Challenges of Data-Driven Technologies for Social Inequality and Privacy: Empirical Research on Context and Public Perceptions

Inaugural dissertation
submitted in partial fulfillment of the requirements
for the degree Doctor of Social Sciences
in the Graduate School of Economic and Social Sciences
at the University of Mannheim

by

Frederic Gerdon

Dean of the School of Social Sciences:

Prof. Dr. Michael Diehl

First supervisor:

Prof. Dr. Frauke Kreuter

Second supervisor:

Prof. Dr. Florian Keusch

Thesis reviewers:

Prof. Dr. Tobias Gummer

Prof. Dr. Markus Strohmaier

Date of defense:

April 24, 2024

General acknowledgments

I would like to thank the many people who supported me in the past few years while working on my dissertation. My supervisors Frauke Kreuter and Florian Keusch are not only great mentors who broadened my thinking about research problems and always gave helpful feedback. They also created a research group and environment with optimal conditions for working on my ideas, a lively and friendly community, and manifold opportunities for collaboration. Thank you for making this dissertation possible!

While I thank every member of the Kreuter-Keusch research group for their feedback and discussions throughout the years, special thanks go to my colleagues and co-authors Ruben Bach and Christoph Kern with whom – along with Frauke and Florian – I learnt and still improve writing papers. Thanks also go to the CAIUS research project team in which I learnt about the value of interdisciplinary research, and the "SoDa Privacy Bubble" consisting of Leah von der Heyde, Carolina Haensch, and Isabela Coelho for great Friday afternoon discussions. Further thanks go to Tobias Gummer and Markus Strohmaier who are willing to review my dissertation; to Henning Silber, whom I productively worked with on several papers besides the dissertation; to Ursula Eckle, Beate Rossi, Hannah Laumann, Wiebke Weber, Yvonne Havel, and Pia Förg for their masterful managerial and administrative work; and to our research group's book club members for insightful discussions about novels from around the world.

Last but not least, I am deeply grateful for my family and friends who support me all this time. Extra special gratitude goes to my parents Benita and Markus Gerdon who always support and encourage me and gave me the freedom to go my own way – thank you so much for everything!

Contents

List of tables

List of figures

1. Introduction

The recent decades have seen an accelerating development and employment of digital technologies that are based on the processing of individual data. Individuals knowingly or unknowingly produce masses of such data when using digital technologies and navigating online services. Organizations can use these data to develop and run a variety of technologies, such as social media algorithms for displaying content and advertisements, personal recommendations in streaming services (Elahi et al., 2022), assistance systems in medical diagnoses (Ngiam & Khor, 2019), or credit scoring (Hurley & Adebayo, 2016). These kinds of technologies can be subsumed under the term "data-driven technologies", with which I refer to digital technologies used by organizations that draw on large amounts of individual data from a large group of people to offer services or products, make decisions about individuals, or conduct research. As the focus is on data about human individuals, the term "human data-driven technologies" might be more accurate, but I will use the term "data-driven technologies" for brevity.

The effects of these data-driven technologies go beyond efficiency gains and service improvement of businesses and governmental agencies. These technologies strongly affect individuals' lives directly in various respects, ranging from opportunities for health tracking with mobile devices (Feng et al., 2021; Sharon, 2017) to showing specific types of products in targeted advertising (Lam et al., 2023). While these innovations may come with several advantages – e.g., promoting health-related behavior (Feng et al., 2021) –, research has long been worried about detrimental effects for individuals and society. Among these worries are privacy concerns when massively performing automated data collections (Nissenbaum, 2019; Nissenbaum, 2010), fairness concerns when automated decisions are being made about individuals based on such data (Mehrabi et al., 2022), and "digital divides" in the access of, use of, and reaping benefits from digital technologies, e.g., depending on socio-economic background (Lutz, 2019).

Accordingly, also the legitimacy of the employment of these data-driven technologies depends not only on their efficiency (Schiff et al., 2022). First, to learn about their actual social impacts, objective effects of these technologies on, among others, social inequality and privacy need to be evaluated. Gauging these effects requires an analytical perspective taking into account the various decisions and processes inherent to the different steps of the employment of data-driven technologies (see below). Second, particularly with respect to technologies based

on machine learning, which is commonly employed for the analysis of large data sets, scholars repeatedly also highlight further ethical requirements in technology design and explanation, such as transparency and accountability (Diakopoulos, 2020). However, third, I argue that policymakers and businesses also need to take into account perceptions of the public to ensure a responsible regulation and use of data-driven technologies. Schiff et al. (2022) point out that public perceptions could express whether a data-driven technology violates important public values, such as relating to fairness. A responsible use of these technologies requires a "social license", a concept that has been introduced for the context of businesses by Gunningham et al. (2004) which means that an organizations' activities face public acceptance and conform with the relevant public's expectations (see details in Chapter 1.2). Thus, knowledge about empirical patterns of acceptance – e.g., depending on the exact design and context of a specific algorithmic decision-making process – is crucial to ensure an ethical employment of novel data-driven technologies. Moreover, public acceptance may itself affect adoption of technologies and discourse, which in turn may influence the development and regulation of such technologies and thereby shape what the actual social impacts of a data-driven technology are.

Before this background, to advance research on data-driven technologies' social impacts and their public acceptance, the key aims of the present dissertation are (1) discussing an analytical perspective on how exactly social impacts can arise from data-driven technologies and how these impacts depend on social context and (2) empirically investigating public acceptance of data-driven technologies and their inherent data uses and showing that acceptance may strongly vary depending on specific (see below) contextual dimensions.

This introductory chapter will present two of the main challenges relating to social impacts of data-driven technologies which this dissertation focusses on: social inequality and privacy. The following subchapters will embed and argue for the relevance of *public acceptance* of these technologies with respect to these challenges and why acceptance needs to be measured *context-specifically*. These considerations conclude with concretized contributions that this dissertation makes. Then, I provide extended summaries of the papers that constitute the subsequent four chapters. The final chapter summarizes and contextualizes key findings and conclusions, with a focus on making a case for context-based research on social impacts and acceptance of data-driven technologies.

1.1. Data-driven technologies: Challenges for social inequality and privacy

The collection and use of individual data for service and product development, policy-making, and decision-making has become a common practice for many tasks and problems. In more basic forms, such a use of data is not a historically new phenomenon. Mennicken and Espeland (2019) outline that population data had systematically and scientifically been used for making social policy already since the seventeenth century, and the first (known) census had taken place as early as in ancient Babylonia. However, technological developments such as the Internet and smartphones have led to unprecedented opportunities for collecting and making use of data. A plethora of examples of data-driven technologies reflect a desire to translate these opportunities into actual benefits – for the organizations and/or for society. The Austrian national employment agency tried to classify job seekers with respect to calculated employability scores and to tailor their offered services to the job seekers accordingly (Allhutter et al., 2020). Online vendors can offer different prices to different individuals based on a different predicted willingness to pay (Lippert-Rasmussen & Aastrup Munch, 2021). Credit scoring algorithms feed into decisions of whether individuals receive a credit or not (Hurley & Adebayo, 2016). Not too long ago, governments used different designs of contact tracing apps to track and contain the spread of COVID-19 (Hogan et al., 2021).

To meaningfully discuss social impacts associated with such and other data-driven technologies, we first need an analytical grasp of the process of data collection and use associated with these technologies. From an analytical perspective, the use of data-driven technologies for taking actions can be divided into three steps (Weyer et al., 2018 and Chapter 2 of the dissertation): (1) data generation, (2) data analysis, and (3) implementation of the technology in a concrete social context. As for (1) data generation, data can come from a variety of sources, including administrative registers and surveys as well as digital technologies, such as digital trace data from Internet use (Keusch & Kreuter, 2021). The latter technologies allow for the collection of "big data", i.e., with higher speed, in larger amounts, and with more variety than traditional approaches (Foster et al., 2021; Gartner, 2024). As for (2) data analysis, machine learning algorithms and other procedures can identify patterns of correlations between individual characteristics in these data. These patterns can be – legitimately or not – used to build predictive models (see Molina & Garip, 2019). As for (3) implementation, the resulting models can be used to take actions, such as making decisions about individuals. Let us consider credit scoring as an example that covers each of these steps: An algorithmically created model of the probability to pay back credits, based on payment behavior data of a large group of

individuals, could assign credit scores to individuals based on their measured and supposedly relevant characteristics, and this score could be used to decide about providing a credit to an individual or not (Hurley & Adebayo, 2016).

In all three steps, researchers, businesses, and governments need to carefully inspect potential detrimental effects of the employment of such data-driven technologies on social inequality. The same rigor is required for considering how to safeguard privacy while also deriving potential benefits from data use. In the following, I briefly discuss these challenges for social inequality and privacy before outlining the acceptance-focused angle that the present dissertation takes to investigate them.

Challenges for social inequality

The first challenge of data-driven technologies that this dissertation focuses on is social inequality and "fairness", which is particularly relevant for algorithmic decision-making (ADM). To analyze the data masses arising from intensified data collection, organizations oftentimes rely on machine learning algorithms for supervised learning tasks (see Molina & Garip, 2019). The resulting models can be used to try to predict future individual behavior or outcomes, as a basis for decisions to be made about an individual (for a more thorough discussion of the workings of machine learning as related to ADM, see Chapter 2).

These procedures come with considerable concerns regarding socially unequal outcomes. Previous research has long identified that ADM can perpetuate biases and discriminate against already disadvantaged groups of the populations (Herzog, 2021). In the above-mentioned credit scoring example, individuals with specific socio-demographic characteristics could spuriously be denied credit based on associations found in the data which may reflect previous discrimination (Hurley & Adebayo, 2016). However, it has been argued that humans may even fare worse than algorithms in terms of bias and accountability (Mayson, 2019). Sunstein (2023) adds further complexity by arguing that while human deciders make mistakes as well and that algorithms can outperform human predictions with respect to accuracy, the latter might still struggle with predictions in complex systems with partly unknown or suddenly changing context characteristics. These considerations highlight that researchers need to pinpoint how exactly a specific ADM system may impact – in whichever direction – social inequalities and to take into account relevant context features (see Chapter 2).

Processes that lead to the production of unequal outcomes in society and specific social contexts can be present in each step of the analytical division of data-driven technology use as outlined above. Many of these processes are researched in the field of "fair machine learning"

(Barocas et al., 2023). Among others, the data might be biased in the first place (Mehrabi et al., 2022), the analysis might disregard fairness or use inadequate fairness metrics (see Makhlouf et al., 2021), the algorithmic recommendations can be sub-optimally adopted, e.g., to over- or under-reliance (Wickens et al., 2015; Zerilli et al., 2019), and single problematic decisions about individuals may aggregate to macro-social impacts in the long run (see Coleman, 1994). Chapter 2 will discuss these and other problems, potential solutions, and resulting research avenues in detail.

Achieving "fairness" and thus mitigating potential negative effects on social inequality is not only a technical challenge, as there may be competing suggestions of what actually is "fair" in specific contexts. Thus, making ADM "fair", or "fairer", requires the involvement of different stakeholders (Rahwan, 2018) and perspectives, such as context-specific public perceptions. These fairness perceptions have become the subject of an own research area (Starke et al., 2022), and previous research found more positive fairness perceptions to also favorably affect other perceptions of and intention to use ADM systems (Aysolmaz et al., 2023). I argue that fairness perceptions and acceptance of ADM are a vital part of the "social license" (Gunningham et al., 2004) to use these kinds of data-driven technologies, as Chapter 1.2 will explain in more detail. Chapter 3 will furthermore empirically underpin the necessity for *context*-based evaluations by using a survey experiment to gauge people's fairness and acceptance ratings of different types of ADM systems.

Challenges for privacy

Privacy is paramount for safeguarding democracy and freedom (Seubert & Becker, 2021). With an ever more encompassing "quantification" of social life (see Mennicken & Espeland, 2019) and more fine-grained and new types of data collections, scholars are concerned whether the privacy of individuals is threatened (e.g., Nissenbaum, 2010; Rubinstein, 2013). The challenge is to protect the privacy of individuals while at the same time being able to make use of data to the profit of individuals and society.

To address this challenge, a holistic view of the data use "lifecycle" is necessary. Following the analytical division from above, privacy issues are not only prevalent in the data generation step, but also in the data analysis and implementation steps. To evaluate whether a specific data use violates privacy from an ethical perspective, we need to assess the different steps in conjunction with each other. This means that organizations have to consider whether the concrete individual data can be legitimately used for a specific purpose or decision after having processed the data in a specific way. For instance, it may be deemed acceptable to collect and

analyze data for one purpose, but analyzing the same data for a different purpose may be questionable (see Chapters 4 and 5). Moreover, a specific data collection endeavor and its purpose could be incrementally, but questionably extended. For instance, scholars warned that digital contact tracing tools that were used to tackle the COVID-19 pandemic might over time become surveillance tools for purposes unrelated to the pandemic (Vitak & Zimmer, 2020). Purpose limitation is enshrined in the General Data Protection Regulation (GDPR) in the European Union, but how personal "big data" is used in practice might conflict with this principle (Forgó et al., 2017). It also has been suggested that purpose limitation should be extended to the use of algorithmically trained models, even when based on anonymous data (Mühlhoff & Ruschemeier, 2024). Going further, the concept of "predictive privacy" (Mühlhoff, 2021) suggests that it may be questionable in the first place to make a decision or infer sensitive information about an individual based on a statistical model trained on other people's data, also even if these data are anonymous.

Given these challenges, organizations should conduct ethical evaluations to make sure that their data collections and uses are acceptable from a legal *and* an ethical perspective. As pointed out above and will be further elaborated below, these ethical evaluations should be informed by the public's perceptions and acceptance. To this end, I will turn to the notion of privacy as "contextual integrity" (Nissenbaum, 2010) that aids in judging data uses in Chapter 1.2. Later, Chapters 4 and 5 will present results from survey experiments that demonstrate the usefulness of the contextual integrity perspective by applying it to context-specific privacy perceptions of the public.

1.2. The relevance of context-specific public acceptance of data-driven technologies

In the following, I will (1) discuss why gauging public acceptance of data-driven technologies is ethically relevant and (2) explain how such investigations on public acceptance benefit from considering contextual factors.

The relevance of public acceptance

Given the challenges outlined in the previous subchapter, from a political and ethical perspective, not everything that can technically be done also *should* be done. Judgments on what *should* be done can be derived from several (ethical) perspectives.

For instance, one can take a consequentialist utilitarian, i.e., outcome-focused perspective to evaluate actions taken with data-driven technologies based on the beneficial and detrimental

effects they entail (Bednar & Spiekermann, 2022). To measure these effects, empirical researchers can investigate how data-driven technologies alter opportunity structures and behaviors and thus aggregate social outcomes (Coleman, 1994), and how effects are distributed across different groups of individuals as to affect social inequality. For instance, such research can theorize or observe effects of data-driven automation tools on productivity, the labor market, and income (Agrawal et al., 2019; Danaher, 2022; Furman & Seamans, 2019), of differences in digital technology use across socio-demographic groups on social inequality (Lutz, 2019), and of surveillance practices on chilling behavior (Büchi et al., 2022). This research can investigate effects on a range of theoretically, practically, or ethically relevant outcomes, such as differences in labor market chances or in the levels of control that different groups of individuals effectively have over their data.

Another perspective that is decisive for practical matters are legal norms that define what actions organizations may lawfully perform with data-driven technologies. The European Union is a key legislator in this field in Europe, which has adopted the GDPR to regulate the use of personal data and, with exceptions, to protect individuals from unconsented completely automated decisions (European Parliament & Council of the European Union, 2016). Another key proposal in this field is the "AI Act" to regulate AI technology with a risk-based approach (Council of the European Union, 2023). However, legal frameworks sometimes allow actions that can still be deemed ethically problematic. For example, the processing of personal data is commonly allowed under the GDPR if the data subject gives consent to the processing purpose (Article 6 of the GDPR). However, there is a debate on the meaningfulness of "notice-and-consent" procedures, e.g., due to a lack of an individuals' full understanding of what is happening with the data or due to "dark patterns" nudging people into consent (Andreotta et al., 2022; Gray et al., 2021; Mills, 2022; Susser, 2019).

Adding to these two perspectives, I argue that what *should* be done is neither merely a matter of impacts on specific outcome measures, nor only of legal conditions. As part of ethical reflections on data-driven technologies, research has suggested that we need to take into account the perspectives of the people, both for issues related to fairness and privacy. For instance, in the context of data science-based health research, Aitken et al. (2019) published a "consensus statement" that research needed public support and to align with public values, and that the public thus took part in providing the benefits to be reaped from such research. With respect to algorithms, Rahwan (2018) suggested to bring "society-in-the-loop" into the employment of algorithmic systems, which included a "social contract" in which societal stakeholders needed to decide on how to navigate value trade-offs and how to distribute costs and benefits. Thus,

taking into account public perspectives is not only relevant in its own right from an ethical deontological perspective that centers "duties" (Bednar & Spiekermann, 2022), but can also involve utilitarian weightings of consequences.

On a more general level, public perceptions in the form of public acceptance are captured by the notion of "social license" that refers to the requirement for organizations to comply with the affected society's expectations, which can be more restrictive than the regulatory environment (Gunningham et al., 2004). Gunningham et al. (2004) initially applied this concept to businesses, but later research also considered governmental organizations (Carter et al., 2015; Shaw et al., 2020). Expectations may vary across groups of individuals, depending on their relatedness to the organization or its action at stake (Gunningham et al., 2004). Organizations may face negative consequences for not complying with societal expectations, such as a loss of reputation and stricter legal regulations (ibd.). Such non-compliance can hence lead to negative consequences for society by missing out on potentially beneficial products or services. One example is the failed *care.data* initiative in England in which general practitioners were to share medical records with a "Health and Social Care Information Centre" to be used for different purposes, with only an opt-out option for patients (Carter et al., 2015). Carter et al. (2015) argue that *care.data* did not obtain a social license as, among others, it deviated from patients' common expectations of the confidentiality of the doctor-patient relationship without taking sufficient additional steps, and because patients might have questioned whether the data will only be used for public benefit purposes. Thus, the concrete implementation failed to receive general acceptance and was suspended, although some benefits could have arisen from this initiative (ibd.).

Given these consequences, social licenses again not only appear as ethically relevant from a deontological perspective, but also from a utilitarian perspective as some outcomes (such as public health) might be improved by appropriately offering "socially licensed" products and services. However, Shaw et al. (2020) point out that there is no consensus among researchers on whether and how such a social license can be accurately measured. Still, I argue that researchers can evaluate the overall acceptance of specific data-driven technologies by taking into account the relevant contextual factors. I will briefly discuss how researchers can approach context-specific acceptance for the two substantive foci of this dissertation – social inequality and fairness in ADM; data flows and privacy – in the following paragraphs. Detailed discussions and empirical applications of these approaches follow in Chapters 3 to 5.

The relevance of context for public acceptance

Acceptance and fairness perceptions (being related to social inequality and acceptance) of ADM systems have shown to be context-specific (Molina & Sundar, 2022; Starke et al., 2022; Wenzelburger et al., 2022), with research findings ranging from algorithmic appreciation (Logg et al., 2019) to algorithmic aversion (Dietvorst et al., 2015). A systematic literature review summarized empirical findings on *fairness* perceptions, concluding that they and some of their predictors depended on the use context Starke et al., 2022). One of the found context-sensitive predictors is the degree of automation of a concrete ADM system, i.e., the extent to which humans and computers are involved in the decision-making process. The survey experiment presented in Chapter 3 shows that the degree of automation, along with other predictors, also affects individuals' *acceptance* ratings of ADM systems. The study found variation in fairness and acceptance ratings across four investigated contexts (finance, hiring/HR, criminal justice, and labor market integration) and depending on the degree of automation, and some context-dependent variation in the relevance of further predictors (see Chapter 3). These further predictors include the relatedness of the used data to the specific task (Dodge et al., 2019; Grgić-Hlača et al., 2018; van Berkel et al., 2019; Waldman & Martin, 2022), and the findings imply that respondents also care about which individual data are used in ADM processes in, with slight variation across contexts.

For the case of privacy, the notion of privacy as "contextual integrity" (Nissenbaum, 2010) suggests that the acceptability of an "information flow" hinges on its compliance with contextual informational norms (or "privacy norms"). Nissenbaum understands privacy norms as one among many types of norms inherent to social contexts. By social contexts, Nissenbaum means social domains such as health, work, and education (Nissenbaum, 2018). Nissenbaum uses the term "appropriateness" to describe the compliance of an "information flow" (or "data flow") with these context-specific privacy norms. Perceived appropriateness can therefore be considered as the social license for the execution of respective data flows by the involved actors. Chapters 4 and 5 explain in detail how exactly these data flows need to be operationalized to meaningfully gauge peoples' acceptance.

To be clear, "context" is not a concept with arbitrary content. By understanding context as *social* context or domain, Nissenbaum grounds this concept within sociological and philosophical theories of society (Nissenbaum, 2010, 2018) and draws on the central concept of norms (see Bicchieri, 2017) to derive a theory that highlights the context-dependency of such norms. Perceptions and acceptance are therefore also context-dependent and can result from a system's or data flow's adherence to or deviation from these norms (Nissenbaum, 2010). Given

this general theoretical foundation, the notion of the ethical relevance of context-specific norms appears applicable to data-driven technologies, including ADM, more broadly. As described above, contextual integrity presents privacy norms as one among many types of norms that are inherent to social contexts (Nissenbaum, 2010). Applied to data-driven technologies more generally, further types of norms may be relevant for acceptance in the different steps of the ADM process, such as notions of fairness or justice (see Chapter 2, and Kuppler et al., 2022). Possibly going a step in this direction, Mulligan and Nissenbaum (2020) sketched an ethical and political analysis of sociotechnical systems that emphasized that changes in a system's components might affect its evaluation in terms of touched on ethical and political values. Other recent research sought to apply contextual integrity to algorithms (Oomen et al., 2024). The relevant point is that *context* (understood as social context) appears as a useful and theoretically grounded comparative perspective for public acceptance relating to data flows *and* ADM more broadly.

The context-specific approach of the present dissertation not only allows researchers to study the specific social impacts of single data-driven technologies within their social contexts (Chapter 2). This approach also opens opportunities for comparative research to systematically compare acceptance of data flows and ADM by contexts depending on further characteristics of a concrete application and of the larger societal circumstances. These characteristics include features of ADM systems as outlined above (more detailed in Chapter 3), details in the definition of data flows (Chapters 4 and 5), the timing of the investigation (Chapter 4), and the studied country (Chapter 5).

In summary, public acceptance is ethically relevant as providing a "social license" for an organizations' employment of data-driven technologies. For both data flows and ADM more broadly, it has become clear that public acceptance is context-dependent. Researchers hence need to measure and evaluate people's acceptance with respect to specific contexts. Measuring rather general perceptions about, e.g., data-driven technologies or AI can also yield important insights, such as what individuals think of risks and benefits of specific applications and which applications they are particularly concerned about (e.g., see the survey by The Ada Lovelace Institute & The Alan Turing Institute, 2023). However, in many cases, acceptance will depend on how a specific data-driven technology is concretely used by whom, as the following chapters will demonstrate.

1.3. Contributions

Given these considerations, in summary, this dissertation makes the following main contributions:

(a) discussing a broadly applicable analytical perspective, current empirical findings, and research avenues on potential social impacts – particularly with respect to social inequality – arising from different steps in the process of data use within the employment of ADM systems as a data-driven technology,

(b) theoretically discussing and empirically showing which context factors are particularly relevant for public acceptance of data-driven technologies and their inherent data flows, and

(c) providing empirical findings on several concrete data-driven technologies and their inherent data flows within their social contexts, including longitudinal and international comparisons, showcasing comparative potentials for systematic context-aware research.

These contributions are scientifically relevant by highlighting key research avenues for social scientists dealing with social impacts of data-driven technologies and by suggesting how comparative research on social context along with further comparative dimensions can enrich empirical research on their public acceptance. These contributions are furthermore practically relevant by providing public agencies and businesses with analytical tools and empirical insights to assess the potential social impacts and the ethical appropriateness of the use of specific data-driven technologies, which can also inform policymakers' regulation efforts to take into account specificities of contexts. More concrete conclusions will be provided in the discussion and conclusion sections of Chapters 2 to 5 and in the final concluding Chapter 6.

The following chapters will present a detailed analytical overview on social impacts of data-driven technologies as well as theoretical elaborations and empirical assessments of context-specific public acceptance of data-driven technologies and their data flows. All of the empirical studies make use of survey experiments, or more concretely: vignette experiments (see Auspurg & Hinz, 2015). Vignette experiments are particularly useful for researching context-specific acceptance as they allow researchers to randomly vary components of descriptions of ADM systems and of data flows. Researchers can then estimate the effects of changes of contextual characteristics on the respondents' acceptance. Based on such findings, researchers can judge the relevance of different contextual characteristics, including which of the characteristics matter in the first place and which constellations are the most and least accepted. How this exactly works will be explained and demonstrated for different applications throughout Chapters 3 to 5.

1.4. Summary of chapters

In the following, I provide extended summaries for the subsequent four chapters (that comprise of three published journal papers and one paper currently under review after revision and resubmission) and contextualize them within the larger aims of this dissertation. Chapter 2 presents an analytical overview and conceptual underpinnings of social impacts of ADM – with a focus on social inequality –, which is in principle applicable to data-driven technologies more generally. This chapter shows that a context-specific study of ADM and other data-driven technologies is necessary to get a firm understanding of their social impacts. The presented analytical distinctions and the emphasis on context characteristics are also applicable to privacy issues with data-driven technologies; later chapters will elaborate on the concept of "contextual integrity" (Nissenbaum, 2010) as a privacy-specific approach to contextuality. The subsequent chapters present empirical studies on context-specific public acceptance of ADM systems (Chapter 3) and data flows (Chapters 4 and 5). While each of the empirical studies focusses on contextual variations, they also take more fundamental comparative dimensions into account: time and space. The study in Chapter 4 emphasizes the relevance of the *temporal* dimension of public acceptance – which has proven to be particularly relevant in times of the COVID-19 pandemic. Chapter 5 not only compares the importance of contextual parameters across time points, but also how their influence on acceptance may vary across *countries*. Figure 1.1 provides an overview over these four chapters of this dissertation. The dissertation concludes with a chapter on theoretical and practical implications and avenues for future research.

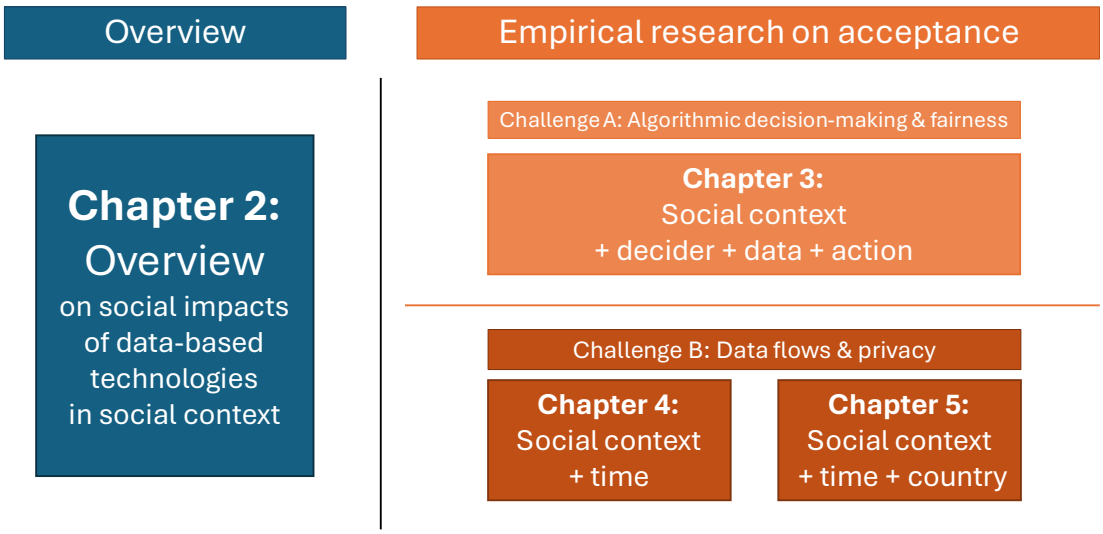


Figure 1.1. Overview of the main chapters of this dissertation.

Chapter 2: "Social impacts of algorithmic decision-making: A research agenda for the social sciences"

Chapter 2 provides a broad overview of the social impacts that can arise from the use of ADM systems, which can be extended to data-driven technologies more generally. Drawing on a "big data process model" proposed by Weyer et al. (2018), the chapter follows an analytical distinction between three key steps of ADM development and use: (1) data generation, (2) data preparation and analysis, and (3) implementation of an ADM system in a concrete social context (also see outline in Chapter 1.1). Each of these steps comes with challenges and decisions that can shape the social impacts of the respective ADM system, which the chapter discusses with a focus on social inequality. By reviewing the literature in this field, for each of these steps, the chapter summarizes key concepts and empirical findings relating to the social impacts of ADM. Based on these summaries, research avenues for social scientists aiming to work in this field are identified. The chapter therefore serves as a general introduction to research on social impacts of data-driven technologies.

Chapter 2 discusses in deeper detail the challenges for each of the three steps as relating to fairness and privacy that have already been outlined in Chapter 1.1. Importantly, a recurring theme is the context-dependency of various sources of social inequality in the different steps of the ADM process. This includes, for instance, how humans react to and rely on ADM applications (Wickens et al., 2015; Zerilli et al., 2019) introduced in specific social contexts. Another challenge is to choose a fairness metric in the data analysis step, where the appropriateness of choice may depend on various factors, including social context (Makhlouf et al., 2021). As argued above, what is acceptable and considered fair is subject to societal scrutiny and also needs to be researched empirically.

A key contribution of the chapter is to show the value of the expertise of social scientists to research how exactly and depending on which factors social impacts of data-driven technologies may evolve in social contexts. Chapter 2 thereby places research on context-specific public acceptance of data-driven technologies within the larger research field of social impacts of these technologies.

Chapter 3: "Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making"

As has been described above, ADM is increasingly being used for various purposes by businesses and public agencies, but the ethical and successful employment of these systems requires researchers and designers of ADM technologies to take into account public acceptance.

Chapter 3 uses a vignette experiment (see details in Chapter 3) to study public acceptance and fairness perceptions of ADM systems in four social contexts that have garnered particular scholarly attention: finance, hiring/HR, criminal justice, and labor market integration. In each of these four contexts, data-driven technologies can be or are used to make decisions about individuals. This study compares the effects of specific characteristics of ADM systems on individual fairness perceptions and acceptance across these contexts. Concretely, the experiment differentiated between a fully automated decision, a human-made decision (with some computational assistance), and a hybrid decision-making scenario where a human decides based on an automated recommendation. The experiment further contained a "punitive" and an "assistive" (for this distinction: Saleiro et al., 2019) use case for each of these contexts (except for "punitive" in the criminal justice context, see Chapter 3). Finally, relating to the issue of privacy, the study varied whether the data used for decision-making were directly context-related or not.

The experiment was placed in the German Internet Panel (Wave 54 in July 2021), a probability-based German online panel survey (Blom et al., 2021). Each respondent received one vignette for each context, i.e., four vignettes in total. A key finding from descriptive analyses and the mixed-effect regression analyses is that respondents accept hybrid decision-making roughly as much as human decision-making – with some variation across contexts –, and preferred both approaches over full automation. These findings imply that people are not necessarily more skeptical when some level of automation is introduced to making decisions about individuals, but they do prefer a human decider to have the final say. This finding can be considered to support endeavors to use ADM to combine efficiency with (human) responsibility (but see Chapter 6). Furthermore, fairness perceptions and acceptance vary across contexts and are overall lower in the criminal justice and hiring contexts than in the other two contexts. Assistive decisions are overall more accepted and perceived more fair than punitive decisions in the hiring and labor market integration contexts (see discussion in Chapter 3). Finally, respondents are more critical of systems that draw on data that are not directly context-related, supporting the notion of contextual integrity. The study therefore adds concrete empirical insights on the cross-contextual importance of ADM system characteristics, particularly regarding the desired level of automation.

Chapter 4: "Individual acceptance of using health data for private and public benefit: Changes during the COVID-19 pandemic"

Chapters 4 and 5 turn in-depth to public acceptance of data use as relating to privacy as the second focused on challenge of this dissertation. The studies presented in these chapters are particularly motivated by the huge potential of using different kinds of individual data not only for personal benefits (such as providing personal health recommendations), but also for public benefits (such as containing the spread of infectious diseases). However, as argued above, an ethical use of such data requires public acceptance. Applying this requirement to privacy considerations, these chapters present the notion of privacy as "contextual integrity" (Nissenbaum, 2010), which suggests that privacy is maintained if a data flow conforms with the applicable contextual privacy norms of a given social context. According to contextual integrity, to analyze whether this is the case, one needs to define the "parameters" of the data flow: the data type, the acting parties (subject, sender, and recipient), and the transmission principle(s) (which are the "conditions" of data transmission).

The key contribution of Chapter 4 is to show that privacy perceptions can context-specifically vary with large societal disruptions, i.e., that acceptance is bound to time. More concretely, the study presented in that chapter empirically shows that the acceptance towards using individual health data collected on smartphones to contain the spread of an infectious disease increased from before to during the COVID-19 pandemic in Germany. The possibility of such a longitudinal comparison arose from a coincidence: In summer 2019, I happened to run a contextual integrity-based online survey experiment on the conditions of acceptance of data use for my Master's thesis, and the experiment featured a public health-vignette as described above. Th vignettes furthermore varied by data type (health, location, energy use), data recipient (company or public agency), and, given the research interest in acceptance of using data personal and public benefit, varied by these two *purposes* of data use as an additional parameter (while transmission principles were held constant; see details in Chapter 4). In March/April 2020, around the first peak of the COVID-19 pandemic in Germany, this experiment was repeated. These new data allowed for performing longitudinal analyses. The results show that acceptance for health-related scenarios indeed clearly increased, while the acceptance for other scenarios (relating to location or household energy use data) did not clearly increase towards spring 2020.

These findings show that people might change their acceptance of data flows context-specifically in times of exceptional challenges. This result should make policymakers aware of temporal and purpose limits of legitimate uses of novel data-driven technologies in times of

crisis (Vitak & Zimmer, 2020). Furthermore, this study adds to the empirical research demonstrating that contextual integrity is a useful lens for gauging people's acceptance towards data use (e.g., Gilbert et al., 2023; Martin & Nissenbaum, 2017; Martin & Shilton, 2016; Utz et al., 2021). Finally, the results suggest that "purpose" can be a useful additional parameter to consider when describing data flows (see Chapter 6).

Chapter 5: "Attitudes on data use for public benefit: Investigating context-specific differences across countries with a longitudinal survey experiment"

While Chapter 4 researched public acceptance of data use based on contextual integrity and added the fundamental dimensions of *time*, Chapter 5 furthermore adds the fundamental dimension of *space*. More concretely, it combines a contextual integrity-based comparison with longitudinal and international comparisons, while also taking account individual-level predictors of privacy perceptions. Chapter 5 argues and empirically shows that an integration of privacy within the Comparative Privacy Research Framework (CPRF, Masur et al., 2021; see Chapter 5) is useful for uncovering how public acceptance of public benefit data use differs internationally, depending on context. A combination of these two theoretical approaches has already been suggested by Masur et al. (2021); Chapter 4 discusses how exactly to theoretically integrate contextual integrity into a larger comparative research program and shows how this integration can be fruitfully empirically applied.

To this end, I conducted a survey experiment on public acceptance of data use for public benefit in which I varied four data types (health, energy use, location, social media), three recipients (researchers at a university, researchers at a company, public authorities), and three transmission principles (opt-in, opt-out, ethics boards with an opt-out option). Furthermore, I included items that measure several attitudes related to privacy and the provision of public benefits. I ran the survey with a non-probability online access panel in three countries that show different levels of individualism (Hofstede Insights, 2023), which could be related to the acceptance of using individual data for public benefit (Li et al., 2017): Germany, Spain, and the United Kingdom. For longitudinal comparisons, I fielded the survey twice: in December 2022 (where colds are more prevalent and issues with energy supplies might have been rather salient) and May 2023.

While health data were overall the most accepted data type to be used (followed by energy use data), there was stronger international variation particularly with respect to data recipients. German respondents stand out to be relatively less accepting of public authorities for public benefit data use (which might contextualize the findings from Chapter 4). The respondents from

the more individualistic United Kingdom do not turn out to be clearly and consistently more restrictive in terms of accepted transmission principles than those from other countries, countering the initial intuition. Longitudinal variations were not particularly more pronounced for health and energy use data than for the other data types, potentially indicating that the COVID-19 pandemic and fears about an energy crisis did not matter (anymore) at both time points. Further longitudinal research on public issues with varying salience might aid in identifying how salient exactly an issue needs to be to affect privacy perceptions (in this direction see, e.g., Goetzen et al., 2022). Some individual-level variables, such as general privacy concerns, were clearly associated with acceptance, which might hint at the usefulness of measuring more general perceptions along with context-based perceptions (see Chapter 5). Extending a similar argument on general versus specific privacy perceptions at the individual level (Martin & Nissenbaum, 2017), the results overall show that context-specific measurements of privacy perceptions enhance international comparisons by showing differences between countries that general measurements cannot reveal.

References

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction. *Journal of Economic Perspectives*, *33*(2), 31–50. https://doi.org/10.1257/jep.33.2.31

Aitken, M., Tully, M. P., Porteous, C., Denegri, S., Cunningham-Burley, S., Banner, N., Black, C., Burgess, M., Cross, L., Van Delden, J., Ford, E., Fox, S., Fitzpatrick, N., Gallacher, K., Goddard, C., Hassan, L., Jamieson, R., Jones, K. H., Kaarakainen, M., … Willison, D. J. (2019). Consensus Statement on Public Involvement and Engagement with Data-Intensive Health Research. *International Journal of Population Data Science*, *4*(1). https://doi.org/10.23889/ijpds.v4i1.586

Allhutter, D., Cech, F., Fischer, F., Grill, G., & Mager, A. (2020). Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. *Frontiers in Big Data*, *3*, 5. https://doi.org/10.3389/fdata.2020.00005

Andreotta, A. J., Kirkham, N., & Rizzi, M. (2022). AI, big data, and the future of consent. *AI & SOCIETY*, *37*(4), 1715–1728. https://doi.org/10.1007/s00146-021-01262-5

Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments*. SAGE.

Aysolmaz, B., Müller, R., & Meacham, D. (2023). The public perceptions of algorithmic decision-making systems: Results from a large-scale survey. *Telematics and Informatics*, *79*, 101954. https://doi.org/10.1016/j.tele.2023.101954

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. https://fairmlbook.org.

Bednar, K., & Spiekermann, S. (2022). Eliciting Values for Technology Design with Moral Philosophy: An Empirical Exploration of Effects and Shortcomings. *Science, Technology, & Human Values*, 016224392211225. https://doi.org/10.1177/01622439221122595

Bicchieri, C. (2017). *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press.

Blom, A. G., Fikel, M., Gonzalez Ocanto, M., Krieger, U., Rettig, T., & SFB 884 ´Political Economy Of Reforms´, Universität Mannheim. (2021). *German Internet Panel, Wave 54 (July 2021)German Internet Panel, Welle 54 (Juli 2021) (1.0.0)*. GESIS Data Archive. https://doi.org/10.4232/1.13835

Büchi, M., Festic, N., & Latzer, M. (2022). The chilling effects of digital dataveillance: A theoretical model and an empirical research agenda. *Big Data & Society*, *9*(1), 205395172110653. https://doi.org/10.1177/20539517211065368

Carter, P., Laurie, G. T., & Dixon-Woods, M. (2015). The social licence for research: Why care.data ran into trouble. *Journal of Medical Ethics*, *41*(5), 404–409. https://doi.org/10.1136/medethics-2014-102374

Coleman, J. S. (1994). *Foundations of Social Theory*. Belknap Press of Harvard University Press.

Council of the European Union. (2023). *Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world*. https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/

Danaher, J. (2022). Automation and the Future of Work. In C. Véliz (Eds.), *The Oxford Handbook of Digital Ethics* (1st edition, pp. 748–768). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198857815.013.37

Diakopoulos, N. (2020). Transparency. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 196–213). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.11

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 275–285. https://doi.org/10.1145/3301275.3302310

Elahi, M., Jannach, D., Skjærven, L., Knudsen, E., Sjøvaag, H., Tolonen, K., Holmstad, Ø., Pipkin, I., Throndsen, E., Stenbom, A., Fiskerud, E., Oesch, A., Vredenberg, L., & Trattner, C. (2022). Towards responsible media recommendation. *AI and Ethics*, *2*(1), 103–114. https://doi.org/10.1007/s43681-021-00107-7

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) (Text with EEA relevance)* (Number 119). http://data.europa.eu/eli/reg/2016/679/2016-05-04

Feng, S., Mäntymäki, M., Dhir, A., & Salmela, H. (2021). How Self-tracking and the Quantified Self Promote Health and Well-being: Systematic Review. *Journal of Medical Internet Research*, *23*(9), e25171. https://doi.org/10.2196/25171

Forgó, N., Hänold, S., & Schütze, B. (2017). The Principle of Purpose Limitation and Big Data. In M. Corrales, M. Fenwick, & N. Forgó (Eds.), *New Technology, Big Data and the Law* (pp. 17–42). Springer Singapore. https://doi.org/10.1007/978-981-10-5038-1_2

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (2021). *Big data and social science: Data science methods and tools for research and practice* (2nd edition). CRC Press.

Furman, J., & Seamans, R. (2019). AI and the Economy. *Innovation Policy and the Economy*, *19*, 161–191. https://doi.org/10.1086/699936

Gartner. (2024). *Big Data*. https://www.gartner.com/en/information-technology/glossary/big-data

Gilbert, S., Shilton, K., & Vitak, J. (2023). When research is the context: Cross-platform user expectations for social media data reuse. *Big Data & Society*, *10*(1), 205395172311641. https://doi.org/10.1177/20539517231164108

Goetzen, A., Dooley, S., & Redmiles, E. M. (2022). Ctrl-Shift: How privacy sentiment changed from 2019 to 2021. *Proceedings on Privacy Enhancing Technologies*, *2022*(4), 457–485. https://doi.org/10.56553/popets-2022-0118

Gray, C. M., Santos, C., Bielova, N., Toth, M., & Clifford, D. (2021). Dark Patterns and the Legal Requirements of Consent Banners: An Interaction Criticism Perspective. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3411764.3445779

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11296

Gunningham, N., Kagan, R. A., & Thornton, D. (2004). Social License and Environmental Protection: Why Businesses Go Beyond Compliance. *Law & Social Inquiry*, *29*(2), 307–341. https://doi.org/10.1111/j.1747-4469.2004.tb00338.x

Herzog, L. (2021). Algorithmic Bias and Access to Opportunities. In C. Véliz (Ed.), *The Oxford Handbook of Digital Ethics* (1st edition, pp. 413–432). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198857815.013.21

Hofstede Insights. (2023). *Country comparison tool*. https://www.hofstede-insights.com/country-comparison-tool

Hogan, K., Macedo, B., Macha, V., Barman, A., & Jiang, X. (2021). Contact Tracing Apps: Lessons Learned on Privacy, Autonomy, and the Need for Detailed and Thoughtful Implementation. *JMIR Medical Informatics*, *9*(7), e27449. https://doi.org/10.2196/27449

Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. *Yale Journal of Law and Technology*, *18*, 148–216.

Keusch, F., & Kreuter, F. (2021). Digital trace data. In U. Engel, A. Quan-Haase, S. X. Liu, & L. Lyberg (Eds.), *Handbook of Computational Social Science, Volume 1* (1st edition, pp. 100–118). Routledge. https://doi.org/10.4324/9781003024583-8

Kuppler, M., Kern, C., Bach, R. L., & Kreuter, F. (2022). From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. *Frontiers in Sociology*, *7*, 883999. https://doi.org/10.3389/fsoc.2022.883999

Lam, M. S., Pandit, A., Kalicki, C. H., Gupta, R., Sahoo, P., & Metaxa, D. (2023). Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 1–37. https://doi.org/10.1145/3610209

Li, Y., Kobsa, A., Knijnenburg, B. P., & Carolyn Nguyen, M.-H. (2017). Cross-Cultural Privacy Prediction. *Proceedings on Privacy Enhancing Technologies*, *2017*(2), 113–132. https://doi.org/10.1515/popets-2017-0019

Lippert-Rasmussen, K., & Aastrup Munch, L. (2021). Price Discrimination in the Digital Age. In C. Véliz (Eds.), *The Oxford Handbook of Digital Ethics* (1st edition, pp. 467–484). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198857815.013.24

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lutz, C. (2019). Digital inequalities in the age of artificial intelligence and big data. *Human Behavior and Emerging Technologies*, *1*(2), 141–148. https://doi.org/10.1002/hbe2.140

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021). On the Applicability of Machine Learning Fairness Notions. *ACM SIGKDD Explorations Newsletter*, *23*(1), 14–23. https://doi.org/10.1145/3468507.3468511

Martin, K., & Nissenbaum, H. (2017). Measuring privacy: An empirical test using context to expose confounding variables. *The Columbia Science & Technology Law Review*, *18*, 176–218. https://doi.org/10.7916/STLR.V18I1.4015

Martin, K., & Shilton, K. (2016). Putting mobile application privacy in context: An empirical study of user privacy expectations for mobile devices. *The Information Society*, *32*(3), 200–216. https://doi.org/10.1080/01972243.2016.1153012

Masur, P. K., Epstein, D., Quinn, K., Wilhelm, C., Baruh, L., & Lutz, C. (2021). *A comparative privacy research framework*. https://doi.org/10.31235/osf.io/fjqhs

Mayson, S. G. (2019). Bias In, Bias Out. *The Yale Law Journal*, *128*(8), 2218–2300.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Mennicken, A., & Espeland, W. N. (2019). What's New with Numbers? Sociological Approaches to the Study of Quantification. *Annual Review of Sociology*, *45*(1), 223–245. https://doi.org/10.1146/annurev-soc-073117-041343

Mills, K. (2022). Consent and the Right to Privacy. *Journal of Applied Philosophy*, *39*(4), 721–735. https://doi.org/10.1111/japp.12592

Molina, M. D., & Sundar, S. S. (2022). Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media & Society*, 146144482211035. https://doi.org/10.1177/14614448221103534

Molina, M., & Garip, F. (2019). Machine Learning for Sociology. *Annual Review of Sociology*, *45*(1), 27–45. https://doi.org/10.1146/annurev-soc-073117-041106

Mühlhoff, R. (2021). Predictive privacy: Towards an applied ethics of data analytics. *Ethics and Information Technology*, *23*(4), 675–690. https://doi.org/10.1007/s10676-021-09606-x

Mühlhoff, R., & Ruschemeier, H. (2024). *Updating Purpose Limitation for AI: A normative approach from law and philosophy*. http://dx.doi.org/10.2139/ssrn.4711621

Mulligan, D. K., & Nissenbaum, H. (2020). The Concept of Handoff as a Model for Ethical Analysis and Design. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 231–251). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.15

Ngiam, K. Y., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, *20*(5), e262–e273. https://doi.org/10.1016/S1470-2045(19)30149-4

Nissenbaum, H. (2010). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Nissenbaum, H. (2018). Respecting Context to Protect Privacy: Why Meaning Matters. *Science and Engineering Ethics*, *24*(3), 831–852. https://doi.org/10.1007/s11948-015-9674-9

Nissenbaum, H. (2019). Contextual Integrity Up and Down the Data Food Chain. *Theoretical Inquiries in Law*, *20*(1), 221–256. https://doi.org/10.1515/til-2019-0008

Oomen, T., Gonçalves, J., & Mols, A. (2024). Rage Against the Artificial Intelligence? Understanding Contextuality of Algorithm Aversion and Appreciation. *International Journal of Communication*, *18*, 609–633.

