is the extent to which the observed results accurately reflect the truth in the population under study and are not the result of a methodological error. Seven of the eight factors listed by Campbell & Stanley, 1963, p. 5 et seq. as influencing the internal validity of a study can be applied to the study of opaque ADM systems:

1. History: “The specific events occurring between the first and second measurement in addition to the experimental variable” (Campbell & Stanley, 1963, p. 5). Since all events that occur between measurements may have an effect on the study, as many of these events as possible should be recorded. This applies not only to interactions with the ADM system and its responses, but also to the socio-technical system, which must be included, as statements from actors or changes in parameters can have a significant impact on the interpretation of the results. When investigating the roll-out of advertisements for unproven stem cell therapies on Google (Section 6), the statement that such advertisements continue to be banned and their roll-out closely monitored was a crucial clue for the subsequent investigation of the data. In addition, when investigating the possibility of personalized pricing in online retail, it was necessary to distinguish personalization from contingent pricing (see Section 7) of products, which is an intentional price change that can occur over time.
2. Maturation: “Processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing more tired, and the like” (Campbell & Stanley, 1963, p. 5). It should always be considered that operators of ADM systems may make adjustments to the system or conduct AB testing almost undetected. If operators actively influence the reactions during the investigation, the situation is analogous to the exhaust emissions scandal in Germany, where automakers actively monitored their vehicles to determine whether they were in an inspection situation, and if one was detected, whether the vehicles displayed abnormal behavior (L. Bovens, 2016).
3. (Repeated) Testing: “The effects of taking a test upon the scores of a second testing” (Campbell & Stanley, 1963, p. 5). The effects of a previous request on the subsequent behavior of an ADM system is a significant issue. For instance, when investigating the personalization of an ADM system, an internal behavior modification based on previous requests may occur. If an account was used by a normal

person with specific characteristics, this profile becomes increasingly distorted with each request made during the actual investigation.

4. Instrumentation: “Changes in the calibration of a measuring instrument or changes in the observers or scorers used may produce changes in the obtained measurements” (Campbell & Stanley, 1963, p. 5). Changes to the study software can have a significant effect on internal validity. In our study during the 2017 German federal elections, software errors corrupted a large portion of the collected data, and the study could only be evaluated by documenting the corrections and deleting the affected data sets (see Section 5).
5. Regression towards the mean “Operating where groups have been selected on the basis of their extreme scores” (Campbell & Stanley, 1963, p. 5). In repeated measurements, this statistical effect indicates that extreme scores tend to move toward the mean. This indicates that a group of subjects selected based on their extreme test scores will have scores that are closer to the mean on a subsequent measurement. The effect arises due to random fluctuations in the measurements. When a test is administered repeatedly, random errors can occur, leading to results that deviate from expectations. In the initial evaluation, some subjects may receive extremely high or low scores due to random errors unrelated to their actual abilities or traits. These random errors are likely to be compensated for in a subsequent measurement, bringing the subjects’ results closer to the mean.
6. Selection or sampling bias: “Biases resulting in differential selection of respondents for the comparison groups” (Campbell & Stanley, 1963, p. 5). This effect, which can compromise the validity of a study, was discussed in greater detail in guiding question 2.14.
7. Experimental mortality: “Differential loss of respondents from the comparison” (Campbell & Stanley, 1963, p. 5). This factor becomes especially significant in studies involving real people, such as non-invasive user audits and crowdsourced audits. Even if the size of the comparison groups is similar, participants may stop submitting results during the course of the study. This is similar to the survivorship bias (Zweig et al., 2021, p. 24), as it is also overestimated if the characteristics of those still participating are more apparent than those of others. Our study during the 2017 German federal elections experienced this issue (see Section 5). However, even with sock puppet audits or scraping audits, the study group can diminish in size, as was the case with bot detection in our study conducted with Rhein-Neckar Fernsehen, for example (T. D. Krafft et al., 2020).

### **Construct validity**

Construct validity is the extent to which a measurement or instrument accurately reflects the theoretical concept or construct it is intended to assess and captures the intended meaning or concept. Construct validity must be established to ensure that the results of a study accurately represent the concepts or phenomena under investigation. In other words, it focuses on what the study actually measures and the plausibility of the argument that the suspected phenomenon is related to the characteristics captured for the study of the hypothesis (see Messick, 1987, p. 8). Since these are always individual aspects, please refer to the respective argumentations in the studies (Section 5.4.3 and 6.6).

### **External validity**

External validity is the extent to which the results of a study can be generalized to other populations, settings, or time periods so that they can be applied to and make predictions about a larger context. When evaluating the results of a study, external validity must be considered because it allows us to determine the applicability of the results to other populations or situations. In short, external validity raises the issue of generalizability of the measured effects (Campbell & Stanley, 1963, p. 5). This kind of threat to validity is referred to as “ecological validity” in certain social sciences (Lahey & Beasley, 2018). In correspondence studies in social sciences, external validity has already been addressed in depth. It is therefore worthwhile referring to the section “Technical Aspects of Correspondence Studies” by Lahey & Beasley, 2018, where a number of the threats to external validity discussed can be applied to the process model presented here and to its evaluation.

## 8.4. Preliminary study

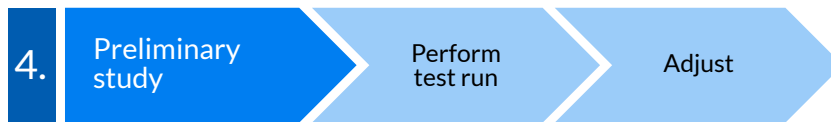


Figure 8.7.: Process step 4: Preliminary study

Due to the fact that a black-box analysis involves interaction with a complex socio-technical system, our research has shown that it is of utmost importance to test the entire study design on the real system in advance, in addition to testing the individual components extensively (T. D. Krafft et al., 2019, 2021). Before the actual study is conducted, it is possible to detect and mitigate the occurrence of new, possibly unforeseen effects and errors based on the interaction of all components and actors by conducting a thorough preliminary study (see Figure 8.7). Consequently, this step of the procedure addresses the guiding questions.

### ?? Guiding Question 4.1 How can a preliminary study be conducted as realistically as possible?

In addition to detecting errors in the software involved, a preliminary study with fewer participants and a shorter duration can, for instance, assist in estimating the magnitude of the effect to be investigated, thereby indicating the minimum number of participants required for reliable statistical analyses. A test run also aids in identifying problems in the actual technical commissioning and setup of individual components prior to the study, which, according to past experience, only occur in the actual application context. During our data donation for the 2017 German federal elections, there were hardly any issues with the installation of the necessary browser plugins among our participants (see Section 5). However, we received a large number of questions when we performed a similar installation for our study on the advertisement of unproven stem cell therapies (Section 6).

In the end, a high level of support was required due to the participants' higher average age, their significantly older technical equipment, and their use of out-of-date browsers. In the 2017 German federal elections study, it was discovered that search queries on Google resulted in the display of Google+ pages, which made data collection difficult. A preliminary study could have prevented this issue. In addition, it was not realized until much later that the participants' preferred language setting may have contributed to anomalies that were overlooked during data analysis. Similarly, if a test had been conducted, problems with the black box's rapidly changing interfaces could have been



identified. For instance, in some studies we encountered the issue that the page layout used as the interface changed frequently, likely as a means for site operators to prevent scraping. In conclusion, it can be stated that a test run can help to uncover study design flaws in advance and defuse unanticipated problems (T. D. Krafft et al., 2021).

## 8.5. Data collection



Figure 8.8.: Process step 5: Data collection

The data collection or actual study execution is likely the most delicate aspect of a black-box analysis. Numerous decisions made in advance manifest their effects at this stage (see Figure 8.8). As a result, this Section concerns the following guiding questions:

### ?? Guiding Question 5.1 What monitoring is available and how are deviations documented?

The precise data collection structure is highly dependent on the specific system and previous design decisions. It is important to monitor the entire data collection process as much as possible in real time. This implies that in addition to routinely validating the ADM system's expected behavior, all collected data should also be validated automatically and, if possible, manually. For instance, if an automated crowdsourced audit collects the individually received results pages of an online search as described above, but the search engine operator changes the structure of the results page, the automated data extraction process from the website fails and must be adjusted quickly (T. D. Krafft et al., 2019, 2021). As mentioned above, Google uses A/B testing in order to continuously refine and optimize the functionality and effectiveness of their search engine (Pansari & Mayer, 2006). In turn, such monitoring necessitates a detailed description of the expected system behavior. To ensure maximum traceability, the entire process, including errors discovered during monitoring, should be properly documented after correction. In order to use the received data correctly in the subsequent preparation and evaluation, it is necessary to have access to all process steps and their modifications, e.g., through modifications to the source code. At this point, the following guiding questions should be asked:

### ?? Guiding Question 5.2 In what timeframe is the black-box analysis conducted?

Not only is the precise time period of data collection essential for planning the implementation, but the collected data can also be examined in relation to modifications to the ADM system or communications about it. In our investigation of unproven stem cell therapy advertising on Google, the publication date of Google's statement to further

prohibit advertising of unproven therapies on their platform was a crucial factor in the assessment and execution of the study.

**?? Guiding Question 5.3 What system behavior can be expected during the study period, and how will this be monitored?**

As stated previously, the significance of a precise description of the expected system behavior cannot be overstated. Due to the adaptability of ADM systems in particular, it is possible that the system will evolve without the users' knowledge. Since this monitoring should be automated at as many points as possible, a description/documentation of the expected system is required. To be able to respond to any change in system behavior, it is advisable to allow for the rapid modification of as many study parameters as possible during its execution. This adaptability should be considered from the very beginning of the study's design.

**?? Guiding Question 5.4 Where is the submitted data collected/stored?**

The question of where and how data will be collected and stored must be clarified beforehand. On the one hand, a data backup strategy can be established at an early stage, while on the other hand, participants may prefer storing their data in a specific country or at a specific institution. If this is defined and communicated, the study's participation rate may also increase. This is particularly true for non-invasive user audits and crowd-sourced audits. If personal data is stored, the location at which it is stored must also comply with legal requirements. The European Data Protection Regulation (European Parliament & Council of the European Union, 2016) and its national implementations provide a significant legal foundation, which must be protected in any case. This raises the subsequent question:

**?? Guiding Question 5.5 Is data collection actively supervised, or is collected data validated manually and automatically?**

Similar to the active monitoring of system behavior and the interface to the black-box system, the collected data should be rigorously validated to ensure that no alterations or errors occurred during the collection process. This validation step should not only be performed manually after a certain amount of data has been gathered, but automated validation should also be performed at various stages. For this reason, it is essential to have a precise documentation of the expected data, both in terms of their individual characteristics (values, size, type) and the quantity of incoming data that can be monitored.

## 8.6. Analysis of the results



Figure 8.9.: Process step 6: Analysis of the results

The final step of the process is to execute the analysis steps outlined in the conception process of the black-box analysis (see Figure 8.9), and thus answer the following guiding question:

**?? Guiding Question 6.1 How exactly is the data analyzed and what is the conclusion of the study?**

From this point forward, the black-box analysis of an ADM system resembles a traditional statistical analysis, so the collected data must be cleansed and structured with extreme caution and a keen awareness of potential validity threats. This step, like any statistical analysis, has both advantages and disadvantages in this regard. The data may contain various types of errors, such as a missing data point due to a connection problem, which may result from problems with the data collection procedure. Consequently, the data must be preprocessed. García et al., 2016 discuss the most important concepts from the extensive literature pertaining to this step. In any case, the preprocessing steps must be meticulously documented for reproducibility purposes.

Again, data analysis requires a high degree of customization. It most likely involves calculating carefully chosen metrics, such as those based on data from a specific distribution measure or a measure that operationalizes a concept of similarity, quality, or fairness. In an oracle-based analysis, for instance, the collected input and/or output pairs must be compared to the expected results. For this comparison, a variety of metrics, such as quality measures in the previously outlined evaluation procedure of AI systems, address different aspects of the comparison (see, for example, Steyerberg et al., 2010). Individual results can be compared, but it is crucial to select both the metric and the comparison partner with care, as incorrect aggregation or comparison can lead to incorrect assumptions. There is relevant research on this type of error that leads to false correlations (e.g., H. A. Simon, 1954; Vigen, 2015). If the results of multiple ADM systems are to be compared, care must be taken to ensure that the structure, data aggregation, and data preparation are identical or as similar as possible, and that all exceptions are documented (see, for example, Hauer et al., 2020).

Following the analysis, the next step is to visualize the results. Improper visualization can result in serious problems, e.g., data being incorrectly represented or even misidentified (Bresciani & Eppler, 2008). Therefore, it is essential to select the visualization with care. Some commonly used visualization techniques are explained in (C.-h. Chen et al., 2007; Yau, 2011).

**?? Guiding Question 6.2 What is the final view on the validity of the study after it has been conducted?**

As previously described for the planning of the study, the validity of a study is central to scientific research because it influences the validity of the results and the conclusions. Before publishing the study's results, it is necessary to conduct a second evaluation of the threats to validity. During the course of the study's execution, new threats to validity not previously considered may have emerged and should be documented. Examples of potential new threats to validity include sampling biases, confounding variables, and unanticipated events. Additionally, the threats to validity already compiled in the preliminary assessment should be reassessed and finalized; in particular, potential threats such as selection bias or measurement error should be re-examined critically to ensure that they have been adequately considered and mitigated.

Prior to publishing the study results, such a re-evaluation of the threats to validity can ensure that the study was conducted correctly in terms of methodology and that the results are valid and reliable.

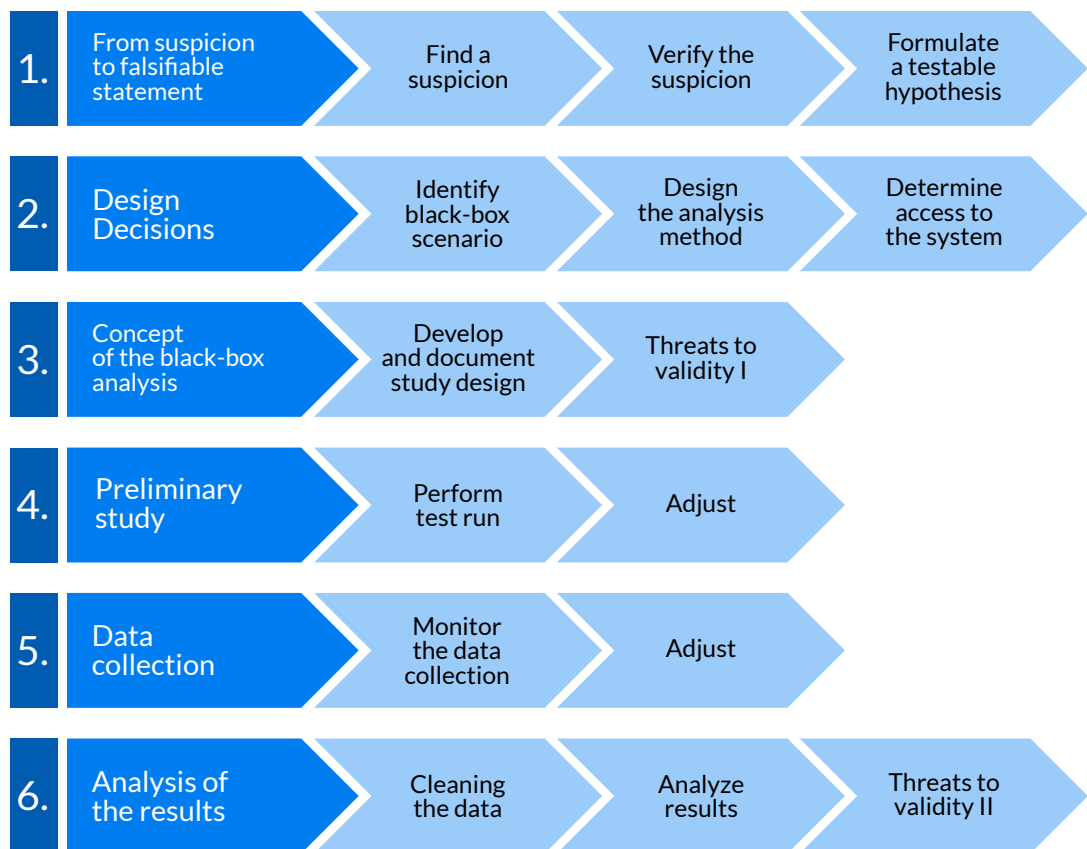


Figure 8.10.: Process steps of a black-box analysis.

Finally, all the steps outlined in this section have culminated in Figure 8.10, which illustrates the result of the entire process. The process model can serve as a reference point for future black-box analyses. Researchers can leverage the model as a guide or template to design and execute similar analyses in different contexts or research domains.

# Limits, political conclusions and outlook

## 9.1. Insights gained from & limits of the three studies

This dissertation has examined in greater depth the method of investigating ADM systems using black-box analysis. The investigation was conducted on the basis of several projects that took place in various settings and demonstrate the range of application areas for black-box analyses.

The process model is a response to the challenge identified at the outset - that there is a need to establish transparency of ADM systems considering that existing regulation does not yet specify how transparency of black-box systems can be realized via audits. The process model addresses this gap with a procedure that is easily comprehensible from different disciplinary perspectives. It is divided into distinct phases that have been discussed separately. For the interdisciplinary reader to be able to assign terms and procedures used in practice, detailed information is provided on each topic. In addition, references are made in each instance to actual implementations, allowing the reader to gain perspective and benefit from past experience. By consolidating different approaches and providing a structured framework, the process model seeks to bridge the knowledge and implementation gap for non-computer science researchers. It empowers individuals from diverse disciplines to engage in black-box analysis, enabling them to comprehend and scrutinize ADM systems with increased confidence and proficiency. This contribution holds the potential to democratize the understanding and examination of these systems, fostering a more inclusive and informed approach among academia and civil society.

Ultimately, by equipping researchers and stakeholders with the necessary tools and knowledge, the process model promotes a more distributed and comprehensive approach to analyzing ADM systems. It empowers individuals to actively participate in the pursuit of transparency and accountability, shaping a society that can effectively navigate the complexities of technology-driven decision-making. The work concludes with a discussion of the limitations of the developed process model and further scientific research opportunities, and a presentation of closing thoughts on the increasing significance of

black-box analyses of ADM systems for society and economy.

In the following, a brief summary of the results of each study and their key insights is presented.

In the **first study**, Google's personalization of search results prior to the 2017 German federal elections was analyzed in order to determine whether results differed based on locality and previous search behavior. Prior to the elections, the results indicated that the space for personalization was limited, which put into question an important pillar from Eli Pariser's filter bubble theory (Pariser, 2011) in this case. Insights gained in the course of this study include:

On the **methodological** level, the large number of participants, the positive responses in social media, and the responses of readers to the project's reports demonstrated that the motivation of citizens to donate data to study opaque algorithms is fundamentally high. However, the difficulties caused by the project's short duration and limited funding demonstrate that a study with a representative sample requires special preparation.