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8

Rubinstein, I. S. (2013). Big Data: The End of Privacy or a New Beginning? *International Data Privacy Law*, *3*(2), 74–87. https://doi.org/10.1093/idpl/ips036

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2019). *Aequitas: A Bias and Fairness Audit Toolkit*. http://arxiv.org/abs/1811.05577

Schiff, D. S., Schiff, K. J., & Pierson, P. (2022). Assessing public value failure in government adoption of Artificial Intelligence. *Public Administration*, *100*(3), 653–673. https://doi.org/10.1111/padm.12742

Seubert, S., & Becker, C. (2021). The Democratic Impact of Strengthening European Fundamental Rights in the Digital Age: The Example of Privacy Protection. *German Law Journal*, *22*(1), 31–44. https://doi.org/10.1017/glj.2020.101

Sharon, T. (2017). Self-Tracking for Health and the Quantified Self: Re-Articulating Autonomy, Solidarity, and Authenticity in an Age of Personalized Healthcare. *Philosophy & Technology*, *30*(1), 93–121. https://doi.org/10.1007/s13347-016-0215-5

Shaw, J. A., Sethi, N., & Cassel, C. K. (2020). Social license for the use of big data in the COVID-19 era. *npj Digital Medicine*, *3*(1), 128. https://doi.org/10.1038/s41746-020-00342-y

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, *9*(2), 205395172211151. https://doi.org/10.1177/20539517221115189

Sunstein, C. R. (2023). The use of algorithms in society. *The Review of Austrian Economics*. https://doi.org/10.1007/s11138-023-00625-z

Susser, D. (2019). Notice After Notice-and-Consent: Why Privacy Disclosures Are Valuable Even If Consent Frameworks Aren't. *Journal of Information Policy*, *9*, 148–173. https://doi.org/10.5325/jinfopoli.9.2019.0148

The Ada Lovelace Institute, & The Alan Turing Institute. (2023). *How do people feel about AI? A nationally representative survey of public attitudes to artificial intelligence in Britain*. https://attitudestoai.uk/

Utz, C., Becker, S., Schnitzler, T., Farke, F. M., Herbert, F., Schaewitz, L., Degeling, M., & Dürmuth, M. (2021). Apps against the spread: Privacy implications and user acceptance of COVID-19-related smartphone apps on three continents. *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 70. https://doi.org/10.1145/3411764.3445517

van Berkel, N., Goncalves, J., Hettiachchi, D., Wijenayake, S., Kelly, R. M., & Kostakos, V. (2019). Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–21. https://doi.org/10.1145/3359130

Vitak, J., & Zimmer, M. (2020). More than just privacy: Using contextual integrity to evaluate the long-term risks from COVID-19 surveillance technologies. *Social Media + Society*, *6*(3), 205630512094825. https://doi.org/10.1177/2056305120948250

Waldman, A., & Martin, K. (2022). Governing algorithmic decisions: The role of decision importance and governance on perceived legitimacy of algorithmic decisions. *Big Data & Society*, *9*(1), 205395172211004. https://doi.org/10.1177/20539517221100449

Wenzelburger, G., König, P. D., Felfeli, J., & Achtziger, A. (2022). Algorithms in the public sector. Why context matters. *Public Administration*, 1–21. https://doi.org/10.1111/padm.12901

Weyer, J., Delisle, M., Kappler, K., Kiehl, M., Merz, C., & Schrape, J.-F. (2018). Big Data in soziologischer Perspektive. In B. Kolany-Raiser, R. Heil, C. Orwat, & T. Hoeren (Eds.), *Big Data und Gesellschaft* (pp. 69–149). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-21665-8_2

Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and Automation Bias in the Use of Imperfect Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(5), 728–739. https://doi.org/10.1177/0018720815581940

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Algorithmic Decision-Making and the Control Problem. *Minds and Machines*, *29*(4), 555–578. https://doi.org/10.1007/s11023-019-09513-7

2. Social impacts of algorithmic decision-making: A research agenda for the social sciences[1]

Abstract. Academic and public debates are increasingly concerned with the question whether and how algorithmic decision-making (ADM) may reinforce social inequality. Most previous research on this topic originates from computer science. The social sciences, however, have huge potentials to contribute to research on social consequences of ADM. Based on a process model of ADM systems, we demonstrate how social sciences may advance the literature on the impacts of ADM on social inequality by uncovering and mitigating biases in training data, by understanding data processing and analysis, as well as by studying social contexts of algorithms in practice. Furthermore, we show that fairness notions need to be evaluated with respect to specific outcomes of ADM systems and with respect to concrete social contexts. Social sciences may evaluate how individuals handle algorithmic decisions in practice and how single decisions aggregate to macro social outcomes. In this overview, we highlight how social sciences can apply their knowledge on social stratification and on substantive domains of ADM applications to advance the understanding of social impacts of ADM.

2.1. Introduction

As the increasing use of algorithmic decision-making (ADM) has raised concerns about its social impacts and particularly about new or reinforced social inequalities, research quantifying consequences of ADM for social inequality remains in demand. Understanding the sources and effects of social inequality is one of the core competencies—and responsibilities—of the social sciences. To facilitate a cross-disciplinary discussion and additional research in this area, we use a process model of automated decision-making to highlight when and where social

[1] This chapter was previously published as a paper in the journal *Big Data & Society:*
Gerdon, F., Bach, R. L., Kern, C., & Kreuter, F. (2022). Social impacts of algorithmic decision-making: A research agenda for the social sciences. *Big Data & Society, 9*(1). https://doi.org/10.1177/20539517221089305.

Only small formal edits (such as the formatting of references) were made in comparison to the published paper version.
This chapter arose on the basis of a literature review worked out and submitted by Frederic Gerdon as part of the doctoral studies program of the Centre of Doctoral Studies in Social and Behavioral Sciences (CDSS) at the Graduate School of Economic and Social Sciences (GESS) at the University of Mannheim, which was subsequently substantially edited.
The Appendix for this chapter is available in Chapter 7.1. References to the Appendix begin with the letter "A".

inequality may arise from ADM systems. Focusing on the data generation, data analysis, and implementation of ADM systems, we suggest a roadmap and research avenues for social scientists interested in answering the increased calls for the study of social impacts of ADM.

ADM is used as an umbrella term for a variety of systems that are used to assist or replace human deciders (see AlgorithmWatch, 2019). For instance, judges may use recidivism risk scores predicted by algorithms trained on decades of criminal records to determine bail decisions (Stevenson, 2018), mortgage lenders can base interest rates on default risks predicted by algorithms (Bartlett et al., 2019), and public social services may draw on algorithmic support to make decisions on financial aids (Lind and Wallentin, 2020).

ADM systems are based on predictions from models that process historical data, which contain both inputs ("predictors," "features," "independent variables," "x") and one or more outputs ("label," "outcome," "dependent variable," "y"). The goal of data processing is to "learn" associations between inputs and output from the past to make predictions where the output is still unknown. Predictions are then used to decide whether some action should be taken or not. While our focus is on ADM systems that draw on some automated learning, these systems can generally vary in the complexity of how inputs determine outputs—including simple threshold rules for single input variables—, as well as in the extent to which humans are involved in the final decisions (see related definitions and surrounding discussions in AlgorithmWatch (2019) and European Parliament, Directorate General for Parliamentary Research Services et al. (2019)).

ADM seems promising as an alternative to (pure) human decision-making, as human decisions may be just as or even more biased than ADM, with ADM potentially having higher efficacy (Miller, 2018), transparency, and accountability (Mayson, 2019). However, concerns have been raised about algorithms exacerbating social inequality and discriminating against certain societal groups, for example, due to learning biases from historical training data (e.g. Zou and Schiebinger, 2018).

A recent example of an ADM system that raised such concerns is a system that has been tested by the Public Employment Service Austria (AMS). This system classifies job seekers into three groups, depending on their predicted chances to find a new employment (Lopez, 2019). The system builds groups of feature combinations based on, for example, gender, age, nationality, education, and previous contact with AMS, and predicts short- as well as long-term chances of integration into the labor market (Gamper et al., 2020). The assignment to a group can influence which kind of assistance is given to an individual: for instance, Kopf (2019) argues that while all job seekers are supported by the employment agency, individuals with low

chances for re-employment would usually profit more from intensified assistance than from, for example, qualification measures. However, concerns arise if, for example, women, with all other characteristics held equal, had lower scores than men. Such concerns sparked discourse regarding the discriminatory potentials of this system (see Kopf, 2019; Lopez, 2019).

While similar decisions have been made without algorithmic assistance in the past, novel ADM systems have specific features that create new and amplify old challenges. First, these systems make use of new technical devices and facilities, unprecedented amounts of data, and novel techniques of data analysis that allow deciders to employ new decision-making strategies to approach old problems. Second, ADM systems constitute socio-technical systems that entail machines and humans (Selbst et al., 2019): they are pervaded by human decisions and cultural notions that we need to scrutinize (Seaver, 2019) to understand potential detrimental effects for society.

Scholars from various disciplines have called for examining algorithmic outcomes to avoid or mitigate undesired consequences of ADM (Kusner and Loftus, 2020; Zou and Schiebinger, 2018). Previous research from computer science (Mehrabi et al., 2019), legal studies (Wachter, 2020), and philosophical (Mittelstadt et al., 2016) perspectives discussed algorithmic, structural, and ethical problems with ADM. Joyce et al. (2021) and Liu (2021) provide a general overview of sociological perspectives on related artificial intelligence.

Drawing on previous literature, our own work on ADM systems, and a previously developed big data processing model (Weyer et al., 2018), we here highlight areas in which social scientists can (and should) use their expertise to contribute to the debate of equitable ADM. We show how a social science perspective on data generating processes, analytical challenges, and implementations can help anticipate (undesired) social impacts of ADM.

2.2. A process model of ADM

To understand how social inequality, here defined as “the unequal distribution of valued resources, opportunities, and positions among the members of a population in a given space and time” (Otte et al., 2021: 362), can arise or be amplified through ADM systems, attention needs to be paid to the distribution of opportunities and restrictions leading into and out of the ADM system. While inequality not always necessarily constitutes injustice, it is oftentimes considered an undesired property of ADM systems (see Kuppler et al. (2021) for a detailed discussion on distributive justice in ADM).

A major path via which algorithms—just like human-made decisions—may affect such distributions is discriminatory behavior. By *discrimination* we mean "an action or practice that excludes, disadvantages, or merely differentiates between individuals or groups of individuals on the basis of some ascribed or perceived trait" (Kohler-Hausmann, 2011), such as gender and race. Computer science research on *Fair Machine Learning* (Fair ML) aims to tackle discrimination by investigating how algorithms can be designed to make predictions *fair*, that is, without "prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics" (Mehrabi et al., 2019: 1).

Social implications of ADM systems do not only arise through biased predictions, but from implementations of decisions within a social environment. Even if fairness on the prediction level is present, disparate impact can occur (Feldman et al., 2015), for example, because impacts of ADM decisions are hard to factor into the preceding data analysis (Kusner et al., 2019), or because human deciders rely disproportionally on the ADM-based recommendations. Recent research extended the notion of fairness to include actual inequality effects resulting from algorithmic discrimination in the social context in which it is placed (see section "Data preparation and analysis—from fairness in algorithmic output to fairness in social impact") and to frame such effects in terms of causal impact (Kasy and Abebe, 2021). To investigate these social impacts of ADM, a social science perspective becomes particularly valuable.

To discuss how ADM systems may impact social inequality, we adapt a "big data process model" (Weyer et al., 2018: 74), by breaking down the ADM process into three steps (based on Weyer et al., 2018). We discuss how social inequalities may be shaped in each step:

- **Data generation:** Data bases may be biased, for example, due to historical discrimination against social groups or incomplete data availability.
- **Data preparation and analysis:** An algorithm may adapt or even reinforce biases that are already present in the data. This includes the choice and construction of variables that serve as the input for the algorithm, the choice of fairness metrics for identifying biases, and the choice of bias mitigation measures.
- **Implementation:** The way ADM systems ultimately affect inequality depends on their implementation within contexts (for contexts, see Weyer et al., 2018). Human decision-makers, if present, might handle algorithmic recommendations differently, and those affected by ADM-based decision might differ in their reactions. This step also includes how single decisions aggregate to social outcomes and how human behavior feeds back into the data.

The different forms of biases may propagate through the ADM process and can be reinforced or mitigated along the way.

We briefly illustrate this three-step model with an example. The data basis of the Austrian AMS model that classifies individuals according to their labor market integration chances likely reflects historical unequal labor market participation rates, for example for women (*data generation*). The data analysis itself potentially manifests this bias, if, for example, the model resulting from training assigned women—all else being equal—lower employability scores than men (*data analysis*). Then, we also need to ask about the actual consequences of the system (*implementation*). For example, Kopf (2019) argues that women were ultimately under-represented in the lowest employability group. However, based on data reported by Gamper et al. (2020) on group assignment at the beginning of unemployment, Allhutter et al. (2020) note that the share of women was roughly double the share of men in the lowest employability group. Allhutter et al. (2020) suspect that varying conclusions may result from considering different models, time frames, or (sub-)populations. We therefore need to carefully scrutinize whether or to which extent women would be effectively disadvantaged by the system in practice. Furthermore, we need to know under which conditions human deciders adopt or disagree with the predicted scores and how job seekers subsequently change their job search behavior (also see Allhutter et al., 2020). In a feedback loop, such behavior may flow into the data basis for future model building. Finally, we need to understand how potentially discriminatory decisions on the individual level will manifest on the macro-level in the long term.

2.3. Sources of bias and social impacts along the ADM process

In the following subsections, we explore each of the potential sources of inequalities that we described in the previous section. We follow the ADM process model step by step and show how the social sciences have already contributed to researching problems related to potentially discriminatory ADM. We also identify promising research questions related to social inequality impacts of ADM that the social sciences could investigate.

2.3.1. Data generation – historical bias and selective participation

Algorithms can be trained on a variety of data, ranging from governmental records, such as individual labor market histories, and survey data to digital data created through individuals' online activities and interactions with digital devices. If bias is present in the data sets used, unfair or discriminatory outputs may result. While using data to make predictions is not

exclusive to algorithms, specific aspects in data generation require heightened attention in ADM. In this section, we focus on exemplary problems in biased data sets (Rodolfa et al., 2021; Sen et al., 2019) and refer to previous literature for more general introductions (e.g. Groves, 2004).

Two major sources for bias can arise in the data generation step (for a detailed overview, e.g. Mehrabi et al., 2019). The first source covers all those cases where data used to develop an ADM system contains historical discrimination. That is, an outcome is unequally distributed between individuals with different characteristics such as gender and race, after controlling for other characteristics of the individuals that cause variation in the outcome. The mechanisms creating such discrimination are manifold and depend on the concrete context, for which social sciences can provide domain-specific knowledge. In the labor market example above, historical labor market records may show that, after controlling for other individual characteristics, women had worse re-employment chances than men after losing their job in the past. Similarly, historical criminal records may insinuate that, all other characteristics of an individual held constant, black offenders had higher risks of recidivism once released from jail than white offenders.

The second source comprises biases due to selective participation and representation of social groups in data generation and collection (see Mehrabi et al., 2019). Selective participation can introduce a mismatch between the data that is used for training a prediction model and the ultimate target population that is affected in its application. If important subgroups are misrepresented in the training step, high error rates (and ultimately incorrect decisions) may result once the model is confronted with the target data in the deployment phase (Daumé III, 2017). Unequal participation in the generation and collection of digital data constitutes a particular challenge for those ADM systems that rely on them. Previous research has shown that the use of information and communication technology is often selective, for example, with respect to digital skills, age, and socio-economic status (Hargittai and Hsieh, 2013; Lutz, 2019). Models trained on such data may thus find relationships that hold only for the group of individuals using such technology. That is, individuals who are already disadvantaged because they do not use specific digital technologies could also be disadvantaged by an ADM system if the system cannot consider their behaviors and preferences (Lerman, 2013).

Social scientists are needed to identify coverage issues due to differences in social characteristics, digital skills, trust, and privacy concerns in the data collection process. Social scientists, and particularly survey researchers, can contribute to tackling representation issues of training data by applying weighting methods or improving data collections. Designing,

conducting, and evaluating various forms of data collection processes such that the acquired sample resembles the target population of interest is a core task of survey research. Recent work in survey research investigates coverage and representation issues in the context of digital data and data collected via smartphones and sensors and introduces methods for adjusting non-random samples (Baker et al., 2013; Japec et al., 2015; Keusch et al., 2020). This includes, for example, pseudo-weighting approaches that allow to correct for biases due to selective participation by leveraging information from an auxiliary reference sample (Elliott and Valliant, 2017). Note that such techniques are closely related to adaptation approaches that have been proposed in computer science to account for covariate shift between training and test data (Daumé III, 2017). Weighting techniques from survey research could similarly be utilized to adjust (survey- and non-survey-based) training data if a suitable reference sample that resembles the target population can be found and both datasets include structural information about the entities of interest (e.g. socio-demographic attributes of individuals or make and type of digital devices). While applying pre-processing techniques such as re-weighting may not be feasible in all ADM contexts, recent work on post-processing predictions exemplifies how ideas from survey research (mass imputation; Yang and Kim, 2020) and computer science (multi-calibration; Hebert-Johnson et al., 2018) can be combined to tackle misrepresentation in training data (Kim et al., 2022).

In addition to historical bias and representation bias, ADM can be adversely affected by using mismeasured variables. Using proxy variables such as healthcare costs as a proxy for health needs can obscure differences in the true outcome of interest when, for example, black individuals generated lower healthcare costs than white individuals once the true health status is held constant (Obermeyer et al., 2019). Such measurement bias can be directly connected to social science work on measurement errors and thus represents one example of how social science already contributes to researching social impacts of ADM (Boeschoten et al., 2020; Jacobs and Wallach, 2021). Moreover, the contextual nature of some individual characteristics and behaviors may not be amenable to quantification and therefore, ADM system cannot cover these characteristics appropriately, such as context-sensitive combinations of protected attributes relating to intersectional discrimination (Mann and Matzner, 2019). These may be only subtly present in social interactions, lead to discrimination, and be insufficiently captured in automated analysis (see section “Data preparation and analysis—from fairness in algorithmic output to fairness in social impact”).

To conclude, biases in datasets gain renewed momentum in the context of ADM for three reasons. First, it is likely that the increased quantity of predictions and decisions that a model

can make compared to a human decider will intensify inequalities that are already present in the data. Second, relying on ADM systems increases the importance of patterns in the data in comparison to the importance of heuristics of human decision-makers (but see section "Implementation—micro-interaction with ADM and macro-social outcomes"). Third, the amount of data produced and used in ADM systems has considerably increased with the advance of digital technologies. Social sciences can apply methodological and domain knowledge (a) to better understand how situation-specific biases may be present already in the data collection stage of ADM processes and (b) to explore how advances in survey methods can be used to correct such biases.

Research avenues:

- How can we utilize methodological advances in survey research to correct biases in data due to selective participation and improve the data input for ADM?
- How can we extend research on digital divides and technical competencies to study inequality in being covered by ADM systems (Lutz, 2019)?

2.3.2. Data preparation and analysis – from fairness in algorithmic output to fairness in social impact

Data preparation and analysis is the step in which developers work with data and construct and refine algorithms. This process entails manifold interpretations and decisions, including, ideally, considerations on how to produce fair outputs. In this section, we outline how algorithms may produce biased predictions due to biased data or decisions during the modeling process. We give a brief survey on the mainly computer scientific research field of Fair ML. Then, we show how a social science perspective can contribute to the identification of meaningful fairness criteria in social contexts, particularly when considering social impacts of ADM systems on macro-level social outcomes and public perceptions of fairness.

Approaches in fair machine learning

Fair ML is a research field that investigates the fairness of machine learning algorithms. This research branch produces important contributions by proposing fairness metrics and improving algorithm design such that individuals are less likely to be assessed by characteristics that should not matter for taking a decision ("protected attributes"). Such steps are necessary as otherwise, algorithms might reproduce existing biases or exacerbate inequalities even when the data sources are unbiased (Aghaei et al., 2019). For example, this is the case when prediction

error rates differ between groups (Rodolfa et al., 2021). Various steps in the construction of variables ("feature engineering"), such as how race is coded, may also introduce biases (Rodolfa et al., 2021).

Fairness definitions oftentimes are formal measures based on rates of correct and incorrect predictions for individuals of different social groups for which non-discrimination should be ensured (Corbett-Davies and Goel, 2018). For instance, an algorithm might be tasked with assigning job seekers into two classes: those with high or low chances of finding a new job. The algorithm could be trained with data that show past job market outcomes of job seekers. The algorithm tries to combine the characteristics of these individuals to build a model that predicts chances of labor market integration as accurately as possible. The prediction outputs can be evaluated by comparing the predicted with the observed outcomes in the data.

Several fairness definitions specify how error rates should be balanced across different groups of individuals. As an example, an algorithm may be considered fair if it results in equal false negative rates (*equal opportunity*; Hardt et al., 2016) or equal false positive rates (*predictive equality*; Rodolfa et al., 2021) between members of different groups (e.g. men and women). A related definition is *equalized odds*, which means that members of different groups experience both false negatives and false positives at the same rate (Hardt et al., 2016). This principle can be applied to various error metrics and their combinations (e.g. false discovery rates, false omission rates, accuracy), resulting in a variety of group-based fairness notions. Furthermore, subgroup fairness (Hebert-Johnson et al., 2018) and individual fairness (Dwork et al., 2012) notions have been proposed that expand beyond comparisons of error rates on the group level (e.g. by considering intersections of gender and race).

Research in Fair ML resulted in various methods and tools that may mitigate biases at different stages of the modeling pipeline (Berk et al., 2018; Mehrabi et al., 2019). *Pre-processing* techniques can be used to eliminate sources of unfairness in the data prior to model training, for example, by removing dependencies between legitimate factors and protected attributes (Johndrow and Lum, 2017). *In-processing* techniques aim at modifying the model building process itself, for example, by introducing fairness constraints in the objective function (Berk et al., 2017). *Post-processing* methods may be used to alter the output of a prediction algorithm after model training, for example, by "nudging" predictions towards the true outcome for subgroups where high errors are observed (Kim et al., 2019). These procedures have been shown to mitigate different notions of unfairness at the prediction stage of the ADM process in several applications (Friedler et al., 2018).

Competing fairness definitions and the importance of social context

Many fairness definitions and correction methods have been proposed, and it may prove difficult to choose the definition and technique that is the most appropriate for the given prediction task (see Makhlouf et al., 2020). Moreover, some fairness definitions were found incompatible with each other and in conflict with overall accuracy (Berk et al., 2018), while Selbst et al. (2019) discuss as "formalism trap" whether an appropriate mathematical definition of the complex concept of fairness was even possible.

One major concern in handling fairness boils down to the question: is it better to ignore specific individual characteristics such as gender or race altogether, or should we try to balance, for example, error rates between groups based on these features (Corbett-Davies and Goel, 2018)? Neglecting group membership may, for instance, lead to aggregation bias, meaning that one model is used for all groups although the model works worse for some of the groups (Suresh and Guttag, 2020). In the case of race, Benthall and Haynes (2019) discuss that ignoring race would still possibly lead to racial discrimination as effects of correlates relevant to race and inequality persisted. However, explicitly including race reified this category. Instead, they propose a third alternative of algorithmically finding latent categories that mirror racial segregation. From a social science perspective, this discussion extends to the question which features are considered protected attributes in the first place. While gender and race represent attributes that are commonly considered sensitive and are protected by legislation, sociological research on the intergenerational transmission of resources and education (e.g. van Doorn et al., 2011) raises questions on which concepts purely measure individual merit and which attributes may constitute "hybrid" characteristics that are (at least partly) socially inherited. Relatedly, using traditional concepts of gender and race for defining protected groups will fail to account for individuals who do not find themselves represented by those categories. Particular attention needs to be paid to intersectional discrimination that may disadvantage individuals based on multiple protected attributes at the same time, for example, gender *and* race: automated analysis of large data bases may contain a plethora of potential protected attributes, suggest new associations between these attributes, and thereby statistically form new groups of people that may then be discriminated against (Mann and Matzner, 2019). Social scientists can scrutinize data, analytical decisions, and outputs with respect to intersectionality in different contexts of ADM applications, and suggest groupings of protected attributes that are contextually relevant.

Eventually, fairness is context-specific. Among others, the choice of a fairness metric may depend on the outcome and which resources will be distributed (Kuppler et al., 2021). The idea of social context is not new in the realm of computer science (Selbst et al., 2019) and is part of

pursuing "algorithmic realism" (Green and Viljoen, 2020). For instance, this entails the question whether a system aims at helping or punishing individuals, which implies an emphasis on disparate distributions towards either false negatives (incorrectly excluded from a positive intervention) or false positives (incorrectly included in receiving a negative intervention) (Saleiro et al., 2019). This discussion extends to the broader question on the just or desired allocation principle in a specific ADM application context. Sociological discourse on distributive justice can enlarge computer science's decision space when it comes to designing allocation systems and selecting bias correction techniques by highlighting which design choices may serve which principle (Kuppler et al., 2021).

Empirical findings on fairness perceptions

Fairness perceptions matter to ADM development for two reasons. First, they are relevant to design socially acceptable ADM systems. Second, the individual evaluation of an algorithm may contribute to how that individual interacts with and acts upon the decision of the ADM system, thereby potentially shaping inequality outcomes.

A comprehensive literature review on fairness perceptions on ADM concludes that perceptions strongly depend on context characteristics, such as the features used by the algorithm and the purpose of the algorithm (Starke et al., 2021). Participants in one study applied some justice principles relevant for human decision-making also to algorithmic decisions, but the concrete style of explaining the algorithm impacted justice perceptions only when the respondent was exposed to multiple styles (Binns et al., 2018). In addition, general trust in ML systems and the features used and *not used* are relevant for fairness judgments (Dodge et al., 2019). Empirically validated frameworks that define process features relevant to fairness perceptions (Grgić-Hlača et al., 2018) can build the basis for practically applicable guidelines for designing contextually fair algorithms, which is why such work is particularly attractive for future work.

Whether and how individual characteristics such as socio-demographic attributes like age and gender interact with, for example, explanation styles and the impact of the decision situation needs further research and possibly depends on individual affectedness (Pierson, 2018). Experimental evidence suggests that fairness ratings depend on whether respondents' characteristics are involved in the algorithmic decision, and conservatives were found to be more accepting of using individual characteristics in computer-assisted bail decisions than liberals (Grgić-Hlača et al., 2020).

In conclusion, building fair algorithms is a prerequisite for arriving at fair predictions and, subsequently, decisions. Software toolkits that assist in assessing the fairness of algorithms are available (Bellamy et al., 2019; Saleiro et al., 2019). To advance Fair ML, we can intensify research on fairness perceptions in concrete ADM processes and strengthen the link between distributive justice principles and (fairness in) automated allocation systems (Kuppler et al., 2021). Moreover, Starke et al. (2021) suggest systematizing situation-specific factors such as whether a decision is high-stake or low-stake and the area of application (e.g. decisions in the criminal justice system or hiring) that may shape fairness perceptions.

Research avenues:

- How do contextual information (the purpose of an algorithm) and explanations of algorithm function shape fairness perceptions of ADM processes? How do individual characteristics influence fairness perceptions?
- How can fairness assessment and mitigation techniques be implemented and extended beyond equalizing error rates towards serving context-specific allocation principles?
- How can social science provide domain-specific knowledge to define appropriate, non-discriminatory outcomes for an ADM system, including the consideration of externalities?

2.3.3. Implementation – micro-interaction with ADM and macro-social outcomes

Researchers from different disciplines have demonstrated that the used data and the data analysis performed do not suffice for explaining the social impacts of algorithms (Cowgill and Tucker, 2020; Kleinberg et al., 2018). The question whether the use of an algorithm will produce fair outcomes is not only a question of the fairness of predictions and decisions, but also of their actual impacts in a social environment (Kusner et al., 2019). In fact, "[…] even fair decisions at the machine learning level may not lead to equitable results in society and the decision-making process may need to compensate for these other inequities" (Rodolfa et al., 2021:304).

The notion of disparate impact helps to understand the difference between the output of an analysis and subsequent societal consequences. Disparate impact refers to effects of practices that result in unintended disadvantages for groups of individuals with certain characteristics (Barocas and Selbst, 2016). For example, even if no discrimination is intended, individuals may be affected differently due to their characteristics. Implementing notions of disparate impact in algorithms is one step to practically achieve fairer results (e.g. Feldman et al., 2015), and

computer scientific research developed and applied such extended notions of fairness. Among those are suggestions to ascertain fairness by optimizing how an outcome of interest is expected to be affected in the long term (Liu et al., 2019), choosing fairness metrics that satisfy specific policy goals (Rodolfa et al., 2020), and engaging with the needs of affected population groups to adjust analyses in feedback loops (Noriega-Campero et al., 2018).

These outcome-oriented approaches seem most promising for the development of an encompassing understanding of fairness that contributes to contextually appropriate assessments. A social science perspective can help to analyze the implementation process of ADM in social contexts and to understand interaction processes at the micro-level between algorithms, affected individuals and, in some cases, human deciders, and their macro-social outcomes.

Human versus algorithmic predictions: Empirical evidence from real-life cases

Studying impacts of ADM systems in real-life cases faces the same challenge as other observational social research: it remains unclear *what the outcome would have been* had a decision been taken without an algorithm (see Holland, 1986). Although methods for tackling such problems of causal inference are well known to social scientists, there is so far only little research applying them to the study of social impacts of ADM (such as Cowgill and Tucker, 2017). One notable exception are recidivism prediction algorithms in the USA, where studies find mixed effects regarding the reduction of crime and racial disparity through algorithms (Berk, 2017; Kleinberg et al., 2018; Stevenson, 2018). Stevenson (2018) suggests that even if algorithms made better predictions, they might not necessarily improve relevant outcomes, and judges' own biases could lead to a sub-optimal use of algorithmic predictions, emphasizing the need to study how human decision-makers rely on algorithms.

Previous research reports mixed findings regarding differences in the accuracy of predictions between algorithms and human deciders. Some find that humans perform worse than algorithms (Green and Chen, 2019), while others find comparable accuracy and fairness in predictions (Bansak, 2019; Dressel and Farid, 2018; Tan et al., 2018). In addition, in the context of recidivism prediction tasks, it is likely that the characteristics of the defendants will matter: given a risk assessment, human deciders deviated more strongly to unfavorable predictions for black defendants than for white defendants (Green and Chen, 2019).

Overall, the question whether an algorithm can outperform a human decider will have to be evaluated on a case-by-case basis. In those cases where the final decision remains in the

hands of a human decider, we also must consider whether and how a human decider is involved and *influenced* by an algorithmic decision or recommendation.

From "automation bias" to "algorithmic aversion" – How human deciders (do not) adopt algorithmic recommendations

Many ADM systems, particularly those in which the stakes are high, involve a human decider who may consider algorithmic predictions in her decisions. While a machine-assisted decision may deviate from a purely human decision, human deciders will not always follow the algorithmic recommendation. Therefore, potential biases inherent in the algorithmic prediction may be alleviated or corrected by human deciders, but humans may also introduce or reinforce discrimination in the process. The interaction of a human decider with an ADM system is likely complex and requires detailed investigation. Research on "human factors" and human-computer interaction provides valuable work that can be applied to the study of ADM systems (Zerilli et al., 2019). The communication between algorithmic recommendations, human deciders, and affected individuals is likely shaped by the complexity of the underlying model. If we want the involved individuals to understand how ADM systems arrive at decisions and to uphold accountability, we need algorithms that can be explained—either by making use of inherently interpretable methods or by employing post-hoc interpretation techniques (Molnar, 2019). Differential social impacts may arise, for example, if explanations are differently effective for social groups and shape the reliance on or compliance with algorithmic recommendations.

Here, we focus on the specific problem of circumstances under which a human decider will be more likely to adopt (or override) an algorithmic recommendation. Two central phenomena characterize human reliance on algorithmic predictions: automation bias and algorithmic aversion. Automation bias refers to errors stemming from human reliance on automated systems such as ADM: while errors of omission refer to cases where someone relies on a flawed algorithmic prediction (false negatives), errors of commission refer to falsely assuming an error (false positives) (Wickens et al., 2015).

Empirical evidence for automation bias has been found, for example, in clinical decision support systems (Goddard et al., 2012). Research shows that factors such as trust and own experience shape reliance on automated systems (Burton et al., 2019; Cepera et al., 2018; Lee and See, 2004; Logg et al., 2019; Weyer et al., 2018). Moreover, Parasuraman and Manzey (2010) note that errors of commission are lower when a system serves information integration and analysis as compared to providing concrete recommendations for actions.

There also is evidence for algorithmic aversion, that is, individuals becoming less likely to rely on algorithmic predictions after experiencing false predictions (Burton et al., 2019). Experimental studies show that confidence in algorithms is lowered when algorithms make a mistake (Dietvorst et al., 2015). Moreover, humans tend to adjust their predictions more often based on human advice than based on statistical forecasting (Önkal et al., 2009). However, Grgić-Hlača et al. (2019) report that machine advice does affect participants' predictions in the case of criminal recidivism and Araujo et al. (2020) even find evidence for algorithmic *appreciation*, that is, a preference for automated decisions compared to human decisions.

Empirical studies of actual adoptions of algorithmic recommendations and consequences for inequality are scarce and mostly investigate the judicial context. Results show that higher recidivism scores lead to longer sentences but judges also seem to rely less on risk scores over time (Stevenson and Doleac, 2019). If risk scores are transformed into a categorical scale (low, medium, and high risk), individuals who are placed just above a threshold value receive on average one to four additional weeks of detention before trial compared to those placed just below the threshold (Cowgill, 2018). Again, individual characteristics seem to play an important role as this effect was more pronounced for black defendants than for white defendants.

Social sciences add domain-specific knowledge and tools for understanding macro-level outcomes of human-ADM interactions

Social sciences contribute to developing fair ADM systems by bringing in their domain-specific expertise on individual behavior and social practices across social environments. A thorough analysis of ADM impacts requires such domain-specific knowledge, for example, on labor market behavior. For instance, social sciences can help to answer questions such as: how will an individual adjust her behavior when an employment agency employee decides for a specific (or no) training program based on an ADM recommendation? Could this decrease motivation as an individual feels more constrained by algorithmic decisions than by human decisions? First research documents how individuals evaluate algorithmic decisions compared to human decisions, finding both similarities and differences (Araujo et al., 2020; Binns et al., 2018; Plane et al., 2017; see section "Data preparation and analysis—from fairness in algorithmic output to fairness in social impact"). Due to potential context-dependency, more research is needed to gain a better understanding of human interpretations of algorithmic decisions.

Social sciences can help to predict and to assess outcomes of ADM processes by providing domain-specific knowledge in the fields of the ADM application (Bertelsmann Stiftung, 2020).

This includes knowledge on which goals human decision-makers may follow, which factors they consider, and how these differ from the purely ADM process (see Kleinberg et al., 2018). This also entails how characteristics of the individual and its environment shape the severity of the impact of a decision derived by an ADM (see Abebe et al., 2020). Moreover, as institutional and organizational contexts may react to the implementation of ADM systems in (unintended) ways (Selbst et al., 2019), social sciences also provide methods and previous research to understand established practices in specific contexts and anticipate potential reactions. These methods can also be used to investigate established practices that shape ADM implementation. In the case of comparing algorithmic and human decisions, understanding the goals programmed into an algorithm and analyzing the goals human deciders consider when taking a decision is crucial (Kleinberg et al. 2018; Stevenson 2018).

Additional to in-depth case studies that investigate ADM in concrete contexts (e.g. Elish and Watkins, 2020), experimental research and observational studies along the lines of the research presented in this section improve our understanding of interactions within ADM systems. To show whether and how algorithmic literacy and subsequent behavior impact social inequality, we need to study how these competencies, awareness (Gran et al., 2021), and knowledge related to algorithms are distributed across social groups—for example, by age and education (Fischer and Petersen, 2018)—, and then how this knowledge translates into behavior (e.g. adjusting to the algorithm's "preferences," see Freeman Engstrom et al. 2020).

Furthermore, social scientists can contribute to investigating how individual decisions made by ADM systems influence inequality and discrimination on the societal macro-level, that is, how single decisions accumulate to overall patterns of inequality in a population akin to the micro-macro model of sociological explanation (Coleman, 1994). Agent-based modeling (ABM) is a promising method to study how interaction on the micro-level produces macro-outcomes as it allows researchers to simulate, for example, interactions of technical and social elements of an ADM process (Gilbert, 2008). ABM could be used to model an interactional setting with three types of agents: affected individuals, algorithms, and deciders. Each affected individual has, for example, certain demographic characteristics and attitudes towards technology. Results of algorithmic predictions based on different fairness strategies can be presented to the decider. The human decider—if applicable—may consider the algorithmic decision and the affected individual's characteristics to arrive at a decision and weigh both according to her own experience, for example. The affected individual may then adapt her behavior according to her characteristics and the decision.

ABM presents many advantages as it allows researchers to represent the interplay of human and machine actors (see Calero Valdez and Ziefle, 2018) in ADM systems and dynamics over time. For example, fairness implications may only show when considering *long-term* effects on macro-outcomes in the population (Heidari et al., 2019; Liu et al., 2019), and simulations can be run for hundreds or thousands of rounds. Furthermore, ABM responds to calls for a stronger integration of the social environment of ADM systems to grasp their impact appropriately. ABM has already been used to study the governance of socio-technical systems (Adelt et al., 2018), and Cruz Cortés and Ghosh (2019: 3), for example, apply ABM in the context of criminal recidivism risk for a "[s]ystematic analysis […] [which] implies analyzing the data generating process, the decision-making stage, and its consequences all under the same framework." In conclusion, simulations are promising tools to assess macro-level outcomes of ADM applications from a social science perspective.

Research avenues:

- How do individuals adapt behavior preemptively or as a reaction towards an algorithmic decision? Which individual resources affect interactional behavior?
- Which situational and individual characteristics determine reliance on ADM systems across social contexts?
- How do these individual decisions and interactions aggregate to macro-social inequality outcomes, and how can researchers study such impacts using simulation techniques?

2.4. Conclusion

Synthesizing several theoretical and empirical advances in the research on the consequences of ADM systems for social inequality, this paper provides an overview geared towards social science research, with a focus on data generation, analysis, and implementation challenges. For each part of the ADM pipeline, we highlighted possible inequality issues and how social sciences can contribute to their study. Put briefly, (1) the data used may be biased, (2) the algorithm itself might rely on contextually problematic conceptualizations and formalizations of fairness—or may not consider fairness at all—, and (3) the inequality outcomes depend on concrete interactional settings that can result in cumulative disadvantages, particularly for those who have been historically disadvantaged. We summarize potential sources of inequality, related social science topics, example papers, and research avenues in Table A2.1 (in Appendix 7.1).

Social sciences can draw on established research to contribute to these efforts by bringing in expertise on methods, concrete social contexts, and human (inter)action to investigate how ADM systems affect (macro-)social inequality outcomes. To study algorithmic bias, social scientists can contribute to developing context-aware fairness notions, and to evaluate the scale of actual impacts that ADM systems produce in practice.

Social science research on inequality and ADM systems as well as interactions between algorithms and humans goes far beyond what we were able to cover here (e.g. Joyce et al., 2021; Liu, 2021). Other challenges range from, for example, accounting for the agency of algorithms (Lange et al., 2019), social and political challenges with respect to regulation (Mittelstadt, 2019), privacy (Anthony et al., 2017), and governance (Danaher et al., 2017), to artificial intelligence shifting power relationships (Kalluri, 2020), or other social impacts beyond inequality outcomes. Moreover, we need to understand the contexts in which ADM are applied, including established practices and interactions between human and technical elements. To this end, researchers can draw on a variety of qualitative approaches, such as ethnography (see, e.g. Lange et al., 2019) within the respective social contexts or expert interviews with individuals involved in ADM implementation.

Finally, "impacts" of ADM on social inequality do not necessarily equal to *increases* in disparities. Human decisions are oftentimes also biased and flawed, and algorithmic decisions could potentially display *less* bias than humans (Mayson, 2019) and reduce social inequality overall. However, social implications need to be thought of when designing and implementing ADM applications. We hope that this paper will assist in the development of a research framework and that it will help to enhance concrete guidelines for creating socially responsible ADM systems. Such guidelines are currently discussed and urgently needed as the supervision, assessment, and even necessity of approval of ADM is an ongoing policy debate (e.g. AlgorithmWatch, 2019).

Acknowledgments

We thank the participants of the doctoral colloquium of the Center of Doctoral Studies in the Social and Behavioral Sciences (Sociology) at the University of Mannheim, the members of the Kreuter-Keusch research lab, as well as the anonymous reviewers for their helpful feedback and comments on the paper. We acknowledge funding from the VolkswagenStiftung for the project "Consequences of Artificial Intelligence for Urban Societies" (CAIUS) and the Baden-Württemberg Stiftung for the project "Fairness in Automated Decision making" (FairADM).

The publication of this article was funded by the Mannheim Centre for European Social Research (MZES). This work was supported by the University of Mannheim's Graduate School of Economic and Social Sciences. FG led the development of the paper, RB and CK contributed to research and writing, FK conceptualized the underlying research projects and contributed to writing.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Baden-Württemberg Stiftung, Volkswagen Foundation; The publication of the article was funded by Mannheim Centre for European Social Research (MZES).

References

Abebe, R., Kleinberg, J., & Weinberg S. M. (2020). Subsidy allocations in the presence of income shocks. *Proceedings of the AAAI Conference on Artificial Intelligence 34*(5), 7032–7039. https://doi.org/10.1609/aaai.v34i05.6188

Adelt, F., Weyer, J., Hoffmann, S., & Ihrig, A. (2018). Simulation of the governance of complex systems (SimCo): basic concepts and experiments on urban transportation. *Journal of Artificial Societies and Social Simulation, 21*, 2. https://doi.org/10.18564/jasss.3654

Aghaei, S., Azizi, M. J., & Vayanos, P. (2019). *Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making*. Available at: https://arxiv.org/abs/1903.10598 (accessed 10 May 2021).

AlgorithmWatch. (2019). *Atlas of Automation. Automated Decision-Making and Participation in Germany*. Available at: https://atlas.algorithmwatch.org/en (accessed 10 May 2021).

Allhutter, D., Mager, A., Cech, F., Fischer, F., & Grill, G. (2020). *Der AMS Algorithmus. Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS)*. Available at: https://doi.org/10.1553/ITA-pb-2020-02 (accessed 18 February 2022).

Anthony, D., Campos-Castillo, C., & Horne, C. (2017). *Toward a sociology of privacy. Annual Review of Sociology 43*(1), 249–269. https://doi.org/10.1146/annurev-soc-060116-053643

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, *35*, 611–623. https://doi.org/10.1007/s00146-019-00931-w

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, *1*(2), 90–143. https://doi.org/10.1093/jssam/smt008

Bansak, K. (2019). Can nonexperts really emulate statistical learning methods? A comment on "the accuracy, fairness, and limits of predicting recidivism". *Political Analysis*, *27*(3), 370–380. https://doi.org/10.1017/pan.2018.55

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671–732.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2019). *Consumer-Lending Discrimination in the FinTech Era*. Available at: https://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf (accessed 10 May 2021).

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1-4:15. https://doi.org/10.1147/JRD.2019.2942287

Benthall, S., & Haynes, B. D. (2019). Racial Categories in Machine Learning. *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29-31 January 2019*, 289–298. Association for Computing Machinery. https://doi.org/10.1145/3287560.3287575

Berk, R. (2017). An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, *13*(2), 193–216. https://doi.org/10.1007/s11292-017-9286-2

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2017). *A Convex Framework for Fair Regression*. Available at: https://arxiv.org/abs/1706.02409 (accessed 10 May 2021).

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in Criminal Justice Risk Assessments. *Sociological Methods & Research*, *104*(6), 1–42. https://doi.org/10.1177/0049124118782533

Bertelsmann Stiftung. (2020). *Praxisleitfaden zu den Algo.Rules. Orientierungshilfen für Entwickler:innen und ihre Führungskräfte*. Available at: https://www.bertelsmann-stiftung.de/fileadmin/files/alg/Algo.Rules_Praxisleitfaden.pdf (accessed 10 May 2021).

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). It's Reducing a Human Being to a Percentage. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18, Montreal QC, Canada, 21-26 April 2018*, 1–14. Association for Computing Machinery. https://doi.org/10.1145/3173574.3173951

Boeschoten, L., van Kesteren, E.-J., Bagheri, A., & Oberski, D. L. (2020). *Fair Inference on Error-Prone Outcomes*. Available at: https://arxiv.org/abs/2003.07621 (accessed 10 May 2021).