On the **technical** level, it has been demonstrated that extensive software testing, adaptable software, and efficient monitoring are of great importance. Due to the limited time available for the software development process, tests could not be conducted adequately, which led to problems with the usability of the collected data. One key discovery from this research is the realization that extensive preparations and trialing of the technical setup are crucial to eliminating different types of mistakes. Communication with the study participants regarding updates was also difficult, resulting in the submission of numerous unusable results. Furthermore, due to changes at Google, adaptations in the parsing of the search results page were required. Observers of the market have noted that Google regularly implements improvements to the search engine in response to shifting search patterns. A functional update process would have been beneficial here. Because the software development process was outsourced and server access was restricted, the monitoring of the search results lists was also compromised.

The **second study** examined the extent to which AdSense, the advertising platform of Alphabet, enforces the ban on direct-to-consumer advertising for unproven stem cell therapy offers. During the study period, these advertisements continued to be displayed. An extremely vulnerable population benefits from this study as it provides concrete evidence of the ongoing roll-out of advertisements for unproven stem cell therapy offers, thereby enabling further in-depth research. This study enables a rich digital ethnography that would not have been possible otherwise. Moreover, the data sets led to further findings regarding the secondary commercial activities that play out through the platform mechanisms of Google Search around health topics. This research has led to additional actions by the biomedical research and patient advocacy community around health information procurement on large ADM systems accessible to the public (see Couturier, 2023). Improved communication with study participants and an active update process are also outcomes of this research that can help to reduce errors and enhance the quality of the findings.



**Methodologically**, a more nuanced understanding of dealing with real participants, of automated investigations using bots, and of international studies was gained during the course of this investigation. Crowdsourced audits offer benefits, such as natural interaction with the web service by participants with real profiles and the ability to obtain a wide range of input configurations from a variety of users. However, this study also highlighted the inherent limitations of a crowdsourced audit. When examining the system as a black box, the sample size of the population under study was limited and geographically scattered, posing challenges in identifying individuals affected by the disease.<sup>1</sup> The average age at diagnosis of Parkinson’s disease presented additional difficulties. A more complex study design and additional support measures, such as training or instructional videos, might have increased participation. Although these enhancements were not implemented due to the study’s limited analysis of the target group, it is important to recognize that individual characteristics within affected groups can significantly impact the design and implementation of a crowdsourced audit in the context of a black-box analysis. This finding opens up opportunities for future studies to tailor their approaches and maximize their effectiveness.

Using virtual private servers (VPS) as benchmarks in a crowdsourcing audit has advantages but also presents risks. Overhead associated with VPS hosts can be a challenge due to the increased risk of IP ranges being blocked by Internet services. This can lead to complications when automated processes, such as making requests to social media platforms, are conducted using VPS servers. For instance, requests to Google Web Search were continuously blocked at a specific VPS server location. Manual intervention may be required to disable captchas, further hindering automated processes.

Large online service providers like Google have the capability to detect automated audits, resulting in some VPSs being blocked due to increased traffic. This suggests the presence of mechanisms for handling suspicious activity. In order to ensure reliable studies, it is essential to closely mimic user behavior and simulate usage scenarios that align with the specific audience of the crowdsourced audit. Previous research has highlighted the discrepancy between real interactions and pure API requests (McCown & Nelson, 2007), emphasizing the need for realistic simulations. Additionally, demographic factors, such as Internet and media literacy, should be considered when designing the simulation.

The **third study** explored the functionality and potential impact of automated price management systems used for pricing in online stores in order to assess their impact on e-commerce. The investigation focused on the limitations of the methodology in particular. In summary, studying potentially personalized pricing in online commerce is difficult due to a number of obstacles in existing research. Because of a lack of information on the pricing systems of online shop operators and the vast variety of differences between online shops, it is challenging to identify all input variables. This makes it challenging

---

<sup>1</sup>However, Couturier received exceptionally favorable responses from her community. It is not uncommon for patient studies to only have ten to twenty participants (Couturier, 2023).

to create accurate user profiles and simulate their behavior on a website. The inability of a black-box analysis to differentiate between personalized and dynamic pricing is a further issue. Any detected personalization can be easily justified by the vendor by pointing to unknown variables or contexts, rendering this accusation invalid. Therefore, it is difficult to refute the claim that personalization does not exist or that pricing is merely a response to dynamic inventory changes without a deeper understanding of the pricing algorithms used. Future research on personalized pricing in online retail should take these constraints into account.

The last section of this dissertation has already examined the three initial research questions of this dissertation, namely: the types of questions that can be addressed through a black-box analysis; how the structure of a black-box analysis for an ADM system is determined; and which approaches can be employed for different forms of black-box analysis based on access to an ADM system. Having addressed these three fundamental research questions, the next Section focuses on critically examining the methodological limitations of black-box analysis. This critical analysis involves identifying and discussing the inherent constraints and challenges that researchers may encounter when employing this analytical approach. By highlighting these limitations, this dissertation aims to provide a balanced perspective on the capabilities and boundaries of black-box analysis, ensuring that researchers and practitioners are cognizant of its limitations and potential pitfalls.

## 9.2. General limitations of black-box analysis

Although black-box analysis can be used effectively to study algorithmic decision-making systems, it is important to note that it cannot be used to answer all types of questions (Seaver, 2019). While it can be a powerful tool for studying ADM systems, some limitations have been discovered (T. D. Krafft et al., 2021).

First, the fact that algorithmic decision-making systems are embedded in complex sociotechnical systems creates challenges for systematic audits. The complexity of these systems is due to the heterogeneous assortment of various types of social and technical entities that all interact with the system. In addition, algorithms in sociotechnical systems are “contingent, ontogenetic, and performative in nature” (Kitchin, 2016, p. 3). Also, ADM systems are frequently subject to a continuous improvement process that must be considered. Consequently, studying a stable representation of such systems is nearly impossible because “understanding the work and effects of algorithms needs to be sensitive to their contextual, contingent unfolding across situation, time and space.” (Kitchin, 2016, p. 8) Therefore, it is difficult to have a research object that is consistently stable (T. D. Krafft et al., 2021).

Second, the investigator and the investigated object form a system with feedback. This means that the inspection process can affect the black box’s inner workings and

thus make it more difficult to replicate experiments (Seaver, 2019, pp. 413 - 414). The problem of an investigation exerting an impact on the phenomenon being studied was already emphasized in the work by Ashby, 1956 and should be taken into account when planning investigations. Gillespie, 2014 states that the intertwining of algorithms with their audiences is a constantly shifting structure, and the relationships are in a constant state of flux. Moreover, the emergent effects of an algorithm can only be evaluated in the context in which it is executed (Introna, 2015). This implies that it may change over time, influencing the outcomes of future research.

Third, the user experience, which is not always the same for all users, is an additional factor that can influence data collection. It can be modified through A/B testing and personalization.

Fourth, websites can be cached (Wessels, 2001), which can influence data collection. In the study conducted on the 2017 German federal elections, links were frequently collected from advertisements or search results; however, these links were invalid at the time of analysis. It would have been preferable to search for these links at the time of data collection and save the resulting pages for later examination. This would have resulted in the participant's browser opening not only the web page of the ADM instance being evaluated, but also any other web page advertised or displayed on the results page. However, this would have presented difficult-to-resolve security and privacy issues, and might have also violated the ADM provider's terms and conditions.

The most important limitations, problems, and open questions raised above are summarized briefly in the following:

1. When conducting research on ADM systems, it is likely that there may be no consistent and unchanging object of study that maintains its stability and behavior throughout the investigation period due to the interactions of social and technical system (Kitchin, 2016), for example A/B testing (Kohavi & Longbotham, 2017) and personalization.
2. The inspection itself may influence the investigation (Ashby, 1956).
3. Depending on the network architecture, different system results can be delivered simultaneously if web pages are cached between the operator (Wessels, 2001) and the user and if these caches are updated at different times (see Section 7).

As discussed in section 2, accountability for problematic algorithmic outcomes can only be enforced if the behavior of the actor, i.e., the provider of the algorithm-supported service, is questioned. In recent decades, this has proven challenging due to a lack of dependable, large-scale, quantifiable evidence and reliance on anecdotal evidence or speculation. Therefore, I conclude that experimental studies such as the ones conducted in the preceding sections are insufficient, at least when it comes to questions concerning,

for instance, the fundamental rights of citizens or the protection of vulnerable individuals such as the patients in our study.

Despite the limitations mentioned, black-box analysis is a useful tool for assessing and evaluating the social consequences of using an algorithm without requiring a comprehensive understanding of how the algorithm works (Diakopoulos, 2014a). A “critical understanding of the mechanisms and operational logic” (Bucher, 2016, p. 86) is adequate so long as it takes into account the necessary conditions to comprehend a phenomenon (Grunwald, 2002).

### 9.3. Further research

The objective of developing the presented process model for black-box analysis was to provide stakeholders from a variety of disciplines with a concrete and workable framework for designing an investigation of an opaque ADM system, even if they lack in-depth computer science knowledge. The model, which is applicable across disciplines, is intended to encourage anyone to investigate existing inconsistencies associated with the use of ADM systems and dispel any doubts or concerns. According to Wieringa, 2020, the results of a black-box analysis represent a significant algorithmic accountability relationship between those who can access its results, namely the forum and the algorithm provider, i.e., the accountable actor. The presented options for black-box analyses of (opaque) ADM systems should be researched further in terms of their implementation and carried out expeditiously through funded scientific projects.

Furthermore, the study of ADM systems as black boxes should be an ongoing endeavor. As technology evolves and new challenges arise, it becomes increasingly important to continue investigating and analyzing these systems. The process model serves as a foundation for further scientific research in this domain, providing researchers with well-defined starting points for their investigations.

It is worth noting that conducting trials in different domains can greatly contribute to refining and improving the individual steps of the process model. By applying the model to various real-world contexts, researchers can gather valuable insights and identify specific challenges and considerations that may arise in different domains. This iterative approach ensures that the process model remains adaptable and relevant to the evolving landscape of ADM systems.

In an earlier version of the process model, the presentation of the results of the data analysis were recognized as significant steps in the black-box analysis process (T. D. Krafft et al., 2020). However, effective knowledge transfer in this step requires collaboration with domain experts. Due to the specific expertise required, this aspect was deliberately excluded from the scope of the current work. Recognizing the importance of interdisciplinary collaboration, future research endeavors can incorporate the involvement of domain experts to ensure comprehensive analysis and understanding of ADM

systems in different contexts.

## 9.4. Conclusion, value added and political implications

The presented method for evaluating opaque ADM systems is becoming more pertinent in the light of recent events. Political agendas such as the EU’s Digital Services Act (European Parliament and European Council, 2022), the risk-based regulatory framework of the EU’s AI Regulation (European Commission, 2020; European Parliament & European Council, 2021), and the U.S. Algorithmic Accountability Act of 2022 (United States House of Representatives, 2022) demonstrate the increasing political and societal interest in enhanced verifiability, greater transparency, and general accountability of algorithmic systems. In this context, the process model as a practical, interdisciplinary tool for the general public that can help to answer the fundamental question of how these goals can be implemented to ensure accountability of algorithms.

Clearly, AI regulation is a complex task that cannot be solved by a single actor or type of regulatory instrument, but rather requires a diverse mix. In addition to responsible research at the earliest stages of ADM system development and design, self-regulation, co-regulation, and perhaps even strict government oversight are required. Other actors, including critical journalists, researchers, and non-governmental organizations, also have a role in holding ADM systems and their operators accountable. Black-box analyses assume a central role in this regard, as they enable third parties to monitor and evaluate ADM systems without having to look inside the black box of the system in question. To accomplish this task, however, actors and stakeholders, regardless of whose interest they represent, require concrete guidelines and resources. Consequently, **it may be necessary to enact additional regulations that enable or facilitate basic forms of black-box analyses in the first place.** An investigation of an ADM system as a black box still poses a number of legal and technical challenges to the investigating entity. This research has yielded the following specific action recommendations to protect consumers in the context of ADM systems with increased criticality (see Section 2):

### 1. Strengthening the rights of data subjects and their enforcement

Current information practices regarding the rights of data subjects are sub-optimal. Relevant information is expressed imprecisely, or data subjects are presented with so much data at once that they rarely read it. There is a lack of concepts that provide concise and targeted information to consumers regarding the use of their data, for example for Internet personalization. Moreover, the rights of data subjects must be reinforced on multiple fronts. Consumers should be made aware of the existence of an algorithmic decision, along with “meaningful information” about the logic involved, as well as the scope and intended effects of such processing. If, as a result of an algorithm, a product is only accessible in a restricted manner, such as at a higher price or with a more limited selection of payment options, the

reason for this should be explained. An example of a reason worth knowing would be a suspicion of ineligibility for credit based on residence. Lastly, a “reasonable action” should be provided to allow the data subject to take corrective action on the decision, such as having a third party verify the accuracy of the information collected or rejection of the use of the data subject’s personal information. One possibility would be to additionally provide access to non-personalized services, perhaps via a profile setting.

**2. Retention obligation**

Due to the impossibility to trace the results of the ADM systems currently in use, it is nearly impossible to pursue a claim of unequal treatment. In certain regions, it would be prudent to consider mandating the retention of ADM system results so that, in the event of damage, it is possible to determine who received which results and when.

**3. Establishing a suitable interface**

Black-box analyses of ADM systems could be greatly simplified by establishing an appropriate application interface (for example an API). Due to the fierce price competition in online retailing in particular and the fact that many websites actively oppose comparison portals, the creation of a public interface is improbable. A preferable alternative would be to create privileged access open only to authorized researchers and auditors working on behalf of the state or a regulatory body. With the knowledge of what data is being used for personalization and the results of a black-box analysis, initial allegations could be confirmed or denied.

**4. Allowing conditional use of bots**

For a platform that employs an ADM system, automated monitoring by accredited researchers (both from academia and NGOs), which may involve the use of bots, should be enabled (T. D. Krafft et al., 2020). In order to set up representative monitoring of such a platform, it is necessary to generate a large number of scientific accounts fully or at least partially automatically. These accounts must be able to navigate the provider’s website without restriction and without their access being restricted or made more difficult due to bot detection. If bots are authorized by a provider and are therefore known to the provider, great care must be taken to ensure that they are treated exactly like regular site visitors. Otherwise, a situation comparable to the Dieselgate scandal could arise (L. Bovens, 2016).

**5. Legal certainty for black-box analyses**

It should not be prohibited by the terms of service or other regulations to audit an ADM system for scientific or regulatory purposes. Frequently, terms of service for online platforms prohibit the automatic downloading of information from a website, even if the information is publicly available. Under the U.S. Computer

Fraud and Abuse Act (CFAA), which criminalizes unauthorized computer access, exploiting security vulnerabilities to raise public awareness can result in legal consequences (Zettlers, 2010). The interpretation of “authorization” has expanded to include terms of service violations like web scraping (A. Robertson, 2019). This could pose legal risks for people assessing ADM systems. Although researchers have previously challenged the CFAA in the Sandvig v. Barr case (ACLU, 2016), resulting in a ruling that terms of service violations do not violate the CFAA, this still highlights the ongoing problem and potential legal complications.

Currently, this legal basis would also apply when a scientist conducts a black-box analysis, despite the fact that these two types of actions are fundamentally distinct and should therefore be treated differently. When examining ADM systems for undesirable or unlawful behavior, scientists should be permitted to conduct research within a secure legal framework. Otherwise, there is no way to rectify existing imbalances of information and power.

#### 6. Watchdog approach

The studies conducted as part of this research were based on specific suspicions and were limited in terms of duration. It would be desirable to adopt a regular and independent review of ADM systems as black boxes, incorporating a heightened level of critical assessment. This would be referred to as a watchdog approach because it would involve the (institutionalized) continuous testing of an algorithm. Implementing such a procedure, where there is ongoing scrutiny and monitoring of ADM systems, would ensure accountability and transparency. By regularly reviewing these systems, independent of any specific suspicion, potential issues and biases could be identified and addressed proactively. Additionally, suspicion-driven research plays a crucial role in investigating specific concerns raised by stakeholders or instances of suspected algorithmic bias, and would facilitate a more comprehensive evaluation of ADM systems. By combining both regular, suspicion-independent reviews and targeted, suspicion-driven research, a more robust and thorough assessment of ADM systems could be achieved. This approach would not only promote accountability and transparency but also help to instill trust and confidence in the use of these systems across various domains and applications.

The results presented in this dissertation are intended to increase the transparency and accountability of ADM systems, thereby increasing trust in such systems and allowing them to fully realize their potential to improve decision-making in a variety of application domains. Consumer advocates, charities/NGOs, regulators, social science and biomedical research centers, as well as patients are working in a variety of settings to uncover and address allegations against ADM systems and their operators. This means that, as seen in section 6, there is specific expertise regarding the lived and contextual realities of ADM systems, but few data-generating methodological resources exist to em-

power these experts to engage in meaningful research and action. The objective is for our society to derive advantages by making it easier for diverse parties to scrutinize and comprehend the content produced in progressively pervasive and frequently semi-public digital resources.

This objective keeps gaining relevance, as more and more black-box systems are developed and given increasing societal power. This fact becomes evident by the inclusion of the third and fourth requests in the existing framework of the European Union's Digital Service Act (European Parliament and European Council, 2022). The rising ubiquity of AI, in particular, demands careful consideration both from civil society and regulators: Which decisions are we as a society willing to let ADM systems make for us? Which amount of "unfairness" and "opacity" are we willing to accept? These debates cannot be held without evidence of how the current systems actually perform on those accounts. The rapid rise of Large Language Models (e.g., ChatGPT) has catapulted these issues to the public stage, and so far, both political and regulatory actors are lacking adequate responses to the new challenges these systems pose.

Without regulation, however, the negative implications these systems can have can wreak unseen havoc on both individuals and groups. As we see more organizations and people use opaque, uncontrolled black-box systems, the number of cases where "unintended side-effects" occur will continue to rise. Society needs norms and standards that regulate both the application and the accountability of those systems, making them as transparent as possible not only ex-post but as the design default, and providing both tools and resources for black-box analyses where transparency is not an option.

Black-box analysis can be a powerful tool to gain transparency and distribute accountability for ADM systems. Five years ago, the obvious gap between the methods available and the tools used suggested that this tool is not yet easy enough to implement for non-computer science researchers. By compiling different methods and decisions into this process model, I hope to contribute to a more empowered academia and civil society.