Burton, J. W., Stein, M., & Jensen, T. B. (2019). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *27*(11), 1309. https://doi.org/10.1002/bdm.2155

Calero Valdez, A., & Ziefle, M. (2018). Human factors in the age of algorithms. Understanding the human-in-the-loop using agent-based modeling. In G. Meiselwitz (Ed.), *Social Computing and Social Media. Technologies and Analytics: 10th International Conference, SCSM 2018, Held as Part of HCI International 2018, Proceedings, Part II* (Vol. 10914, pp. 357–371). Springer International Publishing. https://doi.org/10.1007/978-3-319-91485-5_27

Cepera, K., Konrad, J., & Weyer, J. (2018). Trust in algorithms. An empirical study of users' Willingness to change behaviour. In Getzinger, Günter (Ed.), *Critical Issues in Science, Technology and Society Studies: Conference proceedings of the 17th STS Conference Graz 2018, Graz, Austria, 7-8 May 2018* (pp. 38–47). Verlag der Technischen Universität Graz. https://doi.org/10.3217/978-3-85125-625-3

Coleman, J. S. (1994). *Foundations of Social Theory*. Belknap Press of Harvard University Press.

Corbett-Davies, S., & Goel, S. (2018) *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*. Available at: https://arxiv.org/abs/1808.00023 (accessed 10 May 2021).

Cowgill, B. (2018). *The Impact of Algorithms on Judicial Discretion: Evidence from regression discontinuities*. Available at: http://www.columbia.edu/~bc2656/papers/RecidAlgo.pdf (accessed 10 May 2021).

Cowgill, B., & Tucker, C. E. (2017). *Algorithmic bias: A counterfactual perspective.* Available at: https://bitlab.cas.msu.edu/trustworthy-algorithms/whitepapers/Bo%20Cowgill.pdf (accessed 10 May 2022).

Cowgill, B., & Tucker, C. E. (2020). *Algorithmic fairness and economics*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3361280 (accessed 10 May 2021).

Cruz Cortés, E., & Ghosh, D. (2019). *A Simulation based dynamic evaluation framework for system-wide Algorithmic Fairness*. Available at: https://arxiv.org/abs/1903.09209 (accessed 10 May 2021).

Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., De Paor, A., Felzmann, H., Haklay, M., Khoo, S.-M., Morison, J., Murphy, M. H., O'Brolchain, N., Schafer, B., &

Shankar, K. (2017). Algorithmic governance: developing a research agenda through the power of collective intelligence. *Big Data & Society*, *4*(2): 1–21. https://doi.org/10.1177/2053951717726554

Daumé III, H. (2017). *A Course in Machine Learning*. Available at: http://ciml.info/ (accessed 10 May 2021).

Dietvorst, B.J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them Err. *Journal of Experimental Psychology. General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces - IUI '19 the 24th International Conference, Marina del Ray, California, USA, 17-20 March 2019*, 275–285. Association for Computing Machinery. https://doi.org/10.1145/3301275.3302310

Dressel, J., Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through Awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12, Cambridge, MA, USA, 8-10 January 2012*, 214–226. Association for Computing Machinery. https://doi.org/10.1145/2090236.2090255

Elish, M.C., & Watkins, E. A. (2020). *Repairing Innovation: A Study of Integrating AI in Clinical Care*. Available at: https://datasociety.net/pubs/repairing-innovation.pdf (accessed 16 February 2022).

Elliott, M.R., & Valliant, R. (2017) Inference for nonprobability samples. *Statistical Science*, *32*(2), 249–264. https://doi.org/10.1214/16-STS598

European Parliament, Directorate General for Parliamentary Research Services, Castelluccia, C., & Le Métayer, D. (2019) *Understanding Algorithmic Decision-Making: Opportunities and Challenges*. Luxembourg: Publications Office. Available at: https://data.europa.eu/doi/10.2861/536131 (accessed 8 February 2022).

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 10-13 August 2015*, 259–268. Association for Computing Machinery. https://doi.org/10.1145/2783258.2783311

Fischer, S., & Petersen, T. (2018). *Was Deutschland über Algorithmen weiß und denkt: Ergebnisse einer repräsentativen Bevölkerungsumfrage*. Available at: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/Was_die_Deutschen_ueber_Algorithmen_denken.pdf (accessed 10 May 2021).

Freeman Engstrom, D., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*. Available at: https://www.acus.gov/sites/default/files/documents/Government%20by%20Algorithm.pdf (accessed 16 February 2022).

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2018). *A Comparative Study of Fairness-Enhancing Interventions in Machine Learning*. Available at: https://arxiv.org/abs/1802.04422 (accessed 10 May 2021).

Gamper, J., Kernbeiß, G., & Wagner-Pinte, M. (2020). *Das Assistenzsystem AMAS. Zweck, Grundlagen, Anwendung*. Available at: https://www.ams-forschungsnetzwerk.at/downloadpub/2020_Assistenzsystem_AMAS-dokumentation.pdf (accessed 18 February 2022).

Gilbert, G. N. (2008). *Agent-Based Models*. Sage.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Gran, A.-B., Booth, P., & Bucher, T. (2021). To be or not to be algorithm aware: A question of a new digital divide? *Information, Communication & Society*, *24*(12), 1779–1796. https://doi.org/10.1080/1369118X.2020.1736124

Green, B., & Chen, Y. (2019). Disparate interactions. An algorithm-in-the-loop analysis of fairness in risk assessments. *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29-31 January 2019*, 90–99. Association for Computing Machinery. https://doi.org/10.1145/3287560.3287563

Green, B., & Viljoen, S. (2020). Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency - FAT* '20, Barcelona, Spain, 27-30 January 2020*, 19–31. Association for Computing Machinery. https://doi.org/10.1145/3351095.3372840

Grgić-Hlača, N., Engel, C., & Gummadi, K. P. (2019). Human decision making with machine assistance. An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction 3(CSCW)*, 1–15. https://doi.org/10.1145/3359280

Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). Human perceptions of fairness in algorithmic decision making. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18, Lyon, France, 23-27 April 2018*, 903–912. Association for Computing Machinery. https://doi.org/10.1145/3178876.3186138

Grgić-Hlača, N., Weller, A., & Redmiles, E. M. (2020). *Dimensions of Diversity in Human Perceptions of Algorithmic Fairness*. Available at: https://arxiv.org/abs/2005.00808 (accessed 10 May 2021).

Groves, R. M. (2004). *Survey Errors and Survey Costs*. Wiley.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems 29, Barcelona, Spain, 5–10 December 2016*, 3315-3323. Curran Associates, Inc.

Hargittai, E., & Hsieh, Y. P. (2013). Digital Inequality. In W. H. Dutton (Ed.), *The Oxford Handbook of Internet Studies* (pp. 129–150). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199589074.013.0007

Hebert-Johnson, U., Kim, M. P., Reingold, O., & Rothblum, G. (2018). Multicalibration: Calibration for the (Computationally-Identifiable) Masses. *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, 10-15 July 2018*. PMLR.

Heidari, H., Nanda, V., & Gummadi, K. P. (2019). On the long-term impact of algorithmic decision policies: Effort unfairness and feature, segregation through social learning. *Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019*, 2692–2701. PMLR.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Canada, 3-10 March 2021*, 375–385. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445901

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., & Usher, A. (2015). Big data in survey research. *Public Opinion Quarterly*, *79*(4), 839–880. https://doi.org/10.1093/poq/nfv039

Johndrow, J. E., & Lum, K. (2017). *An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction*. Available at: https://arxiv.org/abs/1703.04957 (accessed 10 May 2021).

Joyce, K., Smith-Doerr, L., Alegria, S., Bell, S., Cruz, T., Hoffman, S. G., Noble, S. U., & Shestakofsky, B. (2021). Toward a sociology of artificial intelligence: A call for research on inequalities and structural change. *Socius: Sociological Research for a Dynamic World*, *7*, 1–11. https://doi.org/10.1177/2378023121999581

Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, *583*(7815), 69. https://doi.org/10.1038/d41586-020-02003-2

Kasy, M., & Abebe, R. (2021) Fairness, equality, and power in algorithmic decision-making. *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Canada*, 576–586. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445919

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., & Trappmann, M. (2020). Coverage error in data collection combining mobile surveys with passive measurement using apps: data from a German national survey. *Sociological Methods & Research*, 0049124120914924. https://doi.org/10.1177/0049124120914924

Kim, M. P., Ghorbani, A., & Zou, J. (2019). Multiaccuracy: Black-box post-processing for fairness in classification. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019*, 247–254. Association for Computing Machinery. https://doi.org/10.1145/3306618.3314287

Kim, M. P., Kern, C., Goldwasser, S., Kreuter, F., & Reingold, O. (2022). Universal adaptability: target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, *119*(4), e2108097119. https://doi.org/10.1073/pnas.2108097119

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Kohler-Hausmann, I. (2011). *Discrimination*. Available at: https://www.oxfordbibliographies.com/view/document/obo-9780199756384/obo-9780199756384-0013.xml (accessed 10 May 2021).

Kopf, J. (2019) *Ein kritischer Blick auf die AMS-Kritiker*. Available at: https://www.derstandard.de/story/2000109032448/ein-kritischer-blick-auf-die-ams-kritiker (accessed 10 May 2021).

Kuppler, M., Kern, C., Bach, R. L., & Kreuter, F. (2021). *Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There?* Available at: https://arxiv.org/abs/2105.01441 (accessed 1 December 2021).

Kusner, M., Russell, C., Loftus, J., & Silva, R. (2019). Making decisions that reduce discriminatory impacts. *Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9-15 June 2019*, 3591–3600. PMLR.

Kusner, M. J., & Loftus, J. R. (2020). The long road to fairer algorithms. *Nature*, *578*(7793), 34–36. https://doi.org/10.1038/d41586-020-00274-3

Lange, A.-C., Lenglet, M., & Seyfert, R. (2019). On studying algorithms ethnographically: making sense of objects of ignorance. *Organization*, *26*(4), 598–617. https://doi.org/10.1177/1350508418808230

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lerman, J. (2013). Big data and its exclusions. *Stanford Law Review Online*, *66*, 55–63.

Lind, K., & Wallentin, L. (2020). *Central Authorities Slow to React as Sweden's Cities Embrace Automation of Welfare Management*. Available at: https://algorithmwatch.org/en/story/trelleborg-sweden-algorithm/ (accessed 10 May 2021).

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2019). *Delayed Impact of Fair Machine Learning*. Available at: https://arxiv.org/abs/1803.04383 (accessed 10 May 2021).

Liu, Z. (2021). Sociological perspectives on artificial intelligence: A typological reading. *Sociology Compass*, *15*(3), 1–13. https://doi.org/10.1111/soc4.12851

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: people prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lopez, J. (2019). Reinforcing intersectional inequality via the AMS algorithm in Austria. *Conference Proceedings of the 18th STS Conference Graz 2019: Critical Issues in Science, Technology and Society Studies, Graz, Austria, 6-7 May 2019*, 289–309. Verlag der Technischen Universität Graz. https://doi.org/10.3217/978-3-85125-668-0-16

Lutz, C. (2019). Digital inequalities in the age of artificial intelligence and big data. *Human Behavior and Emerging Technologies*, *1*(2), 141–148. https://doi.org/10.1002/hbe2.140

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2020). *On The Applicability of ML Fairness Notions*. Available at: https://arxiv.org/abs/2006.16745 (accessed 10 May 2021).

Mann, M., & Matzner, T. (2019). Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination. *Big Data & Society*, *6*(2), 1–11. https://doi.org/10.1177/2053951719895805

Mayson, S. G. (2019). Bias in, bias out. *The Yale Law Journal*, 128(8), 2218–2300.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). *A Survey on Bias and Fairness in Machine Learning*. Available at: https://arxiv.org/abs/1908.09635 (accessed 11 May 2021).

Miller, A. P. (2018). *Want Less-Biased Decisions? Use Algorithms*. Available at: https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms (accessed 10 May 2021).

Mittelstadt, B. D. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: mapping the debate. *Big Data & Society*, *3*(2), 1–21. https://doi.org/10.1177/2053951716679679

Molnar, C. (2019). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/ (accessed 2 December 2021).

Noriega-Campero, A., Bakker, M. A., Garcia-Bulle, B., & Pentland, A. (2018). *Active Fairness in Algorithmic Decision Making*. Available at: https://arxiv.org/abs/1810.00031 (accessed 10 May 2021).

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Otte, G., Boehle, M., & Kunißen, K. (2021). Social Inequalities—Empirical Focus. In B. Hollstein, R. Greshoff, U. Schimank, & A. Weiß (Eds.), *Soziologie—Sociology in the German-Speaking World* (pp. 361–380). De Gruyter. https://doi.org/10.1515/9783110627275-025

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, *22*(4), 390–409. https://doi.org/10.1002/bdm.637

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Pierson, E. (2018). *Demographics and Discussion Influence Views on Algorithmic Fairness*. Available at: https://arxiv.org/abs/1712.09124 (accessed 10 May 2021).

Plane, A. C., Redmiles, E. M., & Mazurek M. L. (2017). Exploring User Perceptions of Discrimination in Online Targeted Advertising. *Proceedings of the 26th USENIX Security Symposium, Vancouver, BC, Canada, 16-18 August 2017*, 935–951. USENIX Association.

Rodolfa, K., Saleiro, P., & Ghani, R. (2021). Bias and fairness. In I. Foster I, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big Data and Social Science. Data Science Methods and Tools for Research and Practice* (pp. 281–312). CRC Press.

Rodolfa, K. T., Salomon, E., Haynes, L., Mendieta, I. H., Larson, J., & Ghani, R. (2020). Case Study: Predictive Fairness to Reduce Misdemeanor Recidivism through Social Service Interventions. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency - FAT* '20, Barcelona, Spain, 27–30 January 2020*, 142–153. Association for Computing Machinery. https://doi.org/10.1145/3351095.3372863

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2019). *Aequitas: A bias and fairness audit toolkit*. Available at: https://arxiv.org/abs/1811.05577 (accessed 10 May 2021).

Seaver, N. (2019). Knowing algorithms. In J. Vertesi, & D. Ribes (Eds) *DigitalSTS: A Field Guide for Science & Technology Studies* (pp. 412–422). Princeton University Press.

Selbst, A. D., boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29-31 January 2019*, Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598

Sen, I., Floeck, F., Weller, K., Weiss, B. & Wagner, C. (2019). *A Total Error Framework for Digital Traces of Humans*. Available at: https://arxiv.org/abs/1907.08228 (accessed 10 May 2021).

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2021). *Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature*. https://arxiv.org/abs/2103.12016 (accessed 10 May 2021).

Stevenson, M. T. (2018). Assessing risk assessment in action. *Minnesota Law Review*, *103*(1), 303–384.

Stevenson, M. T., & Doleac, J. L. (2019). *Algorithmic Risk Assessment in the Hands of Humans*. Available at: http://ftp.iza.org/dp12853.pdf (accessed 10 May 2021).

Suresh, H., & Guttag, J. V. (2020). *A Framework for Understanding Unintended Consequences of Machine Learning*. Available at: https://arxiv.org/abs/1901.10002 (accessed 10 May 2021).

Tan, S., Adebayo, J., Inkpen, K., & Kamar, E. (2018). *Investigating Human + Machine Complementarity: A Case Study on Recidivism*. Available at: https://arxiv.org/abs/1808.09123 (accessed 10 May 2021).

van Doorn, M., Pop, I., & Wolbers, M. H. J. (2011). Intergenerational transmission of education across European countries and cohorts. *European Societies*, *13*(1), 93–117. https://doi.org/10.1080/14616696.2010.540351

Wachter, S. (2020). Affinity profiling and discrimination by association in online behavioural advertising. *Berkeley Technology Law Review*, *35*(2), 1–74.

Weyer, J., Delisle, M., Kappler, K., Kiehl, M., Merz, C., & Schrape, J.-F. (2018). Big data in soziologischer perspektive. In B. Kolany-Raiser, R. Heil, C. Orwat, & T. Hoeren (Eds.), *Big Data und Gesellschaft: Eine Multidisziplinäre Annäherung* (pp. 69–149). Springer VS. https://doi.org/10.1007/978-3-658-21665-8_2

Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and automation bias in the use of imperfect automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(5), 728–739. https://doi.org/10.1177/0018720815581940

Yang, S., & Kim, J. K. (2020). *Statistical Data Integration in Survey Sampling: A Review*. Available at: https://arxiv.org/abs/2001.03259 (accessed 10 May 2021).

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Algorithmic decision-making and the control problem. *Minds and Machines*, *29*(4): 555–578. https://doi.org/10.1007/s11023-019-09513-7

Zou, J., & Schiebinger, L. (2018). AI Can be sexist and racist – it's time to make it fair. *Nature*, *559*(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8

3. Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making[2]

Abstract. Human perceptions of fairness in (semi-)automated decision-making (ADM) constitute a crucial building block toward developing human-centered ADM solutions. However, measuring fairness perceptions is challenging because various context and design characteristics of ADM systems need to be disentangled. Particularly, ADM applications need to use the right degree of automation and granularity of data input to achieve efficiency and public acceptance. We present results from a large-scale vignette experiment that assessed fairness perceptions and the acceptability of ADM systems. The experiment varied context and design dimensions, with an emphasis on who makes the final decision. We show that automated recommendations in combination with a final human decider are perceived as fair as decisions made by a dominant human decider and as fairer than decisions made only by an algorithm. Our results shed light on the context dependence of fairness assessments and show that semi-automation of decision-making processes is often desirable.

3.1. Introduction

Automated decision-making (ADM) is increasingly used in many critical domains that affect individuals' life chances. This includes the use of machine learning (ML) to support public employment services (Körtner & Bonoli, 2021), algorithmic decision-making in human resources (HR) management (Köchling & Wehner, 2020), and (infamous) examples of automated risk assessments in criminal sentencing (Angwin et al., 2016). Against this backdrop, research on fairness in ML has recognized that fairness of ADM systems needs to be evaluated within the social contexts in which they are placed (Selbst et al., 2019). The successful

[2] This chapter was previously published as a paper in the journal *Patterns:*
Kern, C.*, Gerdon, F.*, Bach, R. L., Keusch, F., & Kreuter, F. (2022). Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. *Patterns*, *3*(10), 100591. https://doi.org/10.1016/j.patter.2022.100591
*: shared first authors.

Only small edits were made in comparison to the published paper version.
The Appendix for this chapter is available in Chapter 7.2. References to the Appendix begin with the letter "A". The vignette texts ("supplemental experimental procedures") are available at: https://www.cell.com/cms/10.1016/j.patter.2022.100591/attachment/7bc378cb-6980-46ef-a5b3-879c603d315b/mmc1

implementation of ADM in a given setting requires public support and support of the affected individuals. Beyond risk assessments (Krafft et al., 2020), fairness and acceptability evaluations critically guide discussions on whether and how ADM solutions should be employed in a given context (Skirpan & Gorelick, 2017). Likewise, fairness perceptions inform developers in designing socially accepted ADM systems and policy-makers in considerations on which application contexts are deemed sensitive and need particular (legal) attention.

Multiple design features of the ADM system may affect acceptance. ADM outputs may constitute the final decision or may be used as a recommendation for an action. In other instances, computer programs may simply provide data without suggesting a recommendation or classification. It is likely that context and other characteristics of the concrete ADM system influence whether people deem it acceptable if an ADM actually decides on its own or to which extent human supervision and intervention are desired. People are also likely to vary in their perceptions of the ADM system depending on their own experiences, understanding, and likelihood of being affected by these systems. People from groups who have been discriminated against in the past may particularly worry about unfair or otherwise biased decisions.

Previous research has examined fairness perceptions with respect to selected application contexts, fairness metrics, and explanation styles (see "Background and related work"). The study presented here aims to connect the different findings and lines of previous research. Our focus is on perceptions toward the system as a whole, i.e., whether ADM is perceived to be fair and acceptable to be applied for a specific purpose and in a specific context. Novel is the measurement of fairness assessments in a survey experiment that considers three degrees of human involvement in decision-making across several application contexts, while varying further design features within each context. This set-up allows the examination of interactions between application contexts and characteristics of the ADM approach. Novel is also the combined analysis of fairness ratings in interaction with characteristics of the evaluating individuals, where individuals are drawn from the population at random with known selection probabilities, improving the external validity of our findings.

More specifically, we compare perceptions and acceptance of the use of ADM systems across four different contexts (banking, HR, criminal justice, and employment agencies). We experimentally research scarcely investigated differences in acceptance between mainly human decision-making, semi-ADM, and fully ADM. We furthermore elucidate whether assistive decisions are deemed fairer than punitive decisions, and we explore inter-individual heterogeneity in responses. The main questions we answer are: first, which degree of automation is more accepted/perceived fairer across scenarios and situations? Second, do

individual characteristics interact with context and design characteristics in affecting acceptance/perceived fairness?

We find that semi-ADM is perceived as fairer than fully ADM and roughly as fair as mainly human decision-making. In addition, the preference for human oversight varies by context. These results not only suggest that ADM systems need to be evaluated on a case-by-case basis, but they also provide directions for initial design choices that increase the chance of public acceptance according to specific design categories of interest. In summary, we provide the following contributions to research on public perceptions toward ADM:

- Comparison of perceptions toward different levels of automation in decision-making processes across contexts, providing implications for how to design ADM applications depending on context
- Insights into acceptance of assistive and punitive types of decisions across contexts, showing in which cases human involvement should be particularly considered in ADM design
- Data based on an experimental approach within a nationally representative probability-based sample with known selection probabilities and a larger sample size than (most) previous research, thus providing a high-quality sample

3.1.1. Background and related work

Research on fairness in ML and ADM focused so far primarily on important technical aspects of fairness, such as defining and choosing fairness metrics, evaluating existing ADM applications with respect to their fairness implications, and correcting unfair systems (see, e.g., Barocas et al. (2019) for an overview on fair ML). Other studies have investigated the legal preconditions of using algorithmic systems (Wachter et al., 2021), provided philosophical perspectives on fairness in algorithmic decision-making (Barocas et al., 2019; Wachter et al., 2021), or investigated trust in algorithmic systems in human-machine interactions (Zerilli et al., 2022). However, over the past years, a strand of literature has emerged that investigates human perceptions on fairness in ADM, i.e., how individuals from the populations potentially affected by ADM systems evaluate their use.

A literature review by Starke et al. (2021) identified several papers that investigated humans' perceptions of algorithmic fairness. We focus on four key dimensions that have been investigated with respect to perceptions of algorithmic fairness: (1) the context in which an ADM system is applied and the type of impact the system makes, (2) the degree of human

involvement in decision-making, (3) the features used by an algorithm, and (4) the characteristics of the individual that may influence perceptions of algorithmic fairness.

The first dimension is concerned with the contexts in which ADM systems are used and the impact of a decision for an individual's life (Koene et al., 2017; Starke et al., 2021). Previous research highlighted that empirical results on perceptions in specific ADM contexts may not translate into other contexts, cautioning researchers against over-generalizations (Zerilli et al., 2022). Although each context comes with myriads of idiosyncrasies, it appears likely that the stakes of the decision-making context are one crucial differentiating factor. In an exploratory study, Smith et al. (2020) found that fairness of ADM systems matters less to individuals when the decisions to be made have relatively little impact, such as in music and movie recommendations, while fairness plays a much larger role when the decisions have relatively large impact, such as in job recommendations. Likewise, recent advances in fair ML emphasize that specific types of prediction error may matter more for some kinds of decisions than for others: for assistive actions, avoiding false negatives might be viewed as critical; for punitive actions, avoiding false positives might be considered most important (Makhlouf et al., 2020; Saleiro et al., 2019). Translating this notion into fairness perceptions by drawing on insights from economics, individuals may attribute higher weight to potential losses following from decisions than to potential gains (Kahneman & Tversky, 1979).

Relating to the second dimension, some research exists on direct comparisons between human and (purely) ADM for specific contexts (Starke et al., 2021) and concludes that there is great variation in relative perceived fairness across contexts and that characteristics of the task impact fairness perceptions. In a series of survey experiments, Nagtegaal (2021) found that public sector employees perceived human decision-makers as procedurally fairer for tasks with high complexity, and that adding an algorithm to a human in the decision-making process may increase justice perceptions. In another experiment, participants deemed human decisions as fairer than algorithmic decisions with tasks that particularly required human skills (hiring and work evaluations), while no difference was found for perceived fairness relating to "mechanical" skills (work assignment and scheduling) (Lee, 2018). Research that compares hybrid decision-making (which involves both algorithmic and human decision-making) with solely algorithmic or human decision-making across contexts is scarcer. For instance, Gonzalez et al. (2022) find that combined decision-making is preferred over completely ADM in hiring decisions, but this also depends on the familiarity of the respondent with artificial intelligence (AI). Similarly, another study in the HR context finds that individuals have negative attitudes to purely ADM because of the limited use of information by ADM systems (Newman et al.,

2020). With an Amazon MTurk sample, Waldman and Martin (2022) found that ADM decisions overseen by a "privacy professional" increased perceived legitimacy of the decision compared with purely algorithmic or human decisions. Overall, a literature review by Langer and Landers (2021) suggests that hybrid decision-making is preferred over fully ADM, at least in specific contexts. However, the review study by Starke et al. (2021) finds no clear public preference for whether solely human decision-making or a hybrid process involving humans and algorithms was preferred and conclude that no general statement on the preference for either human or ADM could be made. The literature may therefore profit from a systematic comparison of degrees of automation in several major ADM applications contexts with a large and probability-based sample.

The third dimension is concerned with which features, i.e., which variables and therefore also individual characteristics, an algorithm draws on. Dodge et al. (2019), for example, find in a qualitative study that, among others, the appropriateness of the data basis and the features used and not used by the algorithm matter to people's fairness perceptions. Grgic-Hlaca et al. (2018) suggest, based on their reading of the literature, eight feature properties (e.g., reliability and privacy sensitivity) that may be relevant for fairness perception. Using a survey, the study also finds that most of these properties matter for fairness perceptions, and survey respondents agreed that the use of reliable, relevant, or private information was fair. Furthermore, previous studies have shown that the fairness of data use depends on the proximity of the type of data to the system's purpose in the context of crime (Grgic-Hlaca, Zafar, et al., 2018; van Berkel et al., 2019), and that the legitimacy of ADM is higher when purpose-specific rather than general data in the form of individual online browsing behavior are used (Waldman & Martin, 2022), supporting the idea that the normative appropriateness of using personal data is context dependent (Nissenbaum, 2019).

The fourth dimension focuses on the often-neglected perspective of evaluating individuals and their characteristics and experiences. Particularly, the perceived fairness of the use of specific individual characteristics in an ADM application for bail decisions has been shown to correlate with the characteristics of the evaluating individual. For example, women deemed it less fair for the ADM to rely on gender in this case (Grgic-Hlaca et al., 2020). Similarly, women are less likely to accept automated university course recommendations that use gender when the results disadvantage women for science course recommendations (Pierson, 2018). However, a review found no conclusive evidence for general direct effects of gender on fairness perceptions (Starke et al., 2021).

Beyond protected attributes, inter-individual differences in perceptions may arise from differing attitudes and knowledge. For instance, higher general privacy concerns may lower the acceptance of data regarded irrelevant for decision-making. Additionally, knowledge about algorithms may increase positive evaluations of the employment of algorithms in decision-making processes (Bertelsmann Stiftung, 2019).

Our research aims at connecting the different dimensions and lines of previous research by investigating them within a single framework, thereby enabling us to draw conclusions that may hold beyond a single context. In addition, we advance the literature by focusing on the perceived fairness of three degrees of human involvement in decision-making across several contexts with an experimental approach. We compare several application contexts for decision-making between each other, while also investigating preferences within contexts. Because perceptions may strongly differ between contexts, any variation caused by specific characteristics within contexts does not necessarily imply that this specific characteristic will matter for all contexts. Furthermore, we analyze fairness ratings in interaction with characteristics of the evaluating individual. Moreover, in addition to fairness perceptions, we measure acceptance ratings of ADM use cases. We compare responses to both questions, which allows us to learn whether they measure a common latent construct or whether respondents clearly differentiate between fairness perceptions and overall acceptance.

3.1.2. Data

To investigate the impact of specific characteristics of computationally supported decision-making on people's acceptance and perceived fairness, we conducted a factorial survey experiment, or "vignette" experiment (Auspurg & Hinz, 2015), in July 2021 (Wave 54) using the German Internet Panel (GIP), a probability-based longitudinal online survey (Blom et al., 2015). GIP covers both the online and the offline population living in private households in Germany aged 16–75 years, and participants were recruited face-to-face (in 2012 and 2014) and via postal mail (in 2018). People without a computer and/or no access to the Internet in the first two recruitment waves were provided with a basic laptop/tablet computer to participate. Panel members are invited on a bimonthly basis to participate in web surveys on political and economic attitudes and reform preferences (Blom et al., 2015). The Wave 54 questionnaire of the GIP included a rider with our vignette experiment that was specifically developed for this study. A total of 4,108 GIP panel members participated in the Wave 54 survey with a completion rate for GIP Wave 54 of 65.8% (COMR; see American Association for Public Opinion

Research, American Association for Public Opinion Research, 2016). Excluding participants who broke off the survey or did not provide answers to our vignette questions leaves us with 3,930 respondents with valid fairness assessments and 3,972 respondents with complete acceptance ratings.

Being a probability-based survey, the GIP is based on random sampling from a sampling frame from the target population with known selection and known inclusion probabilities (Blom et al., 2015). Several studies found that, in general, probability-based online panels outperform non-probability samples, which are commonly used in research on ADM fairness perceptions, such as Amazon MTurk, in terms of data quality (Cornesse et al., 2020). As such, the sample of the GIP is a very good representation of the general population in Germany (Cornesse et al., 2021; Cornesse & Schaurer, 2021). Our study design is thus strong in both internal validity, because of the experimental design, and the representativity of the sample, that is, in external validity.

3.1.3. Vignette experiment

In the vignette experiment, respondents are presented with 4 of 42 text descriptions of hypothetical scenarios on decision-making that suggest different degrees of automation, among others (see below). The descriptions vary by characteristics (or dimensions) that can take on different specified levels; by randomly assigning vignettes to respondents, researchers may estimate the causal effects of changes in single-vignette dimensions on responses (Auspurg & Hinz, 2015). We created 42 descriptions that were blocked into four groups that each refer to one specific context of ADM applications (representing the dimension *context*). We investigate four contexts that we chose because they have been extensively discussed in academic literature on ADM and, partly, in public discourse, and therefore are of particular relevance. These contexts vary by the potential severity of decisions, i.e., how strongly they may affect citizens' lives: (1) "Bank," bank credits and products (Bartlett et al., 2022; Peachey, 2019; Weber et al., 2020); (2) "Job," HR decision-making (Köchling & Wehner, 2020); (3) "Prison," criminal justice (Angwin et al., 2016; López-Molina, 2021; H. Wang et al., 2019); and (4) "Unemployment," actions of employment agencies (Lopez, 2019).

Each respondent received one randomly drawn vignette for each context in random order. The vignettes further contained the following dimensions: *action*, *data*, and *decision-maker*. Although we argued that an important difference between contexts is the severity of the decision, previous literature points to the importance of whether effects of decisions on citizens' lives are produced by punitive or assistive actions. This distinction has been recently identified

as a crucial factor in the selection of fairness notions for ML applications (Makhlouf et al., 2020; Saleiro et al., 2019) and because individuals may differ in their perception of the severity of these types of decisions (see "Background and related work"). This distinction allows us to investigate different kinds of decisions within identical contexts. The kinds of *data* used for decision-making have been a key concern of previous empirical research on fairness perceptions. Although previous studies usually focus on specific kinds of information to be used, we follow the notion of contextual integrity (Nissenbaum, 2019), which suggests that the crucial question is whether the use of the data is contextually appropriate (see "Background and related work"). We distinguish between contextually close and contextually remote kinds of data for each context. For instance, contextually close data in the hiring context may be data on performance in previous jobs. Across all investigated contexts, contextually remote data may be data from Internet searches about a person who, e.g., applies for credit. The latter data might improve the accuracy of decisions, but privacy concerns about the appropriateness of their use may arise, particularly if the data in question are not necessarily related to the decision problem at hand. For our purposes, it does not matter which exact kind of additional (Internet) data is considered, what is important is that these data are potentially considered as out of context by respondents but may still improve the accuracy of predictions. Finally, we vary the degree of human involvement in the decision-making process (*decision-maker*) to learn about its optimal levels across different contexts, which represents one of the most crucial design decisions for computationally supported decision-making systems. The concrete levels for each of the dimensions are as follows:

1. Type of action the decision affects (dimension: *action*)
    - Assistive action
        - Bank: provision of exclusive financial products
        - Job: hiring of employees
        - Prison: early release from prison
        - Unemployment: offering support services to unemployed individuals
    - Punitive action
        - Bank: regulating access to credits
        - Job: termination of work in probation period
        - Unemployment: shortening financial assistance for unemployed individuals
        - No punitive action was defined for the justice context because we deemed this case too problematic to confront respondents

2. Type of data used to inform decision (dimension: *data*)
   - Only data that have been produced in the social context of the decision task or closely related contexts ("no Internet data")
   - Additionally using data found on the Internet that may stem from various contexts ("Internet data")

3. Who makes the decision (dimension: *decision-maker*)
   - Solely ADM (fully automated: "Algorithm")
   - Human decision-making based on an automated recommendation (automated recommendation: "Both")
   - Solely human decision-making, assisted by information from computer programs (mainly human: "Human")

For instance, the vignette with the levels employment agency, assistive action, additional Internet data, and mainly human decision-making reads: "A local employment agency has developed a computer program for assigning support measures to job seekers. This program uses data about the person's past periods of employment and unemployment, as well as information about the person available on the Internet. A staff member at the employment agency compares this information with that of other job-seeking individuals who have successfully participated in a measure. The employee decides whether the person is to receive a support measure" (translated from German).

In the vignette with the levels employment agency, assistive action, and additional Internet data, but automated recommendation, the last two sentences above are changed as follows: "The program compares this information with that of other job-seeking individuals who have successfully participated in a measure. The program gives an employee a recommendation whether the person is to receive a support measure. The final decision is made by the employee."

In the corresponding vignette with fully ADM, the last two sentences read: "The program compares this information with that of other job-seeking individuals who have successfully participated in a measure. The program determines automatically whether the person is to receive a support measure."

All vignettes are presented in the data documentation of Wave 54 of the GIP (Blom et al., 2021) and in the supplemental experimental procedures (see footnote 2).

After each vignette, we asked respondents in two separate questions how fair and how acceptable they perceive this way of decision-making ("How fair do you find it is to make a

decision in this way?" "How acceptable do you find it is to make a decision in this way?") using a fully labeled four-point rating scale ("Not at all fair/acceptable," "A little fair/acceptable," "Somewhat fair/acceptable," or "Very fair/acceptable"). We ask about both fairness perceptions and acceptability because the former may be only one among various factors that affect acceptance. In addition to fairness, individuals may consider accountability, transparency, and explainability in their overall assessment of algorithmic decision-making, next to their evaluation of the systems utility (Shin, 2020). Thus, individuals may think that a system is prone to producing unfair results but still be convinced that the system is transparent or more efficient and therefore acceptable. Note that we do not force individuals into a specific role in the ADM process (such as a decider or an affected individual) to learn about citizens' evaluations of the systems as such.

Note that we refrained from pre-defining fairness (or acceptability) for the respondents in our survey instrument. Our aim was to measure respondents' personal perception of the general appropriateness of the presented way of decision-making, without priming and limiting them toward a specific (technical) fairness notion that they might not even consider in real-world evaluations of ADM.

3.1.4. Respondent characteristics

In addition to fairness and acceptability evaluations, we collected information on respondents' socio-demographic characteristics and further background information. We are therefore able to study how fairness perceptions depend on respondents' gender (male and female) and age (older than 60 years versus 60 years or younger). Similar to other countries, these two individual attributes are oftentimes connected to discrimination in Germany (Beigang et al., 2017). In line with the treatment of these characteristics as protected attributes in the fairness literature, this allows us to investigate whether historical disadvantages may be associated with differential fairness evaluations of ADM systems across social groups. We further constructed a "privacy" index that summarizes respondents' concerns toward sharing personal data on a five-point scale (labeled from "not at all concerned" to "very concerned"), one measure that aims at capturing general affinity toward technology (via the total number of digital devices owned) and one measure to assess respondents' knowledge of algorithmic decision-making (via the total number of specific technical and statistical terms known; see Table A3.2 in Appendix 7.2 for details). These variables allow us to investigate whether ADM design features are evaluated differently given individuals' privacy attitudes and technical experience.

3.1.5. Analysis

We conduct our analysis in three steps. First, we present descriptive findings of the fairness evaluations by vignette dimensions. Second, we show results of mixed-effects ordinal probit regressions that model the effects of the ADM's application context and design dimensions on fairness and acceptability assessments. Third, we present context-specific regression models that investigate the effects of respondent characteristics. We use mixed-effects models to account for the hierarchical structure of our data, because multiple (four) vignettes are nested within respondents (Raudenbush & Bryk, 2002). For our fairness measure, e.g., this gives us 15,525 observations based on 3,930 respondents. Given the four ordered response categories of the outcome variables, we follow an ordinal probit approach by linking the observed outcome to an unobserved, continuous response variable via a set of threshold functions (Scott Long, 1997). In our mixed-effects models, we include random intercepts on the respondent level and specify different model variations, including random slopes, to test our assumptions about the mechanisms of fairness perceptions. All regression models control for the order of vignettes shown to respondents to eliminate ordering effects.

3.2. Results

3.2.1. Distribution of fairness evaluations

We first present average fairness ratings depending on vignette characteristics to provide a straightforward overview of the main results. For interpretation purposes, we collapse the four-point response scale into two categories: "Fair" ("Somewhat fair" and "Very fair") and "Not fair" ("A little fair" and "Not at all fair") and show the relative frequencies of respondents that rated a scenario as "Fair" in Figure 3.1. A tabular presentation of relative frequencies for both fairness and acceptance ratings by vignette levels is provided in Table A3.1. Overall summary statistics for fairness and acceptance evaluations, as well as for respondent characteristics, are provided in Table A3.2. A comparison of fairness ratings across vignettes allows the following four conclusions.

First, the highest response categories ("Somewhat fair" and "Very fair") were less frequently chosen than "A little fair" and "Not at all fair," indicating some, although not strong, levels of skepticism against computationally supported decision-making on average. Nonetheless, the level of perceived unfairness strongly depends on the specific vignette characteristics.

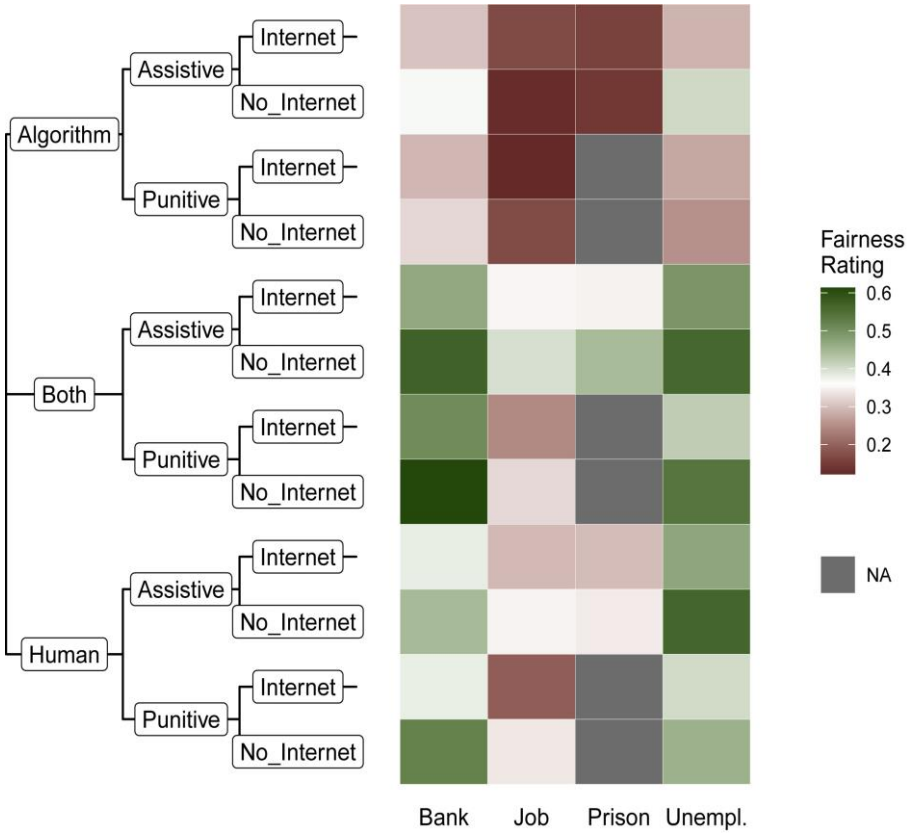


Figure 3.1. Average fairness rating by vignette levels. The heatmap shows relative frequencies of respondents that rated a scenario as ‘‘Fair’’ (i.e., either ‘‘Somewhat fair’’ or ‘‘Very fair’’). The color scale is centered at the average fairness rating over all vignettes.

Second, fairness evaluations vary by application *context*. In particular, the use of ADM in HR contexts (vignette level “Job”) and criminal justice settings (“Prison”) is often evaluated as “Not at all fair” or “A little fair,” whereas ADM applications in the banking sector (“Bank”) or by employment agencies (“Unemployment”) are perceived as less troubling.

Third, decisions performed without any kind of human intervention (“Algorithm”) are perceived as less fair than decisions that include human supervision (“Both” and “Human”). These differences along the dimension *decision-maker* are strongly pronounced for the HR and judicial context, considering their low baseline levels.

Fourth, within contexts, respondents do not appear to strongly distinguish between punitive and assistive *actions*. However, a slight shift toward higher perceived fairness is observable for ADM scenarios that do not use Internet *data*.

We present descriptive results of both the (complete) fairness and acceptance evaluations, including all response categories in Figures A3.1 and A3.2. Overall, the acceptance evaluations show very similar patterns as the fairness ratings, indicating that respondents evaluated fairness primarily with respect to whether they find the presented way of decision-making appropriate

(in a given context). This result may also mean that a common latent construct underlies these two measures. We can, however, notice that respondents are somewhat more restrictive in their acceptability ratings, because the highest response category ("Very acceptable") was rarely chosen across vignettes.

3.2.2. Mixed-effects regression models

We fitted three mixed-effects regression models for each outcome variable, i.e., respondents' fairness evaluations and acceptance ratings: a random-intercept model with main effects of all vignette dimensions (R-I Main), a random-intercept model with additional interactions between the dimensions *decision-maker* and *context* (R-I Interaction), and a random-intercept-random-slope model that allows the effects of *decision-maker* to vary between respondents (R-I-R-S). Focusing on the interactions between *decision-maker* and *context* allows us to shed light on how crucial ADM design decisions drive contextual fairness evaluations and add to the (in part inconclusive) research on publicly accepted degrees of automation in different application settings. Because the interactions are of most substantive interest, we present the R-I Interaction model for both outcome variables in Figure 3.2. Model fit statistics and tests for all models are summarized in Table A3.3.

The results of the R-I Interaction model predicting fairness evaluations (Figure 3.2A) point to the following conclusions: computationally supported decision-making systems that inform assistive *actions* are perceived as fairer than their punitive counterparts. Applications that make additional use of Internet *data* are perceived as less fair, compared with systems that only draw on contextually related data. The conditional main effects of *decision-maker* show that automated recommendation ("Both") is perceived as fairer and fully ADM ("Algorithm") as less fair compared with mainly human decision-making (in the "Bank" context). We further see that respondents valued a stronger human component in the "Job," "Prison," and "Unemployment" context as indicated by the negative interaction effects of *decision-maker* with *context*. Strong negative interactions for fully ADM with the "Job" and "Prison" *context* can be observed ("Algorithm∗Job", "Algorithm∗Prison"). Starting from already negative conditional main effects, the results for "Job" and "Prison" show that ADM is perceived as particularly problematic in these settings.

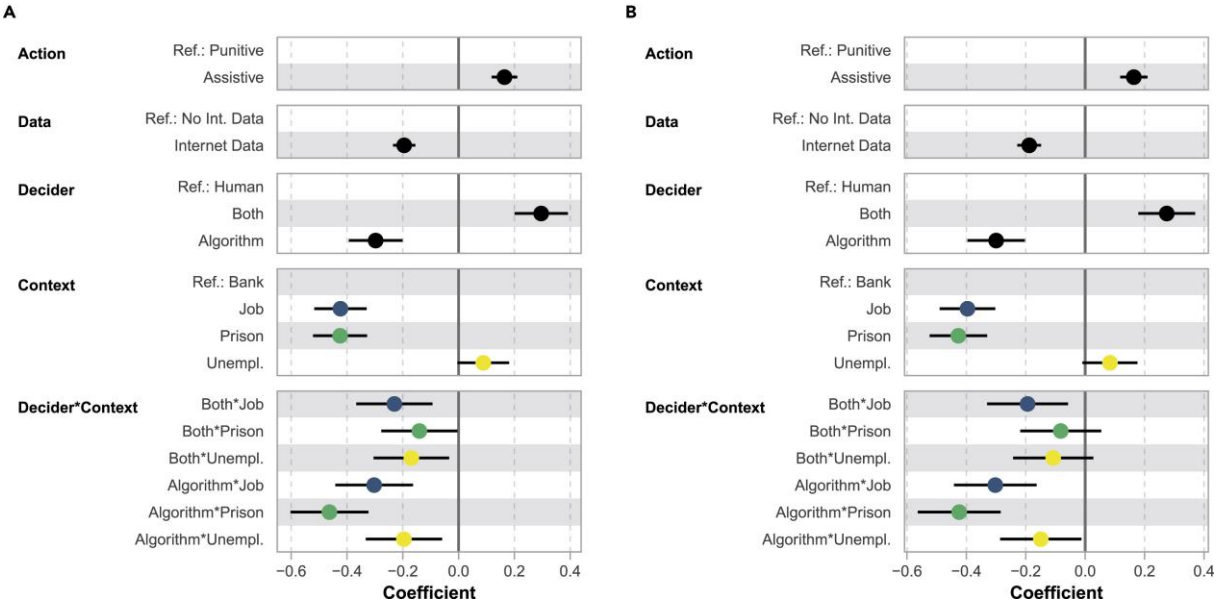


Figure 3.2. Coefficients (with 95% confidence intervals) of mixed-effects ordinal probit regression models predicting fairness evaluations and acceptance ratings with interactions between vignette dimensions decision-maker and context (R-I Interaction). (A) Outcome: fairness (nObs = 15,525). (B) Outcome: acceptance (nObs = 15,566)

To ease interpretation, we present average predicted probabilities for all outcome categories based on the R-I Interaction model across vignette dimensions in Table A3.4. We see that differences in the predicted probabilities of a positive fairness assessment ("Somewhat fair" and "Very fair") are driven by the vignette dimensions *context* and *decision-maker*, with considerably higher average predicted probabilities of both (highest) outcome categories for automated recommendation and the "Bank" and "Unemployment" settings. Focusing on the interaction effects, Table A3.5 shows how differences in the predicted probabilities across levels of *decision-maker* vary by *context*, highlighting that the distance between "Algorithm" versus "Human" is particularly strong in the "Job," "Prison," and "Unemployment" context (for the response categories "Somewhat fair" and "Not at all fair").

Comparing the outlined model with interactions against a model that includes only main effects underlines the context dependency of fairness perceptions, because the former model results in a considerably better model fit (likelihood ratio test of R-I Interaction versus R-I Main; see second column in Table A3.3). An increase in model fit can also be observed when specifying random slopes for *decision-maker*, indicating that the effects of this vignette dimension vary between respondents (likelihood ratio test of R-I-R-S versus R-I Main; see last column in Table A3.3). These findings motivate the specification of context-specific regression models that include interactions between the dimension decision-maker and respondent characteristics.

The results of the mixed-effects models predicting acceptance ratings mirror the above findings. The corresponding R-I Interaction model (Figure 3.2B) shows almost identical effect patterns: computationally supported decision-making is deemed less acceptable in the "Job" and "Prison" context (compared with "Bank") and respondents particularly object to fully ADM in these settings. We also note that for both outcomes we observe intra-class correlations (ICCs) between 0.45 and 0.51, highlighting that there is considerable clustering of vignette ratings within respondents (Table A3.3 again).

3.2.3. Context-specific regressions

We present two sets of context-specific regression models that include both vignette and respondent characteristics in Figure 3.3. The first set includes respondents' age and gender, in interaction with the vignette dimension *decision-maker*. The second set of models includes measures of respondents' privacy concerns, the number of digital devices owned, and the number of technical terms known (reflecting familiarity with AI and ML), all in interaction with *decision-maker*. Each set consists of four regression models that were fitted separately to fairness evaluations of each *context*. Corresponding models for the outcome acceptance are shown in Figure A3.3.

The results of the first model set (Figure 3.3A) show a negative conditional main effect of age in the "Bank" context, indicating that, in this case, older respondents perceive computationally supported decision-making as less fair than younger respondents. We generally observe little effect differences regarding the vignette dimension decision-maker between older and younger respondents. A notable exception is the more positive evaluation of automated recommendation of older respondents ("Both∗ >. 60 Years") in the "Job" context. We do not observe strong differences in the evaluation of either type of decision-making based on gender. At most, a modestly lower fairness evaluation of computationally supported decision-making of female respondents can be observed in the "Job" context (conditional main effect of gender).

Model set two (Figure 3.3B) shows negative conditional main effects of respondents' privacy concerns in the "Bank", "Job," and "Unemployment" contexts. Computationally supported decision-making is particularity viewed as problematic by people with higher privacy concerns. For the "Prison" context, stronger worries about privacy coincide with a more negative evaluation of fully ADM ("Algorithm∗Privacy"). Respondents' affinity toward technology seems to play a minor role in shaping fairness evaluations of ADM systems. Nonetheless, we can observe positive conditional main effects of the number of digital devices owned by respondents on fairness evaluations in the "Bank" and "Job" contexts and negative

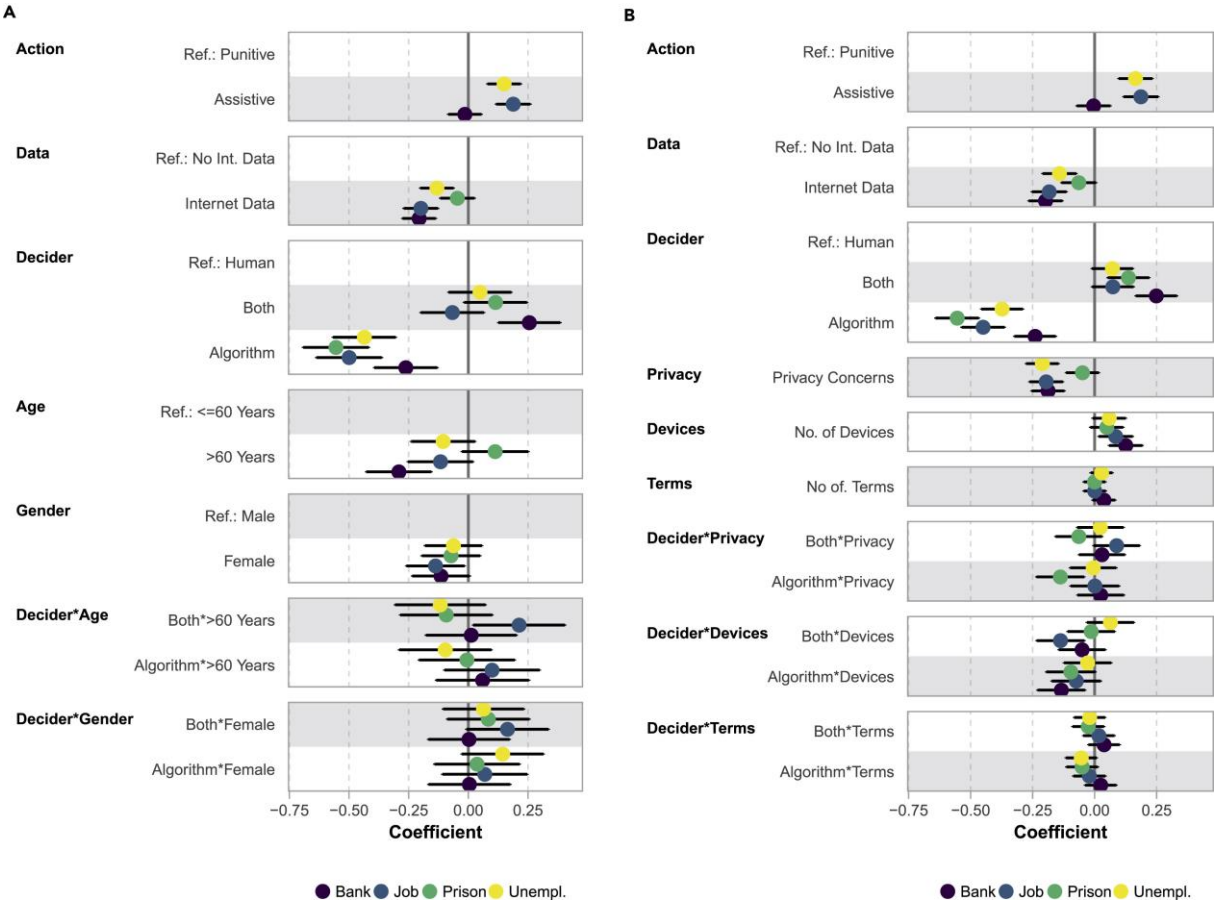


Figure 3.3. Coefficients (with 95% confidence intervals) of ordinal probit regression models predicting fairness evaluations of each context with interactions between the vignette dimension decision-maker and respondent characteristics. (A) Context-specific Interactions 1 (nBank = 3,653, nJob = 3,660, nPrison = 3,652, nUnempl = 3,654). (B) Context-specific Interactions 2 (nBank = 3,854, nJob = 3,858, nPrison = 3,855, nUnempl = 3,851)

interactions between devices and fully ADM ("Algorithm∗Devices") and automated recommendation ("Both∗Devices") in selected settings.

The results of the context-specific regression models predicting acceptance ratings show similar results, although with some exceptions, particularly in the first model set (Figure A3.3A). This includes an additional negative conditional main effect of age in the "Job" context and higher acceptance ratings of fully ADM of female compared with male respondents in the "Unemployment" context.

3.3. Discussion

In this research, we set out to advance our understanding of perceptions of fairness of ADM systems. Specifically, we sought to measure how design decisions, such as the level of human involvement in making the final decision and characteristics of the decision itself (assistive versus punitive), as well as the type of scenario, impact acceptance of various ADM systems

and their perceived fairness. Our results provide implications for how to design ADM applications depending on context. Furthermore, they offer insights into acceptance of assistive and punitive types of decisions across contexts, showing in which cases human involvement should be particularly considered in ADM design. A variation in the scenarios considered, in combination with a nationally representative probability-based sample of the German population, allows us to draw conclusions that future research may use as a starting point to understand the mechanisms causing variation in fairness evaluations across contexts.

Context dependency

Overall, the perceived fairness of computationally supported decision-making varies across contexts of application. Fairness ratings are lower in the "Job" and "Prison" contexts than in the "Bank" and "Unemployment" contexts. We believe that individuals may be particularly sceptical about automation in high-stake contexts (such as the "Prison" scenario) and in settings that may both eventually affect themselves and can have considerable impact (as in the "Job" context) as theories of subjective expected utility (Savage, 1972) suggest. However, we note that we did not measure subjective evaluations of impact; thus, we can only speculate that the perceived impact of a decision (e.g., high stakes versus low stakes) may cause the differences between these contexts.

Furthermore, we find that *assistive* decisions are deemed fairer than punitive decisions in the "Job" and "Unemployment" context, while no such difference is found in the "Bank" context. Following prospect theory (Kahneman & Tversky, 1979), individuals may weigh potential losses higher than potential gains and therefore be more open to assistive decisions. In our vignettes, the change in stakes from assistive to punitive decision-making in contexts that are related to hiring and the labor market are potentially perceived higher than in the "Bank" context. Regarding the implications of this finding for the design of ADM systems, we believe that fairness should be a major concern when the impact of the decision is high and the decision is rather punitive than assistive. However, future research will have to dig deeper into the underlying dimensions of contexts that affect human perceptions of ADM systems.

Human involvement

A second central finding concerns the comparison of fairness ratings for different degrees of human involvement in decision-making: respondents on average deemed automated recommendations as fairer than fully ADM and as similarly fair as mainly human decision-

making. This finding suggests that individuals do not consider the use of algorithms to inform decision-making as necessarily problematic per se. However, at the same time, respondents value the involvement of humans in the decision-making process. Therefore, human oversight appears to be an important element to ameliorate fairness perceptions of the population. While previous literature has shown such tendencies in specific contexts (Langer & Landers, 2021), we show how this effect varies across contexts. In our data, this is particularly true for the "Job" and "Prison" contexts, which are the two contexts in which computationally supported decision-making is generally perceived to be less fair than in the other contexts (see above). That is, ADM applications that may already be perceived as requiring special attention may deserve more human involvement in the decision-making process in order to be perceived as fair. Challenges with trust in novel technologies and misperceptions of the technological risk (e.g., to be treated unfair) may be important drivers for a desire of human oversight. Therefore, designing ADM systems that are perceived as fair may require effective communication of a basic understanding of the underlying technology. Moreover, individuals may feel more comfortable if high-stake decisions, especially in punitive contexts, involve a certain degree of human involvement or oversight in the decision-making process. Finally, if the automated element in decision-making itself is given a human appearance, it may enjoy increased acceptance, as previous research on chatbots suggests (Shin, 2021).

Previous research suggests that higher complexity of the decision task is connected to higher fairness ratings for human versus algorithmic decision-making (Nagtegaal, 2021). Our finding that human involvement is particularly desired in the hiring context aligns with a previous study in which respondents on average deemed human managers as fairer decision-makers for hiring decisions than algorithms (Lee, 2018). Lee (2018) also draws on open-ended responses, showing that this result may be based on expectations of human managers' skills and the concern that algorithms took a too standardized approach to evaluate candidates. It is possible that decisions relating to banking and unemployment are considered to be more amenable to standardization than decisions relating to hiring and prisons.

Data used in ADM

In our study, respondents perceived systems that draw on additional Internet data for decision-making less fair than systems that relied only on data that are close to the respective context. This finding is in line with previous research on feature use in ADM systems (see "Background and related work"). It confirms the importance of appropriate information flow, central to the privacy theory of contextual integrity (Nissenbaum, 2019). Contextual integrity emphasizes

that social contexts shape privacy norms, i.e., whose and which data are appropriate to be transmitted under which conditions.

Individual characteristics

As for the impact of individual socio-demographic characteristics, general fairness ratings of the "Bank" context decrease with higher age, and ratings are lower for women than for men in the "Job" context. Although the uncertainty in the estimated coefficients should make us cautious in over-interpreting these findings, they may hint to the presence of self-interest and/or social identity effects in fairness perceptions and could be worth exploring further. Previous research suggests that there appears to be self-interest involved in the individual evaluation of ADM processes and feature use (Grgic-Hlaca et al., 2020; Wang et al., 2020). Another potential theoretical explanation follows the idea of social identity theory (Tajfel, 1978). That is, individuals may not accept those decisions that may harm their in-group (Everett et al., 2015). Applied to the present study, these perspectives would imply that older people and women may consider that they or their in-group may be particularly disadvantaged in bank- or job-related contexts, respectively. This finding appears to be unrelated to the degree of human involvement. Furthermore, as previous research suggests, placing respondents into a specific position in the described decision-making process (such as decider or being affected by the decision oneself) may lead to different responses (Rieger et al., 2022).

Fairness versus acceptance

The regression results for the second investigated outcome variable "acceptance" mostly mirror the findings on fairness perceptions, although with some exceptions in the context-specific regressions. Indeed, the Spearman rank correlation coefficient for these two variables is 0.907. Although we cannot rule out that these similarities are a result of problematic respondent behavior (i.e., it could be possible that some respondents use satisficing strategies (Krosnick, 1991) when responding to the survey questions), it is conceivable that fairness and acceptance presuppose each other in evaluations of ADM systems, or that they measure a common latent construct. This latent construct may reflect an overall notion that using the respective ADM system is "okay" or desirable.

Limitations of the study

The study presented here draws on a very carefully selected sample of the German population. However, the vignette task used here for measurement is complex, and it is possible that not all

respondents fully understood all questions and settings. Ideally, we would have been able to add on qualitative interviews to capture why people responded the way they did and what exactly they thought about when reading about algorithms. Such probing questions are uncommon in fully standardized interviews and would have not been possible in this data collection instrument.

We also note that in measuring respondents' fairness perceptions, we cannot infer which notion(s) of fairness they operationalize in their evaluations. Respondents may consider notions of disparate treatment or impact with respect to attributes that they may perceive as sensitive or protected, or they may envision differential prediction (and thus decision) errors (Mitchell et al., 2021) as a result of a specific ADM design. Most likely, fairness assessments are the result of a (weighted) combination of multiple dimensions, which also are dependent on the presented ADM application context. Additional research is needed to probe which fairness concepts respondents may consider as most relevant in a given context.

Although we tried to capture a set of relevant contexts and settings, the study does not cover all possibly varying design characteristics of ADM systems. Previous studies have drawn on a plethora of potentially relevant characteristics, and these should also be considered when designing concrete ADM systems. Our intention was not to evaluate concrete ADM systems in detail but to compare crucial design elements within and between contexts of application, with an emphasis on the particularly important element of who makes the final decision and which kind of decision (assistive or punitive) is to be taken. Although we believe that the potential impact of a decision plays an important role in fairness evaluations, we did not directly manipulate whether a decision is high or low stakes. Therefore, we can only speculate that the potential impact of a decision will be a decisive element in individuals' fairness evaluations of ADM systems.

Future work

To expand the generalizability of our findings, future research may consider additional contexts and more nuances of the decision-making process. This may include a systematic variation of the complexity and the potential impact (high versus low stakes) of a decision, as well as the degree to which a decision is perceived to require human skills, such as subjective and intuitive judgment (see also "Background and related work"). Furthermore, previous research has shown that the exact wording with which the computerized components of ADM systems are described affect perceptions (Langer et al., 2022), which may be particularly interesting to compare across further contexts. This may also include surveying populations in other countries than Germany

and a focus on specific, potentially disadvantaged populations. This would allow researchers to investigate the impact of further protected attributes, such as ethnicity, on fairness evaluations. Such research could be conducted in real-life settings or with more immediate, real scenarios to verify the external validity of our findings.

More importantly, however, future work may put special emphasis on cleanly identifying the underlying dimensions that affect human perceptions of ADM systems. For example, a generalizable model of the influence of dimensions on fairness evaluations would allow policy-makers to estimate the degree to which a planned ADM system will meet society's normative expectations. Such a model should include understanding the mechanisms that cause variation in fairness perceptions, and integrate them in a theoretical model, a point also raised by Langer and Landers (2021). Right now, we can only speak to the dimensions that we experimentally varied in our study. In summary, we recommend that applications used to inform punitive decisions, applications with no human involvement, and applications that are not fully transparent regarding the data used should be carefully designed because fairness concerns among individuals seem to be highest in these scenarios.

Conclusion

In conclusion, our study showed that respondents perceive a combination of human and algorithmic decision-making as acceptable as decisions made by a human decider only. Solely algorithmic decisions are less accepted in the instances examined here. Human oversight is therefore deemed a desirable element of ADM systems. Overall, we found fairness perceptions not to be very high but to vary notably across context and design features.

There is a variety of decision tasks we did not touch on. Neither did we investigate perceptions of biometric mass surveillance, drones, and related situations with even higher stakes, nor did we investigate very low-stakes decisions such as algorithm-based navigation suggestions. Even within this narrower scope we see variation in perceptions, driven by context and type of decision, the used data, and individual characteristics. These attitudes are likely to shift with societies becoming more exposed to a variety of ADM systems. For now we want to re-emphasize that context matters, and individual preferences should be taken into consideration when designing these systems. Mapping novel ADM systems along the dimensions that we tested in this study may inform ADM designers beforehand when and where fairness concerns may arise among those impacted by the decisions.

Experimental procedures

Resource availability

Lead contact

For any questions regarding the paper and resources, please contact Dr. Christoph Kern (c.kern@uni-mannheim.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The questionnaire and the data have been deposited at data archive GESIS: https://doi.org/10.4232/1.13835 and are publicly available as of the date of publication. Application and written permission are needed prior to data access through the archive. All original code has been deposited at OSF: https://doi.org/10.17605/OSF.IO/W645F and is publicly available as of the date of publication. Any additional information required to reanalyze the data reported in this paper is available from the lead contact on request.

Acknowledgments

This work was supported by Volkswagen Foundation, grant "Consequences of Artificial Intelligence for Urban Societies (CAIUS)" and Baden-Württemberg Foundation grant "FairADM – Fairness in Algorithmic Decision Making." This work was also supported by the German Research Foundation (DFG) under grant 139943784, "Collaborative Research Center SFB 884 Political Economy of Reforms (Project A8)," and by the University of Mannheim's Graduate School of Economic and Social Sciences. We thank the members of the Kreuter-Keusch research group, the CAIUS project team, and the anonymous reviewers for helpful comments on this paper.

Author contributions

Conceptualization, C.K., F.G., R.L.B., F. Keusch, and F. Kreuter; methodology, C.K., F.G., R.L.B., F. Keusch; formal analysis, C.K.; writing – original draft, C.K., F.G., R.L.B., F. Keusch, and F. Kreuter; writing – review and editing, C.K., F.G., R.L.B., F. Keusch, and F. Kreuter; visualization, C.K.; funding acquisition, C.K., R.L.B., F. Keusch, and F. Kreuter.

Declaration of interests

The authors declare no competing interests.

References

American Association for Public Opinion Research. (2016). *Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys, 9th edition (AAPOR)*. https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine Bias*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments*. SAGE.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech Era. *Journal of Financial Economics*, *143*(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

Beigang, S., Fetz, K., Kalkum, D., & Otto, M. (2017). *Experiences of discrimination in Germany Initial results of a representative survey and a survey of the people affected.* Nomos. https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/Expertisen/expertise_diskriminierungserfahrungen_in_deutschland.pdf?__blob=publicationFile&v=6

Bertelsmann Stiftung. (2019). *What Europe Knows and Thinks About Algorithms. Results of a Representative Survey*. https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WhatEuropeKnowsAndThinkAboutAlgorithm.pdf

Blom, A. G., Fikel, M., Gonzalez Ocanto, M., Krieger, U., Rettig, T., & Universität Mannheim. (2021). *SFB 884 'political economy of reforms'. German internet panel, wave 54 (july 2021) (GESIS Data Archive). Cologne. ZA7762 Data file Version 1.0.0. 44*. https://doi.org/10.4232/1.13835

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting up an online panel representative of the general population: The German internet panel. *Field Methods*, *27*(4), 391–408. https://doi.org/10.1177/1525822X15574494

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1), 4–36. https://doi.org/10.1093/jssam/smz041

Cornesse, C., Krieger, U., Sohnius, M.-L., Fikel, M., Friedel, S., Rettig, T., Wenz, A., Juhl, S., Lehrer, R., Möhring, K., Naumann, E., Reifenscheid, M., & Blom, A. G. (2021). From German internet panel to mannheim corona study: Adaptable probability-based online panel infrastructures during the pandemic. *Journal of the Royal Statistical Society: Series A*, *185*, 773–797. https://doi.org/10.1111/rssa.12749

Cornesse, C., & Schaurer, I. (2021). The long-term impact of different offline population inclusion strategies in probability-based online panels: Evidence from the German internet panel and the GESIS panel. *Social Science Computer Review*, *39*(4), 687–704. https://doi.org/10.1177/0894439320984131

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces (Marina Del Ray California)*, 275–285. https://doi.org/10.1145/3301275.3302310

Everett, J. A. C., Faber, N. S., & Crockett, M. (2015). Preferences and beliefs in ingroup favoritism. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00015

Gonzalez, M. F., Liu, W., Shirase, L., Tomczak, D. L., Lobbe, C. E., Justenhoven, R., & Martin, N. R. (2022). Allying with AI? reactions toward human-based, AI/ML-based, and augmented hiring processes. *Computers in Human Behavior*, *130*, 107179. https://doi.org/10.1016/j.chb.2022.107179

Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 903–912. https://doi.org/10.1145/3178876.3186138

Grgic-Hlaca, N., Weller, A., & Redmiles, E. M. (2020). *Dimensions of Diversity in Human Perceptions of Algorithmic Fairness*. https://arxiv.org/abs/2005.00808

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11296

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2). https://doi.org/10.2307/1914185

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, *13*, 795–848. https://doi.org/10.1007/s40685-020-00134-w

Koene, A., Perez, E., Ceppi, S., Rovatsos, M., Webb, H., Patel, M., Jirotka, M., & Lane, G. (2017). Algorithmic fairness in online information mediating systems. *Proceedings of the 2017 ACM on Web Science Conference (Troy New York USA*, 391–392. https://doi.org/10.1145/3091478.3098864

Körtner, J., & Bonoli, G. (2021). *Predictive algorithms in the delivery of public employment services* (SocArXiv j7r8y). Center for Open Science. https://doi.org/10.31219/osf.io/j7r8y

Krafft, T. D., Zweig, K. A., & König, P. D. (2020). How to Regulate Algorithmic Decision-Making: A Framework of Regulatory Requirements for Different Applications. *Regulation & Governance*, 1748–5991. https://doi.org/10.1111/rego.12369

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Langer, M., Hunsicker, T., Feldkamp, T., König, C. J., & Grgić-Hlača, N. (2022). '"Look! It's a computer program! It's an algorithm! It's ai!"': Does terminology affect human perceptions and evaluations of algorithmic decision-making systems? *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3491102.3517527

Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior*, *123*, 106878. https://doi.org/10.1016/j.chb.2021.106878

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 205395171875668. https://doi.org/10.1177/2053951718756684

Lopez, P. (2019). Reinforcing Intersectional Inequality via the AMS Algorithm in Austria. *Conference Proceedings of the 18th STS Conference Graz 2019: Critical Issues in Science, Technology and Society Studies*, 289–309. https://doi.org/10.3217/978-3-85125-668-0-16

López-Molina, N. B. (2021). *In Catalonia, the RisCanvi Algorithm Helps Decide whether Inmates Are Paroled*. https://algorithmwatch.org/en/riscanvi/

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2020). *On the Applicability of ML Fairness Notions*. http://arxiv.org/abs/2006.16745

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Nagtegaal, R. (2021). The impact of using algorithms for managerial decisions on public employees' procedural justice. *Government Information Quarterly*, *38*(1), 101536. https://doi.org/10.1016/j.giq.2020.101536

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, *160*, 149–167. https://doi.org/10.1016/j.obhdp.2020.03.008

Nissenbaum, H. (2019). Contextual integrity up and down the data food chain. *Theoretical Inquiries in Law*, *20*(1), 221–256. https://doi.org/10.1515/til-2019-0008

Peachey, K. (2019). *Sexist and Biased? How Credit Firms Make Decisions*. https://www.bbc.com/news/business-50432634

Pierson, E. (2018). *Demographics and discussion influence views on algorithmic fairness*. https://doi.org/10.48550/arXiv.1712.09124

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed). Sage Publications.

Rieger, T., Roesler, E., & Manzey, D. (2022). Challenging presumed technological superiority when working with (artificial) colleagues. *Scientific Reports*, *12*(1), 3768. https://doi.org/10.1038/s41598-022-07808-x

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2019). *Aequitas: A Bias and Fairness Audit Toolkit*. http://arxiv.org/abs/1811.05577

Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.

Scott Long, J. (1997). *Regression models for categorical and limited dependent variables*. Sage Publications.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized ai system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, *64*(4), 541–565. https://doi.org/10.1080/08838151.2020.1843357

Shin, D. (2021). The Perception of Humanness in Conversational Journalism: An Algorithmic Information-Processing Perspective. *New Media & Society*. https://doi.org/10.1177/1461444821993801

Skirpan, M., & Gorelick, M. (2017). *The Authority of 'fair' in Machine Learning*. http://arxiv.org/abs/1706.09976

Smith, J., Sonboli, N., Fiesler, C., & Burke, R. (2020). *Exploring User Opinions of Fairness in Recommender Systems*. http://arxiv.org/abs/2003.06461

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2021). *Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature*. http://arxiv.org/abs/2103.12016

Tajfel, H. (1978). Social categorization, social identity and social comparison. In H. Tajfel (Ed.), *Differentiation between social groups: Studies in the social psychology of intergroup relations* (pp. 61–76). Academic Press.

van Berkel, N., Goncalves, J., Hettiachchi, D., Wijenayake, S., Kelly, R. M., & Kostakos, V. (2019). Crowdsourcing perceptions of fair predictors for machine learning: A Recidivism case study. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–21. https://doi.org/10.1145/3359130

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, *41*, 105567. https://doi.org/10.1016/j.clsr.2021.105567

Waldman, A., & Martin, K. (2022). Governing algorithmic decisions: The role of decision importance and governance on perceived legitimacy of algorithmic decisions. *Big Data & Society*, *9*(1), 205395172211004. https://doi.org/10.1177/20539517221100449

Wang, H., Grgic-Hlaca, N., Lahoti, P., Gummadi, K. P., & Weller, A. (2019). *An empirical study on learning fairness metrics for compas data with human supervision*. https://doi.org/10.48550/ARXIV.1910.10255

Wang, R., Harper, F. M., & Zhu, H. (2020). Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376813

Weber, M., Yurochkin, M., Botros, S., & Markov, V. (2020). *Black Loans Matter: Fighting Bias for AI Fairness in Lending*. https://mitibmwatsonailab.mit.edu/research/blog/black-loans-matter-fighting-bias-for-ai-fairness-in-lending/

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*, *3*(4), 100455–100510. https://doi.org/10.1016/j.patter.2022.100455

4. Individual acceptance of using health data for private and public benefit: Changes during the COVID-19 pandemic[3]

Abstract. While the COVID-19 pandemic has been devastating, data collected in this context has unprecedented opportunities for data scientists. The stunning breadth of data obtained through new gathering systems put in place to manage the pandemic offers a richly textured view of a transformed world. Looking forward, privacy researchers worry that these new data-gathering systems risk running afoul of societal norms regarding the flow of information. Looking back at pre-pandemic public preferences with respect to data sharing may provide us some idea of what to expect in the future. In July of 2019, we happened to conduct a vignette study in Germany to examine the public's willingness to share data for fighting an outbreak of an infectious disease. In April of 2020, during the first peak of the pandemic, we repeated the study to examine crisis-driven changes in respondents' willingness to share data for public health purposes with three different samples. Public acceptance of the use of individual health data to combat an infectious disease outbreak increased notably between the two measurements, while acceptance of data use in several other scenarios barely changed over time. This shift aligns with the predictive framework of contextual integrity theory, and the data presented here may serve as a good reminder for policymakers to carefully consider the intended purpose of and appropriate limitations on data use.

4.1. Introduction

While the COVID-19 pandemic has been devastating for individuals, global health, and the economy, it has created unprecedented opportunities for data scientists. The stunning breadth of data, collected through new systems installed to manage the pandemic, offers a richly textured window into a transformed world (e.g., COVID-19 Data Exchange, 2020). These new

[3] This chapter was previously published as a paper in the journal *Harvard Data Science Review:*
Gerdon, F., Nissenbaum, H., Bach, R. L., Kreuter, F., & Zins, S. (2021). Individual Acceptance of Using Health Data for Private and Public Benefit: Changes During the COVID-19 Pandemic. *Harvard Data Science Review*, *Special Issue 1*. https://doi.org/10.1162/99608f92.edf2fc97

Only small formal edits were made in comparison to the published paper version.
The main Appendix (additional figures and tables) for this chapter is available in Chapter 7.3. References to the Appendix begin with the letter "A". Further material (questionnaires and vignettes) is available in an Online Appendix on OSF: https://osf.io/ehmpt/?view_only=c57e5d52475941199e7d36e7e958d5ef

systems repurpose data from familiar services and platforms, such as phone companies, operating system providers, and social media platforms, and deploy them in the service of efforts to increase information about people's movements and predict the spread of COVID-19 (e.g., Apple, 2020; Google, 2020). New smartphone applications track patterns of actions relevant to the spread of disease, and people are donating data from other digital devices (e.g., data4life, 2020; Ferretti et al., 2020; O'Neill et al., 2020; Robert-Koch-Institut, 2020; Whittaker, 2020).

Predictably, and understandably, privacy researchers have thrown up red flags concerning these developments, given they will likely persist long after immediate threats pass (Morley et al., 2020; Sanfilippo et al., 2020). Researchers worry that existing norms regarding privacy and data sharing in the population are being ignored, and state the public's willingness to accept data transmission, far from signifying widespread assent to the sacrifice of privacy across the board, is, in fact, confined to specific purposes (e.g., Martin & Nissenbaum, 2016).

In recent years, the framework of "contextual integrity" (CI) (Nissenbaum, 2010, 2018) has been proposed as a rubric with which to best judge—or encourage others to judge—the conditions under which a data-handling practice is appropriate. Contextual integrity posits that data transmissions meet privacy expectations when they conform with privacy norms, contingent upon the types and circumstances of information collected, as well as the actors involved.

While we cannot predict people's future preferences with respect to sharing their data, we can gather some insights from attitudes expressed prior to the COVID-19 outbreak that may help us clarify what is at stake in this area. In the summer of 2019, we happened to conduct a vignette study in Germany, the primary purpose of which was to test public willingness to share data for a public purpose vs. a private purpose through a survey experiment. Serendipitously, one of the public purpose examples was fighting an infectious disease.

We repeated the experiment in April of 2020 during the first wave of the pandemic with three samples. Once equipped with the set of additional experimental data collections, we addressed our original questions from the 2019 study: "Are people willing to share their individual data for a public purpose or are they more willing to share their data to benefit privately?" and "Are people equally willing to share their data for a public purpose across different areas such as public health, energy consumption, or traffic infrastructure?" We addressed a new question as well: "Did the public's attitude towards sharing individual information for the purpose of promoting public health change due to the COVID-19 pandemic?" While looking back at a potential attitude shift can only provide us with suggestive

insights regarding a possible post-pandemic attitude shift, such a comparison between past and future shifts, when seen through the lens of contextual integrity theory, may enrich the debate about the incorporation of sunset clauses into new technical developments for data collection.

We start out with a brief review of the contextual integrity framework, before describing the pre-COVID-19 experimental data collection, as well as our efforts to replicate the study and to collect additional data for bias assessments. After presenting cross-sectional and longitudinal analyses of the data, we discuss the political importance of this study, as well as implications for future research.

4.2. Contextual integrity and shifts in acceptance

Technological innovation has enabled an unprecedented advance in our capacity to acquire, analyze, communicate, and disseminate data. This advance has forced us to rethink our shifting understandings of and expectations concerning privacy. The concept of privacy, of course, has a complicated history, but many contemporary accounts of privacy reflect a focus on two dominant notions: namely, privacy as control and privacy as secrecy. Given the historical background of notions of privacy (Mulligan et al., 2016), this is not surprising. Yet arguably, the venerable notions of privacy as secrecy and control fail to capture what privacy means in a world of widely adopted digital information systems.

The theory of contextual integrity (Nissenbaum, 2010) offers a new way to think about privacy in our current situation. This approach defines privacy as *appropriate flow of data* where appropriateness is a function of conformity with contextual informational norms. These norms are derived from particular social domains, or contexts, where they attain legitimacy by prescribing flows that judiciously serve stakeholder interests and promote the purposes and values of the respective social domains (Nissenbaum, 2018). Contextual informational norms prescribe flow in terms of five key parameters: (1) the sender of the information, (2) the recipient of the information, (3) the attribute or type of information, (4) the subject of the information, and (5) a transmission principle that states the condition under which the information flow is permitted.

In order to assess whether a given practice respects or violates privacy, information flows associated with that practice are described by assigning values to each of these five parameters. For example, in the health care context, it is commonly accepted that patients (sender and subject) provide their doctors (recipient) with health information (attribute) in confidence (transmission principle). A practice that generates conforming data flows is unproblematic. However, if a practice diverts medical information to a different recipient, such as a patient’s

employer, a red flag is raised, even if all other factors remain the same. Equally critically, if any of the parameters is left unspecified, the description is ambiguous.

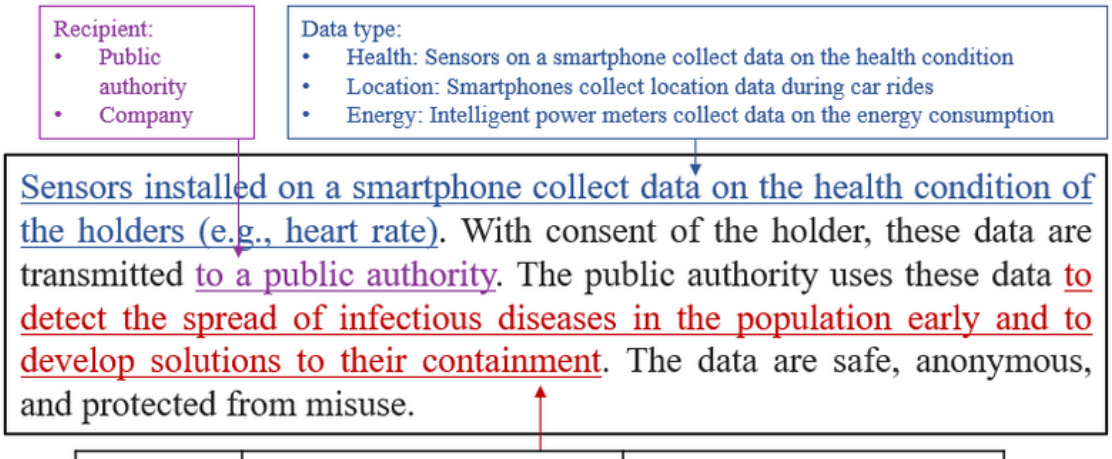
A series of empirical studies in which respondents were presented with different descriptions of data-sharing scenarios demonstrated that the approval of data sharing is contingent on situational parameters (Martin & Nissenbaum, 2017a, 2017b; Martin & Shilton, 2016). Martin and Shilton (2016), for instance, show that secondary use of tracking data for commercial purposes has a large negative impact on perceived appropriateness of data sharing, and Martin and Nissenbaum (2016) find that secondary data use driven by commercial interests meets individuals' privacy expectations less than the use of data in other contexts in which they were collected (for example, the use of information entered into a search platform to improve the search results vs. the use of this information to decide on advertisement shown when visiting other sites).

Over the past few decades, tremendous shifts in data collection practices on digital devices and online platforms have contributed to significant discontinuity between those practices and user privacy expectations. The COVID-19 pandemic adds to this misalignment, requiring quick decisions under intense conditions. Here, CI provides a useful analytic framework, allowing us to first fine-tune multiple factors influencing privacy perception and tailor necessary adjustment.

To empirically investigate the factors that influence the acceptance of data-sharing scenarios, we draw on the situational parameters suggested by CI to design descriptions of situations in which data are being shared. We focus on comparing the acceptance of public purposes and private purpose uses for different data types. Next, we provide details on our data-collection procedure and survey questionnaire.

4.3. A vignette study to measure public's willingness to share data

In 2019, we designed a vignette study or *factorial survey experiment* (Auspurg & Hinz, 2015) to experimentally test the public's willingness to share data for a public purpose vs. a private purpose. Each participant in this survey experiment was asked to rate one randomly chosen data-sharing scenario ('vignette') out of a total of twelve scenarios regarding the acceptability of data collection and use. Each scenario was followed by the question: "How acceptable is it to you to use these data for this purpose?" The answer scale had five points, ranging from 1 (Not acceptable) to 5 (Very acceptable) (see Online Appendix B). The answer to this question serves as the outcome in our analyses.



Depending on data type:	Private purpose	Public purpose
Health	… personal recommendations on health behavior	…to detect outbreaks of infectious diseases early and to develop solutions to their containment.
Location	… personal recommendations on driving behavior and routes	…to develop improvements of the local infrastructure
Energy	… personal recommendations on the optimization and reduction of the own energy consumption	… to develop a more efficient energy distribution system

Figure 4.1. Example vignette as well as dimensions and levels of the other vignettes. The vignettes varied along the indicated data type, recipient, and data use.

The descriptions presented to respondents were structured according to the theory of contextual integrity, that is, we specified values for the five key parameters (i.e., the data sender, data subject, data recipient, information type, and transmission principles; Nissenbaum, 2018). The vignettes varied along two of the five CI parameters: the information type to be transmitted and the recipient of the data. In addition, we varied the purpose of the data. Regarding information types, we investigated health, location, and energy consumption (see Horne & Kennedy, 2017) data. The recipient was either a company or public administration. For each data type, we constructed a public purpose and a private purpose (to the data recipient) of the data. For example, the suggested purpose in the health data vignettes was either personal recommendations for health behavior (private purpose) or contribution to the containment of infectious diseases (public purpose). We held the remaining three CI parameters constant across vignettes. The sender and the data subject were referred to as an unspecified individual (e.g., the "holder" of a smartphone or the "driver" of a car). The transmission principle was described as "with consent" and defined that the "data are safe, anonymous, and protected from misuse." The focus on the parameters we experimentally varied follows our substantive interests and the practical requirement to limit the number of total vignettes. Presenting all respondents with

relatively safe and cautious transmission principles should reduce effects of situation-specific privacy breach concerns.

For illustrative purposes, Figure 4.1 shows the survey vignette that asked about health data used by a public authority for a public health purpose (translated from German). In addition, the survey asked for respondents' age and gender, as well as information on their general privacy concerns. We also collected additional variables in the survey that we do not analyze in this article, such as the perceived sensitivity of several data types, and how much respondents trusted companies and public authorities. The latter variables were placed after the vignette in the questionnaire. The full questionnaires in English and German as well as a list of the vignettes are available in the Online Appendix (Appendices B and C).

4.4. Sample design and data collection in 2019

We implemented the factorial survey experiment in a cross-sectional survey fielded from July 9 to July 18, 2019, among individuals of age 18 to 69 in Germany (*cross-section 2019*). This was the original study we designed to experimentally test public's willingness to share data for a public purpose vs. a private purpose. A total of 1,401 people[4] participated in this study and responded to all questions.

The sample for this first study was drawn from an opt-in panel maintained by respondi AG, a survey vendor that maintains a pool of individuals interested in participating in market and social research studies. Individuals registered in such panels are usually recruited through banner ads placed on websites or on social media, and participation is usually open to everyone interested. For this reason, such panels are often referred to as *nonprobability online panels*. Researchers can buy access to a sample of participants from the survey vendor and ask them questions through online surveys. The survey vendor remunerates participants who successfully complete surveys with small financial incentives.

Samples from these nonprobability online panels are often drawn using river or quota sampling, hoping that the sample will mimic the population of a country. They offer a fast, cheap, and increasingly popular method for conducting experimental studies with high internal validity (Cornesse et al., 2020). Nonprobability online panels face a number of challenges, though. For example, when interested in obtaining accurate estimates of public opinion, bias may arise because people without internet access are not covered in participant pools and because samples consist of volunteers who self-select into participation in these panels

[4] All sample sizes refer to those respondents who responded to all questions and for which we have information on all weighting variables

(Bethlehem, 2017). Therefore, it is difficult to infer population totals from such data without relying on strong additional and often untestable assumptions regarding the data-generating process (Kohler et al., 2019). The focus of the 2019 study was thus on comparing the acceptance of public purpose and private purpose uses for different data types, and our experimental design allows us to obtain results with high internal validity. Due to the nonprobability sample, we cannot guarantee that our findings also represent broader public opinion in Germany, that is, that they have high external validity.

Nevertheless, to achieve a sample of respondents that represents the German adult population with regard to several predefined characteristics, we selected our sample from the vendor's pool using quota sampling. Quotas were based on age and gender population benchmarks for Germany, provided by Eurostat for 2018. Quotas were applied separately and not crossed. In addition, we weighted the final analysis sample using raking (Deville et al., 1993) to population benchmarks obtained from the German micro census for 2019. Age, gender, and state were used in the weighting procedure. While weighting procedure can reduce some of the bias that arises from using a sample from a nonprobability online panel, it is likely that more factors exist that influenced participation in our study.

4.5. Three additional surveys to study the effect of the 2020 pandemic

After the outbreak of the COVID-19 pandemic, we replicated the 2019 study to investigate the question we raised in the introduction (whether the public's attitude toward sharing individual information for the purpose of promoting public health changed as a result of the pandemic). For an ideal research design, we would have interviewed all of the 2019 respondents for a second time in 2020. Ignoring attrition, such a longitudinal sample would have allowed us to eliminate bias due to differences in the composition of the 2019 and the 2020 samples and to unobserved individual heterogeneity, for example, by using fixed effects regression modeling. Unfortunately, we planned the 2019 study as a single cross-sectional survey as, at the time, the pandemic was not contemplated. Therefore, we took several sampling approaches to combat potential biases. We selected a second cross-sectional quota sample from the nonprobability online access panel that we also used in 2019. This second survey was fielded from March 31 to April 5, 2020 (*cross-section 2020*), and we collected responses from 970 respondents who were not selected for the cross-section 2019 survey. We used the same experimental survey design and asked respondents the set of questions that we described here. In order to achieve a maximum of comparability of the two surveys over time, we selected the cross-section 2020

survey with the same quotas. However, we cannot exclude the possibility that the two surveys differ in their composition as the age and gender quotas were in both surveys applied separately and not crossed. We also weighted the cross-section 2020 survey using again the raking approach, but we note that differences remain in the distribution of age and gender between the cross-section 2019 and the cross-section 2020 surveys (see Table A4.1 in Appendix 7.3).

There may also be unobserved confounders that could result in bias when we use the two surveys to study change in acceptability of data collection and use between 2019 and 2020. For example, the pool of potential participants maintained by the survey vendor may have changed over time, and the factors driving individuals into participation may have changed from 2019 to 2020.

To address biases resulting from unobserved differences between the 2019 and the 2020 cross-section samples, we ran a third survey on the respondi survey platform (*longitudinal sample*). The survey vendor was able to identify and reinterview 627 participants of the 2019 survey. These respondents were still registered in the vendor's participant pool in 2020. Identification was based on unique participant IDs assigned to each participant by the vendor. We interviewed these participants for a second time in 2020, parallel to the cross-section 2020 survey using the experimental survey design and the set of questions described in the previous section. Each of these respondents received the same vignette they received in the survey of 2019. These 627 respondents who were interviewed in both 2019 and 2020 form a true longitudinal sample, which we used to assess the robustness of our analyses with respect to both observed and unobserved individual heterogeneity.

Furthermore, we collected responses to a fourth online survey that we ran with a different survey vendor (forsa) between April 2 and April 7, 2020. Forsa runs a similar online panel of participants interested in answering survey questions. The design of the panel is, however, fundamentally different (Baker et al., 2010). Forsa panelists are originally recruited through a probability-based telephone survey. Therefore, it should be less affected by bias due to individuals self-selecting into the participant pool, but we note that it may still be affected by biases due to differential nonresponse, for example. We refer to this sample as *benchmark 2020*. We used the experimental design and the set of questions described in the previous section also in the benchmark 2020 survey.

We used a similar quota-sampling approach to select the benchmark sample (N = 801). Crossed age-gender quotas that mimic the German adult population were provided by forsa. We also weighted the benchmark 2020 sample using the raking procedure and the population benchmarks mentioned here.

Survey:	**Cross-section 2019**	**Cross-section 2020**	**Longitudinal sample**	**Benchmark 2020**
Purpose	1. Sharing individual data for a public purpose vs. benefitting privately 2. Sharing individual data for a public purpose across data types	Changes in sharing individual data for a public purpose (public health) in response to COVID-19 pandemic	Assess robustness of results with respect to sample composition over time	Assess robustness of results with respect to sample recruitment
Field period	7/9 – 7/18 2019	3/31 – 4/5 2020	7/9 – 7/18 2019 and 3/31 – 4/5 2020	4/2 – 4/6 2020
Number of complete responses (unweighted)	1,401	970	1,254 (627 respondents)	801
Recruitment of participant pool	Quota based sample from nonprobability online access panel	Quota based sample from nonprobability online access panel	Quota based sample from nonprobability online access panel	Quota based sample from probability online panel with initial phone recruitment

Table 4.1. Characteristics of the analysis samples.

We collected the benchmark 2020 sample to assess the robustness of the estimates obtained from the nonprobability survey cross-section 2020. While there is no guarantee that using a quota sample selected from a probability sample and weighting the data will remove bias due to, for example, differential nonresponse, using a probability-based online survey weighted to census data is backed by statistical theory that provides justification for confidence, and continuously performed well when compared to population benchmarks (Cornesse et al., 2020). Table 4.1 presents a summary of the characteristics of our data collections and indicates which questions we answer with each survey.