In light of the growing attempts by major technology companies to obscure the inner workings of their systems, such as restricting public access to APIs and withholding their own analyses, the importance of conducting black-box analyses has become increasingly significant. With the implementation of recent regulatory initiatives, there is a pressing need for individuals who are willing and capable of independently examining these opaque systems. It is encouraging to anticipate a rise in the number of such projects employing black-box analysis, as it serves as a proactive measure of societal self-defense, ensuring transparency, accountability, and the protection of public interests.



# Bibliography

- ACLU. (2016). *Sandvig v. Barr - first amendment challenge to federal computer fraud and abuse act* [last accessed on 08.06.2023]. <https://www.acludc.org/en/cases/sandvig-v-barr-first-amendment-challenge-federal-computer-fraud-and-abuse-act>
- Ada Lovelace Institute. (2020). *Examining the Black Box: Tools for assessing algorithmic systems* [last accessed on 31.12.2022]. <https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/>
- Adler, P., Falk, C., Friedler, S. A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., & Venkatasubramanian, S. (2018). Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1), 95–122. <https://doi.org/10.1007/s10115-017-1116-3>
- Aikins, J. S., Kunz, J. C., Shortliffe, E. H., & Fallat, R. J. (1982). PUFF: An expert system for interpretation of pulmonary function data. *Computers and Biomedical Research*, 16(3), 199–208. [https://doi.org/10.1016/0010-4809\(83\)90021-6](https://doi.org/10.1016/0010-4809(83)90021-6)
- Akinwale, A., & Niewiadomski, A. (2015). Efficient Similarity Measures for Texts Matching. *Journal of Applied Computer Science*, 23(1), 7–28. <https://doi.org/10.34658/jacs.2015.23.2.7-28>
- Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through Optimization: How Facebook’s Ad Delivery Can Lead to Biased Outcomes. 3(CSCW, Article 199). <https://doi.org/10.1145/3359301>
- Amdur, R. J., & Bankert, E. A. (2010). *Institutional Review Board - Member Handbook*. Jones and Bartlett Publishers.
- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. <https://doi.org/10.1177/1461444816676645>
- Andrade, L., & Silva, M. J. (n.d.). Relevance Ranking for Geographic IR. In R. Purves & C. B. Jones (Eds.), *Proceedings of the 3rd ACM Workshop On Geographic Information Retrieval, GIR 2006, Seattle, WA, USA, August 10, 2006*. Department of Geography, University of Zurich.
- Andreou, A., Venkatadri, G., Goga, O., Gummadi, K. P., Loiseau, P., & Mislove, A. (2018). Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook’s Explanations. *NDSS 2018 - Network and Distributed System Security Symposium*, 1–15. <https://hal.science/hal-01955309>

- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine Bias - There's software used across the country to predict future criminals. And it's biased against blacks* [last accessed on 30.05.2023]. ProPublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Ashby, W. R. (1956). *An introduction to cybernetics*. Chapman & Hall Ltd.
- Badmaeva, T., & Hüllmann, J. A. (2019). Investigating Personalized Price Discrimination of Textile-, Electronics-and General Stores in German Online Retail [last accessed on 31.12.2022]. Association for Information Systems. <https://aisel.aisnet.org/wi2019/specialtrack01/papers/1/>
- Baer, T. (2019). *Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists*. Apress.
- Bandy, J. (2021). Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1). <https://doi.org/10.1145/3449148>
- Barford, P., Canadi, I., Krushevskaja, D., Ma, Q., & Muthukrishnan, S. (2014). Adscape: Harvesting and Analyzing Online Display Ads. *Proceedings of the 23rd International Conference on World Wide Web*, 597–608. <https://doi.org/10.1145/2566486.2567992>
- Barocas, S., Guo, A., Kamar, E., Krones, J., Morris, M. R., Vaughan, J. W., Wadsworth, W. D., & Wallach, H. (2021). Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 368–378. <https://doi.org/10.1145/3461702.3462610>
- Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities* [<http://www.fairmlbook.org>]. fairmlbook.org.
- Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, 104(3), 671–732. Retrieved April 8, 2023, from <http://www.jstor.org/stable/24758720>
- Barr, E. T., Harman, M., McMinn, P., Shahbaz, M., & Yoo, S. (2014). The oracle problem in software testing: A survey. *IEEE Transactions on Software Engineering*, 41(5), 507–525. <https://doi.org/10.1109/TSE.2014.2372785>
- Bautista, R. (2012). An Overlooked Approach in Survey Research: Total Survey Error. In L. Gideon (Ed.), *Handbook of Survey Methodology for the Social Sciences* (pp. 37–49). Springer New York. [https://doi.org/10.1007/978-1-4614-3876-2\\_4](https://doi.org/10.1007/978-1-4614-3876-2_4)
- Belleflamme, P., & Peitz, M. (2010). *Industrial Organization: Markets and Strategies*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511757808>
- Ben-Israel, I., Cerdio, J., Ema, A., Friedman, L., Ienca, M., Mantelero, A., Matania, E., Muller, C., Shiroyama, H., & Vayena, E. (2020). *Towards regulation of AI systems - Global perspectives on the development of a legal framework on Artificial Intelligence (AI) systems based on the Council of Europe's standards on human rights, democracy and the rule of law* [online: <https://edoc.coe.int/en/artific>

- ial-intelligence/9656-towards-regulation-of-ai-systems.html; last accessed on 07.06.2023]. Council of Europe - CAHAI.
- Bennett, P. N., White, R. W., Chu, W., Dumais, S. T., Bailey, P., Borisyuk, F., & Cui, X. (2012). Modeling the Impact of Short- and Long-Term Behavior on Search Personalization. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 185–194. <https://doi.org/10.1145/2348283.2348312>
- Berger, I., Ahmad, A., Bansal, A., Kapoor, T., Sipp, D., & Rasko, J. E. (2016). Global Distribution of Businesses Marketing Stem Cell-Based Interventions. *Cell Stem Cell*, 19(2), 158–162. <https://doi.org/https://doi.org/10.1016/j.stem.2016.07.015>
- Bertrand, M., & Duflo, E. (2017). Field Experiments on Discrimination. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of Field Experiments* (pp. 309–393, Vol. 1). North-Holland. <https://doi.org/10.1016/bs.hefe.2016.08.004>
- Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4), 991–1013. <https://doi.org/10.1257/0002828042002561>
- Bhat, A., & Quadri, S. M. K. (2015). Equivalence class partitioning and boundary value analysis - A review. *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1557–1562.
- Biddings, A. (2019). *A new policy on advertising for speculative and experimental medical treatments* [last accessed on 16.06.2023]. Google Support. <https://support.google.com/google-ads/answer/9475042>
- Biehl, J. K., & Russell, B. (2009). Introduction to stem cell therapy. *The Journal of cardiovascular nursing*, 24(2), 98. <https://doi.org/10.1097/JCN.0b013e318197a6a5>
- Binns, R. (2018). Algorithmic Accountability and Public Reason. *Philosophy & Technology*, 31(4), 543–556. <https://doi.org/10.1007/s13347-017-0263-5>
- Bonczek, R. H., Holsapple, C. W., & Whinston, A. B. (1981). *Foundations of Decision Support Systems* (J. W. Schmidt, Ed.). Elsevier.
- Bovens, L. (2016). The Ethics of Dieselgate. *Midwest Studies In Philosophy*, 40(1), 262–283. <https://doi.org/10.1111/misp.12060>
- Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal*, 13(4), 447–468. <https://doi.org/10.1111/j.1468-0386.2007.00378.x>
- Braithwaite, V. (2020). Beyond the bubble that is Robodebt: How governments that lose integrity threaten democracy. *Australian Journal of Social Issues*, 55(3), 242–259. <https://doi.org/10.1002/ajs4.122>
- Brauneis, R., & Goodman, E. P. (2018). Algorithmic Transparency for the Smart City. *Yale J.L. & Tech.*, 20, 103. <https://doi.org/10.2139/ssrn.3012499>

- Bresciani, S., & Eppler, M. J. (2008). The Risks of Visualization - A Classification of Disadvantages Associated with Graphic Representations of Information. *Identität und Vielfalt der Kommunikationswissenschaft (2009), ICA Working Paper(1)*, 165–178.
- Brewer, M. B., & Crano, W. D. (2014). Research Design and Issues of Validity. In T. H. Reis & M. C. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (pp. 47–79). Cambridge University Press.
- Bridle, J. (2017). *Something is wrong on the internet* [last accessed on 31.12.2022]. Medium. <https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2>
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine [Proceedings of the Seventh International World Wide Web Conference]. *Computer Networks and ISDN Systems, 30(1)*, 107–117. [https://doi.org/10.1016/S0169-7552\(98\)00110-X](https://doi.org/10.1016/S0169-7552(98)00110-X)
- Brkan, M. (2019). Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond. *International Journal of Law and Information Technology, 27(2)*, 91–121. <https://doi.org/10.1093/ijlit/ey017>
- Bryson, J. J., & Theodorou, A. (2019). How Society Can Maintain Human-Centric Artificial Intelligence. In M. Toivonen & E. Saari (Eds.), *Human-Centered Digitalization and Services* (pp. 305–323). Springer Nature Singapore. [https://doi.org/10.1007/978-981-13-7725-9\\_16](https://doi.org/10.1007/978-981-13-7725-9_16)
- Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project*. Addison Wesley Longman Publishing Co.
- Bucher, T. (2016). Neither Black Nor Box: Ways of Knowing Algorithms. In S. Kubitschko & A. Kaun (Eds.), *Innovative Methods in Media and Communication Research* (pp. 81–98). Springer International Publishing. [https://doi.org/10.1007/978-3-319-40700-5\\_5](https://doi.org/10.1007/978-3-319-40700-5_5)
- Buchholz, S., & Ganser, A. (2009). Hämatopoetische Stammzelltransplantation. *Der Internist, 50(5)*, 572–580. <https://doi.org/10.1007/s00108-008-2273-y>
- Bundesministerium der Justiz und für Verbraucherschutz. (2020). *Referentenentwurf: Entwurf eines Gesetzes zur Änderung des Bürgerlichen Gesetzbuchs und des Einführungsgesetzes zum Bürgerlichen Gesetzbuche in Umsetzung der EU-Richtlinie zur besseren Durchsetzung und Modernisierung der Verbraucherschutzvorschriften der Union* [last accessed on 30.05.2023]. [https://www.bmj.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/RefE\\_BereitstellungdigitalerInhalte\\_2.pdf](https://www.bmj.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/RefE_BereitstellungdigitalerInhalte_2.pdf)
- Buolamwini, J., & Gebru, T. (2018). *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* (S. A. Friedler & C. Wilson, Eds.; Vol. 81). PMLR.

- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1–12. <https://doi.org/10.1177/2053951715622512>
- Burt, R. S. (1991). Measuring age as a structural concept. *Social Networks*, 13(1), 1–34. [https://doi.org/10.1016/0378-8733\(91\)90011-H](https://doi.org/10.1016/0378-8733(91)90011-H)
- Bygrave, L. A. (2019, September). Minding the Machine v2.0: The EU General Data Protection Regulation and Automated Decision Making. In *Algorithmic Regulation*. Oxford University Press. <https://doi.org/10.1093/oso/9780198838494.003.0011>
- Cambazoglu, B. B., & Altıngövdü, I. S. (2012). Impact of Regionalization on Performance of Web Search Engine Result Caches. In L. Calderón-Benavides, C. González-Caro, E. Chávez, & N. Ziviani (Eds.), *String Processing and Information Retrieval* (pp. 161–166). Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-34109-0\\_17](https://doi.org/10.1007/978-3-642-34109-0_17)
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Wadsworth Publishing.
- Caulfield, T., Sipp, D., Murry, C. E., Daley, G. Q., & Kimmelman, J. (2016). Confronting stem cell hype. *Science*, 352(6287), 776–777. <https://doi.org/10.1126/science.1244620>
- CDU, CSU and SPD. (2018). *Ein neuer Aufbruch für Europa. Eine neue Dynamik für Deutschland. Ein neuer Zusammenhalt für unser Land* [last accessed on 09.06.2023]. Koalitionsvertrag zwischen CDU, CSU und SPD - 19. Legislaturperiode. [https://www.bundestag.de/resource/blob/543200/9f9f21a92a618c77aa330f00ed21e308/kw49\\_koalition\\_koalitionsvertrag-data.pdf](https://www.bundestag.de/resource/blob/543200/9f9f21a92a618c77aa330f00ed21e308/kw49_koalition_koalitionsvertrag-data.pdf)
- Chen, C.-h., Härdle, W. K., & Unwin, A. (2007). *Handbook of Data Visualization*. Springer Science & Business Media.
- Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). Investigating the Impact of Gender on Rank in Resume Search Engines. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3173574.3174225>
- Chen, L., Mislove, A., & Wilson, C. (2015). Peeking Beneath the Hood of Uber. *Proceedings of the 2015 Internet Measurement Conference*, 495–508. <https://doi.org/10.1145/2815675.2815681>
- Clark, R., & Vincent, N. (2012). Capacity-contingent pricing and competition in the airline industry. *Journal of Air Transport Management*, 24, 7–11. <https://doi.org/10.1016/j.jairtraman.2012.04.005>
- Coffman, L. C., & Niederle, M. (2015). Pre-analysis Plans Have Limited Upside, Especially Where Replications Are Feasible. *Journal of Economic Perspectives*, 29(3), 81–98. <https://doi.org/10.1257/jep.29.3.81>
- Cohen, C. B., & Cohen, P. J. (2010). International Stem Cell Tourism and the Need for Effective Regulation: Part I: Stem Cell Tourism in Russia and India: Clinical

- Research, Innovative Treatment, or Unproven Hype? *Kennedy Institute of Ethics Journal*, 20(1), 27–49. <https://doi.org/10.1353/ken.0.0305>
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates Publishers.
- Connolly, R., O'Brien, T., & Flaherty, G. (2014). Stem cell tourism—a web-based analysis of clinical services available to international travellers. *Travel medicine and infectious disease*, 12(6 Pt B), 695–701. <https://doi.org/10.1016/j.tmaid.2014.09.008>
- Cook, T., & Campbell, D. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Houghton Mifflin.
- Cortez, P., & Embrechts, M. J. (2011). *Opening black box Data Mining models using Sensitivity Analysis*. IEEE. <https://doi.org/10.1109/CIDM.2011.5949423>
- Couturier, A. (2023). *Google search and the mediation of digital health information: a case study on unproven stem cell treatments* [Doctoral Thesis]. University of Edinburgh. <https://doi.org/10.7488/era/3282>
- Covington, P., Adams, J., & Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations, 191–198. <https://doi.org/10.1145/2959100.2959190>
- Cutrell, E., & Guan, Z. (2007). What Are You Looking for? An Eye-Tracking Study of Information Usage in Web Search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 407–416. <https://doi.org/10.1145/1240624.1240690>
- Daniel, W. W. (1968). *Racial discrimination in England: based on the PEP report* (Vol. S257). Penguin.
- Darolia, R., Koedel, C., Martorell, P., Wilson, K., & Perez-Arce, F. (2015). Do Employers Prefer Workers Who Attend For-Profit Colleges? Evidence from a Field Experiment. *Journal of Policy Analysis and Management*, 34(4), 881–903. <https://doi.org/10.1002/pam.21863>
- Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women* [last accessed on 08.06.2023]. Reuters. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>
- Data Ethics Commission. (2019). *Opinion of the Data Ethics Commission* [last accessed on 30.05.2023]. Bundesministerium des Innern und für Heimat. <https://www.bmi.bund.de/SharedDocs/downloads/EN/themen/it-digital-policy/datenethikkommission-abschlussgutachten-lang.pdf>
- Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1), 92–112. <https://doi.org/10.1515/popets-2015-0007>
- Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., & Sampath, D. (2010). The YouTube Video Recommendation System. *Proceedings of the Fourth ACM Conference on Recommender Systems*, 293–296. <https://doi.org/10.1145/1864708.1864770>

- Delapina, M. (2019). *AK-TEST - Preisdifferenzierung im Online-Handel* [last accessed on 08.06.2023]. Bundesministerium für Arbeit - Österreich. [https://www.arbeit-erkammer.at/beratung/konsument/HandyundInternet/Internet/Online-Handel\\_2019.pdf](https://www.arbeit-erkammer.at/beratung/konsument/HandyundInternet/Internet/Online-Handel_2019.pdf)
- den Boer, A. V. (2015). Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys in Operations Research and Management Science*, 20(1), 1–18. <https://doi.org/https://doi.org/10.1016/j.sorms.2015.03.001>
- Deutscher Bundestag. (2020). *Bericht der Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale* [last accessed on 09.06.2023]. <https://dserver.bundestag.de/btd/19/237/1923700.pdf>
- Diakopoulos, N. (2014a). Algorithmic Accountability Reporting: On the Investigation of Black Boxes. *Tow Center for Digital Journalism*. <https://doi.org/10.7916/D8ZK5TW2>
- Diakopoulos, N. (2014b). Algorithmic Accountability: Journalistic investigation of computational power structures. *Digital Journalism*, 3(3), 398–415. <https://doi.org/10.1080/21670811.2014.976411>
- DIN & DKE. (2020). *German Standardization Roadmap Artificial Intelligence* [last accessed on 09.06.2023]. <https://www.dke.de/resource/blob/2008048/99bc6d952073ca88f52c0ae4a8c351a8/nr-ki-english---download-data.pdf>
- DIN 69901: (2009). *Projektmanagement* (Standard). Beuth Verlag. Berlin.
- Domingos, P. (1999). The Role of Occam's Razor in Knowledge Discovery. *Data Mining and Knowledge Discovery*, 3(4), 409–425. <https://doi.org/10.1023/A:1009868929893>
- Dominique-Ferreira, S., Vasconcelos, H., & Proença, J. F. (2016). Determinants of customer price sensitivity: an empirical analysis. *Journal of Services Marketing*, 30(3), 327–340. <https://doi.org/10.1108/JSM-12-2014-0409>
- Dreyer, S., & Schulz, W. (2018). *Was bringt die Datenschutz-Grundverordnung für automatisierte Entscheidungssysteme?* Bertelsmann Stiftung. <https://doi.org/10.11586/2018011>
- Dreyfus, H., Dreyfus, S. E., & Athanasiou, T. (1986). *Mind over machine - The Power of Human Intuition and Expertise in the Era of the Computer*. Simon; Schuster.
- Drost, E. A. (2011). Validity and reliability in social science research [Place: Crawley, WA, Australia Publisher: University of Western Australia]. *Education Research and Perspectives*, 38(1), 105–123. Retrieved April 8, 2023, from <https://search.informit.org/doi/10.3316/informit.491551710186460>
- Ecke, O. (2016). *Wie häufig und wofür werden Intermediäre genutzt? Die quantitative Perspektive der Zusatzbefragung in der MedienGewichtungsStudie* [last accessed on 31.12.2022]. KANTAR TNS. [https://www.die-medienanstalten.de/fileadmin/user\\_upload/Veranstaltungen/2016/2016\\_11\\_30\\_Intermediaere\\_und\\_Meinu](https://www.die-medienanstalten.de/fileadmin/user_upload/Veranstaltungen/2016/2016_11_30_Intermediaere_und_Meinu)