4.6. Analytical strategy

We use the cross-section 2019 data to answer our first research question (whether people are willing to share their data for a public vs. private purpose) and our second research question (whether people are equally willing to share data for a public purpose across different data types). We examine responses to the 5-point Likert-scale question asking for respondents'

acceptance to use their data. The variable ranges from 1 ("Not acceptable") to 5 ("Very acceptable").

Our analytical strategy to answer the third research question (whether the public's attitude toward sharing individual information for the purpose of promoting public health changed due to the COVID-19 pandemic) is inspired by the difference-in-differences (DiD) approach (Wooldridge, 2010, ch. 6). DiD is a popular technique for evaluating policy interventions in economics and in the social sciences. DiD designs require four groups (see Figure 4.2). First, a treatment group measured prior to treatment, and second, a control group measured prior to treatment. Third, we need a treatment group measured *after* it was treated and, fourth, a control group that did not get the treatment but was also measured *after* treatment was given to the treated.

We think of the pandemic as the treatment, therefore, the cross-section 2019 survey as the pretreatment measurement and the cross-section 2020 survey as the post-treatment measurement. Furthermore, we think of those who were asked about health data as the treated group and those who were asked about non-health data as the control group. The rationale for this is that the health data vignettes described scenarios directly related to the pandemic (sharing health data for personal health behavior recommendations and the detection of an outbreak of an infectious disease), while the non-health data vignettes described scenarios completely unrelated to the pandemic (e.g., sharing data for improving energy-saving measures). We assume that the pandemic influenced privacy attitudes related to health data while leaving attitudes toward sharing other data types mostly unchanged. Of course, it is possible that the pandemic also affected attitudes toward sharing other data types. However, we assume that such effects should be much smaller than the effect of the pandemic on sharing health data.

We apply the same logic to our analysis of the question of whether the pandemic affected respondents' acceptance of health-data sharing *for public purposes*. In two of the four health vignettes, we described a scenario where the transmitted data were used for a public purpose. Specifically, we asked how acceptant respondents were of transmitting their health data to help "detect outbreaks of diseases early and to develop solutions to their containment*"* (see above section for details). We treat these two scenarios as the treated conditions in our analysis of change over time.

The control group conditions are restricted to the two health-data–sharing scenarios with a private purpose ("provide the holders with personal recommendations on their health behavior"). These did not mention a public health crisis. It is not unlikely that the pandemic also affected control-group participants' data-sharing attitudes as the vignette mentioned

recommendations on health behavior. However, we assume that the pandemic had a larger effect on participants' acceptance to share health data for public purposes. That is, we restrict the data to those respondents who answered a health vignette with either public or private purpose (cross-sectional samples: $N = 784$, longitudinal sample: $N = 203$ per wave).

In the traditional DiD logic, we are interested in comparing the difference between the mean outcome of the pretreatment treatment and control groups with the difference in the mean outcomes of the treatment and control groups after treatment has been assigned. Thereby, pretreatment differences between the treatment and the control groups will be removed from post-treatment comparisons of the treatment and control groups.

The key assumption for our design is the parallel trends assumption. That is, we need to assume that had there been no treatment (i.e., had there been no pandemic), the outcomes of the treatment and the control groups would have evolved similarly. In other words, we need to assume that there is no event in Germany between 2019 and 2020 that changed attitudes toward *only one* data type (health data but not non-health data and public purpose but not private purpose and vice versa). In addition, we need to assume that the two cross-sectional samples are truly comparable such that we can attribute any difference in privacy attitudes between the treatment and the control groups in 2020, after adjusting for differences observed between the two groups in 2019, to the pandemic alone. Figure 4.2 illustrates the idea of the design.

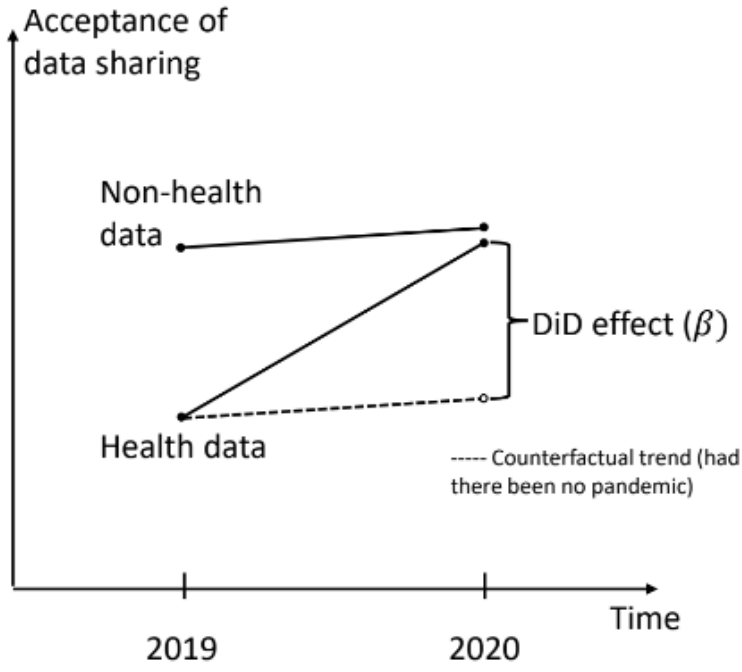


Figure 4.2. Difference-in-differences (DiD) identification strategy. Schematic representation of a mean comparison.

With continuous outcomes, the DiD effect is defined as the difference between the means of the control group outcome and the treatment group outcome *after* treatment has been assigned, subtracted from the difference between the means of the control group outcome and the treatment group outcome *before* treatment has been assigned (Wooldridge, 2010, ch. 6). Athey and Imbens (2006) and Yamauchi (2020) used DiD-like procedures for discrete outcomes for simple random samples. To avoid further assumptions on our outcome variable (treating it as continuous) and to allow for the proper use of survey weights, we conduct a series of Kolmogorov–Smirnov (KS) tests for two discrete samples following the logic described above. The KS test is a nonparametric test that does not require the estimation of standard errors for the test statistic. This is an advantage, as it would be difficult to infer the distribution of most statistics of interest under our survey estimation strategy. Since the distribution of the test statistic of a KS test is also unknown for weighted survey data, we implemented a KS permutation test. We simulate the distribution of the test statistic under the null hypothesis (the data from the two samples are independent and identically distributed, e.g., there is no effect of the pandemic) and we implement the following. In a first step we resample the observations in each sample proportional to their respective weights by sampling from a list of indices. Each index of the list corresponds to one sample element and one element only and is repeated proportional to the weight of the element it corresponds to. Random unbiased rounding is used to coerce noninteger weights into integers. In a second step the indices selected in step 1 are completely randomly permuted. In a third step we calculate the KS test statistic as the maximum distance between the empirical cumulative distribution function (ECDF) of the values corresponding to the first n_1 indices and the last n_2 indices, where n_1 and n_2 are the sizes of the two resamples selected in the first step. Steps 1 to 3 are repeated 1,000 times. We then calculate the proportion of the KS test statistics, calculated in step 3, that are larger than the test statistic based on the original samples and our survey weights. This proportion is the *p*-value for our (one-sided) test. Because the permutation test may tend to reject a null hypothesis too easily for small sample sizes, we compare our test results with those of a more conservative KS test where we estimate the ECDFs using our survey weights. The *p*-values for these tests are obtained from the theoretical distribution of the KS test statistic for two simple random samples. Numerical examples showed that this simple random sample assumption resulted in consistently more conservative *p*-values than with the permutation test. We use these conservative KS tests as robustness checks for our test decisions based on the permutation test.

For our analysis, we use the software *R* (R Core Team, 2020) with the packages *ggpubr* (Kassambara, 2020), *gridExtra* (Auguie, 2017), *sampling* (Tillé & Matei, 2021), *scales*

(Wickham & Seidel, 2020), *srvyr* (Ellis & Schneider, 2020), *survey* (Lumley, 2020), *tidyverse* (Wickham, 2017), and *viridis* (Garnier, 2018). All analyses report weighted estimates.

4.7. Results

In this section, we describe the empirical findings from our four surveys. We first present results from the cross-section 2019 survey and answer the questions regarding differences in sharing data for a public vs. private purpose and sharing data for a public purpose across data types. Second, we report descriptive findings of changes in sharing individual information for a public purpose (public health) in response to the COVID-19 pandemic before turning to results of the KS permutation tests. We conclude this section with several sensitivity and robustness analyses.

4.7.1. Contextual integrity matters for acceptability of data transmission

Figure 4.3 presents acceptance levels for each data type by recipient (public agency vs. private company) and use (public vs. private purpose) using the weighted cross-section 2019 data. We show mean values to provide a quick and simple descriptive impression of the results, while the distributions for all groups are shown in the Appendix (Tables A4.2 and A4.3, Figures A4.2a–e). We find clear evidence that context matters when individuals judge the appropriateness of data transmission. Overall, respondents find the use of health data less acceptable than the use of location or energy data. Furthermore, the figure shows that respondents find it equally acceptable but often more acceptable to transmit data to a company than to a public authority or agency. However, transmission of data seems also to depend on the intended use of the data. Individuals find it in many scenarios more appropriate to transmit data for private purpose to a company than to a public agency. Regarding sharing individual data for a public purpose vs. sharing such data for private benefit, we do not find a consistent pattern across data types.

Looking at each data type separately, we find some evidence that individuals deem it more acceptable to transmit health data for a private purpose (here, personal recommendations on health behavior) to a company than to transmit health data to a public authority or agency for a public purpose (containment of infectious diseases). In fact, rather strikingly, transmitting health data to a public agency for a public purpose is least accepted. For location data, individuals find it equally acceptable to transmit data for a public purpose (here, develop

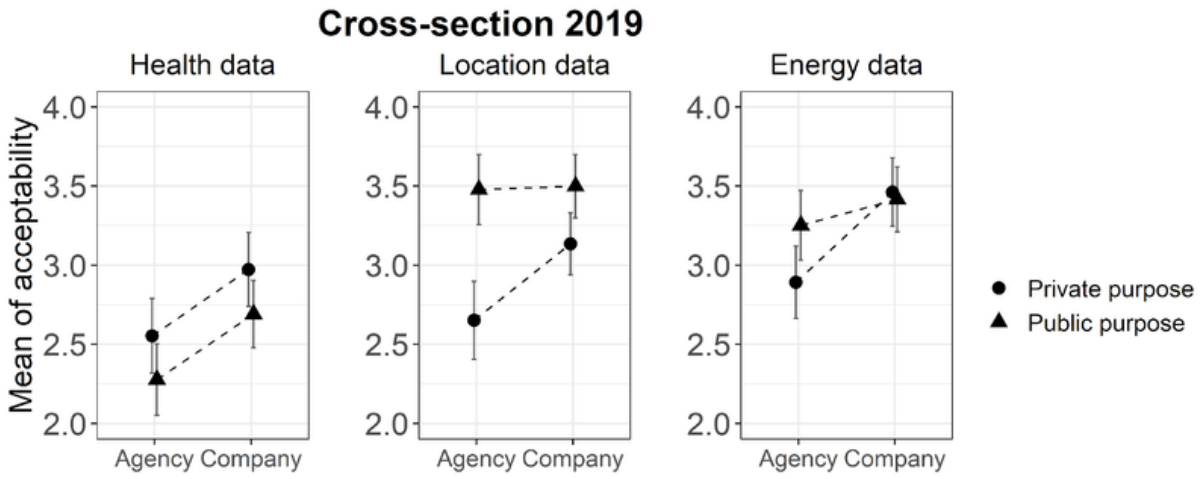


Figure 4.3. Mean acceptability of different data transmissions, depending on data type, data use, and recipient of the data. Vertical bars indicate 95% confidence intervals. N = = 1,401. Weighted analysis.

improvements of the local infrastructure) to an agency or a private company. Transmitting data to an agency for a private purpose (personal recommendations on driving behavior and route) is least accepted. Regarding energy data, differences do not seem as pronounced. It seems that only transmitting data to an agency for a private purpose (personal recommendations on optimization of energy consumption) is less accepted than the other scenarios.

Therefore, regarding differences in sharing data for a public purpose vs. benefitting privately, we find a strong dependency on data type, but also on the recipient of the data.

4.7.2. Longitudinal analysis and the effect of the pandemic on sharing of health data

Next, we compare the distribution of the outcome variables over time and between the groups defined in Section 6. The top row of panels in Figure 4.4 shows that acceptance to transmit data changed for both health and non-health scenarios from 2019 to 2020. Overall, respondents were more likely in 2020 to judge the transmission of health data as acceptable. This effect seems to be mainly driven by fewer respondents choosing the extreme category "1 – Not acceptable" in 2020 than in 2019. At the same time, respondents found it less acceptable to transmit non-health data over time. The KS permutation tests indicate that both changes over time are statistically significant ($p < .05$, see rows three and four in Table A4.4 in the Appendix). The more conservative KS tests indicate insignificant differences in both cases. Visual inspection of the distributions suggests that the change in health data over time is much more pronounced than the change in non-health data over time, however.

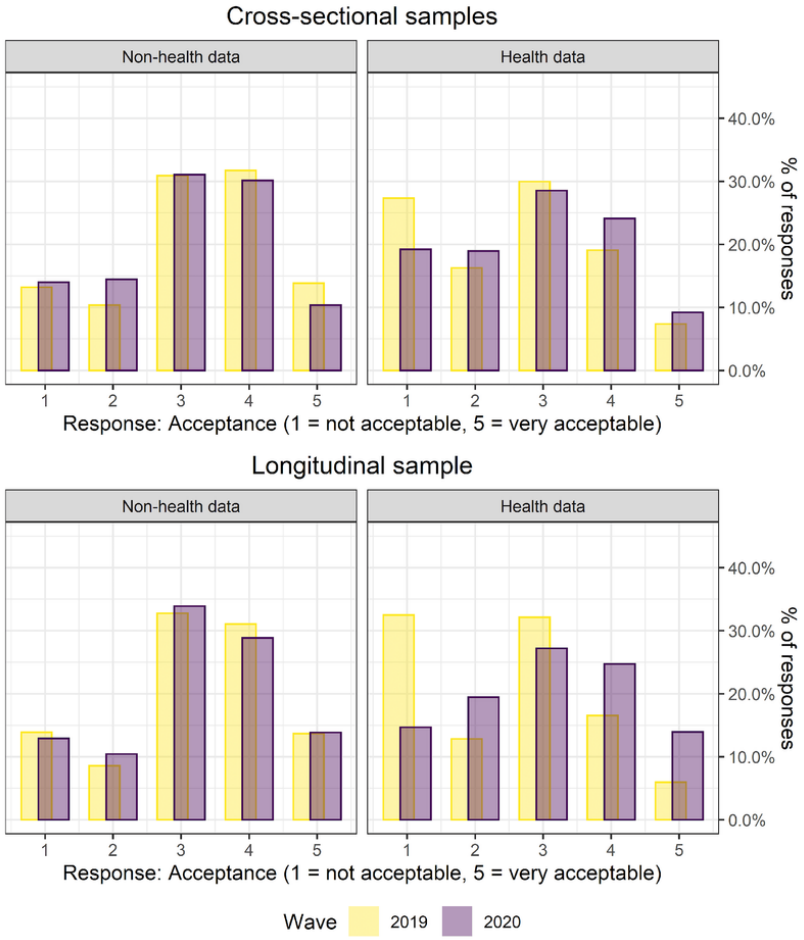


Figure 4.4. Relative frequency of acceptance for respondents shown health or non-health vignettes, by wave. Cross-sectional samples: *N* = 2,371. Longitudinal sample: *N* = 627 per wave. Weighted analysis.

The longitudinal sample confirms this finding (Figure 4.4, bottom row). With this sample, differences between change in health data over time and the change in non-health data are even more pronounced. Transmitting health data became more acceptable, while transmitting non-health data did not change much. Here, the KS permutation tests indicate that the change over time for health data is statistically significant, while it is not statistically significant for non-health data (rows seven and eight in Table A4.4 in Online Appendix A). The conservative KS tests confirm these findings. It is likely that the results obtained with the longitudinal sample are more accurate, as the two cross-sectional samples differ in their compositions while the longitudinal sample does not (see Section 5).

Looking at changes over time *within* health data, we find that the increased levels of acceptance we reported are mainly driven by increased acceptance to share health data for a *public* purpose. Respondents chose the lowest acceptance category less often and the two highest categories more often for public purpose health data (Figure 4.5, top, right panel). At the same time, visual inspection suggests that sharing health data for a private purpose changed to a much smaller degree and in the opposite direction. Indeed, our KS permutation tests show

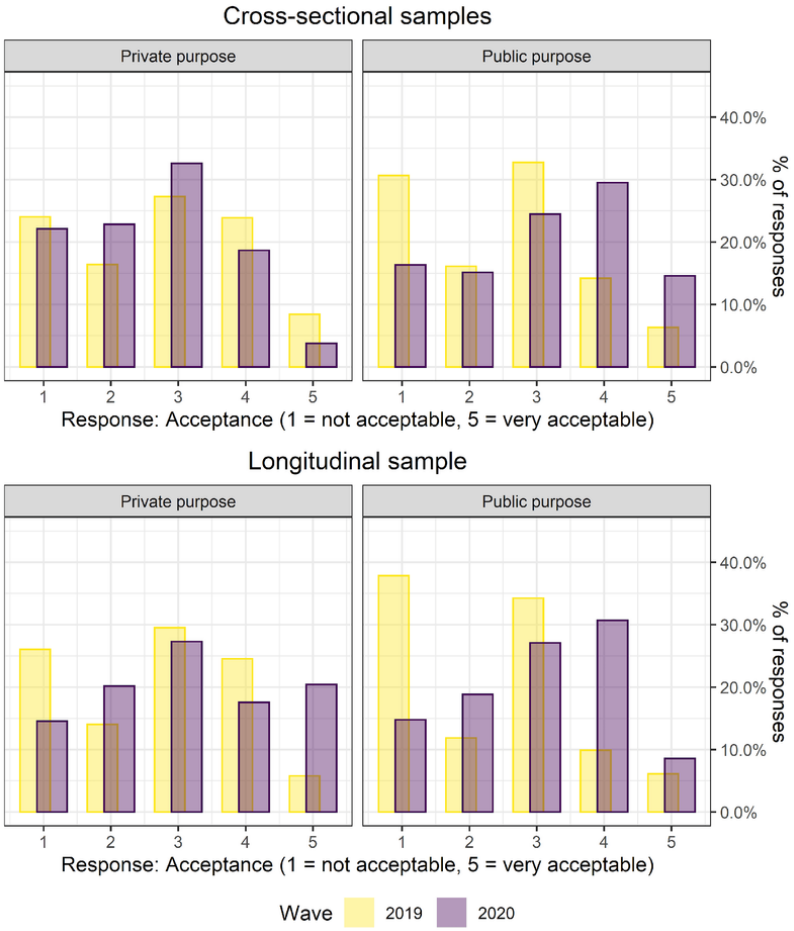


Figure 4.5. Relative frequency of acceptance for respondents shown a health vignette with a public purpose or a health vignette with a private purpose, by wave. Cross-sectional samples: N = 784. Longitudinal sample: N = 203 per wave. Weighted analysis.

that the change over time in acceptance to share health data for a public purpose was significant, while it was not significant for private purpose health data (rows 11 and 12 in Table A4.4 in Appendix 7.3). Overall, these findings are supported by the longitudinal sample. Sharing health data for a public purpose was more accepted in 2020 than in 2019, while sharing health data for a private purpose changed to a smaller degree. This is confirmed by the KS permutation tests, which indicate significant changes over time for a public purpose but not for a private purpose (rows 13 and 14 in Table A4.4 in Appendix 7.3). The conservative KS tests confirm the findings for both groups.

As we discussed, our research design is inspired by the DiD approach. Therefore, one would ideally net out the change over time in the non-health data / private purpose scenarios (our control groups) from the change in the health data / public purpose scenarios (our treatment groups) over time to adjust for baseline shifts. Given that we find substantial changes over time for health data and public purpose health data, respectively, but only mild shifts for non-health

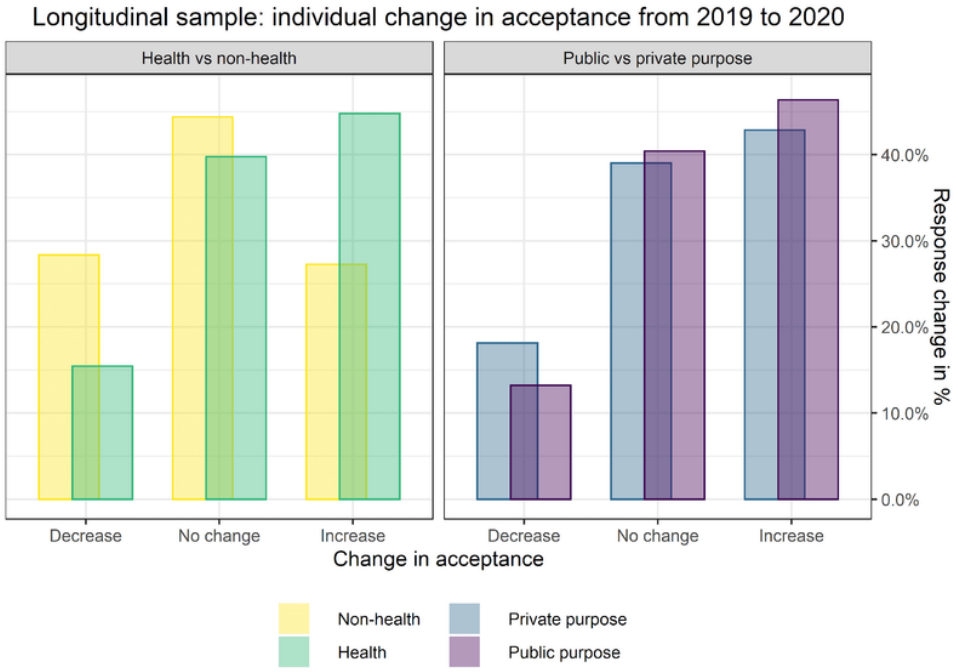


Figure 4.6. Changes in response category chosen by the respondents from 2019 to 2020 in the longitudinal sample. Left panel: Cross-sectional samples: *N* = 2,371. Longitudinal sample: *N* = 627. Right panel: Cross-sectional samples: *N* = 784. Longitudinal sample: *N* = 203. Weighted analysis.

data and private purpose health data, we are confident that the findings reported here would also hold when controlling for baseline shifts.

For the longitudinal sample, we additionally test differences in the number of respondents who changed or did not change their answer from 2019 to 2020. That is, we calculate how many respondents chose a lower response category in 2020 than in 2019, how many did not change their answer, and how many chose a higher response category (Figure 4.6). We then compare the distributions of these three categories (lower in 2020, same answer, higher in 2020) between respondents who answered to a health data scenario and respondents who answered to a non-health data scenario using the KS permutation test. In addition, we conduct this test for the comparison between private purpose health data sharing and public purpose health data sharing. Note that it is not possible to run these analyses with the cross-sectional samples, as we do not observe the same respondents in the two samples.

The left panel of Figure 4.6 shows a clear pattern: the share of respondents choosing a higher acceptance category in 2020 than in 2019 is much larger for health vignette respondents than for non-health vignette respondents. Vice versa, the share of respondents choosing a lower acceptance category in 2020 than in 2019 is much smaller for health vignette respondents than for non-health vignette respondents. The KS permutation test also indicates that the distributions are in fact different between health vignette and non-health vignette respondents

($p = 0$). Regarding differences between the change in acceptance to share public purpose health data and private purpose health data, the right panel of Figure 4.6 shows a similar pattern. The share of respondents changing their response toward a more favorable answer in 2020 compared to 2019 is higher among public purpose respondents. At the same time, the share of respondents who chose a less favorable answer in 2020 than in 2019 is higher among private purpose respondents than among public purpose respondents. The differences between the two groups are less pronounced than those between health and non-health vignette respondents, and our KS permutation does not indicate that the distributions are different in a meaningful way ($p = 1$). The conservative KS tests confirm the results of permutation tests for both cases of public and private purpose use of health data.

4.8. Discussion

When we first designed this study, we set out to empirically investigate the factors that influence the acceptance of data-sharing scenarios through a survey experiment and by drawing on the situational parameters suggested by CI theory. One of the most striking results of this experiment is that individuals in Germany perceive the sharing of health data with a public agency, irrespective of a private purpose or a public purpose, as least acceptable among a series of data types. With this result in mind, the signs for public support of tracking, predicting the spread of, and fighting a pandemic like COVID-19 with data on people's movements and contacts but also information about their health were far from positive.

It may be possible that, back then, the idea of a pandemic such as COVID-19 with its devastating consequences for individuals, global health, and the economy was too abstract for individuals to fully evaluate the potential benefits that sacrificing some privacy might generate. Amid the influence of the COVID-19 pandemic, public opinion toward the acceptability of sharing health data for private purpose but also for a public purpose changed, resulting in increased levels of acceptability. That is, we may conclude that individuals judge the flow of information for fighting a public health crisis as more appropriate when both the devastating consequences of a public health crisis but also the benefits of sharing data become apparent.

We should be careful when considering the question of whether individuals will judge the flow of information as equally appropriate once the pandemic has ended. We suspect, from looking back at pre-pandemic times, it is likely the public's judgment of the appropriateness may decrease again. Future work should replicate our data collection as the pandemic proceeds and eventually ends. Moreover, future data collections may be designed to study additional

questions such as whether individuals' judgment of appropriate data flows is a function of the severity of the pandemic. In addition, more work will be needed to learn whether and how increased levels of acceptance during exceptional times might generalize to other contexts and, more interestingly, to circumstantial changes that might suggest shifts in expectations.

From a policy perspective, our analysis and application of contextual integrity theory suggest the need to reevaluate practices post-pandemic. For these reasons, we call for government policymakers, software developers, and the general public to pay attention to the contextual purposes served by given data practices (sometimes enabled by technical systems) and be ready to adapt data use and storage policies accordingly.

However, we also need to consider that our findings and the implications discussed here are derived from a study that, originally, was never intended to include a longitudinal perspective. In 2019, we could not anticipate that a pandemic would change circumstances in such meaningful ways that we would run a second survey just a few months after the original 2019 study. As a result, several limitations arise. First, we observe that there are differences in the compositions of the two cross-sectional surveys. Although both samples were selected from the same survey platform and with the same specifications, our quota sampling specifications did not cross age and gender quotas but applied them separately, resulting in differences in age and gender compositions. We addressed these differences by weighting both cross-sectional samples to population benchmarks obtained from the German micro census. Unfortunately, weighting could not remove all differences between the two samples. In addition, our analyses rely on the assumption that had there been no pandemic, outcomes of the health data scenarios and the non-health data scenarios would have evolved in a similar way. Unfortunately, we can neither test this assumption itself nor assess its plausibility by, for example, analyzing temporal leads of the outcome variable (see, e.g., Autor, 2003).

Second, it is likely that there are additional unobserved differences between the two cross-sectional samples that may bias our analyses of change in the outcome over time. We did not collect information beyond respondents' age, gender, and state. Since we already observe that there are differences on these two observed confounders, it is likely that additional (unobserved) variables could also differ between the two samples, thereby biasing our analyses of change in data-sharing acceptance.

We addressed these differences by identifying a true longitudinal sample of respondents interviewed in both 2019 and 2020. In general, results obtained with this sample point in a similar direction as the results obtained from the two cross-sectional samples.

Regarding the size of the effects identified, we note that the shift in acceptance to transmit data is small. However, this is not completely unexpected as other studies investigating, for example, the public's willingness to install apps developed to facilitate the tracing of potentially infected people find high levels of support for such apps, but a fair number of individuals not willing to use such apps due to privacy concerns (see, e.g., Altmann et al., 2020). Moreover, uptake of such apps in various countries indicate that actual use of such technologies is likewise far from universal (Mosoff et al., 2020).

Overall, our results indicate a favorable shift toward the idea of using individuals' data for efforts designed to fight the COVID-19 pandemic. This is good news for data scientists and the public health system if these attitudes translate into a high rate of access to the data needed to address the crisis. Whether these attitudes prevail over the course of the pandemic and beyond will be interesting to watch, and we hope research will continue as well. In the meantime, however, public policymakers and researchers should keep in mind that the public's approval of these activities is limited to specific contexts and purposes.

Acknowledgments

We thank the editor, Xiao-Li Meng, Stephanie Eckman, Felix Henninger, Christoph Kern, Florian Keusch, Pascal Kieslich, Johannes Ludsteck, Sonja Malich, Ido Sivan-Sivilia, and Patrick Schenk for helpful comments on earlier versions of this paper, and Ann Sarnak, Suzanne Smith, and Jason McMillan for editing help.

Disclosure Statement

This research was partially funded by the Volkswagen Foundation: "Consequences of Artificial Intelligence for Urban Societies," as well as the Deutsche Forschungsgemeinschaft (DFG, project numbers 396057129 and 139943784, SFB 884). This work was supported by the University of Mannheim's Graduate School of Economic and Social Sciences. Supporting H Nissenbaum, we gratefully acknowledge US National Security Agency (The Science of Privacy: Implications for Data Usage, H98230-18-D-006) and US National Science Foundation (SaTC: CORE: Medium: Collaborative: Contextual Integrity: From Theory to Practice, CNS-1801501).

References

Altmann, S., Milsom, L., Zillessen, H., Blasone, R., Gerdon, F., Bach, R., Kreuter, F., Nosenzo, D., Toussaert, S., & Abeler, J. (2020). Acceptability of app-based contact tracing for COVID-19: Cross-country survey study. *JMIR mHealth and uHealth, 8*(8), Article e19857. https://doi.org/10.2196/19857

Apple. (2020). *Mobility trends reports*. https://covid19.apple.com/mobility

Athey, S., & Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica, 74*(2), 431–497. https://doi.org/10.1111/j.1468-0262.2006.00668.x

Auguie, B., (2017). *gridExtra*: Miscellaneous functions for "grid" graphics (R package version 2.3) [Computer software]. R Foundation. https://CRAN.R-project.org/package=gridExtra

Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments*. SAGE. https://doi.org/10.4135/9781483398075

Autor, D. H. (2003). Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing. *Journal of Labor Economics, 21*(1), 1–42. https://doi.org/10.1086/344122

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D. A., Frankel, M. R., Garland, P., Groves, R. M., Kennedy, C., Krosnick, J., Lavrakas, P. J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R. K., & Zahs, D. (2010). AAPOR report on online panels. *Public Opinion Quarterly, 74*(4), 711–781. https://doi.org/10.1093/poq/nfq048

Bethlehem, J. G. (2017). *Understanding public opinion polls*. CRC Press Taylor & Francis Group. https://doi.org/10.1201/9781315154220

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., de Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology, 8*(1), 4–36. https://doi.org/10.1093/jssam/smz041

COVID-19 Data Exchange. (2020). *Support & contribution*. https://www.covid19-dataexchange.org/support-contributors

data4life. (2020). *COVID-19 survey*. https://www.data4life.care/en/corona/pulsecheck/

Deville, J. C., Särndal, C. E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association, 88*(423), 1013–1020. https://doi.org/10.1080/01621459.1993.10476369

Ellis, G. F., & Schneider, B. (2020). *srvyr*: "*dplyr*"-like syntax for summary statistics of survey data (R package version 0.4.0) [Computer software]. R Foundation. https://CRAN.R-project.org/package=srvyr

Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., Parker, M., Bonsall, D., & Fraser, C. (2020). Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science, 368*(6491), Article eabb6936. https://doi.org/10.1126/science.abb6936

Garnier, S. (2018). *viridis*: Default color maps from "*matplotlib*" (R package version 0.5.1) [Computer software]. R Foundation. https://CRAN.R-project.org/package=viridis

Google. (2020). *COVID-19 community mobility reports*. https://www.google.com/covid19/mobility

Horne, C., & Huddart Kennedy, E. (2017). The power of social norms for reducing and shifting electricity use. *Energy Policy, 107*, 43–52. https://doi.org/10.1016/j.enpol.2017.04.029

Kassambara, A. (2020). *ggpubr*: "*ggplot2*" based publication ready plots (R package version 0.2.5) [Computer software]. R Foundation. https://CRAN.R-project.org/package=ggpubr

Kohler, U., Kreuter, F., & Stuart, E. A. (2019). Nonprobability sampling and causal analysis. *Annual Review of Statistics and Its Application, 6*, 149–172. https://doi.org/10.1146/annurev-statistics-030718-104951

Lumley, T. (2020). *survey*: Analysis of complex survey samples (R package version 4.0) [Computer software]. R Foundation. https://CRAN.R-project.org/package=survey

Martin, K., & Nissenbaum, H. (2017a). Measuring privacy: An empirical test using context to expose confounding variables. *The Columbia Science & Technology Law Revie*w, *18*(1), 176–218. https://doi.org/10.7916/stlr.v18i1.4015

Martin, K., & Nissenbaum, H. (2017b). Privacy interests in public records. An empirical investigation. *Harvard Journal of Law & Technology, 31*(1), 111–143. https://doi.org/10.2139/ssrn.2875720

Martin, K., & Shilton, K. (2016). Putting mobile application privacy in context: An empirical study of user privacy expectations for mobile devices. *The Information Society*, *32*(3), 200–216. https://doi.org/10.1080/01972243.2016.1153012

Morley, J., Cowls, J., Taddeo, M., & Floridi, L. (2020). Ethical guidelines for COVID-19 tracing apps. *Nature, 582*(7810), 29–31. https://doi.org 10.1038/d41586-020-01578-0

Mosoff, R., Friedlich, T., Scassa, T., Bronson, K., & Millar, J. (2020). *Global pandemic app watch (GPAW): COVID-19 exposure notification and contact tracing apps*. GPAW. https://craiedl.ca/gpaw/

Mulligan, D. K., Koopman, C., & Doty, N. (2016). Privacy is an essentially contested concept: A multi-dimensional analytic for mapping privacy. *Philosophical Transactions of the Royal Society Series A, 374*(2083), Article 20160118. https://doi.org/10.1098/rsta.2016.0118

Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press. https://www.sup.org/books/title/?id=8862

Nissenbaum, H. (2018). Respecting context to protect privacy: Why meaning matters. *Science and Engineering Ethics, 24*(3), 831–852. https://doi.org/10.1007/s11948-015-9674-9

O'Neill, P. H., Ryan-Mosley, T., & Johnson, B. (2020, May 7). A flood of coronavirus apps are tracking us: Now it's time to keep track of them. *Technology Review*. https://www.technologyreview.com/2020/05/07/1000961/launching-mittr-covid-tracing-tracker/

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Robert-Koch-Institut. (2020). *Corona-Datenspende*. https://corona-datenspende.de/science/en/

Sanfilippo, M. R., Shvartzshnaider, Y., Reyes, I., Nissenbaum, H., & Egelman, S. (2020). Disaster privacy/privacy disaster. *Journal of the Association for Information Science and Technology, 59*(9), 1–13. https://doi.org/10.1002/asi.24353

Tillé, Y., & Matei, A. (2021). *sampling*: Survey sampling (R package version 2.9) [Computer software]. R Foundation. https://CRAN.R-project.org/package=sampling

Whittaker, J. (2020, September 22). Data from your FitBit could help predict COVID: Research suggests wearable devices could help in virus fight. *Cayman Compass*.

https://www.caymancompass.com/2020/09/22/data-from-your-fitbit-could-help-predict-covid

Wickham, H. (2017). *tidyverse*: Easily install and load the "*tidyverse*." (R package version 1.2.1) [Computer software]. R Foundation. https://CRAN.R-project.org/package=tidyverse

Wickham, H., & Seidel, D. (2020). *scales*: Scale functions for visualization (R package version 1.1.1) [Computer software]. R Foundation. https://CRAN.R-project.org/package=scales

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press.

Yamauchi, S. (2020). Difference-in-differences for ordinal outcomes: Application to the effect of mass shootings on attitudes toward gun control. *arXiv*. https://doi.org/10.48550/arXiv.2009.13404

5. Attitudes on data use for public benefit: Investigating context-specific differences across countries with a longitudinal survey experiment[5]

Abstract. With technological advances, governments and companies gain opportunities to collect data to provide public benefits. However, such data collections and uses need to fulfill ethical standards and comply with citizens' privacy preferences. These privacy preferences may vary by social context, between countries and individuals, and longitudinally. The Comparative Privacy Research Framework suggests several specific comparative dimensions that may shape such privacy-related perceptions. I propose how to integrate into this framework a specific meso-level perspective for concisely operationalizing specific data uses context-specifically: the privacy theory of contextual integrity, as developed by Helen Nissenbaum. This paper presents an empirical application of this approach by investigating specific data use scenarios across countries, while simultaneously considering temporal and international variations as well as individual-level variables. To this end, an online survey experiment was conducted in three countries (Germany, Spain, and UK) in December 2022 and May 2023. In this experiment, respondents rated the appropriateness of fictitious data use scenarios. The scenarios varied by data type, the data recipients, and the conditions of data use. The results show that the effects of contextual parameters vary across countries to different degrees. Respondents react particularly sensitively to changes in data types, with health data being overall most accepted to be used. The relative acceptance of the data recipients clearly varies across countries. Country-level individualism is not found to be consistently related to the desired level of control over data. These findings highlight the usefulness of contextual integrity to unmask meso-level, context-specific variations in privacy attitudes across countries. A meso-level perspective that operationalizes data uses according to contextual integrity can therefore inform comparative privacy research as well as (international) privacy-related policy-making.

[5] An updated version of this chapter is currently under review (revised and resubmitted after a second round of reviews) as a single-authored paper for a special issue of the journal *Social Media + Society*. The main Appendix (details on methods, additional figures) for this chapter is available in Chapter 7.4. References to the figures in the Appendix begin with the letter "A". Further material (additional tables that are referred to in the text with the letter "T", questionnaires, vignettes) is available in an Online Appendix on OSF: https://osf.io/ehmpt/?view_only=c57e5d52475941199e7d36e7e958d5ef

5.1. Introduction

Numerous data collection practices that aim to provide individual benefits produce data that may be also used for a public benefit. Digital patient records, smartphone movement collections, smart home and smart grid technologies, social media data – all these data can be used to provide immediate benefits to individuals, but also could be used for, e.g., scholarly research or the improvement of public management. Novel data collection efforts and an internationalization of data markets, such as envisioned by Common European Data Spaces (European Commission, 2023), increase the opportunities for such public benefit data uses. However, these opportunities come with privacy concerns, for example voiced by scholars who worry about undue surveillance (Newlands et al., 2020; Vitak & Zimmer, 2020). For ethical data collection and use, we need to design data use practices that are informed by citizens' preferences, among others.

However, one-off privacy surveys on public preferences focusing on specific perceptions at a specific time and place are not readily generalizable across countries and contexts. Given the internationalization and cross-sectoral application of data regulations, more fine-grained comparisons become increasingly important. In the present paper, I combine the comparative perspective of the Comparative Privacy Research Framework (Masur et al., 2021) with the context-based notion of privacy as "contextual integrity" (Nissenbaum, 2010) to show that investigating countries with respect to general privacy attitudes might miss important nuances in people's perspectives on which kinds of data uses are acceptable. Contextual integrity can therefore meaningfully enhance the comparative power of the Comparative Privacy Research Framework by providing a template for context-specific comparisons across countries.

To this end, I empirically investigate how attitudes related to data use for public benefit vary across social contexts within *and* between countries. I conduct an international survey experiment that presents respondents with text descriptions of hypothetical data use scenarios. These scenarios vary by parameters as suggested by contextual integrity (data type, involved actors, conditions of data use), such that effects of changes in parameters on respondents' acceptance can be estimated. The study was fielded in three countries with different levels of individualism, which was previously shown to be related to the acceptance of public benefit data use (Li et al., 2017; see subsection *Sample*): Germany, Spain, and the United Kingdom (UK). Moreover, I conduct the study at two time points as privacy attitudes may change over time with the salience of public issues related to a specific data use (Gerdon et al., 2021), and finally also consider individual-level predictors of acceptance.

In summary, the main research question is: *Under which conditions do people deem data use for the public benefit appropriate, and do such attitudes vary across social contexts and countries over time?* I answer this question with respect to the outlined four components of comparison: contextual, international, interindividual, and longitudinal. The resulting evidence on variations of privacy attitudes along these components allows (1) researchers to learn about the variability of privacy attitudes across countries as depending on social context and over time and thereby (2) policymakers to consider people's preferences for an appropriate regulation of data use for public benefit.

5.2. Attitudes on Data Use for Public Benefit: Comparisons by Four Components

The collection of specific pieces of information about individuals may serve different kinds of purposes. For instance, physicians may collect health data of patients to provide diagnoses and treatments. The very same collected health data may also be used by researchers to study, for example, risk factors for specific diseases. The former purpose provides a direct personal benefit to the individual, while the latter purpose can lead to public benefits (such as better treatment options) that may translate into personal benefits. Another example are mobility data (e.g., from smartphones) that can be used by companies for the personal benefit of drivers by suggesting optimal routes to destinations. These mobility data could also be used by researchers or public agencies to learn about mobility behavior on a fine-grained level for infrastructure planning that benefits the local population.

More generally, following the definition by the National Data Guardian for Health and Social Care in England (2022), a "public *benefit*" arises from data use if the achieved benefits are not outweighed by risks, while benefits can also be of indirect nature. Furthermore, according to this definition, to be a "*public* benefit", the broader public or a subsection of the public need to benefit, such that exclusively commercial benefit does not fall under this definition. Additionally, the legitimacy of a public benefit data use hinges on whether it has a "social license" (Carter et al., 2015; Shaw et al., 2020) granted by the population. We therefore need to learn whether and under which conditions the (re-)use of individual data use for public benefit purposes is deemed acceptable by the public. With the internationalization of regulations on data collection and use, it becomes increasingly important to also know how populations of different countries differ in their acceptance of specific data uses. Such knowledge could aid in formulating policies by showing how populations might react differently to specific data use

endeavors (e.g., differences in likelihoods of opting in or out of sharing data from digital patient records with researchers).

In the present paper, I argue that such international comparisons can be substantially enriched by comparing perceptions relating to *specific* data use contexts, additional to rather general privacy perceptions. To this end, I draw on the Comparative Privacy Research Framework (Masur et al, 2021) and extend it with the perspective of "contextual integrity" (Nissenbaum, 2010). The Comparative Privacy Research Framework cautions against the over-generalization of findings from privacy studies that focus on single contexts and offers a structured approach to, among others, international comparisons (Masur et al., 2021). More precisely, Masur et al. (2021) propose to study privacy-related phenomena, such as data use attitudes or behaviors, by comparing at least two structural units (on the macro-, meso-, or micro-level) in which these phenomena occur. They define five types of structures (cultural, social, political, economic, and technological) that may influence phenomena or moderate processes.

Understanding privacy as contextual integrity (as proposed by Nissenbaum, 2010) can enrich the Comparative Privacy Research Framework by drawing attention to the specific configuration of social contexts and their respective privacy norms. Nissenbaum understands social contexts as areas of social life such as health care and work that come with specific, e.g., practices, roles, purposes, and norms (Nissenbaum, 2018). For instance, in the health care context, there may be specific rules, practices, and expectations of how data collected about a patient by a physician may be used and shared. Masur et al. (2021) explicitly refer to contextual integrity to be placed on the meso-level and call for systematically analyzing contextual factors. These meso-level social contexts are embedded in larger structural units such as political systems. What contextual integrity adds to the Comparative Privacy Research Framework is a concrete template to operationalize data uses within meso-level social contexts. Comparisons of attitudes towards data uses in these meso-level contexts can enhance macro-level country comparisons by unmasking context-specific differences in privacy attitudes beyond "general" privacy attitudes in the investigated countries (relatedly, see Martin & Nissenbaum, 2016). At the same time, country comparisons can reveal how the relevance of specific contextual factors varies across countries (e.g., Li et al., 2017). For instance, international differences in the acceptance of digital patient records could in principle deviate from international differences found for general privacy concerns or general perceptions towards health data, and the influence of who exactly would receive these data could vary across countries.

Additional to these macro- and meso-level comparisons, on the micro-level, individuals may display different more general stances towards privacy (Gerber et al., 2018). For instance, social structures (such as age and gender, Masur et al., 2021), general privacy perceptions (Smith et al., 2011), and altruism (Kim & Stanton, 2016; Silber et al., 2022) may shape how acceptable individuals deem data use for public benefit in a given situation. All these structural units may interact with each other to affect privacy attitudes (Masur et al., 2021), such that, e.g., the effects of individual characteristics may vary by country and be more relevant in some social contexts than in others.

Finally, norms and attitudes within units of comparison may change over time due to changes in the societal environment, as previous research has demonstrated with respect to an (potentially temporary) increase of acceptance of use of health data for disease containment early in the COVID-19 pandemic (Gerdon et al., 2021). The Comparative Privacy Research Framework is explicitly open for longitudinal comparisons, due to the potential of major events to affect privacy perceptions (Masur et al., 2021).

Table 5.1 summarizes the comparative approach of the present paper, which I will further explain in the following subsections. In the following, I apply this theoretical background to the concrete case of privacy attitudes towards data use for public benefit.

Level	Structure	Concepts
Macro	Culture (country-level)	Individualism
Meso	Social context	Contextual integrity parameters: data type, actors, transmission principle
Micro	Individual perceptions and behaviors	E.g., privacy concerns and trust in data recipients (see Section *Method*)
	Longitudinal comparison	

Table 5.1. Structures that I simultaneously compare in this paper, based on the Comparative Privacy Research Framework (Masur et al., 2021).

Contextual variation

As argued above, country-level macro comparisons of general privacy attitudes may miss important meso-level contextual differences (relatedly, see Martin & Nissenbaum, 2016). The notion of contextual integrity can enhance such comparisons by asserting that the appropriateness of data flows depends on compliance with privacy norms that are specific to the social contexts in which they are embedded (Nissenbaum, 2010). To assess the

appropriateness of data flows, contextual integrity requires us to define *which* data are at stake under involvement of *which actors* and *under which conditions*. For instance, individuals may be willing to share detailed health information with doctors. At the same time, they might find it outraging if employers requested the voluntary sharing of such data. To concretely analyze and assess the appropriateness of a data flow, Nissenbaum provides a data flow template that consists of five parameters: data type, data subject, data sender, data recipient, and transmission principles (i.e., the prerequisites under which the data flow occurs).