- ngsbildung/TNS\_Intermediaere\_und\_Meinungsbildung\_Praesi\_Web\_Mappe.pdf
- Eilam, E. (2011). *Reversing: Secrets of Reverse Engineering* (2005th ed.). John Wiley & Sons.
- Eirinaki, M., & Vazirgiannis, M. (2003). Web Mining for Web Personalization. *ACM Trans. Internet Technol.*, 3(1), 1–27. <https://doi.org/10.1145/643477.643478>
- Ellenberg, J. H. (1994). Selection bias in observational and experimental studies. *Statistics in Medicine*, 13(5-7), 557–567. <https://doi.org/10.1002/sim.4780130518>
- Enserink, M. (2006). Selling the Stem Cell Dream. *Science*, 313(5784), 160–163. <https://doi.org/10.1126/science.313.5784.160>
- Erikainen, S., Couturier, A., & Chan, S. (2020). Marketing Experimental Stem Cell Therapies in the UK: Biomedical Lifestyle Products and the Promise of Regenerative Medicine in the Digital Era. *Science as Culture*, 29(2), 219–244. <https://doi.org/10.1080/09505431.2019.1656183>
- Erikainen, S., Pickersgill, M., Cunningham-Burley, S., & Chan, S. (2019). Patienthood and participation in the digital era [PMID: 31041112]. *DIGITAL HEALTH*, 5, 2055207619845546. <https://doi.org/10.1177/2055207619845546>
- European Commission. (2020). *White Paper on Artificial Intelligence - A European approach to excellence and trust* [last accessed on 09.06.2023]. [https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-intelligence-feb2020\\_en.pdf](https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-intelligence-feb2020_en.pdf)
- European Commission & Council of the European Union. (2019). *Directive (EU) 2019/2161 of the European Parliament and of the Council of 27 November 2019 amending Council Directive 93/13/EEC and Directives 98/6/EC, 2005/29/EC and 2011/83/EU of the European Parliament and of the Council as regards the better enforcement and modernisation of Union consumer protection rules* [last accessed on 28.03.2023]. Official Journal of the European Union. <http://data.europa.eu/eli/dir/2019/2161/oj>
- European Parliament. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)* [last accessed on 09.06.2023]. Official Journal of the European Union. <http://data.europa.eu/eli/reg/2016/679/oj>
- European Parliament & Council of the European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)* [online: <http://data.europa.eu/eli/reg/2016/679/oj>; last accessed on 30.05.2023].



- European Parliament & European Council. (2021). *Commission Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM (2021) 206 (Apr. 21, 2021)* [last accessed on 09.06.2023]. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206>
- European Parliament and European Council. (2022). *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)* [last accessed on 09.06.2023]. <http://data.europa.eu/eli/reg/2022/2065/oj>
- EuroStemCell. (2023a). *Types of stem cells and their uses* [last accessed on 08.06.2023]. <https://www.eurostemcell.org/types-stem-cells-and-their-uses>
- EuroStemCell. (2023b). *What diseases and conditions can be treated with stem cells?* [last accessed on 30.05.2023]. <https://www.eurostemcell.org/what-diseases-and-conditions-can-be-treated-stem-cells>
- Fagin, R., Kumar, R., & Sivakumar, D. (2003). Comparing Top  $k$  Lists. *SIAM Journal on Discrete Mathematics*, 17(1), 134–160. <https://doi.org/10.1137/S0895480102412856>
- Fan, W., Gordon, M. D., & Pathak, P. (2000). Personalization of Search Engine Services for Effective Retrieval and Knowledge Management. *Proceedings of the Twenty First International Conference on Information Systems*, 20–34. <http://aisel.aisnet.org/icsis2000/4>
- Fassnacht, M. (2003). Preisdifferenzierung. In H. Diller & A. Herrmann (Eds.), *Handbuch Preispolitik: Strategien — Planung — Organisation — Umsetzung* (pp. 483–502). Gabler Verlag. [https://doi.org/10.1007/978-3-322-90512-3\\_23](https://doi.org/10.1007/978-3-322-90512-3_23)
- Fisher, R. J. (1993). Social Desirability Bias and the Validity of Indirect Questioning. *Journal of Consumer Research*, 20(2), 303–315. <https://doi.org/10.1086/209351>
- Frey, H., & Patil, S. R. (2002). Identification and Review of Sensitivity Analysis Methods. *Risk Analysis*, 22(3), 553–578. <https://doi.org/https://doi.org/10.1111/0272-4332.00039>
- Fry, H. (2018). *Hello World: How to Be Human in the Age of the Machine*. Doubleday.
- Gabriel, G., Mittelstraß, J., & Carrier, M. (1995). *Enzyklopädie Philosophie und Wissenschaftstheorie: Band 1: A–B*. Verlag JB Metzler.
- Gaddis, S. M. (2018). An Introduction to Audit Studies in the Social Sciences. In S. M. Gaddis (Ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance* (pp. 3–44). Springer International Publishing. [https://doi.org/10.1007/978-3-319-71153-9\\_1](https://doi.org/10.1007/978-3-319-71153-9_1)
- Gallego, G., & Van Ryzin, G. (1994). Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management science*, 40(8), 999–1020. <https://doi.org/10.1287/mnsc.40.8.999>

- Gao, H., Mittal, V., & Zhang, Y. (2020). The Differential Effect of Local–Global Identity Among Males and Females: The Case of Price Sensitivity. *Journal of Marketing Research*, 57(1), 173–191. <https://doi.org/10.1177/0022243719889028>
- García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J., & Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, 1(1), 9. <https://doi.org/10.1186/s41044-016-0014-0>
- Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of Computer-Mediated Communication*, 14(2), 265–285. <https://doi.org/10.1111/j.1083-6101.2009.01440.x>
- Gauch, H. G. (2003). *Scientific method in practice*. Cambridge University Press.
- Geitz, E., Vater, C., & Zimmer-Merkle, S. (2020). Einleitung: Black Boxes: Bausteine und Werkzeuge zu ihrer Analyse. In E. Geitz, C. Vater, & S. Zimmer-Merkle (Eds.), *Interdisziplinäre Perspektiven* (pp. 3–18). De Gruyter. <https://doi.org/doi:10.1515/9783110701319-001>
- Gideon, L. (2012). *Handbook of survey methodology for the social sciences*. Springer.
- Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The Null Ritual: What You Always Wanted to Know About Significance Testing but Were Afraid to Ask. In D. Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 391–408). Sage Publications Ltd.
- Gillespie, T. (2014). The Relevance of Algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 167–194). MIT Press.
- Giralt, S., Bishop, M. R., & Bishop, M. R. (2009). *Hematopoietic Stem Cell Transplantation* (M. R. Bishop, Ed.). Springer.
- Gönsch, J., Klein, R., & Steinhardt, C. (2009). Dynamic Pricing - State-of-The-Art [<https://ssrn.com/abstract=2179225>]. *Zeitschrift für Betriebswirtschaft, Ergänzungsheft 3 'Operations Research in der Betriebswirtschaft'*(3), 1–40.
- Google. (2005, November). *Personalized Search Graduates from Google Labs* [last accessed on 16.06.2023]. Googlepress. [http://googlepress.blogspot.com/2005/11/personalized-search-graduates-from\\_10.html](http://googlepress.blogspot.com/2005/11/personalized-search-graduates-from_10.html)
- Google. (2009, December). *Personalized Search for everyone* [last accessed on 16.06.2023]. Googlepress. Retrieved May 25, 2018, from <https://googleblog.blogspot.com/2009/12/personalized-search-for-everyone.html>
- Google. (2018). *Google Privacy Policy* [last accessed on 16.06.2023]. <https://policies.google.com/privacy/archive/20180525?hl=en>
- Google. (2023a). *About automation with Google Ads* [last accessed on 08.06.2023]. Google Help Center. <https://support.google.com/google-ads/answer/9297584>
- Google. (2023b). *About Discovery campaigns* [last accessed on 16.06.2023]. Google Help Center. <https://support.google.com/google-ads/answer/9176876>
- Google. (2023c). *Determine a bid strategy based on your goals* [last accessed on 08.06.2023]. Google Help Center. <https://support.google.com/google-ads/answer/2472725>

- Google. (2023d). *Display campaigns* [last accessed on 16.06.2023]. Google Help Center. <https://support.google.com/google-ads/topic/10016807>
- Google. (2023e). *How AdSense works* [last accessed on 16.06.2023]. Google Help Center. <https://support.google.com/adsense/answer/6242051>
- Google. (2023f). *How Search algorithms work* [last accessed on 16.06.2023]. Google Help Center. <https://www.google.com/search/howsearchworks/algorithms/>
- Google. (2023g). *Personalized advertising* [last accessed on 16.06.2023]. Google Help Center. <https://support.google.com/adspolicy/answer/143465>
- Google. (2023h). *Targeting your ads* [last accessed on 16.06.2023]. Google Help Center. <https://support.google.com/google-ads/answer/1704368?hl=en>
- Google. (2023i). *Why you're seeing an ad* [last accessed on 16.06.2023]. Google Help Center. <https://support.google.com/accounts/answer/1634057>
- Google Analytics. (2020). *Domainübergreifendes Tracking* [last accessed on 31.12.2022]. <https://support.google.com/analytics/answer/1033876>
- Granka, L. A. (2010). The Politics of Search: A Decade Retrospective. *The Information Society*, 26(5), 364–374. <https://doi.org/10.1080/01972243.2010.511560>
- Granka, L. A., Joachims, T., & Gay, G. (2004). Eye-Tracking Analysis of User Behavior in WWW Search. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 478–479. <https://doi.org/10.1145/1008992.1009079>
- Grimmelmann, J. (2008). The google dilemma. *New York Law School Law Review; NYLS Legal Studies Research Paper*, 53(08/09-2), 939–950. <https://ssrn.com/abstract=1160320>
- Grunwald, A. (2002). *Technikfolgenabschätzung - Eine Einführung* (Vol. 3). Nomos.
- Guha, S., Cheng, B., & Francis, P. (2010). Challenges in Measuring Online Advertising Systems. *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, 81–87. <https://doi.org/10.1145/1879141.1879152>
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5). <https://doi.org/10.1145/3236009>
- Gunawardana, A., & Shani, G. (2009). A Survey of Accuracy Evaluation Metrics of Recommendation Tasks. *Journal of Machine Learning Research*, 10, 2935–2962.
- Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI - Explainable artificial intelligence. *Science Robotics*, 4(37), eaay7120. <https://doi.org/10.1126/scirobotics.aay7120>
- Guo, M.-H., Liu, C.-A., & Huang, S.-F. (2022). Dynamic Co-movement Detection of High Frequency Financial Data. *Journal of Data Science*, 10(3), 345–362. [https://doi.org/10.6339/JDS.201207\\_10\(3\).0001](https://doi.org/10.6339/JDS.201207_10(3).0001)
- Haeri, M., Hartmann, K., König, P., Krafft, T., Sirsch, J., Joisten, K., Wenzelburger, G., & Zweig, K. (2020). *Denkanstöße zum Einsatz von ADM-Systemen in der*

- öffentlichen Verwaltung (tech. rep.). Technische Universität Kaiserslautern, [http://fairandgoodadm.cs.uni-kl.de/res/Denkanstöße\\_final.pdf](http://fairandgoodadm.cs.uni-kl.de/res/Denkanstöße_final.pdf).
- Haeri, M. A., Hartmann, K., Sirsch, J., Wenzelburger, G., & Zweig, K. A. (2022). Promises and Pitfalls of Algorithm Use by State Authorities. *Philosophy & Technology*, 35(2), 33. <https://doi.org/10.1007/s13347-022-00528-0>
- Haim, M., Graefe, A., & Brosius, H.-B. (2018). Burst of the Filter Bubble?: Effects of personalization on the diversity of Google News. *Digital Journalism*, 6(3), 330–343. <https://doi.org/10.1080/21670811.2017.1338145>
- Hallensleben, S., Hustedt, C., Fetic, L., Fleischer, T., Grünke, P., Hagendorff, T., Hauer, M. P., Hauschke, A., Heesen, J., Herrmann, M., Hillerbrand, R., Hubig, C., Kaminski, A., Krafft, T. D., Loh, W., Otto, P., & Puntschuh, M. (2020). From Principles to Practice - An interdisciplinary framework to operationalise AI ethics. *iRights. Lab, Tech. Rep.* [https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO\\_2020\\_final.pdf](https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf).
- Hamby, D. M. (1994). A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment*, 32(2), 135–154. <https://doi.org/10.1007/BF00547132>
- Hannák, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., & Wilson, C. (2013). Measuring Personalization of Web Search. *Proceedings of the 22nd International Conference on World Wide Web*, 527–538. <https://doi.org/10.1145/2488388.2488435>
- Hannák, A., Soeller, G., Lazer, D., Mislove, A., & Wilson, C. (2014). Measuring Price Discrimination and Steering on E-Commerce Web Sites. *Proceedings of the 2014 Conference on Internet Measurement Conference*, 305–318. <https://doi.org/10.1145/2663716.2663744>
- Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., & Wilson, C. (2017). Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1914–1933. <https://doi.org/10.1145/2998181.2998327>
- Harvey, M., Crestani, F., & Carman, M. J. (2013). Building User Profiles from Topic Models for Personalised Search. *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, 2309–2314. <https://doi.org/10.1145/2505515.2505642>
- Hauer, M. P., Hofmann, X. C., Krafft, T. D., & Zweig, K. A. (2020). Quantitative analysis of automatic performance evaluation systems based on the h-index. *Scientometrics*, 123(2), 735–751. <https://doi.org/10.1007/s11192-020-03407-7>
- Hauer, M. P., Kevekordes, J., & Haeri, M. A. (2021). Legal perspective on possible fairness measures –A legal discussion using the example of hiring decisions. *Computer Law & Security Review*, 42, 105583. <https://doi.org/10.1016/j.clsr.2021.105583>

- Hauer, M. P., Krafft, T. D., & Zweig, K. (2023). Overview of transparency and inspectability mechanisms to achieve accountability of AI systems [The paper is already accepted and in publication.]. *Data and Policy*.
- Hays, C. L. (1999). *Variable-Price Coke Machine Being Tested* [last accessed on 08.06.2023]. The New York Times. <https://www.nytimes.com/1999/10/28/business/variable-price-coke-machine-being-tested.html>
- Heckman, J. J. (1998). Detecting Discrimination. *Journal of Economic Perspectives*, 12(2), 101–116. <https://doi.org/10.1257/jep.12.2.101>
- Heckman, J. J. (2010). Selection Bias and Self-Selection. In S. N. Durlauf & L. E. Blume (Eds.), *Microeconometrics* (pp. 242–266). Palgrave Macmillan UK. [https://doi.org/10.1057/9780230280816\\_29](https://doi.org/10.1057/9780230280816_29)
- Heise. (2017). *Verbraucherzentralen fordern „Algorithmen-Tüv“* [author: dpa; last accessed on 30.12.2022]. <https://www.heise.de/newsticker/meldung/Verbraucherzentralen-fordern-Algorithmen-Tuev-3691265.html>
- Herberts, C. A., Kwa, M. S., & Hermsen, H. P. (2011). Risk factors in the development of stem cell therapy. *Journal of Translational Medicine*, 9(1), 29. <https://doi.org/10.1186/1479-5876-9-29>
- Heßler, M. (2015). Das Öffnen der black box. Perspektiven der Genderforschung auf Technikgeschichte. In *{Gender; Technik; Museum} - Strategien für eine geschlechtergerechte Museumspraxis* (pp. 29–38). Zentrum für Interdisziplinäre frauen- und Geschlechterforschung.
- Hildebrandt, M. (2016). Law as Information in the Era of Data-Driven Agency. *The Modern Law Review*, 79(1), 1–30. <https://doi.org/10.1111/1468-2230.12165>
- Hines, M. (2004, March). *Google takes searching personally* [last accessed on 31.12.2022]. CNET. <https://www.cnet.com/news/google-takes-searching-personally/>
- Hirsch, T., Rothoefl, T., Teig, N., Bauer, J. W., Pellegrini, G., De Rosa, L., Scaglione, D., Reichelt, J., Klaussegger, A., Kneisz, D., Romano, O., Secone Seconetti, A., Contin, R., Enzo, E., Jurman, I., Carulli, S., Jacobsen, F., Luecke, T., Lehnhardt, M., ... De Luca, M. (2017). Regeneration of the entire human epidermis using transgenic stem cells. *Nature*, 551(7680), 327–332. <https://doi.org/10.1038/nature24487>
- HLEG on AI. (2019). *Ethics Guidelines for Trustworthy AI* [online: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>; last accessed on 07.06.2023]. High-Level Expert Group on Artificial Intelligence; European Commission.
- Hoffmann, H., Vogt, V., Hauer, M. P., & Zweig, K. (2022). Fairness by awareness? On the inclusion of protected features in algorithmic decisions. *Computer Law & Security Review*, 44, 105658. <https://doi.org/https://doi.org/10.1016/j.clsr.2022.105658>
- Horchert, J. (2017). *Googeln für die Wissenschaft* [last accessed on 31.12.2022]. Spiegel Netzwerk. <https://www.spiegel.de/netzwelt/web/datenspende-forschungsprojekt-ueber-google-sucht-freiwillige-helfer-a-1156060.html>

- Hornung, G., & Engemann, C. (2016). *Der digitale Bürger und seine Identität*. Nomos.
- Houston, T. K., & Allison, J. J. (2002). Users of Internet health information: differences by health status. *Journal of medical Internet research*, 4(2), E7. <https://doi.org/10.2196/jmir.4.2.e7>
- Howden, W. E. (1978). Theoretical and Empirical Studies of Program Testing. *IEEE Transactions on Software Engineering*, SE-4(4), 293–298. <https://doi.org/10.1109/TSE.1978.231514>
- Hunter, M. C. W., et al. (1989). *Establishing the new science: the experience of the early Royal Society*. Boydell & Brewer Ltd.
- IEEE Std 1028. (2008). *IEEE Standard for Software Reviews and Audits* (Standard). IEEE. New York. <https://doi.org/10.1109/IEEESTD.2008.4601584>
- Illyes, G. (2017). *Exclusive Q&A with Google's Gary Illyes at BrightonSEO 2017* [last accessed on 31.12.2022]. Youtube Video; Brighton SEO 2017. <https://www.youtube.com/watch?v=QJMZILz7WyU>
- Imana, B., Korolova, A., & Heidemann, J. (2021). Auditing for Discrimination in Algorithms Delivering Job Ads. *Proceedings of the Web Conference 2021*, 3767–3778. <https://doi.org/10.1145/3442381.3450077>
- Ingold, D., & Soper, S. (2016). *Amazon Doesn't Consider the Race of its Customers. Should it?* [last accessed on 28.03.2023]. Bloomberg. <https://www.bloomberg.com/graphics/2016-amazon-same-day/>
- Introna, L. D. (2015). Algorithms, Governance, and Governmentality: On Governing Academic Writing. *Science, Technology, & Human Values*, 41(1), 17–49. <https://doi.org/10.1177/0162243915587360>
- Iooss, B., & Saltelli, A. (2017). Introduction to Sensitivity Analysis. In R. Ghanem, D. Higdon, & H. Owhadi (Eds.), *Handbook of Uncertainty Quantification* (pp. 1103–1122). Springer International Publishing. [https://doi.org/10.1007/978-3-319-12385-1\\_31](https://doi.org/10.1007/978-3-319-12385-1_31)
- ISO 19011: (2018, October). *Guidelines for auditing management systems* (Standard). Beuth Verlag. Berlin.
- ISO 21500: (2012). *Guidance on project management* (Standard). Beuth Verlag. Berlin.
- ISO/IEC 17065: (2012, December). *Conformity assessment - Requirements for bodies certifying products, processes and services* (Standard). Beuth Verlag. Berlin.
- Zweig, K. A., Fischer, S., & Lischka, K. (2018). *Wo Maschinen irren können - Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung* (Arbeitspapier). Bertelsmann Stiftung. Gütersloh. <https://doi.org/10.11586/2018006>
- ISO/IEC 2382: (2015). *Information technology — Vocabulary* (Standard). Beuth Verlag.
- ISO/IEC/IEEE 24765: (2017). *Systems and software engineering — Vocabulary* (Standard). Beuth Verlag.
- ISSCR. (2023). *How to Report False Marketing Claims and Adverse Events from Clinics Offering Unapproved Stem Cell “Therapies”* [last accessed on 08.06.2023]. <https://www.isscr.org/wp-content/uploads/2023/06/ISSCR-2023-How-to-Report-False-Marketing-Claims-and-Adverse-Events-from-Clinics-offering-Unapproved-Stem-Cell-Therapies.pdf>

- [//www.closerlookatstemcells.org/patient-resources/how-to-report-false-marketing-claims-and-adverse-events-from-clinics-offering-unapproved-stem-cell-therapies](http://www.closerlookatstemcells.org/patient-resources/how-to-report-false-marketing-claims-and-adverse-events-from-clinics-offering-unapproved-stem-cell-therapies)
- Iyer, G. R., Miyazaki, A. D., Grewal, D., & Giordano, M. (2002). Linking Web-based segmentation to pricing tactics. *Journal of Product & Brand Management*, 11(5), 288–302. <https://doi.org/10.1108/10610420210442175>
- Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs [Formal Methods for Information Retrieval]. *Information Processing & Management*, 42(1), 248–263. <https://doi.org/10.1016/j.ipm.2004.10.007>
- Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4), 305–360. [https://doi.org/10.1016/0304-405X\(76\)90026-X](https://doi.org/10.1016/0304-405X(76)90026-X)
- Jentzsch, N. (2017). *Wohlfahrts- und Verteilungswirkungen personalisierter Preise und Produkte*. Friedrich-Ebert-Stiftung, Abteilung Wirtschafts- und Sozialpolitik, WISO Diskurs. <https://library.fes.de/pdf-files/wiso/13457-20170704.pdf>
- Jowell, R., & Prescott-Clarke, P. (1970). Racial discrimination and white-collar workers in Britain. *Race*, 11(4), 397–417. <https://doi.org/10.1177/030639687001100401>
- Just, A. (2019). *Profilbildungstheorien und ein praktischer Ansatz für Reverses A/B-Testing zur Analyse von Personalisierten und Dynamischen Preisen in Online-Shops* [Bachelor's Thesis]. Technische Universität Kaiserslautern, Department of Computer Science.
- Just, N., & Latzer, M. (2017). Governance by algorithms: reality construction by algorithmic selection on the Internet. *Media, Culture & Society*, 39(2), 238–258. <https://doi.org/10.1177/0163443716643157>
- Kanji, G. K. (2006). *100 statistical tests*. Sage.
- Karst, M. (1992). *Methodische Entwicklung von Expertensystemen*. Deutscher Universitätsverlag Wiesbaden.
- Kaul, S., Yates, R., & Gruteser, M. (2012). Real-time status: How often should one update? *Proceeding, 31st Annual IEEE International Conference on Computer Communications*, 2731–2735. <https://doi.org/10.1109/INFCOM.2012.6195689>
- Keane, M. T., O'Brien, M., & Smyth, B. (2008). Are People Biased in Their Use of Search Engines? *Commun. ACM*, 51(2), 49–52. <https://doi.org/10.1145/1314215.1314224>
- Kewley, R., Embrechts, M., & Breneman, C. (2000). Data strip mining for the virtual design of pharmaceuticals with neural networks. *IEEE Transactions on Neural Networks*, 11(3), 668–679. <https://doi.org/10.1109/72.846738>
- Kirchgaessner, S. (2017). *Cambridge Analytica used data from Facebook and Politico to help Trump* [last accessed on 08.06.2023]. The Guardian. <https://www.theguardian.com/technology/2017/oct/26/cambridge-analytica-used-data-from-facebook-and-politico-to-help-trump>

- Kitchin, R. (2016). Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1), 14–29. <https://doi.org/10.1080/1369118X.2016.1154087>
- Klein, R., & Steinhardt, C. (2008). *Revenue Management: Grundlagen und Mathematische Methoden*. Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-540-68845-7\\_2](https://doi.org/10.1007/978-3-540-68845-7_2)
- Kliman-Silver, C., Hannák, A., Lazer, D., Wilson, C., & Mislove, A. (2015). Location, Location, Location: The Impact of Geolocation on Web Search Personalization. *Proceedings of the 2015 Internet Measurement Conference*, 121–127. <https://doi.org/10.1145/2815675.2815714>
- Knapton, S. (2014). *First stem-cell therapy approved for medical use in Europe* [last accessed on 08.06.2023]. The Telegraph. <https://www.telegraph.co.uk/news/science/science-news/11304926/First-stem-cell-therapy-approved-for-medical-use-in-Europe.html>
- Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). *A governance framework for algorithmic accountability and transparency*. European Parliament and Directorate-General for Parliamentary Research Services. <https://doi.org/10.2861/59990>
- Kohavi, R., & Longbotham, R. (2017). Online Controlled Experiments and A/B Testing. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 922–929). Springer US. [https://doi.org/10.1007/978-1-4899-7687-1\\_891](https://doi.org/10.1007/978-1-4899-7687-1_891)
- König, D., Pascal, & Krafft, T. D. (2021). Evaluating the evidence in algorithmic evidence-based decision-making: the case of US pretrial risk assessment tools. *Current Issues in Criminal Justice*, 33(3), 359–381. <https://doi.org/10.1080/10345329.2020.1849932>
- Koops, B.-J. (2013). On decision transparency, or how to enhance data protection after the computational turn. In M. Hildebrandt & K. de Vries (Eds.), *Privacy, due process and the computational turn* (pp. 196–220). Routledge.
- Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., Ahamat, G., Leutner, F., Goebel, R., Knight, A., et al. (2021). Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms. *SSRN*. <https://doi.org/10.2139/ssrn.3778998>
- Krafft, R. (2018). *Entwicklung einer Plattform zur Erstellung und Überwachung von partizipativen Studien zur Messung von Preisdiskriminierung im Online-Handel* [Bachelor's Thesis]. Technische Universität Kaiserslautern, Department of Computer Science.
- Krafft, T. D., Gamer, M., Marcel, L., & Zweig, K. A. (2017). *Filterblase geplatzt? Kaum Raum für Personalisierung bei Google-Suchen zur Bundestagswahl 2017*. [https://www.blm.de/files/pdf/1\\_zwischenbericht\\_\\_final.pdf](https://www.blm.de/files/pdf/1_zwischenbericht__final.pdf)



- Krafft, T. D., Gamer, M., & Zweig, K. A. (2018a). *Wer sieht was? Personalisierung, Regionalisierung und die Frage nach der Filterblase in Googles Suchmaschine* (tech. rep.). <https://www.blm.de/files/pdf2/bericht-datenspende---wer-sieht-was-auf-google.pdf>
- Krafft, T. D., Gamer, M., & Zweig, K. A. (2018b). What did you see? Personalization, regionalization and the question of the filter bubble in Google’s search engine. *arXiv*. <https://doi.org/10.48550/arXiv.1812.10943>
- Krafft, T. D., Gamer, M., & Zweig, K. A. (2018c, February). *Personalisierung auf Googles Nachrichtenportal während der Bundestagswahl 2017* (tech. rep.). Algorithm Watch. <https://doi.org/10.13140/RG.2.2.29139.07203>
- Krafft, T. D., Gamer, M., & Zweig, K. A. (2019). What did you see? A study to measure personalization in Google’s search engine. *EPJ Data Science*, 8(1), 38. <https://doi.org/10.1140/epjds/s13688-019-0217-5>
- Krafft, T. D., Hauer, M. P., & Zweig, K. A. (2020). Why Do We Need to Be Bots? What Prevents Society from Detecting Biases in Recommendation Systems. In L. Boratto, S. Faralli, M. Marras, & G. Stilo (Eds.), *Bias and Social Aspects in Search and Recommendation* (pp. 27–34). Springer International Publishing. [https://doi.org/10.1007/978-3-030-52485-2\\_3](https://doi.org/10.1007/978-3-030-52485-2_3)
- Krafft, T. D., Krafft, R., Wölki, M., Rahe, M., & Zweig, K. A. (2023). *Algorithmische Governance von personalisierten Preisen im Online-Handel* (tech. rep.). Ministerium für Familie, Frauen, Kultur und Integration des Landes Rheinland-Pfalz, in publication.
- Krafft, T. D., Reber, M., Krafft, R., Coutrier, A., & Zweig, K. A. (2021). Crucial Challenges in Large-Scale Black Box Analyses. In L. Boratto, S. Faralli, M. Marras, & G. Stilo (Eds.), *Advances in Bias and Fairness in Information Retrieval* (pp. 143–155). Springer International Publishing. [https://doi.org/10.1007/978-3-030-78818-6\\_13](https://doi.org/10.1007/978-3-030-78818-6_13)
- Krafft, T. D., & Zweig, K. A. (2017). Ein Faktencheck - Ließ ein Algorithmus Trump triumphieren? *Informatik-Spektrum*, 40(4), 336–344. <https://doi.org/10.1007/s00287-017-1052-3>
- Krafft, T. D., & Zweig, K. A. (2018). Wie Gesellschaft algorithmischen Entscheidungen auf den Zahn fühlen kann. In R. Mohabbat Kar, B. E. P. Thapa, & P. Parycek (Eds.), *(Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft* (pp. 471–492). Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS, Kompetenzzentrum Öffentliche IT (ÖFIT).
- Krafft, T. D., & Zweig, K. A. (2019). Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse | Ein Regulierungsvorschlag [Verbraucherzentrale Bundesverband. Online verfügbar unter [https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22\\_zweig\\_krafft\\_transparenz\\_adm-n eu.pdf](https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-n eu.pdf); aufgerufen am 31.12.2022].

- Krafft, T. D., & Zweig, K. A. (2020). Ethische Herausforderungen digitalen Wandels in bewaffneten Konflikten Herausforderungen bei der Nutzung von KI in militärischen Anwendungsgebieten - Sozioinformatische Perspektive. In M. Rogg, S. Scheidt, & H. von Schubert (Eds.). German Institute for Defence; Strategic Studies.
- Krafft, T. D., Zweig, K. A., & König, P. D. (2022). How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. *Regulation & Governance*, 16(1), 119–136. <https://doi.org/10.1111/rego.12369>
- Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24), 8788–8790. <https://doi.org/10.1073/pnas.1320040111>
- Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and psychological measurement*, 30(3), 607–610. <https://doi.org/10.1177/001316447003000308>
- Kroll, J. A., Joanna, H., Solon, B., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Harlan, Y. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165(3), 633–705. [https://scholarship.law.upenn.edu/penn\\_law\\_review/vol165/iss3/3](https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3)
- Kucklick, C. (2014). *Die granulare Gesellschaft: Wie das Digitale unsere Wirklichkeit auflöst*. Ullstein Ebooks.
- Lahey, J., & Beasley, R. (2018). Technical Aspects of Correspondence Studies. In S. M. Gaddis (Ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance* (pp. 81–101). Springer International Publishing. [https://doi.org/10.1007/978-3-319-71153-9\\_4](https://doi.org/10.1007/978-3-319-71153-9_4)
- Lander, E., & Nelson, A. (2011). *Americans Need a Bill of Rights for an AI-Powered World* [last accessed on 29.05.2023]. WIRED. <https://www.wired.com/story/opinion-bill-of-rights-artificial-intelligence/>
- Langstrup, H. (2011). Interpellating Patients as Users: Patient Associations and the Project-Ness of Stem Cell Research. *Science, Technology, & Human Values*, 36(4), 573–594. <https://doi.org/10.1177/0162243910368397>
- Lanzing, M. (2019). “Strongly Recommended” Revisiting Decisional Privacy to Judge Hypernudging in Self-Tracking Technologies. *Philosophy & Technology*, 32(3), 549–568. <https://doi.org/10.1007/s13347-018-0316-4>
- Latour, B. (1994). On Technical Mediation. *Common Knowledge*, 3(2), 29–64.
- Lazarsfeld, P. F., Merton, R. K., et al. (1954). Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 18(1), 18–66.
- Leite, J. C. S. d. P., & Cappelli, C. (2010). Software Transparency. *Business & Information Systems Engineering*, 2(3), 127–139. <https://doi.org/10.1007/s12599-010-0102-z>

- Leonhardt, D. (2005). *Why Variable Pricing Fails at the Vending Machine* [last accessed on 08.06.2023]. The New York Times. <https://www.nytimes.com/2005/06/27/business/why-variable-pricing-fails-at-the-vending-machine.html>
- Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology*, 31, 611–627. <https://doi.org/10.1007/s13347-017-0279-x>
- Levinson, R. M. (1975). Sex Discrimination and Employment Practices: An Experiment with Unconventional Job Inquiries. *Social Problems*, 22(4), 533–543. <https://doi.org/10.2307/799750>
- Li, J., Theng, Y.-L., & Foo, S. (2016). Predictors of online health information seeking behavior: Changes between 2002 and 2012. *Health informatics journal*, 22(4), 804–814. <https://doi.org/10.1177/1460458215595851>
- Lin, W., & Green, D. P. (2016). Standard Operating Procedures: A Safety Net for Pre-Analysis Plans. *PS: Political Science and Politics*, 49(3), 495–500. <https://doi.org/10.1017/S1049096516000810>
- Liu, F., Yu, C., & Meng, W. (2004). Personalized Web search for improving retrieval effectiveness. *IEEE Transactions on Knowledge and Data Engineering*, 16(1), 28–40. <https://doi.org/10.1109/TKDE.2004.1264820>
- Lü, L., Medo, M., Yeung, C. H., Zhang, Y.-C., Zhang, Z.-K., & Zhou, T. (2012). Recommender systems [Recommender Systems]. *Physics Reports*, 519(1), 1–49. <https://doi.org/10.1016/j.physrep.2012.02.006>
- Lucas, P., & van der Gaag, L. (1991). *Principles of Expert Systems*. Addison-Wesley Longman Publishing Co., Inc.
- Lysaght, T., Kerridge, I. H., Sipp, D., Porter, G., & Capps, B. J. (2017). Ethical and Regulatory Challenges with Autologous Adult Stem Cells: A Comparative Review of International Regulations. *Journal of Bioethical Inquiry*, 14(2), 261–273. <https://doi.org/10.1007/s11673-017-9776-y>
- Lysaght, T., Lipworth, W., Hendl, T., Kerridge, I., Lee, T.-L., Munsie, M., Waldby, C., & Stewart, C. (2017). The deadly business of an unregulated global stem cell industry. *Journal of medical ethics*, 43(11), 744–746. <https://doi.org/10.1136/mjedethics-2016-104046>
- Maslow, A. H. (1943). A theory of human motivation. *Originally Published in Psychological Review*, 50(4), 370.
- Master, Z., Robertson, K., Frederick, D., Rachul, C., & Caulfield, T. (2014). Stem Cell Tourism and Public Education: The Missing Elements. *Cell stem cell*, 15(3), 267–270. <https://doi.org/10.1016/j.stem.2014.08.009>
- Matthijs, N., & Radlinski, F. (2011). Personalizing Web Search Using Long Term Browsing History. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 25–34. <https://doi.org/10.1145/1935826.1935840>

- Mayer, J. R., & Mitchell, J. C. (2012). Third-Party Web Tracking: Policy and Technology. *2012 IEEE Symposium on Security and Privacy*, 413–427. <https://doi.org/10.1109/SP.2012.47>
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: die Revolution, die unser Leben verändern wird*. Redline.
- McAfee, R. P., & Te Velde, V. (2006). Dynamic pricing in the airline industry. In B. Sulim, D. Wenjing, D. Xianjun, G. Alok, R. H.R., S. Raghu T., & Z. Han (Eds.), *Economics and Information* (pp. 527–569, Vol. 1). Elsevier.
- McCown, F., & Nelson, M. L. (2007). Agreeing to Disagree: Search Engines and Their Public Interfaces. *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, 309–318. <https://doi.org/10.1145/1255175.1255237>
- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27, 415–444. <https://doi.org/10.1146/annurev.soc.27.1.415>
- Mendick, R., & Hall, A. (2011). *Europe's largest stem cell clinic shut down after death of baby* [last accessed on 08.06.2023]. The Telegraph. <https://www.telegraph.co.uk/news/worldnews/europe/germany/8500233/Europes-largest-stem-cell-clinic-shut-down-after-death-of-baby.html>
- Mertens, P., Borkowski, V., & Geis, W. (1988). *Betriebliche Expertensystem-Anwendungen: Eine Materialsammlung*. Springer.
- Messick, S. (1987). VALIDITY. *ETS Research Report Series*, 1987(2), i–208. <https://doi.org/10.1002/j.2330-8516.1987.tb00244.x>
- Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., & Sandvig, C. (2021). Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human-Computer Interaction*, 14(4), 272–344. <https://doi.org/10.1561/11000000083>
- Mikians, J., Gyarmati, L., Erramilli, V., & Laoutaris, N. (2012). Detecting Price and Search Discrimination on the Internet. *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*, 79–84. <https://doi.org/10.1145/2390231.2390245>
- Mikians, J., Gyarmati, L., Erramilli, V., & Laoutaris, N. (2013). Crowd-Assisted Search for Price Discrimination in e-Commerce: First Results. *Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies*, 1–6. <https://doi.org/10.1145/2535372.2535415>
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2). <https://doi.org/10.1177/2053951716679679>
- Moe, H., & Syvertsen, T. (2007). Media institutions as a research field: Three Phases of Norwegian Broadcasting Research. *Nordicom Review*, 28, 149–167. <https://hdl.handle.net/1956/2332>
- Monopolkommission. (2018). Algorithms and collusion [last accessed on 09.06.2023]. In *Biennial Report of the Monopolies Commission, XXII* (pp. 62–88). <https://www.monopolkommission.de>