Previous research has repeatedly shown that individual evaluations of data flows are sensitive towards changes in the specifications of data flow parameters (e.g., Martin & Nissenbaum, 2017; Terpstra et al., 2023; Utz et al., 2021). I now turn to discussing the contextual integrity parameters in more detail with respect to data use for public benefit and develop hypotheses and research questions.

With respect to data types, contextual integrity does not suggest that any data type is *as such* necessarily more sensitive than another, since sensitivity depends on context (Martin & Nissenbaum, 2017). Empirically, relatively much research has been dedicated to the specific case of health data use. For several kinds of health data, literature reviews have identified that public benefit uses are overall acceptable if the data are safe, the recipients are deemed trustworthy, and commercial interests are not the main focus, among others (Aitken et al., 2016; Hutchings et al., 2020; Kalkman et al., 2022). For social media data, research found that the acceptance of research uses depends on factors such as the research purpose, with a preference towards context-specific user experience research (Gilbert et al., 2021), but the acceptance of research may vary across social media platforms (Gilbert et al., 2023).

Acceptance of public benefit data use may further be affected by salient societal issues. One useful theoretical perspective is provided by Büchi et al. (2022) who draw on the theory of planned behavior and argue that privacy-related scandals may, in the long run, lead to more chilled digital communication behavior. I apply this argumentation to other societal events that make specific issues salient and could therefore (temporarily) affect individual's attitudes on data use for issue-related contexts. Furthermore, I argue that salience may also lead to more *appreciative* attitudes towards data use, depending on the specific salient event. For instance, previous research has demonstrated that the COVID-19 pandemic increased the acceptance for the use of health data collected on smartphones for public benefit (Gerdon et al., 2021). However, such effects may be temporary (see below). For example, several attitudes relating to surveillance for public security were more favorable in the US right after 9/11, but this attenuated in the following months and years (Best et al., 2006).

To test this relationship, I compare the acceptance of health data use with the acceptance towards other data types that vary in their relatedness to currently debated public issues, such as the COVID-19 pandemic. This includes energy use data which became a potentially salient issue in 2022 in the face of increasing energy prizes and the need to save energy. As less immediately salient data types, I investigate the yet important and frequently debated types of location data (e.g., see the critical discussion by Walsh, 2023) and social media data (Proferes & Walker, 2020).

> H1: *The use of health data for public benefit is more accepted than the use of other data types for public benefit that are less directly related to the pandemic.*

With respect to data recipients, relatively much comparative research is available for health data. Studies show that researchers or associated institutions appear as more accepted health data recipients than government agencies, while companies are least accepted (Kim et al., 2015) and, for Germany, that pharmaceutical companies are less accepted than researchers associated with universities or research-related public agencies (Haug et al., 2023). While these results suggest relatively high acceptance of health data use by public entities, not all studies share this finding (Gerdon et al., 2021; or for public benefit purposes: Deruelle et al., 2023), and the findings may further vary by concrete data recipient and consent procedure. Some studies suggest that individuals could find the private sector using specific types of health data acceptable if public benefits stood above profit (see Aitken et al., 2016).

From a contextual integrity perspective, as I focus on public benefit purposes, a tendency towards higher acceptance for public recipients can be expected. Public recipients usually more frequently take part in contexts that have the explicit goal to foster public welfare than private recipients and therefore be deemed appropriate. Within public recipients, I expect that respondents consider researchers affiliated with public institutions to be the least likely expected to use data for out-of-context purposes and therefore may enjoy the highest acceptance rates. These relationships may vary by concrete data type and the conditions of data use.

> H2: *Public actors, and particularly public researchers, are more accepted than private actors as recipients of data to use for public benefit. The effect of recipient interacts with data type and transmission principles.*

With respect to transmission principles, several conditions to share data for public benefit exist. In opt-in scenarios, data are only used after the individual explicitly consents to data use. In the context of health data use for research, a review study found opt-in as the most favored

approach, while results varied when de-identified data were to be shared (Stockdale et al., 2019). The review also concludes that individuals may change their opinions upon learning more about the benefits for research. However, review studies found that consent for use of medical records correlates with individual characteristics and that data sets that only include consented data may be biased (De Man et al., 2023; Kho et al., 2009). Opt-out approaches partially diminish this problem as data would be used as long as individuals do not explicitly indicate that they do not want their data to be used. A third option is to rely on data access regimes that include oversight bodies (Ausloos et al., 2020), such as *Findata* in Finland (ibd.).

However, we know little about which of these transmission principles are more accepted across public benefit contexts (for context-specific research see, e.g., on using phone data during the COVID-19 pandemic: Office of the Australian Information Commission & Lonergan Research, 2020). Given the scarcity of cross-context research, I formulate an open research question on this parameter.

> RQ1: *Which modes of consent do individuals accept more than other modes for data use for public benefit?*

International variation

As argued above, evaluations of social contexts, and thereby the effectiveness and acceptance of international policies surrounding these contexts, may vary by country. Previous privacy research has paid particular attention to cultural differences between countries by drawing on Hofstede's cultural dimensions (see Hofstede et al., 2010), which may also shape acceptance of data use for public benefit. However, one should be cautious to assume that there was a "national culture" permeating all domains of social life consistently (Masur et al., 2021). Different social contexts may have their very own privacy-related norms (Nissenbaum, 2010) that are not fully determined by general cultural orientations.

Among the cultural dimensions that pertain to the Hofstede approach, some scholars assess the individualism dimension to be the most central dimension for privacy by which to compare cultures (as discussed in Liu, 2022). Empirical research frequently identified effects of individualism on privacy-related phenomena. For instance, an international survey experiment found that public benefit uses of data are relatively more accepted than other uses by individuals with a more collectivist cultural background as compared to individuals with a more individualist cultural background (Li et al., 2017). The study also found that individualism is related to stronger effects of the option of "notice and control" methods on acceptance, and to lower acceptance of government as data collector. In a similar vein, another study on contact

tracing apps found higher use willingness in China (which is considered rather collectivist) than in Germany and the US (which are comparatively more individualist) (Utz et al., 2021). However, other studies provided an opposite relationship or null findings for individualism (see Trepte & Masur, 2016; Engström et al., 2023). Beyond individualism, less consistent or null effects have been found for the dimension of uncertainty avoidance (Engström et al., 2023; Schumacher et al., 2023; Trepte et al., 2017).

Given these findings, higher levels of individualism in a country may be associated with a higher desire of transmission principles that allow the affected individual more control over data flows. However, international differences may be hard to pinpoint to individualism with few countries of comparison, as countries may differ in further respects. Under this circumstance and partly contradictory findings, I approach international differences with an exploratory research question.

> RQ2: *Do countries with higher levels of individualism desire higher levels of control over data flows?*

Interindividual variation

Additional to the macro- and the meso-level, there may also be variation on the micro-level of interindividual in assessing data use for public benefit, i.e., individual differences within and between countries. Attitudes on data use for public benefit can relate to either of its constitutive elements *data use* (i.e., privacy attitudes) and *public benefit*. Concretely, I distinguish between four types of relevant individual-level variables: (1) general attitudes and perceptions related to privacy, (2) perceptions with respect to specific elements of data flows, (3) general attitudes and perceptions related to the provision of public benefits, and (4) affinity towards technology and socio-demographic variables. First, individuals may differ with respect to privacy concerns, for instance due to personality characteristics and own privacy-related experiences (Smith et al., 2011), and with respect to how they value privacy. The acceptance of data use scenarios may vary with individual general privacy concerns – possibly mediated by scenario-specific perceptions – (Kehr et al., 2015) and, second, with general perceptions relating to the parameters of the specific scenario: trust in data recipients (Kao & Sapp, 2022; Trein & Varone, 2023) and perceived sensitivity of data types (Mothersbaugh et al., 2012). Third, data use specifically for public benefit may be more accepted among individuals who value such public benefits higher more generally (relatedly for issue importance: Trein & Varone, 2023) and who have a more positive relationship to or picture of society, e.g., who are more altruistic (Kim & Stanton, 2016; Silber et al., 2022) and have higher interpersonal trust. Fourth, given the focus

on digital data collection in the present paper, familiarity with digital technologies may also affect privacy perceptions (e.g., Park, 2013). Finally, shared experiences and interests of social groups such as age and gender groups, i.e., the "social" structures of the Comparative Privacy Research Framework (Masur et al., 2021), may shape privacy perceptions (Schomakers et al., 2019).

RQ3: *Does the overall level of acceptance of data use scenarios vary with age, gender, general privacy concerns, perceptions relating to specific flow parameters, perceptions on public benefit uses of data, altruism, interpersonal trust, and with affinity towards technology?*

Longitudinal variation

Finally, comparisons on all three levels (macro, meso, and micro) are contingent on the specific time point of the comparison. Perceptions on the importance and the salience of privacy may change with major societal events (Büchi et al., 2022). Given that privacy has been intensely discussed during the COVID-19 pandemic, e.g., due to contact tracing, privacy perceptions related to public health may have changed. A previous study found that acceptance of use of health data from smartphones for disease containment has increased from 2019 to Spring 2020, but not for other non-pandemic-related data use scenarios, which supports the notion of context-dependent effects of societal developments (Gerdon et al., 2021). This notion has also found support with a longitudinal study on privacy attitudes in the US which has shown that acceptance to use fitness tracker data for medical research increased from 2019 to 2020 and then stayed higher, while this was not the case with government collecting data to counter terrorism, towards which acceptance decreased (Goetzen et al., 2022). However, salience or its effect may wane over time: Wnuk et al. (2021) conducted a longitudinal study in Poland and found that acceptance for (partly rather intrusive) surveillance technologies decreased between May and December 2020, i.e., during the COVID-19 pandemic. This tendency did not change even when the pandemic threat was particularly high in the second wave of the pandemic.

Research therefore suggests that privacy attitudes related to health data may vary with the severity of the pandemic situation. However, it is unclear how severe the shifts in societal circumstances need to be to affect privacy attitudes. I compare developments in attitudes towards health data use with attitudes towards the use of other data types that are either also affected by current public issues (energy use data) or that are less immediately affected by current public issues (location and social media data).

RQ4: *Does the acceptance of health data use, relative to the acceptance of using other data type, change with the pandemic situation?*

5.3. Method

5.3.1. Experimental design and questionnaire

To compare a variety of data use scenarios within multiple social contexts, an online survey experiment ("vignette experiment", see Auspurg & Hinz, 2015) was conducted. In this experiment, people's attitudes towards several hypothetical scenarios in which data are used for the purposes of research and public management were measured. This experiment entails 33 text descriptions (so-called "vignettes") of hypothetical data use scenarios and allows researchers to estimate how changes in scenario characteristics affect acceptance. The vignettes vary by factors that can take on different levels (ibd.), with the factors representing contextual integrity parameters. Table 5.2 shows the factors and levels. The full list of vignettes is available in the Online Appendix (*Section V*).

The vignettes are constructed around four data types: health data (digital patient records), location data (smartphone location), household energy use data (see Horne et al., 2015), and social media data. Particularly health data have been very salient during the COVID-19 pandemic. To some extent, this may be true for location data as well as contact tracing apps sparked discussions on how to store information on contacts of individuals. Energy usage is unrelated to the pandemic but may have been particularly salient in winter 2022 due to public discussions about energy supply. Finally, social media data are of particular interest to researchers, but not directly related to the pandemic.

As for data recipients, individuals may vary by their concerns about public and private data recipients, which is why public agencies get compared to researchers from universities and companies. Note that for the social media data type, I exclude public agencies as a data recipient as this might appear as a too intrusive scenario to respondents. Finally, I differentiate between three transmission principles under which individual data may be shared with the recipients: opt-in, opt-out, and a combination of opt-out with an ethics board (see subsection *Contextual variation*).

Factors	Levels	Text
Data type[1]	*Health*	"a person's health, diseases, and treatments"
	Location	"location of smartphones"
	Energy use	"the energy consumption of household appliances"
	Social media	"a person's social media usage (for example Facebook and Twitter)"
Data recipient[2]	*University researchers*	"university researchers"
	Researchers at an Internet company	"researchers at an Internet company"
	Public agency	Further specified according to data type, e.g., "local public planning agency"
Transmission principle[3]	*Opt-in*	"…[recipient] may use this information for this purpose only if [data subject] agrees…"
	Opt-out	"…[recipient] must not use this information for this purpose in any case if [data subject] rejects…"
	Ethics board	"These data are stored at a national data storage centre. The [recipient] need[s] to request these data from this centre. A committee of independent ethics experts working at this centre decides on the request." + opt-out text

Table 5.2. Experimental design: vignette factors and levels.
[1] The data subject is described depending on the data type (e.g., "resident" for energy use data).
[2] Company recipients were always defined as "researchers at an Internet company", as these can be associated with handling different types of data for various purposes. As there is no public agency recipient that could be directly associated with such a multitude of data uses, the vignettes refer to different specific public agencies depending on data type. The two public and the one private recipient come with different data use purposes to create realistic scenarios. The purpose for university and company researchers always is research, and for agencies it is planning. For instance, researchers (both public and private) may use health data to study diseases. Public agencies are presented to use data for planning and control purposes, e.g., location data for infrastructure planning.
[3] To describe the "ethics board", I refer to "committee of independent ethics experts" and "data centres" as simplifications that work across data types. The exact means to either accept or reject the data use is adjusted to data type to increase plausibility. The basis for all descriptions of the transmission principles is that individuals are informed about the data use.

Combining four data types, three recipients, and three transmission principles, and excluding governmental agencies as a recipient for social media data use, results in a total of 33 vignettes. For example, the vignette with the combination health data (data type), university researchers (recipient), and opt-in (transmission principle) reads:

> Information about persons' health, diseases, and treatments can be stored in a digital patient record. Each person is informed in a doctor's office about the possibility that university researchers could use this information in anonymous form to study diseases. This information does not contain the person's name or address.
> Researchers may use this information for this purpose only if the person agrees verbally or in writing via a form after being informed.

Each respondent was presented with four vignettes, receiving exactly one random vignette for each of the four data types. The order of shown data types was random, with one exception: I treat the social media vignette as a separate experiment in order to maintain a fully factorial experimental design (see Auspurg & Hinz, 2015) for the other vignettes. Thus, it was always presented in the last position such that it does not affect the ratings of the other vignettes.

Respondents were asked to evaluate each presented scenario. To capture approval in general terms for scenarios that may not yet exist, or which respondents may not be aware of, the question was: "To what extent would you say that the use of the information as described above is appropriate or not appropriate?", with a fully labelled seven-point scale ranging from "Completely appropriate" to "Not at all appropriate".

Afterwards, respondents were shown items that measure relevant concepts for the research questions on interindividual variation. This includes, in the following order: general privacy concerns (adopted and slightly edited version from Trepte, 2020); trust in all possible vignette data recipients (based on ESS Round 9: European Social Survey, 2021); perceived sensitivity of all possible vignette data types (based on Pew Research Center, 2014); two items on agreement with statements relating to control over and public use of individual data (from Trepte, 2020); whether respondents think that their concerns about and importance of privacy changed from before to during the COVID-19 pandemic (loosely based on Office of the Australian Information Commission & Lonergan Research, 2020); interpersonal trust measured by an index based on three variables as presented in the *Interpersonal Trust Short Scale* (KUSIV3), where the sum is divided by the number of items (Nießen et al., 2021); altruism (from SOEP-IS Group, 2021); technical affinity, measured by the number of regularly used communication devices (based on Bauer et al., 2022); and additive indices of two subscales (*General* and *Safe application*) of the *Information and Communication Technology Self-Concept Scale* (ICT-SC25) to measure affinity towards technology (Schauffel et al., 2021).

The questionnaires are available in *Section Q* and information on cognitive pre-tests and a pilot study is available in *Section M* of the (Online) Appendix.

5.3.2. Sample

The vignette experiment was conducted as part of an online survey. The survey was fielded in three countries that varied in their levels of individualism according to the Hofstede cultural values index (Hofstede Insights, 2023): Germany, Spain, and the UK. According to Hofstede's dimensions, the UK is a state with a high level of individualism, while Germany displays medium values (ibd.). Spain is one of the countries with the lowest individualism scores in Europe (ibd.).

Respondents were invited to participate via a commercial non-probability online panel provider (*Bilendi*). Respondents can self-select into the respondent pool and are then invited and incentivized by the provider to participate in specific surveys. For this survey, crossed age and gender quotas were applied that correspond to the respective population distributions based on Eurostat data from 2020. While inference with non-probability samples is oftentimes problematic (Elliott & Valliant, 2017), the goal is to estimate effects of experimental stimuli, which non-probability samples can be useful for (Jamieson et al., 2023; Kohler & Post, 2023), and to explore associations with individual-level variables.

The survey was fielded at two time points: December 14 to 21, 2022 (*Wave 1*) and May 11 to 22, 2023 (*Wave 2*). Based on these two cross-sectional samples, I also constructed a longitudinal data set of respondents who participated in both waves. In the second wave, each recurring respondent received the same vignettes in the same order as in the first wave. The societal environment with respect to health and energy data perceptions potentially differs between these two time points, as December is a colder month where infections and the energy crises overall may affect people's lives more than in spring. The two time points are also not too far separated, which makes it less likely that other major events happen that systematically affect perceptions related to the vignette scenarios between the surveys.

The sample sizes for analysis are as follows: Wave 1: 1,682 respondents (Germany: 562; Spain: 564; United Kingdom: 556); Wave 2: 1,795 respondents (Germany: 594; Spain: 603; United Kingdom: 598). Wave 2 comprises 1,110 respondents who already participated in wave 1; the remaining number of participants was newly recruited. To exclude potentially inattentive respondents from the analyses, these samples do not include "speeders", i.e., respondents who completed the questionnaire in less than 60% of the country- and wave-specific median response time (Roßmann, 2010). A small number of other respondents have been excluded for, e.g., not agreeing to reading the questions carefully (see Conrad et al., 2017). Details on the sample and exclusion criteria are available in Appendix 7.4 (*Section M*). The final distribution of age and gender for each country and wave is available in the Online Appendix (Table T1).

The age distribution shifts towards a higher prevalence of older age groups in wave 2 and the longitudinal sample as compared to wave 1.

5.4. Results

To address the hypotheses and research questions on contextual and international variation, I use the data from wave 1 and regress perceived appropriateness on the vignette variables and on country dummy variables, while controlling for age and gender. To this end, I run linear mixed-effects models with random intercepts for the respondent-level. I investigate further context-dependencies of effects of data recipients by adding respective interaction terms. To answer the research question related to interindividual variation, I add individual-level variables to the regression models. For the longitudinal comparisons, I inspect changes in acceptance (note that I will use the terms "acceptance" and "perceived appropriateness" interchangeably for easier text flow) of specific scenarios from wave 1 to wave 2. Finally, I use the wave 2 data for a replication of wave 1 results by running the same regression analyses as in wave 1 and inspecting whether substantive changes occur.

All models include only those respondents who have no missing or "Don't know" (or similar) values for any of the individual-level variables that are introduced in the later extended models. However, I do not remove all vignette responses by one respondent from the data set if the respondent has missing or "Don't know" values only for single vignettes. Note that I analyze responses to the "social media" vignettes as a separate experiment (see *Method* section) using OLS regression.

I furthermore ran two additional types of models to check whether the findings are robust to model and data choices: (a) models that include respondents that I defined as speeders and (b) logistic cumulative link mixed models (Christensen, 2019) that treat the outcome variables as ordinal. I focus on the interpretation of the linear mixed-effects models and highlight important substantive differences compared to the two other types of models. Note that I cannot directly compare effect sizes of the ordinal models with other models (Mood, 2010). Table 5.3 summarizes the models that will be shown in the *Results* section.

Note that although I draw on a non-probability sample, I calculate standard errors as an orientation using the respectively implemented procedures of the software. However, the focus is on the interpretation of effect strength.

Summary statistics for all individual-level variables are available in the Online Appendix (Table T2a for wave 1 and Table T2b for wave 2).

For data preparation, analysis, and presentation/visualization, I use *R version 4.0.4* (R Core Team, 2021) with the libraries: *ggpubr 0.4.0* (Kassambara, 2020), *grid* (R Core Team, 2021), *lme4 1.1-29* (Bates et al., 2015), *ordinal 2019.12-10* (Christensen, 2019), *sjPlot 2.8.14* (Lüdecke, 2023), *tidyverse 2.0.0* (Wickham et al., 2019), and *viridis 0.6.0* (Garnier et al., 2021).

Model	Predictors
Model 1	vignette levels, vignette positions, country dummies, age, gender; random intercept for the respondents
Model 2	Model 1 + interaction between recipient and data type
Model 3	Model 1 + interaction between recipient and transmission principle
Model 4	Model 1 + individual-level variables
Model types:	*Main model:* Linear mixed-effect models (*lmer*) without speeders *Alternatives:* Linear mixed-effect models (*lmer*) with speeders "Ordinal" logistic cumulative link mixed models (*clmm*) *For social media data:* OLS models (*lm*) and ordinal models (*clmm2*) without random intercepts

Table 5.3. Overview of used regression models.

5.4.1. Contextual and international variation in December 2022

Before turning to the regression analyses, I descriptively inspect the mean values of the vignette scenarios across countries in wave 1. The mean perceived appropriateness varies across countries and vignette factors (Figure 5.1). While the ratings do not vary strongly for some vignettes and between single vignette levels, some patterns are discernable. The highest perceived appropriateness is found for health and energy data, followed by location data and then by social media data. However, the lower acceptance of social media vignettes could be partly driven by always being the last vignette to be shown to respondents (see below). Finally, while not true for each single scenario, respondents from the UK appeared as the overall most accepting country, followed by Spain and then Germany. The (Online) Appendix contain the exact mean and median values (Table T3a) and the full distribution of answers for each vignette (Figure A5.1a in Appendix 7.4).

To answer H1, H2, and RQ1 (the effects of vignette factors on acceptance), I compare the effects of vignette factors on perceived appropriateness ratings across countries. To this end, I run regression analysis with pooled data from all countries as well as separately for each country.

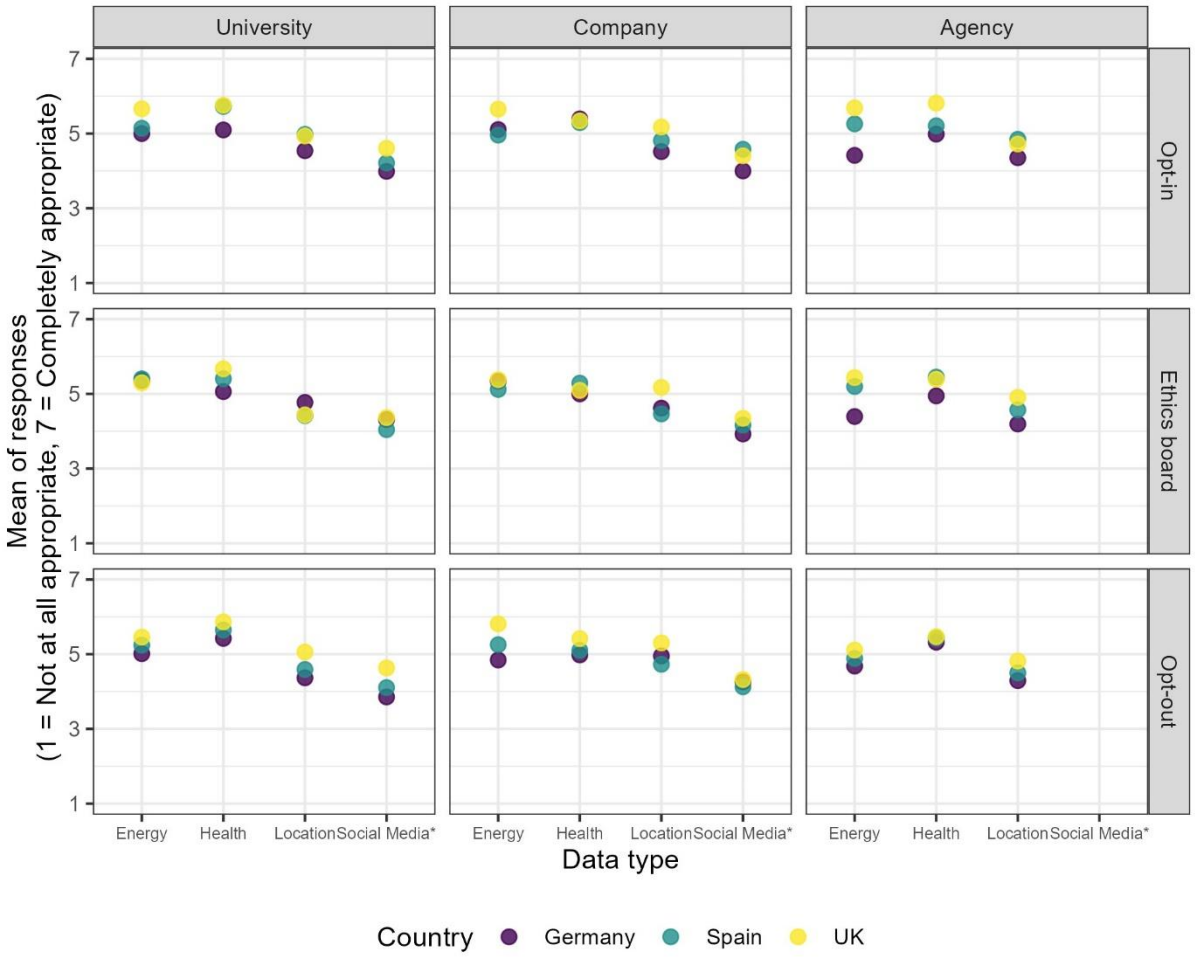


Figure 5.1. Arithmetic mean values of perceived appropriateness of all vignette scenarios in wave 1 (December 2022). Each column represents one data recipient, each row one transmission principle. Each box shows the arithmetic mean values for each data type and for each country. Number of responses per country: Germany: 2,248; Spain: 2,256; UK: 2,224.
*: Note that the social media vignettes were always shown last and are treated as a separate experiment, such that their overall lower acceptance values might at least partly be due to order effects.

I first run linear mixed-effects models that only contain the vignette levels, vignette positions, country dummies, age, gender, and a random intercept for the respondents (Figure 5.2; M1 columns in Table T4a in the Online Appendix; ordinal and speeder models in Tables T4b and T4c). The results show that among vignette factors, the data types have the strongest effects. Scenarios with health data overall appear as more accepted than energy use data, while location data use is rated lower. However, in the UK, there are no meaningful differences between health and energy use data (albeit the effect is slightly stronger in the ordinal model). These findings support H1 ("health data use is more accepted than the use of other data types") in Germany and Spain, while for UK, the health and energy are similarly accepted.

While respondents overall do not strongly differentiate between company and university researchers, there is a slight preference for the latter in Germany and Spain. Moreover, except

for in Spain, respondents rate vignettes with public agencies similar or lower than vignettes with researchers. This rejects H2 ("public actors, and particularly public researchers, are more accepted than private actors") overall, there being only a slight such tendency in Spain. To learn whether the recipient effect varies by data type and transmission principles, I add the respective interactions in two separate models. That is, there is one model with interactions between recipient and data type (Model 2) as well as one model with interactions between recipient and transmission principle (Model 3). All results are available in Table T4a in the Online Appendix (columns M2 and M3).

The results for Model 2 show that many interactions between recipients and data types are likely random – given the oftentimes small effect sizes and large standard errors –, but there are some stronger effects. There is a positive interaction effect between agencies and health data in Germany. Agency recipients are in tendency more accepted for location data than for energy use data. Companies are in tendency less accepted for health data and more accepted for location data, compared to energy use data; however, for companies, there are barely such differences in Germany.

The results for Model 3 show mostly small and likely random interaction effects between recipients and transmission principles. Two of the more consistent findings are that the combination of an agency recipient and opt-out compared to the reference categories is somewhat less accepted in the UK, and that the combination of company and ethics board is somewhat more accepted in Spain, again compared to the reference categories.

In summary, the results confirm the expectation that recipient and data type interact. There is less consistent evidence for strong interactions between recipient and transmission principle, although somewhat stronger effects show for specific combinations.

Finally, I analyze the additional experiment on social media data (Table T5a in the Online Appendix; ordinal and speeder models in Tables T5b and T5c). The social media vignette was always placed in the last position and never contained a public agency as a recipient. The results show that most effects could be random, but there are some tendencies. Again, overall acceptance is higher in UK and Spain than Germany, but the latter difference is smaller than in the previous models. Depending on the country, respondents assess company researchers differently, as compared to university researchers. In Germany, there are no strong differences (except for a somewhat stronger negative effect in the model with speeders), while there is a higher relative acceptance in Spain and a lower relative acceptance in the UK. The effects of transmission principles in tendency vary by data recipient and across countries (see Table T5a for details).

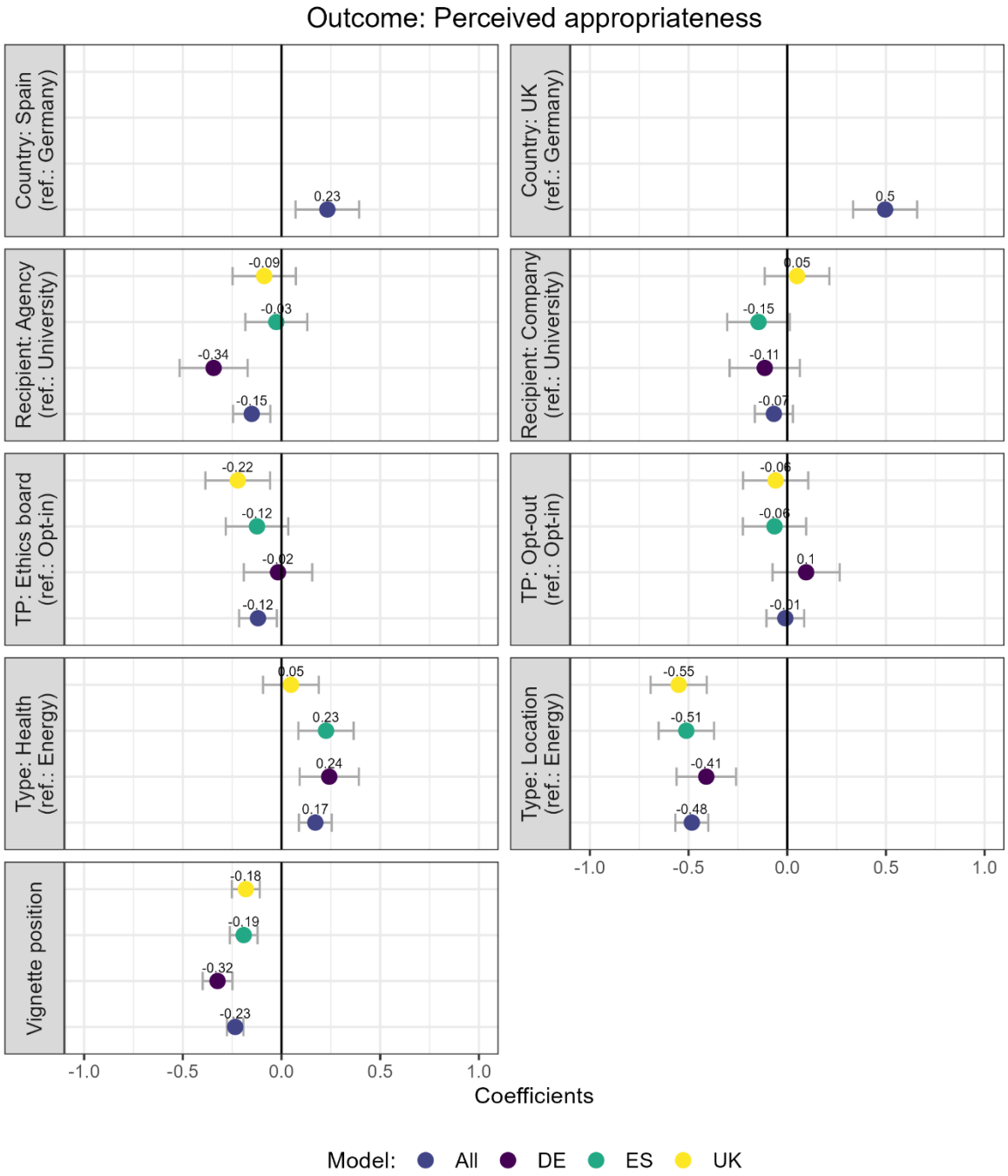


Figure 5.2. Linear mixed-effects model regression coefficients and 95% confidence intervals for effects of vignette levels on perceived appropriateness in wave 1. Based on four separate models. TP means "transmission principle". N: All: 4,562; Germany: 1,510; Spain: 1,549; UK: 1,503. Models further include age, gender, and a random intercept on the respondent-level (not displayed in the figure).

Based on these results, I can answer RQ1 ("Which modes of consent are more accepted?"). The above models show no overall strong differences between opt-in and opt-out procedures (except for a somewhat stronger negative effect for opt-out in the case of social media data in Spain). Ethics boards are in tendency less accepted than opt-in procedures. Particularly the latter effect varies by country.

This leads to RQ2 ("Do countries with higher levels of individualism desire higher levels of control over data flows?"). While acceptance in all three countries barely changes between opt-in and opt-out procedures, UK respondents are somewhat relatively more skeptical about ethics boards. Spanish respondents accept ethics boards slightly less than opt-in. If individualism was responsible for international differences, there should be clearer differences in the acceptance of transmission principles (especially for opt-out versus opt-in) particularly between UK and Spain. Moreover, contrary to expectation, the overall acceptance is highest in the UK. The answer to RQ2 thus is that there is no clear and consistent association of higher desire for control with higher country-level individualism.

5.4.2. Interindividual and international variation in December 2022

To answer RQ3 on the associations of individual-level variables with perceived appropriateness, I add variables to Model 1 that are related to trust, altruism, perceived sensitivity of data types, other privacy-related perceptions, device use, and affinity towards technology.

I focus on the effects of individual-level variables across models (Figure 5.3; Table T6a in the Online Appendix; ordinal and speeder models in Tables T6b and T6c). On average, female respondents report somewhat lower acceptance than male respondents. Four associations are relatively consistent across countries: higher trust in data recipients comes with higher acceptance, while higher perceived sensitivity of data types comes with lower acceptance (note that sensitivity and trust always refer to the specific data type or data recipient shown in the vignette). General privacy concerns are associated with lower acceptance, while agreement with the statement that "The privacy of individuals may be invaded if this results in a greater benefit to society" (Trepte, 2020) comes with higher acceptance. Additional to these consistent associations, more thinking about privacy in times of the pandemic is associated with higher acceptance in the UK. Altruism has a positive association with acceptance, particularly in Germany. The same is true for the number of used devices in Spain. Otherwise, there are mostly small and likely random associations (although single effects appear more meaningful in the alternative model types). With respect to the effects of vignette factors, it is noteworthy that controlling for the individual-level variables, public agencies appear as the most accepted recipient in Spain, while company recipients are most accepted in Germany and UK.

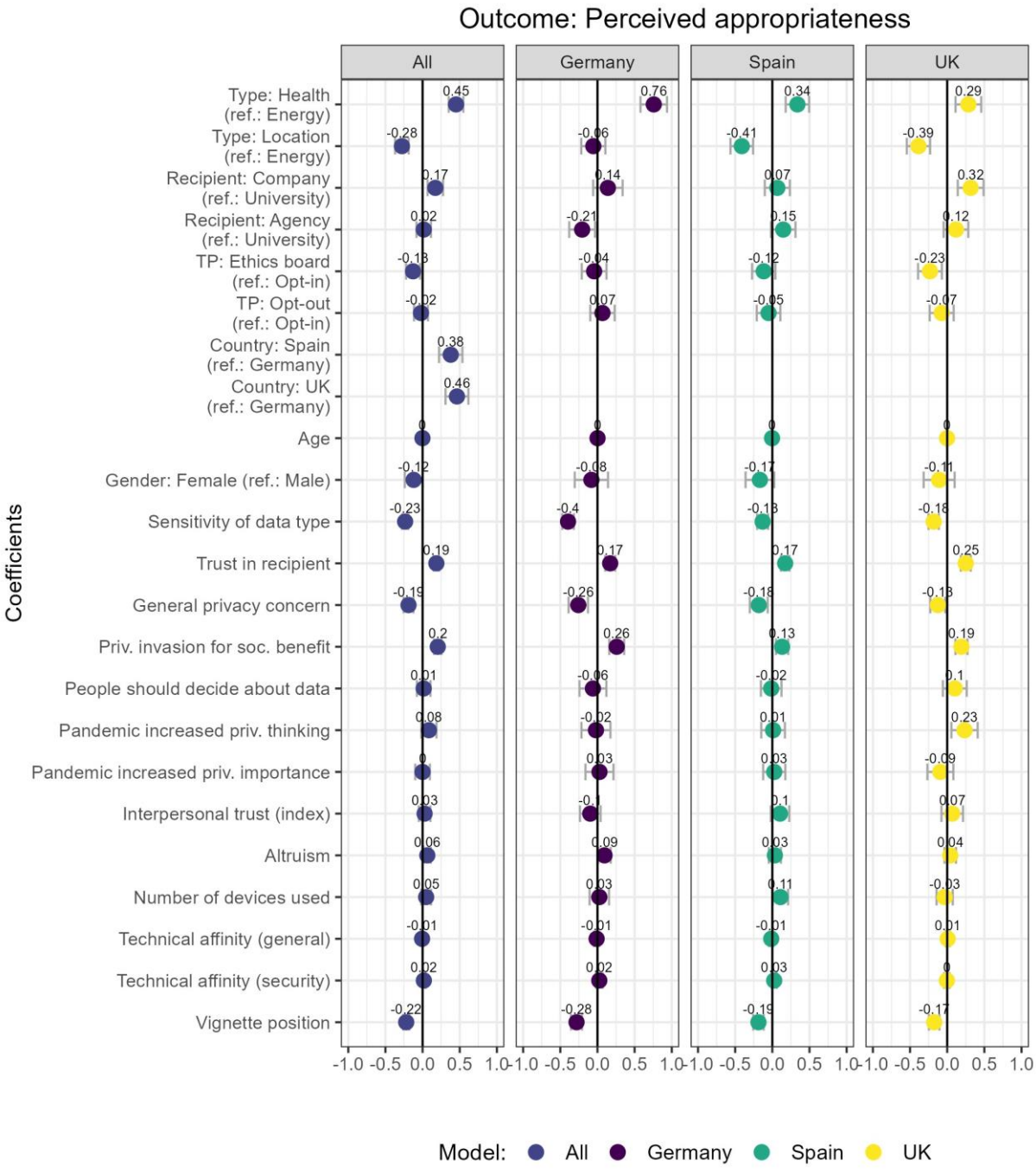


Figure 5.3. Linear mixed-effects model regression coefficients and 95% confidence intervals for effects of vignette levels and individual-level characteristics on perceived appropriateness in wave 1. Based on four separate models. TP means "transmission principle". N: All: 4,562; Germany: 1,510; Spain: 1,549; UK: 1,503. Models further include a random intercept on the respondent-level (not displayed in the figure).

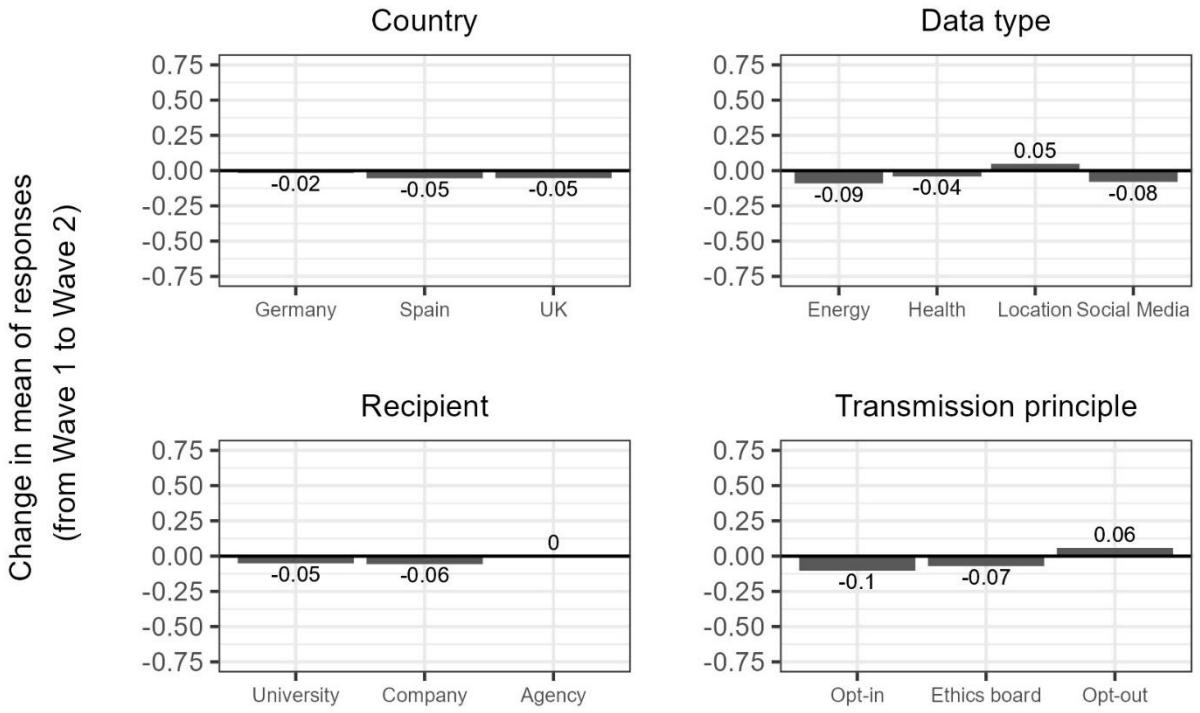


Figure 5.4. Changes in arithmetic means of responses from wave 1 (December 2022) to wave 2 (May 2023) among those respondents who participated in both waves. Aggregated for country, data type, recipient, or transmission principle. Based on 8,880 responses from 1,110 respondents who participated in both waves.

Figure 5.5. Changes in arithmetic means of responses from wave 1 (December 2022) to wave 2 (May 2023) among those respondents who participated in both waves. Differentiated by country, data type, and data recipient. Based on 8,880 responses from 1,110 respondents who participated in both waves.

5.4.3. Longitudinal variation

To answer RQ4 ("Does the acceptance of health data use, relative to the acceptance of using other data type, change with the pandemic situation?"), I first inspect changes in perceived appropriateness from wave 1 (December 2022) to wave 2 (May 2023) across data types and countries. To this end, I use a data set that only contains respondents who participated in both waves (see details on the construction of this sample in Online Appendix *Section M*). The (Online) Appendix contains mean values – plotted (Figure A5.2) and as a table (Table T3b) – and the full distribution (Figure A5.1b) of vignette responses in wave 2.

When analyzing longitudinal changes in acceptance separately for each data flow parameter or by country, only small differences over time are observed (Figure 5.4; with speeders: Figure A5.3 in Appendix 7.4). A comparison of changes between more specific vignette scenarios reveals a more nuanced picture (Figure 5.5; exact values in Table T7a in the Online Appendix). Acceptance changed only slightly for several combinations of data types and recipients. However, in the UK, health data use became less accepted for companies and more accepted for universities. Changes for energy and social media data are almost consistently negative and in some cases relatively small. Moreover, there are some stronger increases in acceptance for location data use. With respect to RQ4, these results still show that the acceptance of health data did not overall decrease (or increase) much more relative to other data types. While there are stronger changes for health data with specific recipients in UK (and partly Spain), there are similarly strong changes for other settings (when including speeders, however, the decrease for company recipients in the UK is particularly strong, but the increase for public agency is less pronounced; see Figure A5.4 and Table T7b in the Appendices). However, some changes vary considerably when further taking into account transmission principles (but note that the number of responses per combination is lower in this more fine-grained analysis). In some cases, the ratings of the same data type and recipient changes into different directions depending on transmission principles (Table T8a and Figure A5.5a without speeders, Table T8b and Figure A5.5b with speeders). Still, ratings of health data vignettes do not stand out to have overall changed particularly more than other ratings.

5.4.4. Replication of December 2022 results with data from May 2023

Finally, I make use of the full data set for wave 2 that comprises respondents who already participated in wave 1 as well as newly recruited respondents. I treat this second wave as a replication of the first wave and check whether the substantive findings with respect to the

hypotheses and research questions hold. However, this approach cannot reveal whether any differences are attributable to changes over time or to differences in sample composition.

I run all models from wave 1 again with data from the second wave and show all regression tables in the Online Appendix (Tables with ending letters *d* to *f*). The finding holds that health data are more accepted than the other data types (H1). In fact, in wave 2, there are somewhat stronger positive effects for health data for the UK, compared to wave 1. Also, public recipients (H2) are again not overall clearly more accepted than company recipients, although the relative acceptance of the latter in tendency is lower. As for transmission principles and their importance across countries (RQ1 and RQ2), opt-out is the overall most accepted transmission principle in Spain in wave 2 (except when using social media data), while opt-in and opt-out are again most accepted in the other countries. The individual-level variables (RQ3) overall display the same tendencies as in wave 1, but women display rather equal acceptance compared to men in the UK. However, the associations with interpersonal trust and of thinking about privacy in times of the pandemic tend more towards zero. Moreover, the differences between recipients tend to be somewhat smaller, a slight exception being a relatively higher acceptance of agency recipients in the UK than in wave 1. Across models, there are changes in effects for further specific constellations – especially for the interaction effects and for the case of social media data – that can be ascertained from the respective tables.

5.5. Discussion

The results demonstrate considerable variation of perceived appropriateness of data use for public benefit across contexts. The contextual effects moreover vary by country and, to some extent, over time. These findings support the notion that contextual integrity is a useful approach for comparative research across countries: Additional to country comparisons with respect to *general* privacy notions, contextual integrity can reveal *context-specific* differences between countries that may otherwise remain unnoticed (while this gap between general and specific perceptions has already been argued for within single countries, Martin & Nissenbaum, 2016). This study replies to the suggestion by Masur et al. (2021) to employ contextual integrity for comparative privacy research and demonstrates that future research can operationalize data uses in meso-level social contexts by drawing on contextual integrity's data flow parameters for such comparative purposes. The results also show that some individual-level variables are mostly consistently associated with higher or lower acceptance. In the following, I discuss more specific implications and research avenues with respect to contextual, international, interindividual, and longitudinal comparisons, before turning to limitations of the study.