- w.monopolkommission.de/images/HG22/Main\_Report\_XXII\_Algorithms\_and\_Collusion.pdf
- Moorstedt, M. (2018). *Youtubes Lügenalgorithmus* [last accessed on 31.12.2022]. Süddeutsche Zeitung. <https://www.sueddeutsche.de/digital/netzkolumne-youtubes-luegenalgorithmus-1.3853777>
- Moz. (2023). *Google Algorithm Update History* [last accessed on 31.12.2022]. <https://moz.com/google-algorithm-change>
- Munsie, M., Lysaght, T., Hendl, T., Tan, H.-Y. L., Kerridge, I., & Stewart, C. (2017). Open for business: a comparative study of websites selling autologous stem cells in Australia and Japan [PMID: 29125016]. *Regenerative Medicine*, 12(7), 777–790. <https://doi.org/10.2217/rme-2017-0070>
- Murnane, T., Reed, K., & Hall, R. (2007). On the Learnability of Two Representations of Equivalence Partitioning and Boundary Value Analysis. *2007 Australian Software Engineering Conference (ASWEC'07)*, 274–283. <https://doi.org/10.1109/ASWEC.2007.35>
- Muthukrishnan, S. (2009). Bidding on Configurations in Internet Ad Auctions. In H. Q. Ngo (Ed.), *Computing and Combinatorics* (pp. 1–6). Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-02882-3\\_1](https://doi.org/10.1007/978-3-642-02882-3_1)
- Myers, G. J., Sandler, C., & Badgett, T. (2011). *The Art of Software Testing*. John Wiley & Sons.
- Nadig, R. R. (2009). Stem cell therapy—Hype or hope? A review. *Journal of conservative dentistry: JCD*, 12(4), 131–138. <https://doi.org/10.4103/0972-0707.58329>
- Namiki, Y., Ishida, T., & Akiyama, Y. (2013). Acceleration of sequence clustering using longest common subsequence filtering. *BMC Bioinformatics*, 14(S7). <https://doi.org/10.1186/1471-2105-14-S8-S7>
- Neave, H. R. (1976). The Teaching of hypothesis-testing. *Journal of Applied Statistics*, 3(1), 55–63. <https://doi.org/10.1080/768371017>
- NHS. (2023). *Overview - Parkinson's disease* [last accessed on 08.06.2023]. <https://www.nhs.uk/conditions/parkinsons-disease/>
- Nidhra, S., & Dondeti, J. (2012). Black box and white box testing techniques - a literature review. *International Journal of Embedded Systems and Applications (IJESA)*, 2(2), 29–50. <https://doi.org/10.5121/ijesa.2012.2204>
- Niedermayer, J., Züfle, A., Emrich, T., Renz, M., Mamoulis, N., Chen, L., & Kriegel, H.-P. (2013). Similarity Search on Uncertain Spatio-temporal Data. In N. Brisaboa, O. Pedreira, & P. Zezula (Eds.), *Similarity Search and Applications* (pp. 43–49). Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-41062-8\\_5](https://doi.org/10.1007/978-3-642-41062-8_5)
- Niklas, J., Sztandar-Sztanderska, K., & Szymielewicz, K. (2015). *Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic Decision Making*. Poland: Panoptikon Foundation.
- Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

- O'Donnell, L., Turner, L., & Levine, A. D. (2016). Part 6: The role of communication in better understanding unproven cellular therapies. *Cytotherapy*, 18(1), 143–148. <https://doi.org/10.1016/j.jcyt.2015.11.002>
- Olken, B. A. (2015). Promises and Perils of Pre-analysis Plans. *Journal of Economic Perspectives*, 29(3), 61–80. <https://doi.org/10.1257/jep.29.3.61>
- O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
- Orwat, Bareis, Folberth, Jahnel, & Wadehul. (2022). In T. Hoeren & S. Pinelli (Eds.), *Künstliche Intelligenz – Ethik und Recht*. C.H.Beck. <https://doi.org/10.5771/9783748929680-255>
- Osmani, A., & Grigorik, I. (2018). *Speed is now a landing page factor for Google Search and Ads* [last accessed on 08.06.2023]. Google. <https://developers.google.com/web/updates/2018/07/search-ads-speed>
- Pagano, G., Ferrara, N., Brooks, D. J., & Pavese, N. (2016). Age at onset and Parkinson disease phenotype. *Neurology*, 86(15), 1400–1407. <https://doi.org/10.1212/WNL.0000000000002461>
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1998, November). *The PageRank Citation Ranking: Bringing Order to the Web*. (Technical Report No. 66) (Previous number = SIDL-WP-1999-0120). Stanford InfoLab. Stanford InfoLab. <http://ilpubs.stanford.edu:8090/422/>
- Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google We Trust: Users' Decisions on Rank, Position, and Relevance. *Journal of Computer-Mediated Communication*, 12(3), 801–823. <https://doi.org/10.1111/j.1083-6101.2007.00351.x>
- Pansari, A., & Mayer, M. (2006). *This is a test. This is only a test*. [last accessed on 16.06.2023]. Google Official Blog. <https://googleblog.blogspot.com/2006/04/this-is-test-this-is-only-test.html>
- Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.
- Park, J. S., & Sandhu, R. (2000). Secure cookies on the Web. *IEEE Internet Computing*, 4(4), 36–44. <https://doi.org/10.1109/4236.865085>
- Pasquale, F. (2015). *The Black box society: the secret algorithms that control money and information*. Harvard University Press.
- Patani, R., & Chandran, S. (2012). Experimental and Therapeutic Opportunities for Stem Cells in Multiple Sclerosis. *International Journal of Molecular Sciences*, 13(11), 14470–14491. <https://doi.org/10.3390/ijms131114470>
- Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Pappalardo, L., Ruggieri, S., & Turini, F. (2018). Open the Black Box Data-Driven Explanation of Black Box Decision Systems. *CoRR*, abs/1806.09936. <http://arxiv.org/abs/1806.09936>
- Petersen, A., Munsie, M., Tanner, C., MacGregor, C., & Brophy, J. (2017). *Stem Cell Tourism and the Political Economy of Hope*. Palgrave Macmillan UK. <https://doi.org/10.1057/978-1-137-47043-0>

- Petersen, A., Seear, K., & Munsie, M. (2014). Therapeutic journeys: the hopeful travails of stem cell tourists. *Sociology of Health & Illness*, 36(5), 670–685. <https://doi.org/10.1111/1467-9566.12092>
- Petersen, A., Tanner, C., & Munsie, M. (2019). Citizens' use of digital media to connect with health care: Socio-ethical and regulatory implications. *Health (London, England : 1997)*, 23(4), 367–384. <https://doi.org/10.1177/1363459319847505>
- Petitti, D. B. (2000). *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*. Oxford University Press.
- Pigou, A. (1920). *The economics of welfare*. Routledge.
- Popescu, I., & Wu, Y. (2007). Dynamic Pricing Strategies with Reference Effects. *Operations Research*, 55(3), 413–429. <https://doi.org/10.1287/opre.1070.0393>
- Popper, K. (2005). *The logic of scientific discovery*. Routledge.
- Portugal, R. L. Q., Engiel, P., Roque, H., & do Prado Leite, J. C. S. (2017). Is There a Demand of Software Transparency? *Proceedings of the XXXI Brazilian Symposium on Software Engineering*, 204–213. <https://doi.org/10.1145/3131151.3131155>
- Puppe, F. (1988). *Einführung in Expertensysteme*. Studienreihe Informatik, Springer-Verlag.
- Raji, I. D., & Buolamwini, J. (2019). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435. <https://doi.org/10.1145/3306618.3314244>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing, 33–44. <https://doi.org/10.1145/3351095.3372873>
- Rashtchy, S., Kessler, A. M., Bieber, P. J., Schindler, N. H., & Tzeng, J. C. (2007). *The User Revolution: The New Advertising Ecosystem And The Rise Of The Internet As A Mass Medium* [last accessed on 08.06.2023]. PiperJaffray, Investment Research. <https://web.archive.org/web/20100713195348/https://people.ischool.berkeley.edu/~hal/Courses/StratTech09/Lectures/Google/Articles/user-revolution.pdf>
- Reber, M. (2020). *Abusive Advertising: Scrutinizing socially relevant algorithms in a Black Box analysis to examine their impact on vulnerable patient groups in the health sector* [Master's Thesis]. Technische Universität Kaiserslautern, Department of Computer Science. <https://doi.org/10.48550/arXiv.2101.02018>
- Reber, M., Krafft, T. D., Krafft, R., Zweig, K. A., & Couturier, A. (2020). Data Donations for Mapping Risk in Google Search of Health Queries: A case study of unproven stem cell treatments in SEM. *Symposium Series on Computational Intelligence (SSCI)*, 2985–2992. <https://doi.org/10.1109/SSCI47803.2020.9308420>

- Regenberg, A. C., Hutchinson, L. A., Schanker, B., & Mathews, D. J. H. (2009). Medicine on the Fringe: Stem Cell-Based Interventions in Advance of Evidence. *Stem Cells*, 27(9), 2312–2319. <https://doi.org/10.1002/stem.132>
- Reichborn-Kjennerud, K., & Vabo, S. I. (2017). Performance audit as a contributor to change and improvement in public administration. *Evaluation*, 23(1), 6–23. <https://doi.org/10.1177/1356389016683871>
- Reinartz, W., Haucap, J., Wiegand, N., & Hunold, M. (2017). *Preisdifferenzierung und -dispersion im Handel* [last accessed on 09.06.2023]. IFH-Förderer. [https://marketing.uni-koeln.de/sites/marketingarea/user\\_upload/171130\\_Whitepaper\\_Preisdifferenzierung\\_und\\_-dispersion\\_im\\_Handel.pdf](https://marketing.uni-koeln.de/sites/marketingarea/user_upload/171130_Whitepaper_Preisdifferenzierung_und_-dispersion_im_Handel.pdf)
- Robertson, A. (2019). *Scraping public data from a website probably isn't hacking, says court* [last accessed on 08.06.2023]. The Verge. <https://www.theverge.com/2019/9/10/20859399/linkedin-hiq-data-scraping-cfaa-lawsuit-ninth-circuit-ruling>
- Robertson, R. E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., & Wilson, C. (2018). Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW, Article 148). <https://doi.org/10.1145/3274417>
- Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5), 582–638. <https://doi.org/10.1017/S0269888913000039>
- Rose, M., & Rahman, M. (2015). Who's Paying More to Tour These United States? Price Differences in International Travel Bookings. *Technology Science*, 16, 1–16. <https://techscience.org/a/2015081105/>
- RStV. (1991). *Staatsvertrag für Rundfunk und Telemedien (Rundfunkstaatsvertrag)* [last accessed on 09.06.2023]. State of Baden-Württemberg, Free State of Bavaria, State of Berlin, State of Brandenburg, Free Hanseatic City of Bremen, Free Hanseatic City of Hamburg, State of Hesse, State of Mecklenburg-Western Pomerania, State of Lower Saxony, State of North Rhine-Westphalia, State of Rhineland-Palatinate, Saarland, Free State of Saxony, State of Saxony-Anhalt, State of Schleswig-Holstein, Free State of Thuringia. [https://www.die-medienanstalten.de/fileadmin/user\\_upload/Rechtsgrundlagen/Gesetze\\_Staatsvertraege/RStV\\_22\\_nichtamtliche\\_Fassung\\_medienanstalten\\_final\\_web.pdf](https://www.die-medienanstalten.de/fileadmin/user_upload/Rechtsgrundlagen/Gesetze_Staatsvertraege/RStV_22_nichtamtliche_Fassung_medienanstalten_final_web.pdf)
- Ryan, K. A., Sanders, A. N., Wang, D. D., & Levine, A. D. (2010). Tracking the rise of stem cell tourism. *Regenerative medicine*, 5(1), 27–33. <https://doi.org/10.2217/rme.09.70>
- Saltelli, A. (2002). Sensitivity Analysis for Importance Assessment. *Risk Analysis*, 22(3), 579–590. <https://doi.org/10.1111/0272-4332.00040>
- Saltelli, A., Ratto, M., Tarantola, S., & Campolongo, F. (2005). Sensitivity Analysis for Chemical Models. *Chemical reviews*, 105(7), 2811–2828. <https://doi.org/10.1021/cr040659d>



- Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M., et al. (2004). *Sensitivity analysis in practice: a guide to assessing scientific models* (Vol. 1). John Wiley & Sons Ltd.
- Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22, 4349–4357.
- Saretsky, G. (1972). The OEO P.C. Experiment and the John Henry Effect. *The Phi Delta Kappan*, 53(9), 579–581. <http://www.jstor.org/stable/20373317>
- Saurwein, F., Just, N., & Latzer, M. (2015). Governance of algorithms: options and limitations. *info*, 19(6), 35–49. <https://doi.org/10.1108/info-05-2015-0025>
- Schiffman, L. G., Kanuk, L. L., & Hansen, H. (2012). *Consumer behaviour—A european outlook*. Harlow: Financial Times Prentice Hall.
- Schleusener, M. (2017). Personalisierte Preise im Handel – Chancen und Herausforderungen. In E. Stüber & K. Hudetz (Eds.), *Praxis der Personalisierung im Handel: Mit zeitgemäßen E-Commerce-Konzepten Umsatz und Kundenwert steigern* (pp. 71–89). Springer Fachmedien Wiesbaden. [https://doi.org/10.1007/978-3-658-16244-3\\_4](https://doi.org/10.1007/978-3-658-16244-3_4)
- Schleusener, M., & Hosell, S. (2015). *Expertise zum Thema „Personalisierte Preisdifferenzierung im Online-Handel“* [last accessed on 10.06.2023]. Berlin: SVRV. [http://www.svr-verbraucherfragen.de/wp-content/uploads/eWeb-Research-Center\\_Preisdifferenzierung-im-Online-%20handel.pdf](http://www.svr-verbraucherfragen.de/wp-content/uploads/eWeb-Research-Center_Preisdifferenzierung-im-Online-%20handel.pdf)
- Schultheiß, S., Sünkler, S., & Lewandowski, D. (2018). We still trust in Google, but less than 10 years ago: an eye-tracking study. *Information Research: An International Electronic Journal*, 23(3), n3.
- Schütte, J. F. (2019). *Steht unsere Gesellschaft auf »Autoplay«?* [Bachelor’s Thesis]. Technische Universität Kaiserslautern, Department of Computer Science.
- Seaver, N. (2019). Knowing Algorithms. In *A Field Guide for Science & Technology Studies* (pp. 412–422). Princeton University Press. <https://doi.org/10.1515/9780691190600-028>
- Shore, J., & Warden, S. (2021). *The art of agile development*. O’Reilly.
- Sieg, A., Mobasher, B., & Burke, R. (2007). Web Search Personalization with Ontological User Profiles. *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, 525–534. <https://doi.org/10.1145/1321440.1321515>
- Siegfried, S., Johannes, K., Gerlitsch, A., Lorscheid, E., & Philippi, M. (2016). *Medienkonvergenzmonitor der DLM; MedienVielfaltsMonitor Ergebnisse 1. Halbjahr 2016 - Anteile der Medienangebote und Medienkonzerne am Meinungsmarkt der Medien in Deutschland* [last accessed on 31.12.2022]. die medienanstalten. [http://www.die-medienanstalten.de/fileadmin/user\\_upload/die\\_medienanstalten/Forschung/Medienvielfaltsmonitor/MedienvielfaltsMonitor\\_2016-1.pdf](http://www.die-medienanstalten.de/fileadmin/user_upload/die_medienanstalten/Forschung/Medienvielfaltsmonitor/MedienvielfaltsMonitor_2016-1.pdf)

- Simon, H. A. (1954). Spurious Correlation: A Causal Interpretation\*. *Journal of the American Statistical Association*, 49(267), 467–479. <https://doi.org/10.1080/01621459.1954.10483515>
- Simon, H., & Fassnacht, M. (2019). *Price Management: Strategy, Analysis, Decision, Implementation*. Springer. <https://doi.org/10.1007/978-3-319-99456-7>
- Singhal, A. (2011). *Some thoughts on personalization* [last accessed on 20.07.2018]. Google Search blog. <https://search.googleblog.com/2011/11/some-thoughts-on-personalization.html>
- Sipp, D., Caulfield, T., Kaye, J., Barfoot, J., Blackburn, C., Chan, S., Luca, M. D., Kent, A., McCabe, C., Munsie, M., Sleeboom-Faulkner, M., Sugarman, J., van Zimmeren, E., Zarzeczny, A., & Rasko, J. E. J. (2017). Marketing of unproven stem cell-based interventions: A call to action. *Science Translational Medicine*, 9(397). <https://doi.org/10.1126/scitranslmed.aag0426>
- Sipp, D., & Turner, L. (2012). U.S. Regulation of Stem Cells as Medical Products. *Science*, 338(6112), 1296–1297. <https://doi.org/10.1126/science.1229918>
- Smith, A. (1776). *The Wealth of Nations: An inquiry into the nature and causes of the Wealth of Nations*. W. Strahan; T. Cadell, London.
- Sokol, K., & Flach, P. (2020). Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56–67. <https://doi.org/10.1145/3351095.3372870>
- Spreckelsen, C., & Spitzer, K. (2009). *Wissensbasen und Expertensysteme in der Medizin: KI-Ansätze zwischen klinischer Entscheidungsunterstützung und medizinischem Wissensmanagement*. Vieweg Teubner Verlag.
- Statista. (2023a). Anzahl der Suchanfragen bei Google weltweit in den Jahren 2000 bis 2016 [last accessed on 28.05.2023]. <https://de.statista.com/statistik/daten/studie/71769/umfrage/anzahl-der-google-suchanfragen-pro-jahr>
- Statista. (2023b). *Marktanteile der führenden Browserfamilien an der Internetnutzung in Deutschland von Januar 2009 bis Juni 2023* [last accessed on 16.06.2023]. <https://de.statista.com/statistik/daten/studie/13007/umfrage/marktanteile-der-browser-bei-der-internetnutzung-in-deutschland-seit-2009/>
- Stehle, A. (2016). *Personalisierte Preise alarmieren Justizministerium* [last accessed on 09.06.2023]. Handelsblatt. <https://www.handelsblatt.com/politik/deutschland/online-shopping-personalisierte-preise-alarmieren-justizministerium-/12966436.html>
- Stepanov, A. (2015). On the Kendall Correlation Coefficient [arXiv: 1507.01427]. <https://doi.org/10.48550/arXiv.1507.01427>
- Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology*, 21(1), 128–138. <https://doi.org/10.1097/EDE.0b013e3181c30fb2>

- Stole, L. A. (2007). Price Discrimination and Competition. In M. Armstrong & R. Porter (Eds.), *Handbook of Industrial Organization* (pp. 2221–2299, Vol. 3). Elsevier. [https://doi.org/10.1016/S1573-448X\(06\)03034-2](https://doi.org/10.1016/S1573-448X(06)03034-2)
- Strathern, M. (2000). *Audit cultures: Anthropological studies in accountability, ethics and the academy*. Taylor & Francis.
- Strauer, B. E., & Kornowski, R. (2003). Stem Cell Therapy in Perspective. *Circulation*, 107(7), 929–934. <https://doi.org/10.1161/01.CIR.0000057525.13182.24>
- Sunstein, C., & Chambers, E. (2001). *Echo Chambers - Bush v. Gore, Impeachment, and Beyond*. Princeton University Press.
- SVRV. (2018). *Verbrauchergerechtes Scoring - Gutachten* [last accessed on 30.05.2023]. Bundesministerium der Justiz und für Verbraucherschutz. [https://www.svr-verbraucherfragen.de/wp-content/uploads/SVRV\\_Verbrauchergerechtes\\_Scoring.pdf](https://www.svr-verbraucherfragen.de/wp-content/uploads/SVRV_Verbrauchergerechtes_Scoring.pdf)
- Sweeney, L. (2013). Discrimination in Online Ad Delivery. *Communications of the ACM*, 56(5), 44–54. <https://doi.org/10.1145/2447976.2447990>
- Takahashi, K., & Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, 126(4), 663–676. <https://doi.org/https://doi.org/10.1016/j.cell.2006.07.024>
- Tanner, C., Munsie, M., Sipp, D., Turner, L., & Wheatland, C. (2019). The politics of evidence in online illness narratives: An analysis of crowdfunding for purported stem cell treatments. *Health (London, England : 1997)*, 23(4), 436–457. <https://doi.org/10.1177/1363459319829194>
- Teevan, J., Karlson, A., Amini, S., Brush, A. J. B., & Krumm, J. (2011). Understanding the Importance of Location, Time, and People in Mobile Local Search Behavior. *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, 77–80. <https://doi.org/10.1145/2037373.2037386>
- Teevan, J., Morris, M. R., & Bush, S. (2009). Discovering and Using Groups to Improve Personalized Search. *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, 15–24. <https://doi.org/10.1145/1498759.1498786>
- Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, 17(1), 168–192. <https://doi.org/10.1016/j.aci.2018.08.003>
- Thorun, C., & Diels, J. (2016). *Was Verbraucherinnen und Verbraucher in NRW über individualisierte Preise im Online-Handel denken* [last accessed on 30.05.2023]. Conpolicy; Aktenzeichen: I-4-2.1-15/085. <https://docplayer.org/54455832-Was-verbraucherinnen-und-verbraucher-in-nrw-ueber-individualisierte-preise-im-online-handel-denken.html>
- Tirole, J. (1989). *The Theory of Industrial Organization*. MIT Press.
- Trielli, D., Mussenden, S., & Diakopoulos, N. (2015, December). *Why Google Search Results Favor Democrats* [last accessed on 31.12.2022]. SLATE. <https://slate.com/technology/2015/12/why-google-search-results-favor-democrats.html>

- Turner, L. (2017). ClinicalTrials.gov, stem cells and 'pay-to-participate' clinical studies. *Regenerative medicine*, 12(6), 705–719. <https://doi.org/10.2217/rme-2017-0015>
- Turner, L. (2018). The US Direct-to-Consumer Marketplace for Autologous Stem Cell Interventions. *Perspectives in biology and medicine*, 61(1), 7–24. <https://doi.org/10.1353/pbm.2018.0024>
- Ulbricht, L., & Yeung, K. (2020). Algorithmic regulation: A maturing concept for investigating regulation of and through algorithms. *Regulation & Governance*, 16(1), 3–22. <https://doi.org/10.1111/rego.12437>
- United States House of Representatives. (2022). *Algorithmic Accountability Act of 2022, H.R. 6580, 117th Cong.* [last accessed on 09.06.2023]. U.S. Government Publishing Office (GPO). <https://www.congress.gov/bill/117th-congress/house-bill/6580>
- van Drunen, M. Z., Helberger, N., & Bastian, M. (2019). Know your algorithm: what media organizations need to explain to their users about news personalization. *International Data Privacy Law*, 9(4), 220–235. <https://doi.org/10.1093/idpl/ipy011>
- Vater, C. (2020). Turings Maschine und Blacks Box – Mechanische Intelligenz nach dem Feedback. In E. Geitz, C. Vater, & S. Zimmer-Merkle (Eds.), *Interdisziplinäre Perspektiven* (pp. 323–350). De Gruyter. <https://doi.org/doi:10.1515/9783110701319-017>
- Vecchione, B., Levy, K., & Barocas, S. (2021). Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies. <https://doi.org/10.1145/3465416.3483294>
- Verbrugge, L. M. (1977). The structure of adult friendship choices. *Social forces*, 56(2), 576–597. <https://doi.org/10.2307/2577741>
- Vigen, T. (2015). *Spurious correlations*. Hachette.
- Vinoth, G. (2017). *Google Algorithm Updates Explained* [last accessed on 31.12.2022]. HackerNoon. <https://hackernoon.com/google-algorithm-updates-explained-f4a4640154ea>
- Visser, T., Nikiforakis, N., Bielova, N., & Joosen, W. (2014). Crying wolf? on the price discrimination of online airline tickets. *7th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2014)*. <https://inria.hal.science/hal-01081034>
- Watt, J., Borhani, R., & Katsaggelos, A. K. (2020). *Machine Learning Refined: Foundations, Algorithms, and Applications*. Cambridge University Press.
- Weare, C. B. (2006, October). System and method for personalized search [US Patent US7599916B2, 26.10.2006]. <https://patents.google.com/patent/US7599916B2>
- Webber, W., Moffat, A., & Zobel, J. (2010). A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.*, 28(4). <https://doi.org/10.1145/1852102.1852106>
- Weber, H. (2017). Blackboxing? – Zur Vermittlung von Konsumtechniken über Gehäuse- und Schnittstellendesign. In *Gehäuse: Mediale Einkapselungen* (pp. 115–136). Brill Fink. [https://doi.org/10.30965/9783846760192\\_007](https://doi.org/10.30965/9783846760192_007)

- Weber, I., & Jaimes, A. (2011). Who Uses Web Search for What: And How. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 15–24. <https://doi.org/10.1145/1935826.1935839>
- Weisberg, J. (2011, June). *Bubble Trouble* [last accessed on 31.12.2022]. SLATE. [http://www.slate.com/articles/news\\_and\\_politics/the\\_big\\_idea/2011/06/bubble\\_trouble.html](http://www.slate.com/articles/news_and_politics/the_big_idea/2011/06/bubble_trouble.html)
- Wessels, D. (2001). *Web caching*. O'Reilly.
- West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, 27(11), 1131–1152. [https://doi.org/https://doi.org/10.1016/S0305-0548\(99\)00149-5](https://doi.org/https://doi.org/10.1016/S0305-0548(99)00149-5)
- White, R. W., Bailey, P., & Chen, L. (2009). Predicting User Interests from Contextual Information. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 363–370. <https://doi.org/10.1145/1571941.1572005>
- White House. (2015). *Big Data and Differential Pricing* [last accessed on 31.12.2022]. [https://obamawhitehouse.archives.gov/sites/default/files/whitehouse\\_files/docs/Big\\_Data\\_Report\\_Nonembargo\\_v2.pdf](https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/docs/Big_Data_Report_Nonembargo_v2.pdf)
- Whitten, A. (2012, February). *Google's new Privacy Policy* [last accessed on 16.06.2023]. Google Search blog. <https://googleblog.blogspot.com/2012/02/googles-new-privacy-policy.html>
- Wiedmann, K.-P., Buxel, H., & Walsh, G. (2002). Customer profiling in e-commerce: Methodological aspects and challenges. *Journal of Database Marketing & Customer Strategy Management*, 9(2), 170–184. <https://doi.org/10.1057/palgrave.jdm.3240073>
- Wiener, N. (1948). *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press.
- Wieringa, M. (2020). What to Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 1–18. <https://doi.org/10.1145/3351095.3372833>
- Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., & Polli, F. (2021). Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666–677. <https://doi.org/10.1145/3442188.3445928>
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques;3rd ed.* Elsevier.
- WMA. (2022). *Declaration of Helsinki - Ethical Principles for Medical* [last accessed on 31.12.2022]. World Medical Association. <https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/>

- Wölki, M. (2018). *Entwicklung einer crowd-basierten Experimentalplattform zur automatisierten Preisfilterung von Studien für die Untersuchung von personalisierten Preisen im Online-Handel* [Bachelor's Thesis]. Technische Universität Kaiserslautern, Department of Computer Science.
- Wormith, J. S. (2017). Automated Offender Risk Assessment. *Criminology & Public Policy*, 16(1), 281–303. <https://doi.org/10.1111/1745-9133.12277>
- Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In J. Tang, M.-Y. Kan, D. Zhao, S. Li, & H. Zan (Eds.), *Natural Language Processing and Chinese Computing* (pp. 563–574). Springer International Publishing. [https://doi.org/10.1007/978-3-030-32236-6\\_51](https://doi.org/10.1007/978-3-030-32236-6_51)
- Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y., & Chen, Z. (2009). How Much Can Behavioral Targeting Help Online Advertising? *Proceedings of the 18th International Conference on World Wide Web*, 261–270. <https://doi.org/10.1145/1526709.1526745>
- Yau, N. (2011). *Visualize this: the FlowingData guide to design, visualization, and statistics*. John Wiley & Sons.
- Yeung, K. (2017a). ‘Hypernudge’: Big Data as a mode of regulation by design. *Information, Communication & Society*, 20(1), 118–136. <https://doi.org/10.1080/1369118X.2016.1186713>
- Yeung, K. (2017b). Algorithmic regulation: A critical interrogation. *Regulation & Governance*, 12(4), 505–523. <https://doi.org/10.1111/rego.12158>
- Yuan, S., Abidin, A. Z., Sloan, M., & Wang, J. (2012). Internet Advertising: An Interplay among Advertisers, Online Publishers, Ad Exchanges and Web Users. *CoRR*, abs/1206.1754. <http://arxiv.org/abs/1206.1754>
- Zarzechny, A., Tanner, C., Barfoot, J., Blackburn, C., Couturier, A., & Munsie, M. (2019). Contact us for more information: an analysis of public enquiries about stem cells [PMID: 31960784]. *Regenerative Medicine*, 14(12), 1137–1150. <https://doi.org/10.2217/rme-2019-0092>
- Zettlers, K. (2010). *Judge Clears CAPTCHA-Breaking Case for Criminal Trial* [last accessed on 08.06.2023]. wired. <https://www.wired.com/2010/10/hacking-captcha/>
- Žliobaitė, I., & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, 24(2), 183–201. <https://doi.org/10.1007/s10506-016-9182-5>
- Zuiderveen Borgesius, F. (2018). *Discrimination, artificial intelligence, and algorithmic decision-making* [online: <https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73>; last accessed on 28.03.2023]. Strasbourg: Directorate General of Democracy © Council of Europe.

- Zuiderveen Borgesius, F., Trilling, D., Moeller, J., Bodó, B., de Vreese, C. H., & Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review*, 5(1). <https://doi.org/10.14763/2016.1.401>
- Zweig, K. A., Deussen, O., & Krafft, T. D. (2017). Algorithmen und Meinungsbildung. *Informatik-Spektrum*, 40(4), 318–326. <https://doi.org/10.1007/s00287-017-1050-5>
- Zweig, K. A., Krafft, T. D., Klingel, A., & Park, E. (2021). *Sozioinformatik – Ein neuer Blick auf Informatik und Gesellschaft*. Carl Hanser Verlag.
- Zweig, K. A., Wenzelburger, G., & Krafft, T. D. (2018). On Chances and Risks of Security Related Algorithmic Decision Making Systems. *European Journal for Security Research*, 3(2), 181–203. <https://doi.org/10.1007/s41125-018-0031-2>





# Curriculum Vitae Tobias Krafft, M.Sc.

<b>Name:</b>	Tobias Krafft, M.Sc.
<b>Pseudonym for publications:</b>	Tobias D. Krafft
<b>Google Scholar:</b>	 Tobias D. Krafft
<b>ORCID:</b>	 0000-0002-3527-1092
<b>2002-2011</b>	Education at the Hohenstaufen Gymnasium, a science-focused high school located in Kaiserslautern, Germany, <b>Baccalaureate</b> April 2011.
<b>10/2011-06/2015</b>	<b>Computerscience, B.Sc.</b> at the Technischen Universität Kaiserslautern (Grade 'gut'); Title of the bachelor thesis: " <i>Vorstellung eines sozioinformatischen Analyseansatzes zur Technikfolgenabschätzung in Anlehnung an Vesters Sensitivitätsmodell am Beispiel des Unternehmens ,Uber‘ als sozio-technisches System</i> " (Grade 'sehr gut'), Supervisor: Prof. Dr. Katharina A. Zweig and Prof. Dr. Paul Lukowicz, Universität Kaiserslautern.
<b>07/2011-07/2012</b>	IHK-certified " <b>Versicherungsfachmann</b> "
<b>06/2015-06/2017</b>	<b>Socioinformatics, M.Sc.</b> at the Technischen Universität Kaiserslautern; (grade 'sehr gut'); Title of the master thesis: " <i>Qualitätsmaße binärer Klassifikatoren im Bereich kriminalprognostischer Instrumente der vierten Generation</i> " (Note 'sehr gut'), Supervisor: Prof. Dr. Katharina A. Zweig, Technische Universität Kaiserslautern and Prof, Dr. Nicholas A. Diakopoulos, University of Maryland.
<b>08/2022-06/2023</b>	<b>Senior Consultant</b> , Partnerschaft Deutschland GmbH.
<b>since 01/2017</b>	<b>Co-head</b> , regional group of the German Informatics Society in Kaiserslautern.
<b>since 07/2017</b>	<b>Research Assistant</b> , RPTU in Kaiserslautern (Prof. Dr. Katharina A. Zweig).
<b>since 02/2019</b>	<b>Founder and CEO</b> , Trusted AI GmbH.
<b>since 01/2023</b>	<b>Technical Assessor</b> , Deutsche Akkreditierungsstelle (DAkkS).
<b>since 06/2023</b>	<b>Mananger / Lead Specialist</b> , Partnerschaft Deutschland GmbH.

## Trainings:

- Rheinland-Pfalz-Zertifikat für Hochschuldidaktik (University didactic certificate of the state of Rhineland-Palatinate)
- PRINCE 2 Foundations

## Honors and Awards

- Weizenbaum-Studienpreis of the Forum InformatikerInnen für Frieden und gesellschaftliche Verantwortung (FIfF) e.V. - 2017
- Head of the working group “Ethics and responsible AI” of the first german standardization roadmap Artificial Intelligence (DIN/DKE) - 2019 – 2020

## Projects

1. **2017-2018** BTW17 - Datenspendeprojekt, together with AlgorithmWatch and Spiegel Online as media partners.
2. **2018-2023** “Deciding about, by, and together with algorithmic decision-making systems”, Volkswagenstiftung

## Supervision of final theses<sup>1</sup>

### 1. Bachelor thesis

- a) ”Profilbildungstheorien und ein praktischer Ansatz für Reverses A/B-Testing zur Analyse von Personalisierten und Dynamischen Preisen in Online-Shops”; Arthur Just, B.Sc
- b) ”Steht unsere Gesellschaft auf »Autoplay«?”, Jan Fiete Schütte, B.Sc
- c) ”Personalization of search results in a medical context”, Frederik Maximilian Stegner, B.Sc.

### 2. Master thesis

- a) ”Abusive Advertising: Scrutinizing socially relevant algorithms in a Black Box analysis to examine their impact on vulnerable patient groups in the health sector”, Martin Reber, 2020, M.Sc.

---

<sup>1</sup>All theses were co-supervised by Prof. Dr. Katharina A. Zweig as first examiner.

## Selected talks and presentations

Throughout my doctoral journey, I have given more than 80 presentations focusing on transferring scientific knowledge, and facilitated more than 100 workshops aimed at disseminating insights from my field of expertise. Below, I provide a brief overview of a selected few of the numerous events I have been fortunate to participate in during my doctoral journey. These events encompass both presentations and workshops, all aimed at the transfer and dissemination of scientific knowledge in my field of expertise.

- |                |                                                                                                                                                                                                                                                  |
|----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>04/2018</b> | <b>Automatisierte Zensur? Technische Lösungen zur Regulation politischer Online-Kommunikation;</b> Karlsruhe Institut für Technologie and Universität Salzburg; Expert workshop                                                                  |
| <b>05/2019</b> | <b>Methodik für ein kontinuierliches Monitoring von Medienintermediäre – Fokus Diskriminierungen;</b> Direktorenkonferenz der Landesmedienanstalten; Workshop of the Direktorenkonferenz                                                         |
| <b>12/2019</b> | <b>Herausforderungen bei der Nutzung von KI in militärischen Anwendungsgebieten - Sozioinformatische Perspektive;</b> German Institute for Defence and Strategic Studies; Ethische Herausforderungen digitalen Wandels in bewaffneten Konflikten |
| <b>06/2020</b> | <b>Forschungsklauseln - verbesserter Datenzugang für Forschung und Wissenschaft,</b> BMBF; BMBF-Fachgespräch                                                                                                                                     |
| <b>01/2021</b> | <b>Ethik/Responsible AI: Vorstellung und Diskussion der Handlungsbedarfe sowie Vorstellung relevanter laufender Normungsaktivitäten;</b> DIN and DKE; Normungsroadmap KI                                                                         |
| <b>11/2021</b> | <b>Mit KI gegen Hatespeech? - Ethische Prinzipien und Verantwortung,</b> Universität Hildesheim; KI gegen Online-Hass II                                                                                                                         |
| <b>06/2022</b> | <b>Faire KI;</b> UNESCO; Workshop with the German Delegation                                                                                                                                                                                     |

Aside from occasional publications during my master, my track record of publishing began when I commenced my doctoral studies in June 2017. When using the Publish or Perish tool to search for either “Tobias Krafft” or “Tobias D. Krafft”, over 35 cited publications were discovered. These publications were written over a period of 6 years and have accumulated more than 430 citations, leading to an h-index of 12. The following compilation provides an overview of the publications I perused and highlights my specific contributions in each instance:

---

### Book

---

- [Zweig et al., 2021]      Zweig, K. A., Krafft, T. D., Klingel, A., and Park, E. (2021). *Sozioinformatik: Ein neuer Blick auf Informatik und Gesellschaft*. Carl Hanser Verlag.  
**Authors contribution:** The book was written by me and Katharina Zweig as the main authors.