Among contextual factors, changes in data types had the strongest effects on acceptance, with health data being more accepted than energy data, followed by location data, and social media data being least accepted to be used. However, as explained above, the social media vignette appeared last, which means that the lower acceptance for social media vignettes may be due to an order effect. The found negative effects of vignette position among non-social-media vignettes suggest the presence of such order effects. The relatively high acceptance of health data use is somewhat striking as one might consider this data type to be particularly sensitive. Indeed, a higher sensitivity of data types is associated with lower acceptance, but the positive estimates for health data remain. This finding supports the notion that sensitivity is a context-dependent concept (Martin & Nissenbaum, 2017). The relatively higher acceptance of health and energy data, compared to location data, could be explained by their particular societal relevance, but further longitudinal research would need to investigate whether this is a temporary or stable difference in preferences.

Supporting the importance of cross-national comparisons when inspecting the parameters, for instance, at least for the simplest presented models in wave 1, the higher acceptance of health data does not show that strongly for the UK. Thus, the respective hypothesis might have been evaluated differently if only the UK had been researched. As another example, public agencies are relatively less accepted data recipients in Germany compared to the other countries, which future research could further elucidate by in-depth studies on Germany as well as further international comparisons (see below). Moreover, recipients and transmission principles tend to be less important, but their combinations do matter depending on the concrete scenario and the country. These findings have two implications for future comparative privacy research. First, country comparisons of meso-level social contexts can identify cross-country differences that are not captured by general differences in privacy perceptions, e.g., in health versus energy use contexts. Second, data use contexts that appear similar across countries may be differently evaluated by the respective populations, calling for further cross-country research using the notion of contextual integrity. For instance, while acceptance towards data use by public agencies may appear overall relatively lower in Germany, the acceptance of this data recipient still varies by used data type within Germany. As data types were particularly influential, separate OLS models for each country and data type in the Online Appendix further illustrate data type- and country-specific differences (Tables T9a to T9f) for interested readers.

Focusing on overall differences in acceptance across countries, I found stable international differences in overall acceptance at both survey time points. Contrary to the initial expectations that respondents from countries with a higher level of individualism have a higher desire to

control data use for public benefit, UK respondents do not clearly and consistently prefer opt-in procedures more strongly than respondents from other countries. The higher acceptance for opt-out in Spain in wave 2 makes the differences in desire to control data between Spain and UK somewhat more in line with the differences in levels of individualism in the respective countries, but the overall differences across all countries are still not consistently and pronouncedly aligning. Moreover, the overall acceptance is highest in UK and lowest in Germany. Future research may further context-specifically investigate cultural dimensions that better explain the found differences. For instance, Germany scores high and the UK scores low on Hofstede's dimension of "uncertainty avoidance", with Spain being in the ranked between these two countries (Hofstede Insights, 2023).

Another explanation for this finding may be that country-level cultural variables cannot capture the complexities that are inherent to international comparisons. Scholars criticized Hofstede's cultural dimensions approach for various reasons, among them methods-related concerns and considering it a too positivistic approach (discussed in Jackson, 2020). As the Comparative Privacy Research Framework suggests, countries may differ with respect to a variety of structures, not only the cultural aspect of individualism (Masur et al., 2021). To learn more about the concrete structures that matter, future research would need to include a large variety of countries that differ with respect to multiple structures at different levels, such as *economic* structures (see Masur et al., 2022). With respect to economic structures, e.g., degrees of free market economy (ibd.) could be related to the establishment of relatively free use of data in the respective countries. As a hint in this direction, UK respondents are on average relatively more skeptical about ethics boards, i.e., an external body taking part in deciding about the use of individual data. However, while such structures may explain some of the differences between countries, the present study suggests that a context-specific view is necessary to avoid undue generalizations to all kinds of data uses.

The analysis further explored associations between individual-level variables and acceptance across countries. However, note that these estimates do not represent causal effects. Mostly consistent findings are that perceived sensitivity of data types and general privacy concerns are associated with lower acceptance across countries. Trust in data recipients and agreeing that privacy may be invaded for public benefit are associated with higher acceptance. In summary, even after taking into account contextual factors, general privacy perceptions still matter. Adding to Martin and Nissenbaum's (2016) suggestion that context-specific preference measurements may partly close the gap between stated privacy concerns and situation-specific data sharing behavior (the so-called "privacy paradox"), these findings imply that the

measurement of general perceptions may still be worthwhile in context-based research. Future research needs to investigate how context-specific alongside general privacy perceptions translate into data sharing behavior.

Some changes in perceived appropriateness over time were found in the longitudinal comparison for specific contextual constellations, but not pronouncedly for the acceptance of health data use. The found partly changes for health data use by companies do not seem to be directly explainable by a change in the pandemic situation, as this also should have affected agency recipients. Instead, it might be that respondents were on average not overly concerned about the pandemic anymore already as of December 2022. As for energy supply problems as another salient public issue, there is a tendency towards decreased acceptance from December to May, but these changes are overall not very strong. When additionally considering transmission principles, some stronger differences show, but these results are based on a smaller number of responses per vignette. However, these results highlight again that privacy research and policy-making need to reflect how the timing of data collection – e.g., during a specific crisis – might affect context-specific results. Momentary assessments of public opinion do not necessarily constitute a "social license" (Shaw et al., 2020) to carry out questionable data uses. Instead, they constitute one element of an assessment of the appropriateness of a data use, along with further legal considerations and, as suggested by contextual integrity, the discussion of context-specific and more general values and goals at stake (Nissenbaum, 2010).

I now turn to limitations of the paper that were not already discussed above. First, as is commonly the case with vignette studies, we need to keep in mind potential limitations with respect to external validity (Eifler & Petzold, 2019). Moreover, this study can only speak about the concrete investigated scenarios. For instance, the company recipients were always defined as "researchers at an Internet company" and it is possible that recipients rate vignettes differently if more context-specific companies are involved. The relatively strong effects found for data types may also be due to the circumstance that the vignettes were structured around data types to make them appear plausible and to not present, e.g., construction planning agencies using patient records. Moreover, some of the investigated scenarios may have been very hypothetical or unknown for respondents. With increasing concreteness or public awareness of these kinds of data uses, attitudes towards these data collection practices may change.

Second, in principle, internationally different response behavior patterns (Kemmelmeier, 2016) – however, less so due to the experimental design – and potential variations in the interpretations of vignettes may account for some of the differences found between countries.

Researchers also need to validate the cross-cultural invariance of privacy-related measurements (Ghaiumy Anaraky et al., 2021). Moreover, a larger number of countries might allow researchers to better disentangle effects of, e.g., cultural and economic differences (see Masur et al., 2021) on privacy attitudes. While this study has detected differences even between three European countries, differences could be further pronounced particularly when extending comparisons to non-WEIRD – i.e., western, educated, industrialized, rich, and democratic (Henrich et al., 2010) – countries.

Third, as explained above, the study is based on a non-probability sample for which inference is only feasible under specific conditions and for specific fields of application (Kohler & Post, 2023). This means that while more confident claims with respect to experimental effects can be possible, e.g., mere mean values are not to be inferred to the general populations of the respective countries. Future research would need to confirm these findings with probability samples.

5.6. Conclusion

The Comparative Privacy Research Framework (Masur et al., 2021) proposes to compare privacy-related phenomena across different levels and types of structures. The present study draws on "contextual integrity" (Nissenbaum, 2010) to enhance comparative privacy research by focusing on social contexts as a meso-level structure. To this end, the contextual integrity data flow parameters offer a useful approach to operationalize data uses in meso-level social contexts. The present study applies this approach by employing a survey experiment in three countries (Germany, Spain, and the UK) and at two time points (December 2022 and May 2023) to compare privacy perceptions related to data use for public benefit along four components: contextual, international, interindividual, and longitudinal variation. The results show that the effects of data flow parameters vary across countries, but to different degrees. The strongest effects are found for the data type, with health data overall being the relatively most accepted data type overall. The effect for data recipients varies such as to lead to different substantive conclusions for the different countries. Country-level individualism was not found to be clearly and consistently associated with desire for control over the data. Interindividually, several general privacy perceptions still matter after considering contextual factors. Finally, longitudinal comparisons show overall minor but context-dependent variation over time.

In conclusion, using the contextual integrity approach can unmask meso-level context-specific differences in the acceptance of data uses within and between countries. These differences could be relevant for ascertaining "social licenses" (see Shaw et al., 2020) regarding

data use practices, for international privacy-related regulation, and for suggestions on sector-specific policies. This study can therefore serve as a call for more deliberately incorporating meso-level contexts in comparative privacy research that can inform privacy-related public decision-making.

Acknowledgments

I thank Ellen Laurischk, Kerstin Fischer, and Tonja Dingerdissen for their support in this research project, and Elsa Peris, Wiebke Weber, and Joshua Fullard for their work regarding questionnaire translation. I further thank Frauke Kreuter, Helen Nissenbaum, and Thomas Fetzer for their comments and ideas, and the Kreuter-Keusch research group and the reviewers for helpful comments on this manuscript.

Funding

This research was funded by Volkswagen Foundation, grant "The Covid-19 Pandemic and Data Sharing for the Public Good: Attitudinal, Ethical, and Legal Approaches to Privacy During the Pandemic and Beyond". Further funding for data collection came from the Ludwig-Maximilians-Universität of Munich. This work was supported by the University of Mannheim's Graduate School of Economic and Social Sciences.

References

Aitken, M., de St. Jorre, J., Pagliari, C., Jepson, R., & Cunningham-Burley, S. (2016). Public responses to the sharing and linkage of health data for research purposes: A systematic review and thematic synthesis of qualitative studies. *BMC Medical Ethics*, *17*(1), 73. https://doi.org/10.1186/s12910-016-0153-x

Ausloos, J., Leerssen, P., & ten Thije, P. (2020). *Operationalizing Research Access in Platform Governance. What to learn from other industries?* https://algorithmwatch.org/en/wp-content/uploads/2020/06/GoverningPlatforms_IViR_study_June2020-AlgorithmWatch-2020-06-24.pdf

Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments*. SAGE.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bauer, P. C., Gerdon, F., Keusch, F., Kreuter, F., & Vannette, D. (2022). Did the GDPR increase trust in data collectors? Evidence from observational and experimental data. *Information, Communication & Society*, *25*(14), 2101–2121. https://doi.org/10.1080/1369118X.2021.1927138

Best, S. J., Krueger, B. S., & Ladewig, J. (2006). Privacy in the Information Age. *Public Opinion Quarterly*, *70*(3), 375–401. https://doi.org/10.1093/poq/nfl018

Büchi, M., Festic, N., & Latzer, M. (2022). The chilling effects of digital dataveillance: A theoretical model and an empirical research agenda. *Big Data & Society*, *9*(1), 205395172110653. https://doi.org/10.1177/20539517211065368

Carter, P., Laurie, G. T., & Dixon-Woods, M. (2015). The social licence for research: Why care.data ran into trouble. *Journal of Medical Ethics*, *41*(5), 404–409. https://doi.org/10.1136/medethics-2014-102374

Christensen, R. H. B. (2019). *Ordinal—Regression models for ordinal data. R package version 2019.12-10* [Computer software]. https://CRAN.R-project.org/package=ordinal

Conrad, F., Tourangeau, R., Couper, M., & Zhang, C. (2017). Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods*, *11*(1), 45–61. https://doi.org/10.18148/SRM/2017.V11I1.6304

De Man, Y., Wieland-Jorna, Y., Torensma, B., De Wit, K., Francke, A. L., Oosterveld-Vlug, M. G., & Verheij, R. A. (2023). Opt-in and opt-out consent procedures for the reuse of routinely recorded health Data in scientific research and their consequences for consent rate and consent bias: Systematic review. *Journal of Medical Internet Research*, *25*, e42131. https://doi.org/10.2196/42131

Deruelle, T., Kalouguina, V., Trein, P., & Wagner, J. (2023). Designing privacy in personalized health: An empirical analysis. *Big Data & Society*, *10*(1), 205395172311586. https://doi.org/10.1177/20539517231158636

Eifler, S., & Petzold, K. (2019). Validity aspects of vignett experiments: Expected "What-If" differencesbetween reports of behavioral intentions and actual behavior. In P. Lavrakas, M. Traugott, C. Kennedy, A. Holbrook, E. De Leeuw, & B. West (Eds.), *Experimental Methods in Survey Research* (1st ed., pp. 393–416). Wiley. https://doi.org/10.1002/9781119083771.ch20

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*(2). https://doi.org/10.1214/16-STS598

Engström, E., Eriksson, K., Björnstjerna, M., & Strimling, P. (2023). Global variations in online privacy concerns across 57 countries. *Computers in Human Behavior Reports*, *9*, 100268. https://doi.org/10.1016/j.chbr.2023.100268

ESS Round 9: European Social Survey (2021): *ESS-9 2018 Documentation Report. Edition 3.1*. Bergen, European Social Survey Data Archive, NSD - Norwegian Centre for Research Data for ESS ERIC. https://doi.org/10.21338/NSD-ESS9-2018

European Commission. (2023). *A European strategy for data*. https://digital-strategy.ec.europa.eu/en/policies/strategy-data

Garnier, S., Ross, N., BoB Rudis, Filipovic-Pierucci, A., Galili, T., Timelyportfolio, Greenwell, B., Sievert, C., Harris, D. J., & JJ Chen. (2021). *sjmgarnier/viridis: Viridis 0.6.0 (pre-CRAN release)* (v0.6.0pre) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.4679424

Gerber, N., Gerber, P., & Volkamer, M. (2018). Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers & Security*, *77*, 226–261. https://doi.org/10.1016/j.cose.2018.04.002

Gerdon, F., Nissenbaum, H., Bach, R. L., Kreuter, F., & Zins, S. (2021). Individual acceptance of using health data for private and public benefit: Changes during the COVID-19 pandemic. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.edf2fc97

Ghaiumy Anaraky, R., Li, Y., & Knijnenburg, B. (2021). Difficulties of measuring culture in privacy studies. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–26. https://doi.org/10.1145/3479522

Gilbert, S., Shilton, K., & Vitak, J. (2023). When research is the context: Cross-platform user expectations for social media data reuse. *Big Data & Society*, *10*(1), 205395172311641. https://doi.org/10.1177/20539517231164108

Gilbert, S., Vitak, J., & Shilton, K. (2021). Measuring Americans' comfort with research uses of their social media data. *Social Media + Society*, *7*(3), 205630512110338. https://doi.org/10.1177/20563051211033824

Goetzen, A., Dooley, S., & Redmiles, E. M. (2022). Ctrl-Shift: How privacy sentiment changed from 2019 to 2021. *Proceedings on Privacy Enhancing Technologies*, *2022*(4), 457–485. https://doi.org/10.56553/popets-2022-0118

Haug, S., Schnell, R., Raptis, G., Dotter, C., & Weber, K. (2023). *Wissen und Einstellung zur Speicherung und Freigabe von Gesundheitsdaten. Ergebnisse einer Bevölkerungsbefragung*. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*. https://doi.org/10.1016/j.zefq.2023.11.001

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hofstede, G. H., Hofstede, G. J., & Minkov, M. (2010). *Cultures and organizations: Software of the mind. Intercultural cooperation and its importance for survival* (3rd ed). McGraw-Hill.

Hofstede Insights. (2023). *Country comparison tool*. https://www.hofstede-insights.com/country-comparison-tool

Horne, C., Darras, B., Bean, E., Srivastava, A., & Frickel, S. (2015). Privacy, technology, and norms: The case of Smart Meters. *Social Science Research*, *51*, 64–76. https://doi.org/10.1016/j.ssresearch.2014.12.003

Hutchings, E., Loomes, M., Butow, P., & Boyle, F. M. (2020). A systematic literature review of health consumer attitudes towards secondary use and sharing of health administrative and clinical trial data: A focus on privacy, trust, and transparency. *Systematic Reviews*, *9*(1), 235. https://doi.org/10.1186/s13643-020-01481-9

Jackson, T. (2020). The legacy of Geert Hofstede. *International Journal of Cross Cultural Management*, *20*(1), 3–6. https://doi.org/10.1177/1470595820915088

Jamieson, K. H., Lupia, A., Amaya, A., Brady, H. E., Bautista, R., Clinton, J. D., Dever, J. A., Dutwin, D., Goroff, D. L., Hillygus, D. S., Kennedy, C., Langer, G., Lapinski, J. S., Link, M., Philpot, T., Prewitt, K., Rivers, D., Vavreck, L., Wilson, D. C., & McNutt, M. K. (2023). Protecting the integrity of survey research. *PNAS Nexus*, *2*(3), pgad049. https://doi.org/10.1093/pnasnexus/pgad049

Kalkman, S., van Delden, J., Banerjee, A., Tyl, B., Mostert, M., & van Thiel, G. (2022). Patients' and public views and attitudes towards the sharing of health data for research: A narrative review of the empirical evidence. *Journal of Medical Ethics*, *48*(1), 3–13. https://doi.org/10.1136/medethics-2019-105651

Kao, Y.-H., & Sapp, S. G. (2022). The effect of cultural values and institutional trust on public perceptions of government use of network surveillance. *Technology in Society*, *70*, 102047. https://doi.org/10.1016/j.techsoc.2022.102047

Kassambara, A. (2020). *Ggpubr: "ggplot2" Based Publication Ready Plots. R package version 0.4.0.*. https://CRAN.R-project.org/package=ggpubr

Kehr, F., Kowatsch, T., Wentzel, D., & Fleisch, E. (2015). Blissfully ignorant: The effects of general privacy concerns, general institutional trust, and affect in the privacy calculus. *Information Systems Journal*, *25*(6), 607–635. https://doi.org/10.1111/isj.12062

Kemmelmeier, M. (2016). Cultural differences in survey responding: Issues and insights in the study of response biases. *International Journal of Psychology*, *51*(6), 439–444. https://doi.org/10.1002/ijop.12386

Kho, M. E., Duffett, M., Willison, D. J., Cook, D. J., & Brouwers, M. C. (2009). Written informed consent and selection bias in observational studies using medical records: Systematic review. *BMJ*, *338*, b866. https://doi.org/10.1136/bmj.b866

Kim, K. K., Joseph, J. G., & Ohno-Machado, L. (2015). Comparison of consumers' views on electronic data sharing for healthcare and research. *Journal of the American Medical Informatics Association*, *22*(4), 821–830. https://doi.org/10.1093/jamia/ocv014

Kim, Y., & Stanton, J. M. (2016). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, *67*(4), 776–799. https://doi.org/10.1002/asi.23424

Kohler, U., & Post, J. C. (2023). Welcher Zweck heiligt die Mittel? Bemerkungen zur Repräsentativitätsdebatte in der Meinungsforschung. *Zeitschrift Für Soziologie*, *52*(1). https://doi.org/10.1515/zfsoz-2023-2001

Li, Y., Kobsa, A., Knijnenburg, B. P., & Carolyn Nguyen, M.-H. (2017). Cross-Cultural Privacy Prediction. *Proceedings on Privacy Enhancing Technologies*, *2017*(2), 113–132. https://doi.org/10.1515/popets-2017-0019

Liu, J. (2022). Social data governance: Towards a definition and model. *Big Data & Society*, *9*(2), 205395172211113. https://doi.org/10.1177/20539517221111352

Lüdecke, D. (2023). *sjPlot: Data visualization for statistics in social science. R package version 2.8.14*. https://CRAN.R-project.org/package=sjPlot

Martin, K., & Nissenbaum, H. (2016). Measuring privacy: An empirical test using context to expose confounding variables. *The Columbia Science & Technology Law Review*, *18*, 176–218. https://doi.org/10.7916/STLR.V18I1.4015

Masur, P. K., Epstein, D., Quinn, K., Wilhelm, C., Baruh, L., & Lutz, C. (2021). *A comparative privacy research framework*. https://doi.org/10.31235/osf.io/fjqhs

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about It. *European Sociological Review*, *26*(1), 67–82. https://doi.org/10.1093/esr/jcp006

Mothersbaugh, D. L., Foxx, W. K., Beatty, S. E., & Wang, S. (2012). Disclosure Antecedents in an Online Service Context: The Role of Sensitivity of Information. *Journal of Service Research*, *15*(1), 76–98. https://doi.org/10.1177/1094670511424924

National Data Guardian for Health and Social Care in England. (2022). *What do we mean by public benefit? Evaluating public benefit when health and adult social care data is used for purposes beyond individual care*. https://www.gov.uk/government/publications/what-do-we-mean-by-public-benefit-evaluating-public-benefit-when-health-and-adult-social-care-data-is-used-for-purposes-beyond-individual-care/what-do-we-mean-by-public-benefit-evaluating-public-benefit-when-health-and-adult-social-care-data-is-used-for-purposes-beyond-individual-care

Newlands, G., Lutz, C., Tamò-Larrieux, A., Villaronga, E. F., Harasgama, R., & Scheitlin, G. (2020). Innovation under pressure: Implications for data privacy during the Covid-19 pandemic. *Big Data & Society*, *7*(2), 205395172097668. https://doi.org/10.1177/2053951720976680

Nießen, D., Beierlein, C., Rammstedt, B., & Lechner, C. M. (2021). *Interpersonal Trust Short Scale (KUSIV3)*. ZIS-The Collection of Items and Scales for the Social Sciences. https://doi.org/10.6102/ZIS292_EXZ

Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

Nissenbaum, H. (2018). Respecting Context to Protect Privacy: Why Meaning Matters. *Science and Engineering Ethics*, *24*(3), 831–852. https://doi.org/10.1007/s11948-015-9674-9

Office of the Australian Information Commission, & Lonergan Research. (2020). *Australian Community Attitudes to Privacy Survey 2020*. https://www.oaic.gov.au/__data/assets/pdf_file/0015/2373/australian-community-attitudes-to-privacy-survey-2020.pdf

Park, Y. J. (2013). Digital literacy and privacy behavior online. *Communication Research*, *40*(2), 215–236. https://doi.org/10.1177/0093650211418338

Pew Research Center. (2014). *Pew Research Center's Internet Project / GfK Privacy Panel. Survey #1 Topline*. https://www.pewresearch.org/wp-content/uploads/sites/9/2015/07/PrivacyPanelTopline_11-12-14.pdf

Proferes, N., & Walker, S. (2020). Researcher views and practices around informing, getting consent, and sharing research outputs with social media users when using their public data. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*. https://aisel.aisnet.org/hicss-53/dsm/critical_and_ethical_studies/2/

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roßmann, J. (2010). *Data quality in web surveys of the German longitudinal election study 2009*. [Conference presentation]. 3rd ECPR Graduate Conference, Dublin, Ireland.

Schauffel, N., Schmidt, I., Peiffer, H., & Ellwart, T. (2021). *ICT Self-Concept Scale (ICT-SC25)*. ZIS-The Collection of Items and Scales for the Social Sciences. https://doi.org/10.6102/ZIS308_EXZ

Schomakers, E.-M., Lidynia, C., Müllmann, D., & Ziefle, M. (2019). Internet users' perceptions of information sensitivity – insights from Germany. *International Journal of Information Management*, *46*, 142–150. https://doi.org/10.1016/j.ijinfomgt.2018.11.018

Schumacher, C., Eggers, F., Verhoef, P. C., & Maas, P. (2023). The effects of cultural differences on consumers' willingness to share personal information. *Journal of Interactive Marketing*, *58*(1), 72–89. https://doi.org/10.1177/10949968221136555

Shaw, J. A., Sethi, N., & Cassel, C. K. (2020). Social license for the use of big data in the COVID-19 era. *Npj Digital Medicine*, *3*(1), 128. https://doi.org/10.1038/s41746-020-00342-y

Silber, H., Gerdon, F., Bach, R., Kern, C., Keusch, F., & Kreuter, F. (2022). A preregistered vignette experiment on determinants of health data sharing behavior: Willingness to donate sensor data, medical records, and biomarkers. *Politics and the Life Sciences*, *41*(2), 161–181. https://doi.org/10.1017/pls.2022.15

Smith, Dinev, & Xu. (2011). Information privacy research: An interdisciplinary review. *MIS Quarterly*, *35*(4), 989–1015. https://doi.org/10.2307/41409970

SOEP-IS Group. (2021). *SOEP-IS 2018 – Questionnaire for the SOEP Innovation Sample (Update Release 2019)* (SOEP Survey Papers 948: Series A – Survey Instruments (Erhebungsinstrumente).). Berlin: DIW Berlin/SOEP. https://www.diw.de/documents/publikationen/73/diw_01.c.821936.de/diw_ssp0948.pdf

Stockdale, J., Cassell, J., & Ford, E. (2019). "Giving something back": A systematic review and ethical enquiry into public views on the use of patient data for research in the United Kingdom and the Republic of Ireland. *Wellcome Open Research*, *3*(6). https://doi.org/10.12688/wellcomeopenres.13531.2

Terpstra, A., De Rooij, A., & Schouten, A. (2023). Online proctoring: Privacyinvasion or study alleviation? Discovering acceptability using Contextual integrity. *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 167. https://doi.org/10.1145/3544548.3581181

Trein, P., & Varone, F. (2023). Citizens' agreement to share personal data for public policies: Trust and issue importance. *Journal of European Public Policy*, 1–26. https://doi.org/10.1080/13501763.2023.2205434

Trepte, S. (2020). *The Privacy Longitudinal Study* (2.0.0) [dataset]. GESIS Data Archive. https://doi.org/10.7802/2117

Trepte, S., & Masur, P. K. (2016). *Cultural differences in media use, privacy, and self-disclosure. Research report on a multicultural survey study*. University of Hohenheim. https://research.vu.nl/en/publications/cultural-differences-in-social-media-use-privacy-and-self-disclos

Trepte, S., Reinecke, L., Ellison, N. B., Quiring, O., Yao, M. Z., & Ziegele, M. (2017). A cross-cultural perspective on the privacy calculus. *Social Media + Society*, *3*(1), 205630511668803. https://doi.org/10.1177/2056305116688035

Utz, C., Becker, S., Schnitzler, T., Farke, F. M., Herbert, F., Schaewitz, L., Degeling, M., & Dürmuth, M. (2021). Apps against the spread: Privacy implications and user acceptance of COVID-19-related smartphone apps on three continents. *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 70. https://doi.org/10.1145/3411764.3445517

Vitak, J., & Zimmer, M. (2020). More than just privacy: Using contextual integrity to evaluate the long-term risks from COVID-19 surveillance technologies. *Social Media + Society*, *6*(3), 205630512094825. https://doi.org/10.1177/2056305120948250

Walsh, T. (2023). Modeling COVID-19 with big mobility data: Surveillance and reaffirming the people in the data. *Big Data & Society*, *10*(1), 205395172311641. https://doi.org/10.1177/20539517231164115

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wnuk, A., Oleksy, T., & Domaradzka, A. (2021). Prosociality and endorsement of liberty: Communal and individual predictors of attitudes towards surveillance technologies. *Computers in Human Behavior*, *125*, 106938. https://doi.org/10.1016/j.chb.2021.106938

6. Conclusion

Researchers and organizations need a sensible notion of how to appropriately make sense of and describe novel data-driven technologies to meaningfully gauge their social impacts and their public acceptance. These notions become increasingly important with data-driven technological advances that potentially affect social inequality and privacy at large scales. From a scientific perspective, empirical research on social impacts and public acceptance should be informative about specific applications and be based on a sound conceptual account of context that allows for systematic comparisons. From an ethical and practical perspective, context is a necessary factor in assessing whether a specific data-driven technology is covered by public acceptance in the form of a "social license" (Gunningham et al., 2004) and who concretely is impacted by such technologies in which way. Perceptions of abstract suggestions for data-driven technologies may be insightful for some purposes, but even technologies that are based on ideas that are in principle acceptable could be undermined by an undesirable constellation of actors and actions involved in the concrete implementation. For instance, while using primary health care data for research purposes could in principle garner acceptance, unclearness regarding the involvement of private actors without sufficient safeguards might face criticism and resistance (Carter et al., 2015).

Before this background, this dissertation (1) discussed how data-driven technologies affect societies with respect to social inequality and privacy and why gauging these effects requires the consideration of social context, (2) argued that public acceptance is a relevant factor for the legitimacy and success of data-driven technologies, and empirically demonstrated why and how public acceptance needs to be measured context-specifically, and (3) provided empirical evidence on the acceptance of concrete data-driven technologies and their inherent data flows, showing which context factors are of particular importance in comparative research.

How can researchers and organizations meaningfully make sense of and describe data-driven technologies and their inherent data flows within their contexts? For privacy, Nissenbaum's (2010) understanding of "context" as "social context", and of privacy norms as one kind of context-specific norms, provides a conceptualization that is soundly grounded in social theory (see Chapter 1.2) while also defining specific parameters that can be concretely operationalized in empirical research. As outlined in Chapter 1.2, this contextual norm-based perspective is in principle applicable to data-driven technologies including ADM as well, while

the different steps in ADM processes come with different challenges and types of norms that may be respectively relevant.

Investigating challenges regarding social impacts in each of these steps, Chapter 2 provided an analytical perspective on ADM systems as data-driven technologies, based on a "big data process model" by Weyer et al. (2018). For each step of ADM employment, the chapter showed which challenges ADM systems pose for the social contexts in which they are placed. With a focus on social inequality, these challenges relate to norms and understandings of what is considered fair or just, a distinction made by Kuppler et al. (2022), who relate the former to the algorithm (i.e., the analysis phase) and the latter to the actual decision (i.e., the implementation phase). Additionally, norms relating to social inequality are relevant when assessing the challenge of adequate representation of societal groups in the used data bases (see Chapter 2). If not properly addressed, these challenges could threaten the legitimacy of potentially advantageous ADM systems. To address these challenges, a context-based view appears necessary to evaluate ADM systems' concrete impacts, given that they are "sociotechnical systems" interwoven with social contexts (Selbst et al., 2019). To this end, researchers from the social sciences have great potential to use methods, general sociological concepts, and context-specific substantive knowledge – not only for measuring the "objective" impacts on specific contexts and society at large, but also the ethically relevant public acceptance of such systems.

Empirically applying a context-based perspective to ADM systems, the findings from the survey experiment presented in Chapter 3 have shown that the importance of specific design features for the acceptance of ADM systems varies across four social contexts in which potentially problematic ADM systems have been or could be used. However, there was the general tendency to deem fair and accept hybrid decision-making (with a human deciding based on an algorithmic recommendation) similarly as or even more than a fully human decision (only with some computational assistance), and both were overall deemed fairer and more accepted than fully automated decisions. Moreover, supporting the notion of contextual integrity, respondents on average rated systems worse if they draw on non-contextual data. Assistive decisions were more accepted than punitive decisions, but not in all contexts. These findings imply that all of the included design features are relevant for a called-for further theoretical integration of research on ADM-related perceptions (Starke et al., 2022, and see below). Of particular importance for practical purposes, the results could encourage increased use of ADM systems with some element of automation responsibly, as long as the final decision is made by a human. However, considering the research described in Chapter 2, the practical interaction of

humans with algorithmic recommendations may be marked by over- or underreliance (Wickens et al., 2015; Zerilli et al., 2019), leading to suboptimal outcomes. The empirical results from Chapter 3 therefore do not pave the way for a "free ride", but they do hint at public acceptance towards responsibly reaping some of the benefits of automation for specific decision-making procedures.

Chapter 4 provided evidence that public acceptance of a specific data use at a specific time point does not constitute a permanent "social license" for this data use. The results from a longitudinal survey experiment on acceptability of data uses in different social contexts showed not only that acceptance is sensitive towards changes in the concrete specifications of the data use. The results also demonstrated that the COVID-19 pandemic, as a large societal crisis, affected public acceptance of data uses that directly relate to this crisis. Concretely, the findings revealed that acceptance of health data use clearly increased from before the pandemic (July 2019) to the first peak of the pandemic in Germany in March/April 2020. These results demonstrate the usefulness of a contextual integrity-based perspective also for longitudinal comparisons of privacy attitudes, as the found longitudinal changes were shown to be confined to this specific data type (and did not appear for location or energy use data). Furthermore, contributing to theory development, "purpose" was shown to be a useful addition to the contextual integrity parameters. Nissenbaum understands "purpose" as defining what social contexts are, along with functions and values (Nissenbaum, 2019). She also points out that one context might have multiple purposes and the relative importance of these purposes might be debated (ibd.). Therefore, an explication of the concrete purpose of a data flow may be worthwhile, and this study demonstrated the purpose's empirical relevance. The results also re-iterate that researchers and policymakers always need to interpret findings on public attitudes with respect to the time that the data were collected. From a regulatory perspective, these results pose the challenge to allow specific data flows to tackle exceptional challenges, while not unduly extending these data flows over time and for questionable purposes (Vitak & Zimmer, 2020). To be clear, societal challenges do not necessarily justify every kind of data use and should be limited to *appropriate* uses in the sense of contextual integrity (ibd.; e.g., note the different solutions for digital contact tracing during the COVID-19 pandemic, see Hogan et al., 2021).

Finally, Chapter 5 placed and empirically applied the perspective of contextual integrity within a larger framework, the Comparative Privacy Research Framework (Masur et al., 2021). While the latter has called for considering contextual integrity for meso-level comparisons, Chapter 5 theoretically elaborated and empirically demonstrated how exactly contextual

integrity can be employed in international comparative privacy research. To this end, this chapter added an international and an interindividual comparative perspective to the context-specific and longitudinal comparisons provided in the previous chapter. The results have shown that the relative importance of contextual parameters varied across countries. Particularly the acceptance of data recipients differed across countries, such that German respondents were relatively less accepting of public authorities as data recipients than respondents from Spain and the UK. For Germany, this mirrors findings for several scenarios of the study presented in Chapter 4 (which was conducted in Germany). There are some more consistent findings across countries, e.g., that respondents react particularly sensitively to a change in data type. Thus, while Chapter 4 highlights the relevance of interpreting findings with respect to *time*, Chapter 5 additionally highlights the relevance of *place*, i.e., countries, in this case understood as cultural structures (Li et al., 2017; Masur et al., 2021). Future research can draw on this integration of contextual integrity within the Comparative Privacy Research Framework, i.e., combining international comparisons with a meso-level context-based approach, and broaden the comparative scope to additional countries and types of structures, such as economic and political systems (see Masur et al., 2021).

Limitations of the presented studies have been discussed in the respective chapters. As for more general limitations, while probability-based samples were used in Chapter 3 and partly in Chapter 4, the results from the other data collections would profit from replication with probability-based samples. The experimental procedures could be further improved, e.g., by using additional means to check respondents' attention (e.g., instructional manipulation checks: Oppenheimer et al., 2009). As an avenue for theoretical development, while Chapter 2 has analytically categorized sources of "objective" impacts of ADM on social inequality, future research could also work towards to a stronger theoretical integration of determinants of ADM fairness perceptions (as suggested by Starke et al., 2022) and general acceptance (see above). This dissertation has demonstrated that such a systematization should take into account not only features of ADM systems, but also features of their social contexts, such as how strongly public benefits are expected to arise in a context. Furthermore, the presented studies only investigated a selection of contexts, countries, and time points. To arrive at more generalizable theories, future research needs to carefully select and research further instances of these comparative components. Combining a systematization of context features with international comparisons, future studies could, for example, include a larger number of countries that cover different levels of individualism (see Hofstede, 1984) while investigating data flows and ADM systems that

vary by how closely they relate to public- versus private-focused contexts (see on privacy and with a smaller scope: Li et al., 2017).

As a more general call, this dissertation has argued and demonstrated that the sometimes-overlooked expertise of social scientists is crucial for the development and regulation of data-driven technologies. As has been discussed above and particularly in Chapter 2, social scientists have relevant knowledge about social contexts and processes. They – together with other disciplines – have theories and empirical evidence on how norms are formed and influence behavior (see, e.g., Gelfand et al. 2024; Horne & Mollborn, 2020). They also know, for instance, how discrimination works in specific social contexts, and they command qualitative and quantitative research methods to learn about contexts for which reliable knowledge is not yet available. Social scientists can furthermore uncover how exactly ADM impacts contexts and societies at the macro-level by paying attention to how the aggregation of individual actions leads to macro-social outcomes (Coleman, 1994) using, e.g., empirically informed agent-based modeling (see Gilbert, 2008), and disentangle these processes by adjusting measurements to account for social spheres being jointly created by humans *and* algorithms (Wagner et al., 2021). Furthermore, longitudinal research, including longitudinal content and discourse analysis, can unveil how public acceptance feeds back to the use and regulation of data-driven technologies. When conducting such research, social scientists have the means to give appropriate account to context-specificities, while also considering how these specificities can be systematized. With these means, social scientists can contribute to a more comprehensive theoretical integration of determinants of impacts and acceptance of data-driven technologies.

As a final summarizing note, the findings presented in this dissertation made clear that individuals react sensitively towards how exactly their data are being used by whom and when. They also pay attention to how exactly data-driven ADM systems make decisions about people. From a scientific perspective, as argued above, such research is relevant for gathering empirical knowledge and building theories on acceptance that appropriately take into account context-specific norms. From a practical perspective, public agencies and businesses should take these individual perceptions seriously to evaluate whether their uses of data-driven technologies obtain "social licenses". However, ethical evaluations of data-driven technologies do not hinge on social licenses and contextual norms alone, and further constraints and requirements need to be considered. These include economic and legal licenses (Gunningham et al., 2004) and, beyond contextual norms and values, according to contextual integrity, also broader societal and political values such as equality (Nissenbaum, 2010). Furthermore, there are different ethical stances on the relevance of "duties" vis-à-vis consequences (see Chapter 1.2 and Bednar

& Spiekermann, 2022). While “social licenses” may seem to emphasize the aspect of “duties”, preferences regarding the objective consequences of the use of data-driven technologies can be part of negotiating the agreements on how such technologies should be employed (Rahwan, 2018).

In conclusion, context-specific public acceptance is not the only, but one important component of ethical evaluations of data-driven technologies. These evaluations should therefore acknowledge public acceptance of data-driven technologies as ethically relevant in its own right.

References

Bednar, K., & Spiekermann, S. (2022). Eliciting Values for Technology Design with Moral Philosophy: An Empirical Exploration of Effects and Shortcomings. *Science, Technology, & Human Values*, 016224392211225. https://doi.org/10.1177/01622439221122595

Carter, P., Laurie, G. T., & Dixon-Woods, M. (2015). The social licence for research: Why care.data ran into trouble. *Journal of Medical Ethics*, *41*(5), 404–409. https://doi.org/10.1136/medethics-2014-102374

Coleman, J. S. (1994). *Foundations of Social Theory*. Belknap Press of Harvard University Press.

Gelfand, M. J., Gavrilets, S., & Nunn, N. (2024). Norm Dynamics: Interdisciplinary Perspectives on Social Norm Emergence, Persistence, and Change. *Annual Review of Psychology*, *75*(1), 341–378. https://doi.org/10.1146/annurev-psych-033020-013319

Gilbert, G. N. (2008). *Agent-Based Models*. Sage.

Gunningham, N., Kagan, R. A., & Thornton, D. (2004). Social License and Environmental Protection: Why Businesses Go Beyond Compliance. *Law & Social Inquiry*, *29*(2), 307–341. https://doi.org/10.1111/j.1747-4469.2004.tb00338.x

Hofstede, G. (1984). Cultural dimensions in management and planning. *Asia Pacific Journal of Management*, *1*(2), 81–99. https://doi.org/10.1007/BF01733682

Hogan, K., Macedo, B., Macha, V., Barman, A., & Jiang, X. (2021). Contact Tracing Apps: Lessons Learned on Privacy, Autonomy, and the Need for Detailed and Thoughtful Implementation. *JMIR Medical Informatics*, *9*(7), e27449. https://doi.org/10.2196/27449

Horne, C., & Mollborn, S. (2020). Norms: An Integrated Framework. *Annual Review of Sociology*, *46*(1), 467–487. https://doi.org/10.1146/annurev-soc-121919-054658

Kuppler, M., Kern, C., Bach, R. L., & Kreuter, F. (2022). From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. *Frontiers in Sociology*, *7*, 883999. https://doi.org/10.3389/fsoc.2022.883999

Li, Y., Kobsa, A., Knijnenburg, B. P., & Carolyn Nguyen, M.-H. (2017). Cross-Cultural Privacy Prediction. *Proceedings on Privacy Enhancing Technologies*, *2017*(2), 113–132. https://doi.org/10.1515/popets-2017-0019

Masur, P. K., Epstein, D., Quinn, K., Wilhelm, C., Baruh, L., & Lutz, C. (2021). *A comparative privacy research framework*. https://doi.org/10.31235/osf.io/fjqhs

Nissenbaum, H. (2010). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Nissenbaum, H. (2019). Contextual Integrity Up and Down the Data Food Chain. *Theoretical Inquiries in Law*, *20*(1), 221–256. https://doi.org/10.1515/til-2019-0008

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29-31 January 2019*, 59–68. https://doi.org/10.1145/3287560.3287598

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, *9*(2), 205395172211151. https://doi.org/10.1177/20539517221115189

Vitak, J., & Zimmer, M. (2020). More than just privacy: Using contextual integrity to evaluate the long-term risks from COVID-19 surveillance technologies. *Social Media + Society*, *6*(3), 205630512094825. https://doi.org/10.1177/2056305120948250

Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N., & Eliassi-Rad, T. (2021). Measuring algorithmically infused societies. *Nature*, *595*(7866), 197–204. https://doi.org/10.1038/s41586-021-03666-1

Weyer, J., Delisle, M., Kappler, K., Kiehl, M., Merz, C., & Schrape, J.-F. (2018). Big Data in soziologischer Perspektive. In B. Kolany-Raiser, R. Heil, C. Orwat, & T. Hoeren (Eds.), *Big Data und Gesellschaft* (pp. 69–149). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-21665-8_2

Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and Automation Bias in the Use of Imperfect Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(5), 728–739. https://doi.org/10.1177/0018720815581940

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Algorithmic Decision-Making and the Control Problem. *Minds and Machines*, *29*(4), 555–578. https://doi.org/10.1007/s11023-019-09513-7

7. Appendices

7.1. Appendix for Chapter 2

Table A2.1. Summary of sources of inequality, related social science topics, example papers, and research avenues.

	Elements that may affect social inequality	Related social science topics	Example papers	Research avenues for social sciences
Data Generation	*Historical bias*	Domain-specific mechanisms of discrimination	E.g., for digital divides: Lutz (2019)	• Describing, e.g., digital divides in competencies, use, and affectedness • Using methodological advances in survey research to correct biases in data input for ADM
	Selective participation	Coverage and misrepresentation in survey research	Kim et al. (2022)	
	Measurement bias	Measurement in survey research; latent variable modeling	Boeschoten et al. (2020); Jacobs and Wallach (2021)	
Data Preparation and Analysis	*Selection and definition of protected attributes*	Social stratification; intersectionality	Mann and Matzner (2019)	• Developing a framework how contextual and individual characteristics shape fairness perceptions • Choosing context-specific fairness metrics • Defining domain-specific non-discriminatory outcomes
	Choice of fairness metrics	Theories of distributive justice	Kuppler et al. (2021)	
	Differential fairness perceptions	Psychology and sociology of justice	Binns et al. (2018); Grgić-Hlača et al. (2018); Starke et al. (2021)	
Implementation	*Differential bias and accuracy compared to human deciders*	Methods: experiments; causal inference in real-life settings	Dressel and Farid (2018); Green and Chen (2019)	• Describing individual preemptive or reactive behavior towards ADM • Explaining reliance on ADM by situational and individual characteristics • Developing methods for investigating macro-social outcomes of ADM implementation
	Differential adoption of algorithmic recommendations	Human factors research, trust, experience; causal inference	Stevenson and Doleac (2019)	
	Macro social outcomes	Micro-macro model of sociological explanation	Adelt et al. (2018); Cruz Cortés and Ghosh (2019)	

Publications referenced in Table A2.1

Adelt, F., Weyer, J., Hoffmann, S., & Ihrig, A. (2018). Simulation of the Governance of Complex Systems (SimCo): Basic Concepts and Experiments on Urban Transportation. *Journal of Artificial Societies and Social Simulation, 21*, 2. https://doi.org/10.18564/jasss.3654

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's Reducing a Human Being to a Percentage'. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18, Montreal QC, Canada, 21-26 April 2018*, 1–14. Association for Computing Machinery. https://doi.org/10.1145/3173574.3173951

Boeschoten, L., van Kesteren, E.-J., Bagheri, A., & Oberski, D. L. (2020). *Fair Inference on Error-Prone Outcomes*. Available at: https://arxiv.org/abs/2003.07621 (accessed 10 May 2021).

Cruz Cortés, E., & Ghosh, D. (2019). *A Simulation Based Dynamic Evaluation Framework for System-wide Algorithmic Fairness*. Available at: https://arxiv.org/abs/1903.09209 (accessed 10 May 2021).

Dressel, J. & Farid, H. (2018). The Accuracy, Fairness, and Limits of Predicting Recidivism. *Science Advances*, 4(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Green, B., & Chen, Y. (2019). Disparate Interactions. An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29-31 January 2019*, 90–99. Association for Computing Machinery. https://doi.org/10.1145/3287560.3287563

Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). Human Perceptions of Fairness in Algorithmic Decision Making. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18, Lyon, France, 23-27 April 2018*, 903–912. Association for Computing Machinery. https://doi.org/10.1145/3178876.3186138

Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Canada, 3-10 March 2021*, 375–385. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445901

Kim, M. P., Kern, C., Goldwasser, S., Kreuter, F., & Reingold, O. (2022). Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, *119*(4), e2108097119. https://doi.org/10.1073/pnas.2108097119

Kuppler, M., Kern, C., Bach, R. L., & Kreuter, F. (2021). *Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There?* Available at: https://arxiv.org/abs/2105.01441 (accessed 1 December 2021).

Lutz, C. (2019). Digital Inequalities in the Age of Artificial Intelligence and Big Data. *Human Behavior and Emerging Technologies*, *1*(2), 141–148. https://doi.org/10.1002/hbe2.140

Mann, M. & Matzner, T. (2019). Challenging Algorithmic Profiling: The Limits of Data Protection and Anti-Discrimination in Responding to Emergent Discrimination. *Big Data & Society*, *6*(2), 1–11. https://doi.org/10.1177/2053951719895805

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2021). *Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature*. https://arxiv.org/abs/2103.12016 (accessed 10 May 2021).