---

### Journal Articles & Peer-Reviewed Conference Proceedings

---

- [Hauer et al., 2023a]      Hauer, M. P., Krafft, T. D., Sesing-Wagenpfeil, A., and Zweig, K. A. (2023a). Quantitative study about the estimated impact of the AI Act. <https://doi.org/10.48550/arXiv.2304.06503>[submitted]  
**Authors contribution:** I initiated and conducted the investigation and research in collaboration with Katharina Zweig, Marc Hauer, and Andreas Sesing. Andreas Sesing authored the legal section.

- [Krafft et al., 2023] Krafft, T. D., Hauer, M. P., and Zweig, K. (2023). Black box testing and auditing of bias in ADM systems. *Minds and Machines*. [submitted]  
**Authors contribution:** The paper is the outcome of a close collaboration between Marc Hauer and me. Throughout the process, we worked together and made substantial contributions that effectively connect our respective research fields. Our individual contributions to the paper exhibit significant similarities in terms of their impact and importance.
- [Hauer et al., 2023b] Hauer, M. P., Krafft, T. D., and Zweig, K. (2023b). Overview of transparency and inspectability mechanisms to achieve accountability of AI systems. *Data and Policy*. [accepted; in publication]  
**Authors contribution:** The paper is a product of close collaboration with Marc Hauer, where we have worked together closely and made significant contributions that bridge our respective research fields. Our contributions to the paper share many similarities in terms of their impact and importance.
- [Krafft et al., 2021] Krafft, T. D., Reber, M., Krafft, R., Coutrier, A., and Zweig, K. A. (2021). Crucial Challenges in Large-Scale Black Box Analyses. In Boratto, L., Faralli, S., Marras, M., and Stilo, G., editors, *Advances in Bias and Fairness in Information Retrieval*, pages 143–155, Cham. Springer International Publishing. [https://doi.org/10.1007/978-3-030-78818-6\\_13](https://doi.org/10.1007/978-3-030-78818-6_13)  
**Authors contribution:** This paper draws heavily on the findings derived from my diverse research results, making me the primary author of the work. My extensive contributions form the foundation of the paper, shaping its content and direction.
- [König and Krafft, 2021] König, P. D. and Krafft, T. D. (2021). Evaluating the evidence in algorithmic evidence-based decision-making: the case of US pretrial risk assessment tools. *Current Issues in Criminal Justice*, 33(3):359–381. <https://doi.org/10.1080/10345329.2020.1849932>  
**Authors contribution:** I was responsible for originating the concept and conducting a technical analysis of various risk assessment tools. Pascal Köning, on the other hand, authored the political science aspect of the paper. We have contributed to the paper in equal parts.

- [Harkens et al., 2020] Harkens, A., Yeung, K., Achtziger, A., Felfel, J., Krafft, T., Koenig, P., Schmees, J., Schultz, W., Wenzelburger, G., and Zweig, K. (2020). The rise of ai-based decision-making tools in the criminal justice: Implications for judicial integrity. *Commonwealth Judicial Journal*, 25(2):18–26.  
**Authors contribution:** The paper draws heavily from the discussions within our research project, with Adam Harkens taking the primary role in compiling the content. My contributions to the paper include providing the technical embedding and contributing to the literary background.
- [Hauer et al., 2020] Hauer, M. P., Hofmann, X. C. R., Krafft, T. D., and Zweig, K. A. (2020). Quantitative analysis of automatic performance evaluation systems based on the h-index. *Scientometrics*, 123(2):735–751. <https://doi.org/10.1007/s11192-020-03407-7>  
**Authors contribution:** This paper is based on Marc Hauer’s master thesis and Xavier Hofmann’s bachelor thesis. Together with Marc Hauer, I have scientifically processed the results and embedded them in the scientific literature.
- [Reber et al., 2020] Reber, M., Krafft, T. D., Krafft, R., Zweig, K. A., and Couturier, A. (2020). Data Donations for Mapping Risk in Google Search of Health Queries: A case study of unproven stem cell treatments in SEM. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2985–2992. IEEE. <https://doi.org/10.1109/SSCI47803.2020.9308420>  
**Authors contribution:** The paper is based on the research project I designed with Anna Couturier, compiling the results from Martin Reber’s master thesis.
- [Krafft et al., 2020a] Krafft, T. D., Hauer, M. P., and Zweig, K. A. (2020a). Why Do We Need to Be Bots? What Prevents Society from Detecting Biases in Recommendation Systems. In Boratto, L., Faralli, S., Marras, M., and Stilo, G., editors, *Bias and Social Aspects in Search and Recommendation*, pages 27–34, Cham. Springer International Publishing. [https://doi.org/10.1007/978-3-030-52485-2\\_3](https://doi.org/10.1007/978-3-030-52485-2_3)  
**Authors contribution:** This chapter is based on a joint project for which I was responsible for the research and Marc Hauer was more responsible for the practical application. Ap-

proximately half of the publication is based on my written contributions.

- [Krafft et al., 2020b] Krafft, T. D., Zweig, K. A., and König, P. D. (2020b). How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. *Regulation & Governance*, 16(1):119–136. <https://doi.org/10.1111/re-go.12369>  
**Authors contribution:** The core concept of this article originated from my dialogue with Pascal König. We collaborated closely to transfer our risk assessment framework from AI to principal-agent theory, with Pascal contributing the political science perspective and myself focusing on the (socio-)technical aspects.
- [Krafft et al., 2019] Krafft, T. D., Gamer, M., and Zweig, K. A. (2019). What did you see? A study to measure personalization in Google’s search engine. *EPJ Data Science*, 8(1):38. <https://doi.org/10.48550/arXiv.1812.10943>  
**Authors contribution:** The research project findings are summarized in the paper. The majority of the analytical work was my design with assistance from Michael Gamer. The programming carried out by both Michael Gamer and myself, and the validity was confirmed by Katharina Zweig. I made the main contribution to the writing and composition of the paper.
- [Krafft and Zweig, 2018b] Krafft, T. D. and Zweig, K. A. (2018b). Wie Gesellschaft algorithmischen Entscheidungen auf den Zahn fühlen kann. In Mohabbat Kar, R., Thapa, B. E. P., and Parycek, P., editors, *(Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft*, pages 471–492. Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS, Kompetenzzentrum Öffentliche IT (ÖFIT), Berlin. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-57621-7>  
**Authors contribution:** The work presented in this project is predominantly built upon my original conception and thorough elaboration. I have invested significant effort in developing and shaping the ideas that form the foundation of this work.
- [Zweig and Krafft, 2018] Zweig, K. A. and Krafft, T. D. (2018). Fairness und Qual-

ität algorithmischer Entscheidungen. In Mohabbat Kar, R., Thapa, B. E. P., and Parycek, P., editors, *(Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft*, pages 204–227. Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS, Kompetenzzentrum Öffentliche IT (ÖFIT), Berlin. <http://nbn-resolving.de/urn:nbn:de:0168-ssolar-57570-1>

**Authors contribution:** In collaboration with Katharina Zweig, I took a lead role in developing the transfer of the fairness debate to the implementation of automated decision-making (ADM) systems in the public sector.

[Zweig et al., 2018b]

Zweig, K. A., Wenzelburger, G., and Krafft, T. D. (2018b). On Chances and Risks of Security Related Algorithmic Decision Making Systems. *European Journal for Security Research*, 3(2):181–203. <https://doi.org/10.1007/s41125-018-0031-2>

**Authors contribution:** All authors made equal contributions to the paper.

[Zweig et al., 2018a]

Zweig, K. A., Krafft, T. D., Muramalla, S., and Sieberts, J. (2018a). Algorithmic literacy. In Biehler, R. and Schulte, C., editors, *Paderborn Symposium on Data Science Education at School Level 2017*, pages 33–36. <http://dx.doi.org/10.17619/UNIPB/1-374>

**Authors contribution:** Following in-depth discussions, I collaborated with Katharina Zweig to document and articulate the ideas we generated. The process involved joint efforts in capturing and expressing our shared insights and perspectives.

[Krafft and Zweig, 2017]

Krafft, T. D. and Zweig, K. A. (2017). Ein Faktencheck - Ließ ein Algorithmus Trump triumphieren? *Informatik-Spektrum*, 40(4):336–344. <https://doi.org/10.1007/s00287-017-1052-3>

**Authors contribution:** The authors have contributed equally to this article.

[Zweig et al., 2017a]

Zweig, K. A., Deussen, O., and Krafft, T. D. (2017a). Algorithmen und Meinungsbildung. *Informatik-Spektrum*, 40(4):318–326. <https://doi.org/10.1007/s00287-017-1050-5>

**Authors contribution:** Katharina Zweig and I were re-



sponsible for structuring the paper and conducting the initial groundwork. However, beyond that, all authors have made equal contributions to the article.

[Groen et al., 2017]

Groen, E. C., Kopczyńska, S., Hauer, M. P., Krafft, T. D., and Doerr, J. (2017). Users-The Hidden Software Product Quality Experts?: A Study on How App Users Report Quality Aspects in Online Reviews. In *Requirements Engineering Conference (RE), 2017 IEEE 25th International*, pages 80–89. IEEE. <https://doi.org/10.1109/RE.2017.73>

**Authors contribution:** Marc Hauer and I shared equal responsibility for pre-processing additional test data and enhancing the regular expressions.

---

## Book chapter

---

[Krafft and Zweig, 2020]

Krafft, T. D. and Zweig, K. A. (2020). Herausforderungen bei der Nutzung von KI in militärischen Anwendungsgebieten - Sozioinformatische Perspektive. In Rogg, M., Scheidt, S., and von Schubert, H., editors, *Ethische Herausforderungen digitalen Wandels in bewaffneten Konflikten*, pages 55–66. German Institute for Defence and Strategic Studies, Hamburg.

**Authors contribution:** The content of the article derives from my lecture given at the German Institute for Defence and Strategic Studies, where I discussed the difficulties associated with employing artificial intelligence in various military domains.

[Klingel et al., 2020]

Klingel, A., Krafft, T. D., and Zweig, K. A. (2020). Mögliche Best-Practice- Ansätze beim Einsatz eines algorithmischen Entscheidungsunterstützungssystems am Beispiel des AMS-Algorithmus. In Hengstschläger, M., editor, *Digital Transformation and Ethics*, pages 190–215. Ecowin, Elsbethen.

**Authors contribution:** Katharina Zweig and I laid the scientific foundation and then I worked out the chapter with Anita Klingel.

---

## Technical reports / grey literatur

---

- [DIN/DKE, 2023] DIN/DKE (2023). German standardization roadmap on artificial intelligence 2.0. *DIN/DKE, Berlin/Frankfurt*. <https://www.din.de/resource/blob/916792/20bf33d405710a703aa26f81362493bb/nrm-ki-deutsch-2022-final-web-250-data.pdf>  
**Authors contribution:** For this version, I mainly participated in the sociotechnical system working group.
- [Krafft et al., 2022] Krafft, T. D., Krafft, R., Wölki, M., Rahe, M., and Zweig, K. A. (2022). Algorithmische Governance von personalisierten Preisen im Online-Handel. Technical report, Ministerium für Familie, Frauen, Kultur und Integration des Landes Rheinland-Pfalz. [in publication]  
**Authors contribution:** In this report, I provide an extensive overview of the algorithmic governance of personalized pricing in online retailing using black box analytics. While incorporating the work of Marcel Wölki and Roman, I have taken the lead in writing the majority of the overview. However, it's important to note that the legal section of the paper was specifically written by Michael Rahe.
- [Heesen et al., 2021] Heesen, J., Müller-Quade, J., Wrobel, S., Dabrock, P., , Decker, M., Damm, W., Grunwald, A., Heine, K., Monnet, J., Houdeau, D., Matzner, T., Rost, P., Schauf, T., Zweig, K. A., Krafft, T. D., and Poretschkin, M. (2021). Kritikalität von KI-Systemen in ihren jeweiligen Anwendungskontexten: ein notwendiger, aber nicht hinreichender Baustein für Vertrauenswürdigkeit : Whitepaper. Technical report, Lernende Systeme - Die Plattform für Künstliche Intelligenz; acatech, München. [https://doi.org/10.48669/pls\\_2021-3](https://doi.org/10.48669/pls_2021-3)  
**Authors contribution:** The article was collectively written through a collaborative effort, making it challenging to determine individual contributions, despite the fact that the content is rooted in our research.
- [DIN/DKE, 2020] DIN/DKE (2020). German standardization roadmap on artificial intelligence 1.0. *DIN/DKE, Berlin/Frankfurt*. <https://www.din.de/resource/blob/772438/6b5ac6680543eff9>

[fe372603514be3e6/normungsroadmap-ki-data.pdf](https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf)

**Authors contribution:** For the first german artificial intelligence standardization roadmap, I was leader of the working group "Ethics and responsible AI" . Here, I largely wrote the texts and coordinated with over 100 experts. As working group leader, I was also involved in the quality assurance of the entire standardization roadmap.

[Hallensleben et al., 2020] Hallensleben, S., Hustedt, C., Fetic, L., Fleischer, T., Grünke, P., Hagendorff, T., Hauer, M. P., Hauschke, A., Heesen, J., Herrmann, M., Hillerbrand, R., Hubig, C., Kaminski, A., Krafft, T. D., Loh, W., Otto, P., and Puntschuh, M. (2020). From principles to practice - an interdisciplinary framework to operationalise ai ethics. Technical report, AI Ethic Impact Group, Bertelsmann Stiftung, [https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO\\_2020\\_final.pdf](https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf).

**Authors contribution:** Marc Hauer and I wrote Chapter 3 with equal participation.

[Haeri et al., 2020] Haeri, M., Hartmann, K., König, P., Krafft, T., Sirsch, J., Joisten, K., Wenzelburger, G., and Zweig, K. (2020). Denkanstöße zum Einsatz von ADM-Systemen in der öffentlichen Verwaltung. Technical report, Technische Universität Kaiserslautern, [https://fairandgoodadm.cs.uni-kl.de/res/Denkanstöße\\_final.pdf](https://fairandgoodadm.cs.uni-kl.de/res/Denkanstöße_final.pdf).

**Authors contribution:** The paper summarizes the final research results of Project FairAndGood ADM. The reflections and discernments expressed within were generated in collaboration with my assistance.

[Krafft et al., 2018c] Krafft, T. D., Gamer, M., and Zweig, K. A. (2018c). What did you see? Personalization, regionalization and the question of the filter bubble in Google's search engine. <https://doi.org/10.48550/arXiv.1812.10943>

**Authors contribution:** I held primary responsibility for conceiving the evaluation methodology, which was subsequently conducted in collaboration with Michael Gamer. The written elaboration of the paper predominantly draws from my work and contributions.

[Krafft and Zweig, 2018a] Krafft, T. D. and Zweig, K. A. (2018a). Transparenz und

Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse - Ein Regulierungsvorschlag. Technical report, Verbraucherzentrale Bundesverband. [https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22\\_zweig\\_krafft\\_transparenz\\_adm-neu.pdf](https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf).

**Authors contribution:** The idea for the paper was developed in collaboration with Katharina Zweig, and I took on the primary responsibility for writing the majority of the paper.

[Krafft et al., 2018b]

Krafft, T. D., Gamer, M., and Zweig, K. A. (2018b). Wer sieht was? Personalisierung, Regionalisierung und die Frage nach der Filterblase in Googles Suchmaschine. Technical report, Bayerischen Landeszentrale für neue Medien (BLM), Algorithm Watch, <https://www.blm.de/files/pdf2/bericht-datenspende---wer-sieht-was-auf-google.pdf>.

**Authors contribution:** Michael Gamer and I took on the responsibility of conducting the evaluations for the study. In addition, I played a crucial role in conceptualizing the study and extensively developing the content of the text.

[Burger et al., 2018]

Burger, J., Krafft, T. D., and Schwikal, A. (2018). Zertifikatsangebot zum Themenfeld Sozioinformatik - Die bedarfsorientierte Entwicklung von wissenschaftlichen Weiterbildungsangeboten an der Technischen Universität Kaiserslautern. Technical report, Technische Universität Kaiserslautern, [http://kluedo.ub.uni-kl.de/frontdoor/deliver/index/docId/5216/file/\\_Burger\\_Krafft\\_Schwikal\\_2018\\_Sozioinformatik.pdf](http://kluedo.ub.uni-kl.de/frontdoor/deliver/index/docId/5216/file/_Burger_Krafft_Schwikal_2018_Sozioinformatik.pdf).

**Authors contribution:** Katharina Zweig and I collaborated to develop the concept of a distance learning course in socioinformatics, which was further refined and formalized by Anita Schwikal.

[Krafft et al., 2018a]

Krafft, T. D., Gamer, M., and Zweig, K. A. (2018a). Personalisierung auf Googles Nachrichtenportal während der Bundestagswahl 2017. Technical report, Algorithm Watch, <https://doi.org/10.13140/RG.2.2.34815.51360>.

**Authors contribution:** Michael Gamer and I played a vital role in evaluating the report, and subsequently, we worked with Katharina Zweig to co-author this report.

[Zweig et al., 2017b]

Zweig, K. A., Krafft, T. D., and Hauer, M. P. (2017b). Dein Algorithmus – meine Meinung! <https://www.blm.de/aktivitaeten/medienkompetenz/materialien/algorithmenbrochure.cfm>

**Authors contribution:** The brochure was written together with Katharina Zweig and Marc Hauer and has been printed and downloaded over 20,000 times.

[Krafft et al., 2017]

Krafft, T. D., Gamer, M., Laessing, M., and Zweig, K. A. (2017). Filterblase geplatzt? Kaum Raum für Personalisierung bei Google-Suchen zur Bundestagswahl 2017. Technical report, Algorithm Watch, <https://doi.org/10.13140/RG.2.2.29139.07203>.

**Authors contribution:** I took the lead in developing the evaluations, and together with Michael Gamer and Katharina Zweig, we conducted the evaluation process collaboratively. Around half of the paper is based on my contribution.