Stevenson, M. T., & Doleac J. L. (2019). *Algorithmic Risk Assessment in the Hands of Humans*. Available at: http://ftp.iza.org/dp12853.pdf (accessed 10 May 2021).

7.2. Appendix for Chapter 3

Figure A3.1. Distribution of fairness evaluations by vignette levels

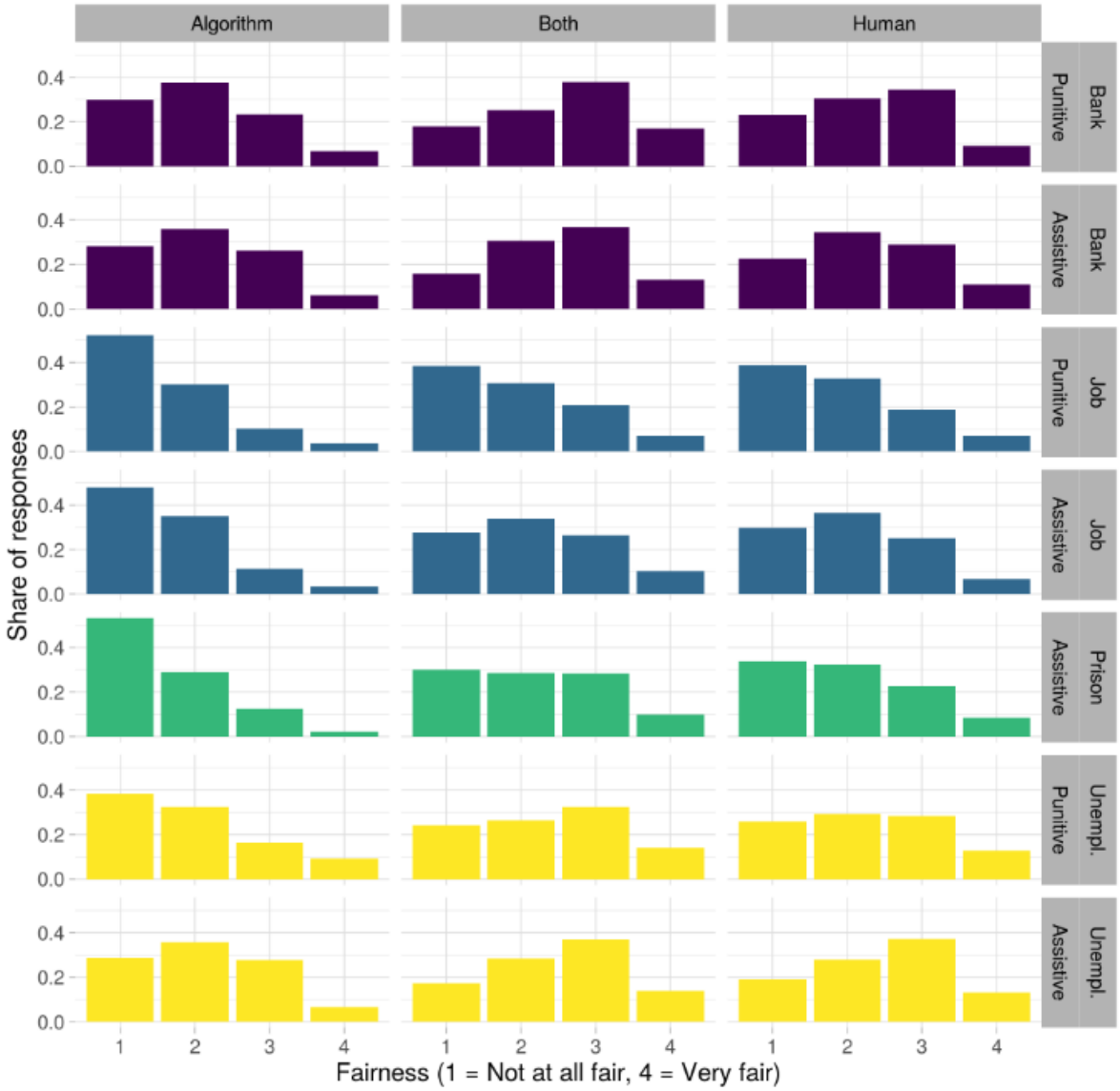


Figure A3.2. Distribution of acceptance ratings by vignette levels.

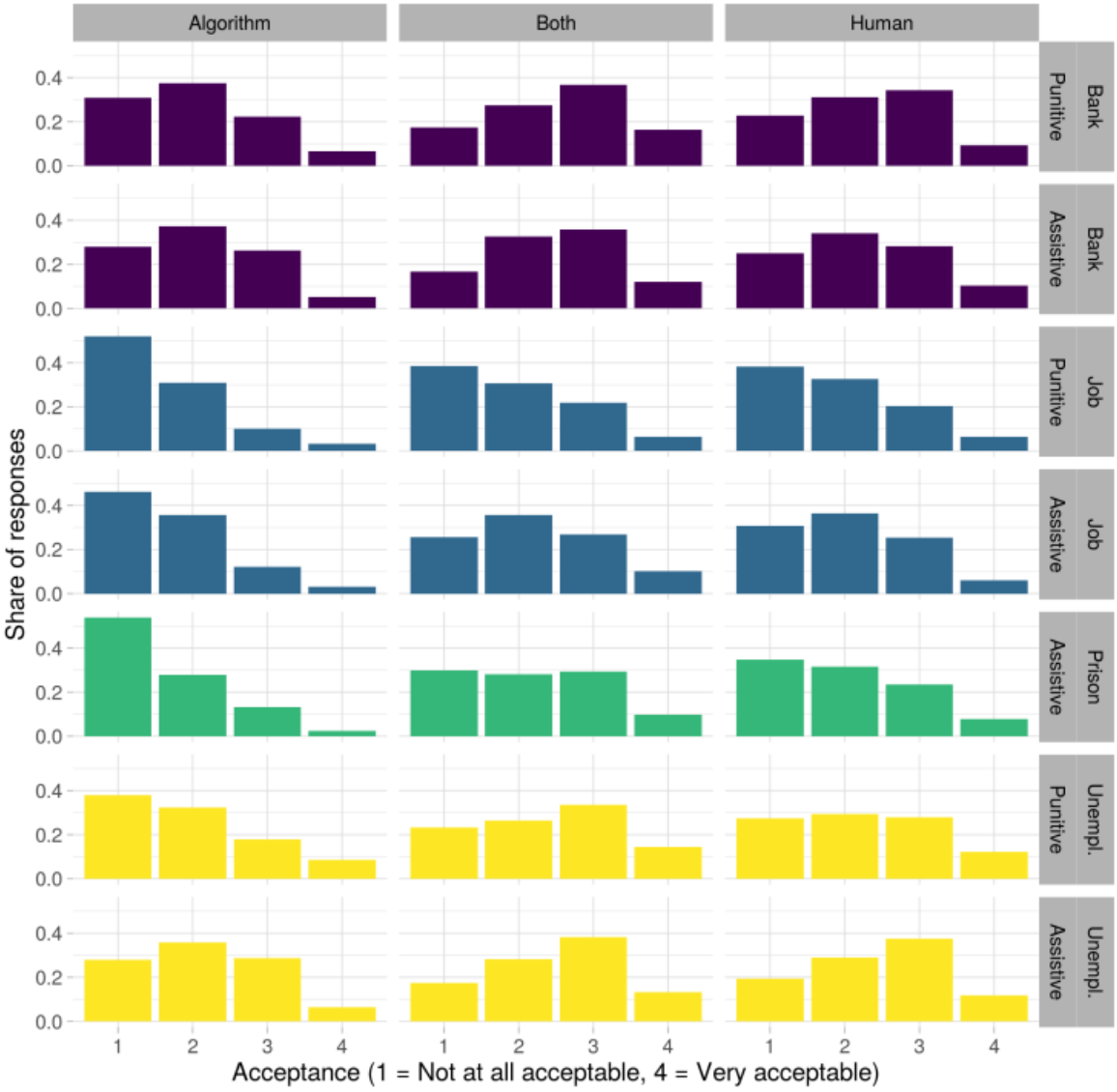
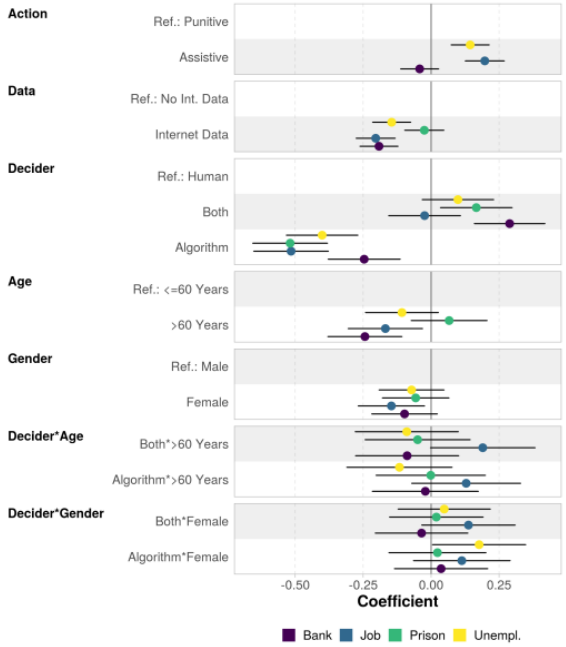


Figure A3.3. Coefficients (with 95% confidence intervals) of ordinal probit regression models predicting acceptance ratings of each context with interactions between the vignette dimension decision-maker and respondent characteristics.

(a) Context-specific Interactions 1 ($n_{Bank} = 3662$, $n_{Job} = 3671$, $n_{Prison} = 3660$, $n_{Unempl.} = 3666$)



(b) Context-specific Interactions 2 ($n_{Bank} = 3852$, $n_{Job} = 3855$, $n_{Prison} = 3852$, $n_{Unempl.} = 3854$)

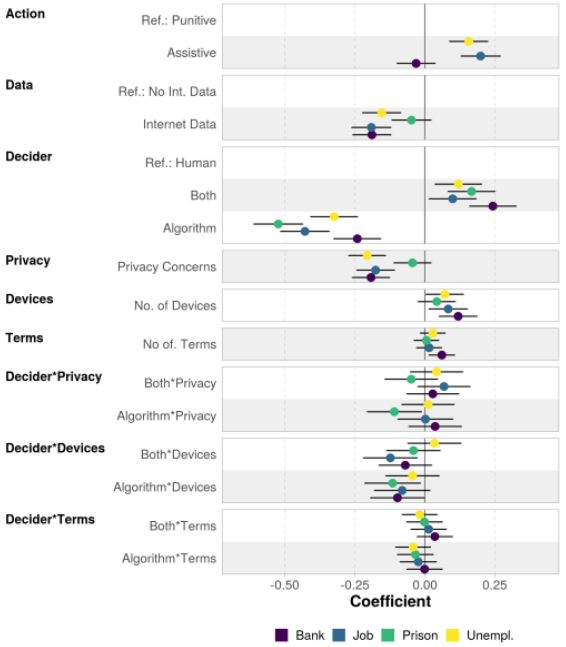


Table A3.1. Average fairness and acceptance ratings by vignette levels.

(a) Relative frequencies of respondents that rated a scenario as "Fair" ("Somewhat fair" or "Very fair").

			Context			
Decision-maker	Action	Data	Bank	Job	Prison	Unempl.
Algorithm	Assistive	Internet	0.31	0.17	0.16	0.29
Algorithm	Assistive	No Internet	0.37	0.13	0.14	0.40
Algorithm	Punitive	Internet	0.29	0.12		0.28
Algorithm	Punitive	No Internet	0.32	0.17		0.25
Both	Assistive	Internet	0.47	0.35	0.35	0.49
Both	Assistive	No Internet	0.57	0.40	0.44	0.56
Both	Punitive	Internet	0.50	0.25		0.42
Both	Punitive	No Internet	0.61	0.33		0.54
Human	Assistive	Internet	0.38	0.30	0.30	0.47
Human	Assistive	No Internet	0.44	0.35	0.34	0.56
Human	Punitive	Internet	0.38	0.20		0.40
Human	Punitive	No Internet	0.52	0.34		0.45

(b) Relative frequencies of respondents that rated a scenario as "Acceptable" ("Somewhat acceptable" or "Very acceptable").

			Context			
Decision-maker	Action	Data	Bank	Job	Prison	Unempl.
Algorithm	Assistive	Internet	0.29	0.15	0.17	0.30
Algorithm	Assistive	No Internet	0.36	0.16	0.15	0.41
Algorithm	Punitive	Internet	0.29	0.10		0.29
Algorithm	Punitive	No Internet	0.31	0.17		0.26
Both	Assistive	Internet	0.44	0.34	0.37	0.49
Both	Assistive	No Internet	0.54	0.41	0.44	0.57
Both	Punitive	Internet	0.50	0.26		0.42
Both	Punitive	No Internet	0.58	0.32		0.56
Human	Assistive	Internet	0.35	0.30	0.30	0.46
Human	Assistive	No Internet	0.43	0.34	0.34	0.55
Human	Punitive	Internet	0.39	0.20		0.39
Human	Punitive	No Internet	0.50	0.34		0.44

Table A3.2. Summary statistics.

(a) Vignettes

Variable	Values	Freqs (% of Valid)	Valid
Fairness	Not at all fair	5058 (32.6%)	15525 (97.2%)
	A little fair	5016 (32.3%)	
	Somewhat fair	4036 (26.0%)	
	Very fair	1415 (9.1%)	
Acceptance	Not at all acceptable	5078 (32.6%)	15566 (97.5%)
	A little acceptable	5044 (32.4%)	
	Somewhat acceptable	4101 (26.3%)	
	Very acceptable	1343 (8.6%)	

(b) Respondents

Variable	Stats/ Values	Freqs (% of Valid)	Valid
Age	≤60 Years	2729 (72.6%)	3760 (94.2%)
	>60 Years	1031 (27.4%)	
Gender	Male	2072 (51.9%)	3991 (99.9%)
	Female	1919 (48.1%)	
Number of Devices†	Mean (sd) : 2.6 (0.9)	0 : 25 (0.6%)	3876 (97.1%)
	min < med < max:	1 : 335 (8.6%)	
	0 < 3 < 5	2 : 1379 (35.6%)	
	IQR (CV) : 1 (0.3)	3 : 1452 (37.5%)	
		4 : 636 (16.4%)	
		5 : 49 (1.3%)	
Number of Terms††	Mean (sd) : 3.3 (1.4)	0 : 184 (4.7%)	3877 (97.1%)
	min < med < max:	1 : 266 (6.9%)	
	0 < 3 < 5	2 : 491 (12.7%)	
	IQR (CV) : 2 (0.4)	3 : 1046 (27.0%)	
		4 : 900 (23.2%)	
		5 : 990 (25.5%)	
Privacy Index†††	Mean (sd) : 3.5 (0.9)		3873 (97.0%)
	min < med < max:		
	1 < 3.5 < 5		
	IQR (CV) : 1 (0.3)		

†Answer categories: Smartphone, Cell Phone, Desktop Computer, Tablet, eBook Reader.

††Answer categories: Artificial Intelligence, Computer Algorithms, Machine Learning, Recommender Systems, Targeted/personalized Ads.

†††Item 1: "I do not mind sharing personal information as nowadays everyone is doing this anyway."

Item 2: "You cannot live in the modern world without sharing personal information."

Item 3: "When you provide personal information you never know who else is going to see it."

Item 4: "I do not mind sharing personal information in return for a product or service that I want."

Table A3.3. Random effects estimates and model fit indices of mixed-effects ordinal probit regression models predicting fairness evaluations and acceptance ratings.

(a) Outcome: Fairness

	R-I Main	R-I Interaction	R-I-R-S
LL	−18 234.01	−18 206.79	−18 196.84
BIC	36 583.83	36 587.28	36 557.73
ICC	0.46	0.46	0.51
Variance: Intercept	0.84	0.84	1.04
Variance: Decider Both			0.09
Variance: Decider Algorithm			0.38
L-R Test		54.45†	74.35†
Num. Observations	15 525	15 525	15 525
Num. Respondents	3930	3930	3930

†: $p \leq 0.001$

(b) Outcome: Acceptance

	R-I Main	R-I Interaction	R-I-R-S
LL	−18 227.38	−18 202.31	−18 188.81
BIC	36 570.59	36 578.36	36 541.71
ICC	0.45	0.45	0.50
Variance: Intercept	0.83	0.83	1.02
Variance: Decider Both			0.07
Variance: Decider Algorithm			0.43
L-R Test		50.14†	77.14†
Num. Observations	15 566	15 566	15 566
Num. Respondents	3972	3972	3972

†: $p \leq 0.001$

Table A3.4. Average predicted probabilities based on the R-I Interaction model. Predictions for a given predictor level are computed while setting the remaining vignette dimensions to their reference level.

(a) Outcome category "Very fair"

Predictor	$\hat{P}(y=4)$	95% CI	
Bank	0.06	0.06	0.07
Hire	0.02	0.02	0.02
Prison	0.02	0.02	0.02
Unempl	0.06	0.05	0.07
Human	0.04	0.04	0.05
Both	0.06	0.06	0.07
Alg.	0.02	0.01	0.02
Punitive	0.03	0.03	0.04
Assistive	0.05	0.04	0.05
No Internet	0.05	0.04	0.05
Internet	0.03	0.03	0.04

(b) Outcome category "Somewhat fair"

Predictor	$\hat{P}(y=3)$	95% CI	
Bank	0.34	0.32	0.35
Hire	0.19	0.18	0.20
Prison	0.19	0.17	0.20
Unempl	0.33	0.31	0.34
Human	0.29	0.27	0.30
Both	0.33	0.31	0.34
Alg.	0.16	0.15	0.17
Punitive	0.24	0.23	0.25
Assistive	0.28	0.27	0.29
No Internet	0.28	0.27	0.29
Internet	0.24	0.23	0.25

(c) Outcome category "A little fair"

Predictor	$\hat{P}(y=2)$	95% CI	
Bank	0.41	0.40	0.42
Hire	0.41	0.40	0.43
Prison	0.40	0.39	0.41
Unempl	0.41	0.40	0.43
Human	0.43	0.42	0.44
Both	0.41	0.40	0.42
Alg.	0.39	0.38	0.40
Punitive	0.41	0.40	0.42
Assistive	0.41	0.40	0.42
No Internet	0.41	0.40	0.42
Internet	0.41	0.40	0.42

(d) Outcome category "Not at all fair"

Predictor	$\hat{P}(y=1)$	95% CI	
Bank	0.19	0.17	0.20
Hire	0.38	0.36	0.40
Prison	0.39	0.37	0.41
Unempl	0.20	0.18	0.21
Human	0.24	0.23	0.25
Both	0.20	0.18	0.21
Alg.	0.43	0.41	0.45
Punitive	0.31	0.30	0.33
Assistive	0.26	0.25	0.28
No Internet	0.26	0.25	0.27
Internet	0.32	0.31	0.33

Table A3.5. Average conditional predicted probabilities based on the R-I Interaction model. Predictions for a given level of decision-maker are computed conditional on different levels of context, while setting the remaining vignette dimensions to their reference level.

(a) Outcome category "Very fair"

Context	Decider	$\hat{P}(y=4)$	95% CI	
Bank	Human	0.07	0.06	0.08
	Both	0.12	0.10	0.13
	Alg.	0.04	0.03	0.04
Job	Human	0.03	0.02	0.03
	Both	0.03	0.03	0.04
	Alg.	0.01	0.00	0.01
Prison	Human	0.03	0.02	0.03
	Both	0.04	0.03	0.05
	Alg.	0.00	0.00	0.00
Unempl.	Human	0.08	0.07	0.09
	Both	0.10	0.09	0.11
	Alg.	0.03	0.02	0.03

(b) Outcome category "Somewhat fair"

Context	Decider	$\hat{P}(y=3)$	95% CI	
Bank	Human	0.36	0.34	0.38
	Both	0.43	0.41	0.45
	Alg.	0.28	0.26	0.30
Job	Human	0.25	0.22	0.27
	Both	0.26	0.24	0.28
	Alg.	0.11	0.09	0.12
Prison	Human	0.24	0.22	0.27
	Both	0.29	0.27	0.31
	Alg.	0.08	0.07	0.09
Unempl.	Human	0.38	0.36	0.40
	Both	0.41	0.39	0.43
	Alg.	0.25	0.23	0.27

(c) Outcome category "A little fair"

Context	Decider	$\hat{P}(y=2)$	95% CI	
Bank	Human	0.41	0.40	0.43
	Both	0.36	0.34	0.38
	Alg.	0.44	0.43	0.45
Job	Human	0.44	0.43	0.45
	Both	0.44	0.43	0.45
	Alg.	0.38	0.36	0.40
Prison	Human	0.44	0.43	0.45
	Both	0.44	0.43	0.45
	Alg.	0.34	0.32	0.36
Unempl.	Human	0.40	0.38	0.41
	Both	0.37	0.36	0.39
	Alg.	0.44	0.43	0.45

(d) Outcome category "Not at all fair"

Context	Decider	$\hat{P}(y=1)$	95% CI	
Bank	Human	0.16	0.14	0.18
	Both	0.10	0.08	0.11
	Alg.	0.24	0.22	0.27
Job	Human	0.28	0.26	0.31
	Both	0.26	0.24	0.29
	Alg.	0.51	0.48	0.54
Prison	Human	0.28	0.26	0.31
	Both	0.23	0.21	0.26
	Alg.	0.57	0.54	0.60
Unempl.	Human	0.14	0.12	0.16
	Both	0.11	0.10	0.13
	Alg.	0.28	0.25	0.30

7.3. Appendix for Chapter 4

Table A4.1. Distribution of age and gender across samples.

Cross-section 2019

	Unweighted		Weighted	
	Female	**Male**	**Female**	**Male**
18-29	164	111	168	113
30-39	113	150	115	157
40-49	128	141	120	132
50-59	171	168	170	170
60-69	129	126	132	134

Cross-section 2020

	Unweighted		Weighted	
	Female	**Male**	**Female**	**Male**
18-29	167	24	171	27
30-39	119	62	121	63
40-49	42	148	38	140
50-59	82	151	81	148
60-69	66	109	63	116

Benchmark 2020

	Unweighted		Weighted	
	Female	**Male**	**Female**	**Male**
18-29	89	76	87	76
30-39	64	80	68	86
40-49	79	84	72	74
50-59	96	94	95	98
60-69	69	70	77	73

Table A4.2. Rounded mean values and standard errors of acceptance for different data-sharing scenarios.

Wave	Data type	Use	Recipient	Mean (weighted)	SE (weighted)	N
Cross-section 2019	Energy	Private purpose	Agency	2.89	0.12	113
	Energy	Private purpose	Company	3.46	0.11	117
	Energy	Public purpose	Agency	3.25	0.11	115
	Energy	Public purpose	Company	3.42	0.10	112
	Health	Private purpose	Agency	2.55	0.12	116
	Health	Private purpose	Company	2.97	0.12	117
	Health	Public purpose	Agency	2.28	0.12	114
	Health	Public purpose	Company	2.69	0.11	119
	Location	Private purpose	Agency	2.65	0.13	120
	Location	Private purpose	Company	3.14	0.10	117
	Location	Public purpose	Agency	3.48	0.11	122
	Location	Public purpose	Company	3.50	0.10	119
Cross-section 2020	Energy	Private purpose	Agency	3.06	0.13	87
	Energy	Private purpose	Company	3.06	0.13	80
	Energy	Public purpose	Agency	3.06	0.14	79
	Energy	Public purpose	Company	3.46	0.12	79
	Health	Private purpose	Agency	2.56	0.13	79
	Health	Private purpose	Company	2.62	0.13	80
	Health	Public purpose	Agency	3.08	0.15	78
	Health	Public purpose	Company	3.14	0.14	81
	Location	Private purpose	Agency	2.57	0.13	80
	Location	Private purpose	Company	3.04	0.14	82
	Location	Public purpose	Agency	3.12	0.13	83
	Location	Public purpose	Company	3.34	0.12	82
Benchmark 2020	Energy	Private purpose	Agency	3.20	0.15	67
	Energy	Private purpose	Company	3.11	0.14	67
	Energy	Public purpose	Agency	3.69	0.13	61
	Energy	Public purpose	Company	3.59	0.13	59
	Health	Private purpose	Agency	2.60	0.14	69
	Health	Private purpose	Company	2.93	0.14	72
	Health	Public purpose	Agency	3.41	0.14	69
	Health	Public purpose	Company	3.31	0.16	64
	Location	Private purpose	Agency	3.04	0.15	68
	Location	Private purpose	Company	3.68	0.15	63
	Location	Public purpose	Agency	3.71	0.13	69
	Location	Public purpose	Company	3.79	0.12	73

Table A4.3. Rounded mean values and standard errors of acceptance for different data-sharing scenarios for longitudinal sample.

Wave	Data type	Use	Recipient	Mean (weighted)	SE (weighted)	N
Longitudinal 2019	Energy	Private purpose	Agency	2.84	0.16	57
	Energy	Private purpose	Company	3.42	0.19	56
	Energy	Public purpose	Agency	3.40	0.15	53
	Energy	Public purpose	Company	3.18	0.18	44
	Health	Private purpose	Agency	2.58	0.19	46
	Health	Private purpose	Company	2.81	0.19	49
	Health	Public purpose	Agency	2.09	0.17	58
	Health	Public purpose	Company	2.63	0.15	50
	Location	Private purpose	Agency	2.75	0.18	59
	Location	Private purpose	Company	3.02	0.15	56
	Location	Public purpose	Agency	3.67	0.20	39
	Location	Public purpose	Company	3.61	0.11	60
Longitudinal 2020	Energy	Private purpose	Agency	2.83	0.18	57
	Energy	Private purpose	Company	3.37	0.13	56
	Energy	Public purpose	Agency	3.38	0.13	53
	Energy	Public purpose	Company	3.18	0.18	44
	Health	Private purpose	Agency	3.30	0.19	46
	Health	Private purpose	Company	2.88	0.20	49
	Health	Public purpose	Agency	2.94	0.15	58
	Health	Public purpose	Company	3.06	0.17	50
	Location	Private purpose	Agency	2.92	0.16	59
	Location	Private purpose	Company	3.22	0.17	56
	Location	Public purpose	Agency	3.35	0.25	39
	Location	Public purpose	Company	3.45	0.14	60

Table A4.4. Rounded p-values of permutation KS tests (see Section 4.6). Weighted analysis.

Group comparison	Permutation KS test (p-value)	Conservative KS test (p-value)
Health vs non-health		
Cross-sectional: 2019, health vs non-health	0	0.000
Cross-sectional: 2020, health vs non-health	0	0.032
Cross-sectional: 2019 vs 2020 (health)	0.022	0.156
Cross-sectional: 2019 vs 2020 (non-health)	0.035	0.267
Longitudinal: 2019, health vs non-health	0	0.000
Longitudinal: 2020, health vs non-health	0.001	0.077
Longitudinal: 2019 vs 2020 (health)	0	0.002
Longitudinal: 2019 vs 2020 (non-health)	1	1
Among health vignettes: public vs private		
Cross-sectional: 2019, public vs private	0.018	0.069
Cross-sectional: 2020, public vs private	0	0.001
Cross-sectional: 2019 vs 2020 (public)	0	0.000
Cross-sectional: 2019 vs 2020 (private)	0.079	0.295
Longitudinal: 2019, public vs private	0.055	0.222
Longitudinal: 2020, public vs private	0.153	0.434
Longitudinal: 2019 vs 2020 (public)	0	0.004
Longitudinal: 2019 vs 2020 (private)	0.081	0.254

Figure A4.1. Mean acceptability of different data transmission scenarios across samples, excluding respondents of age 70+ in the benchmark sample. Cross-section 2020: N = 970. Benchmark 2020: N = 801. Weighted analysis.

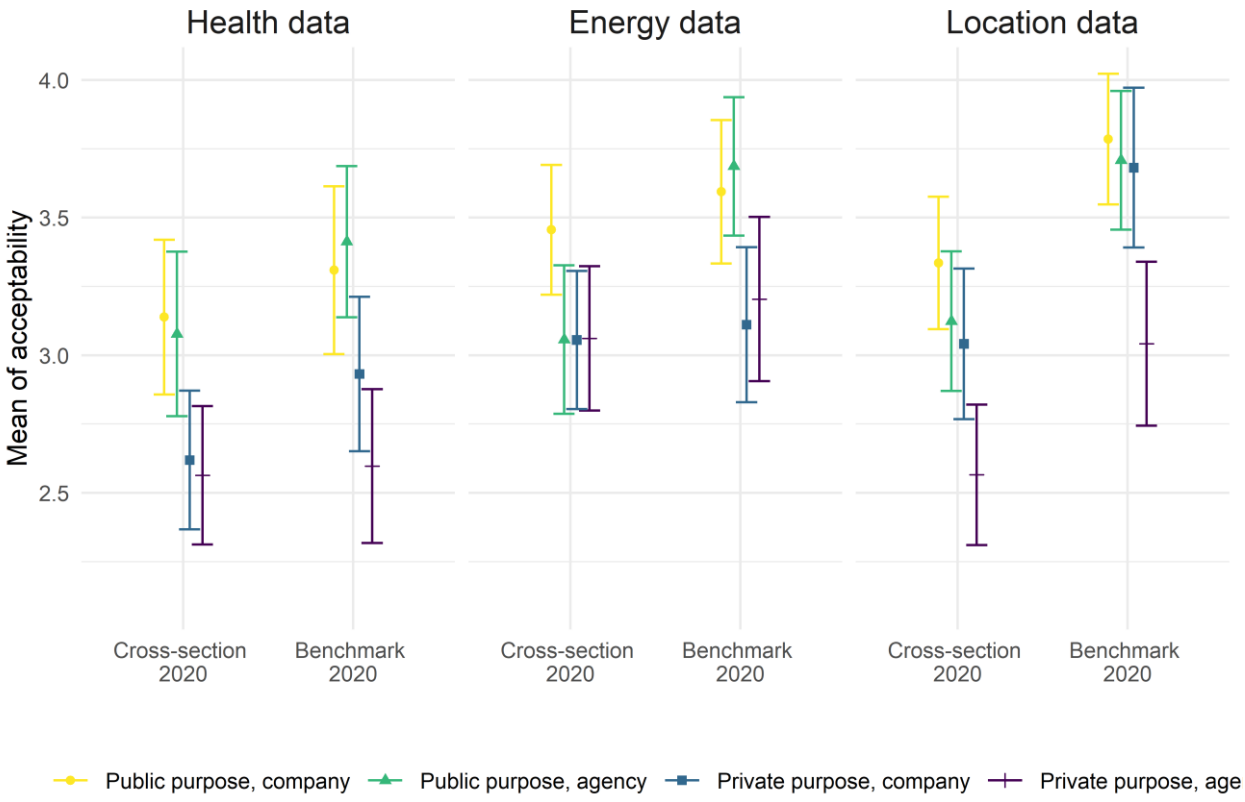


Figure A4.2a. Distribution of responses in the cross-sectional sample 2019. Weighted analysis.

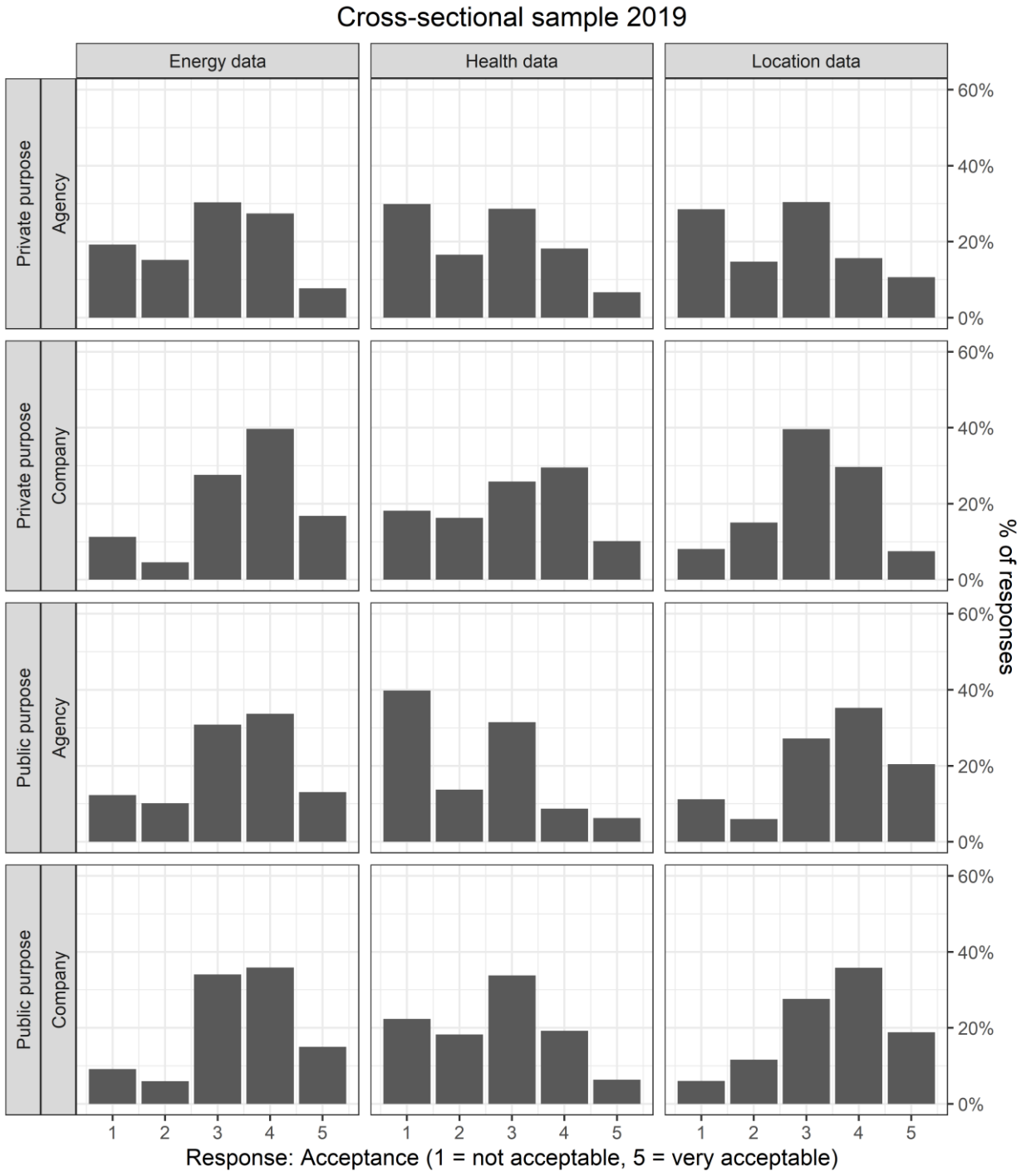


Figure A4.2b. Distribution of responses in the cross-sectional sample 2020. Weighted analysis.

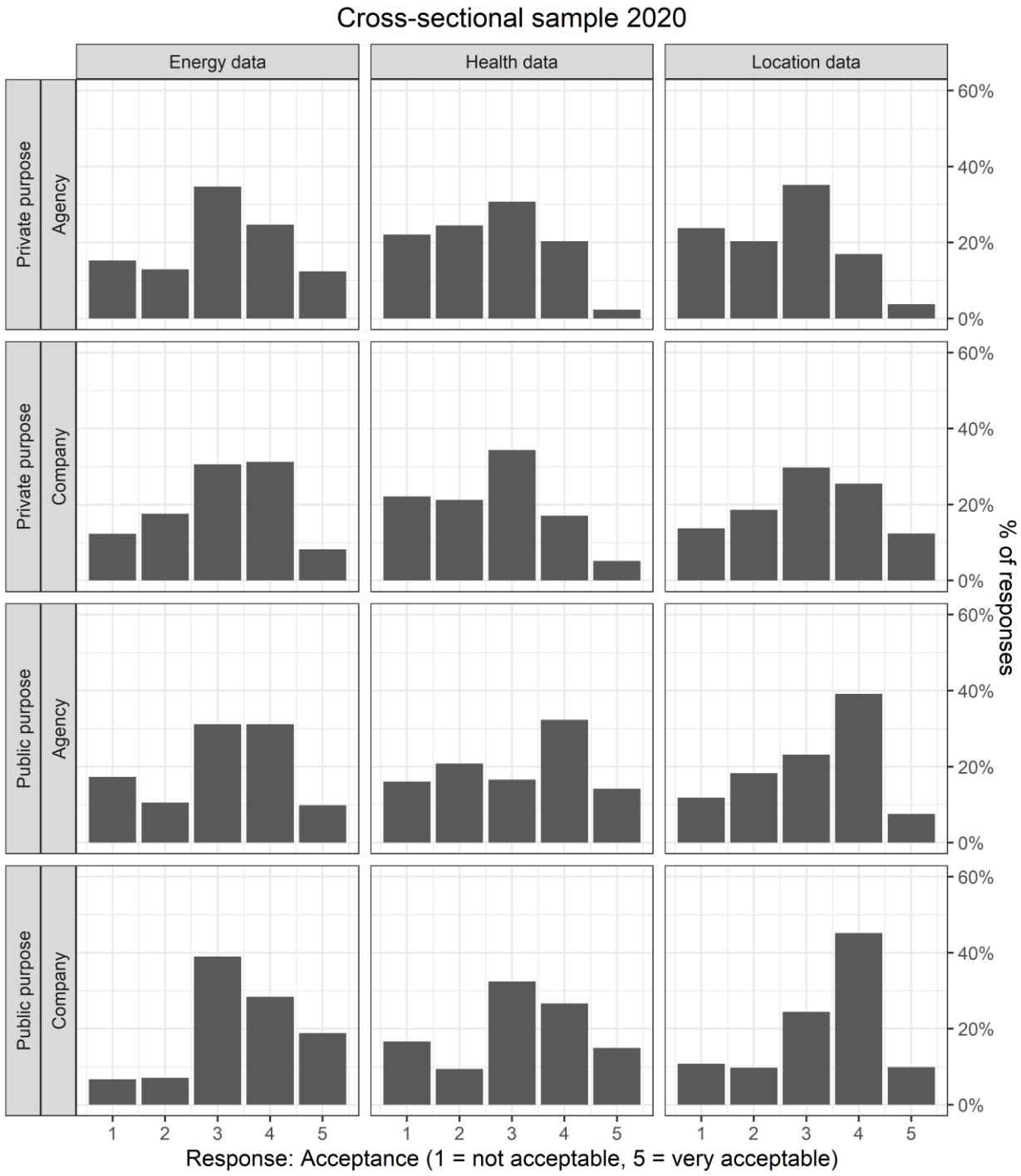


Figure A4.2c. Distribution of responses in the longitudinal sample 2019. Weighted analysis.

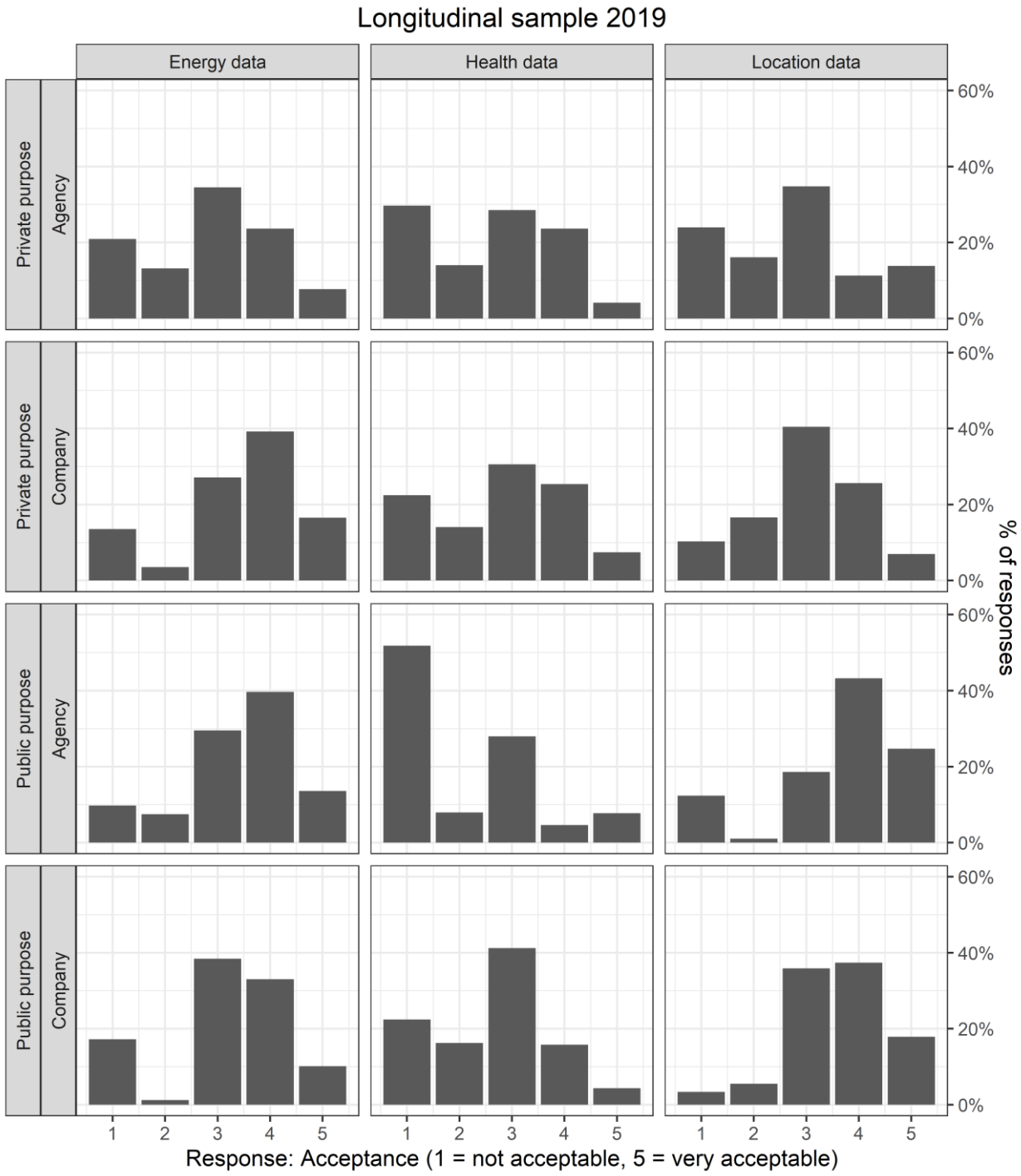


Figure A4.2d. Distribution of responses in the longitudinal sample 2020. Weighted analysis.

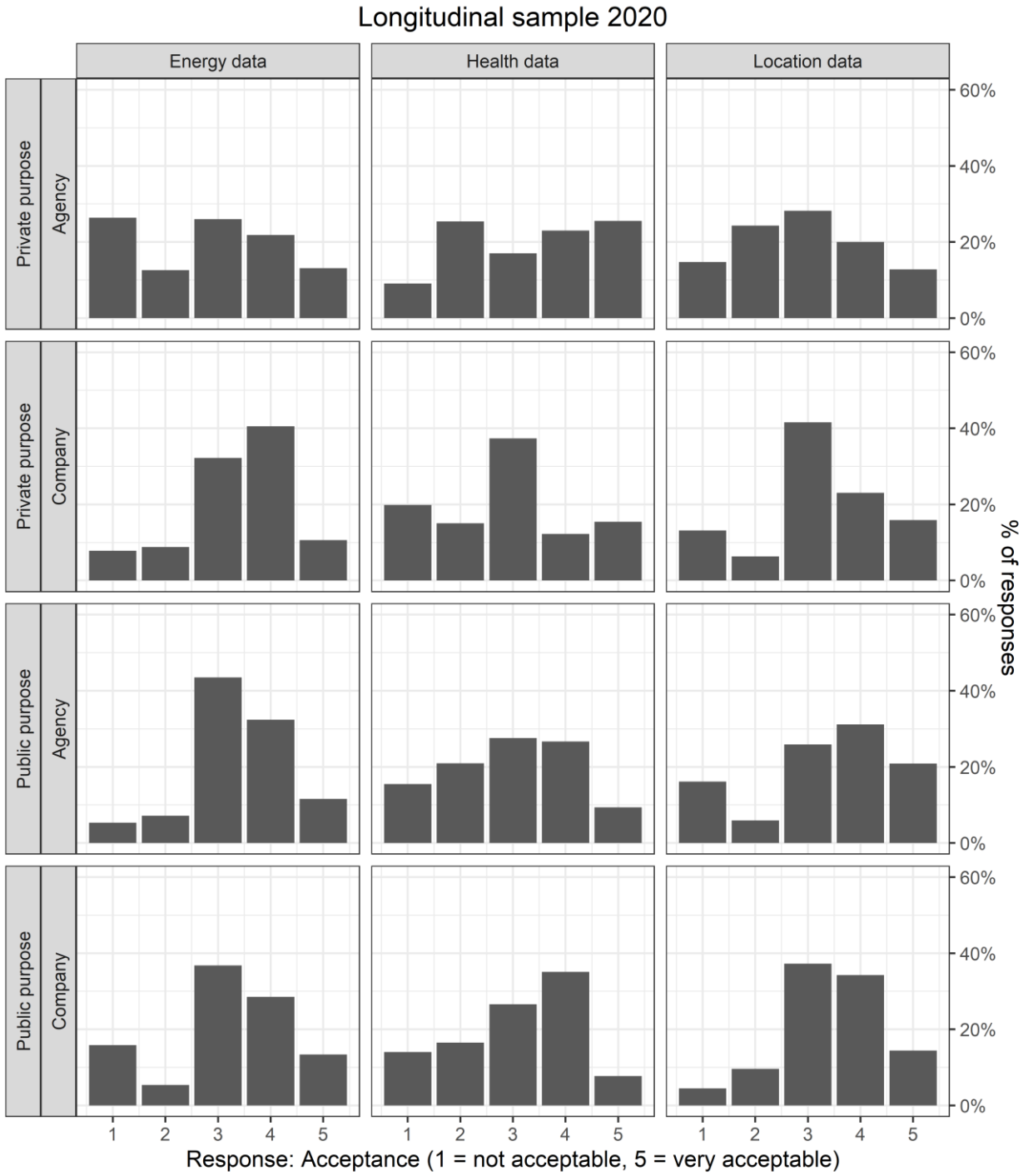
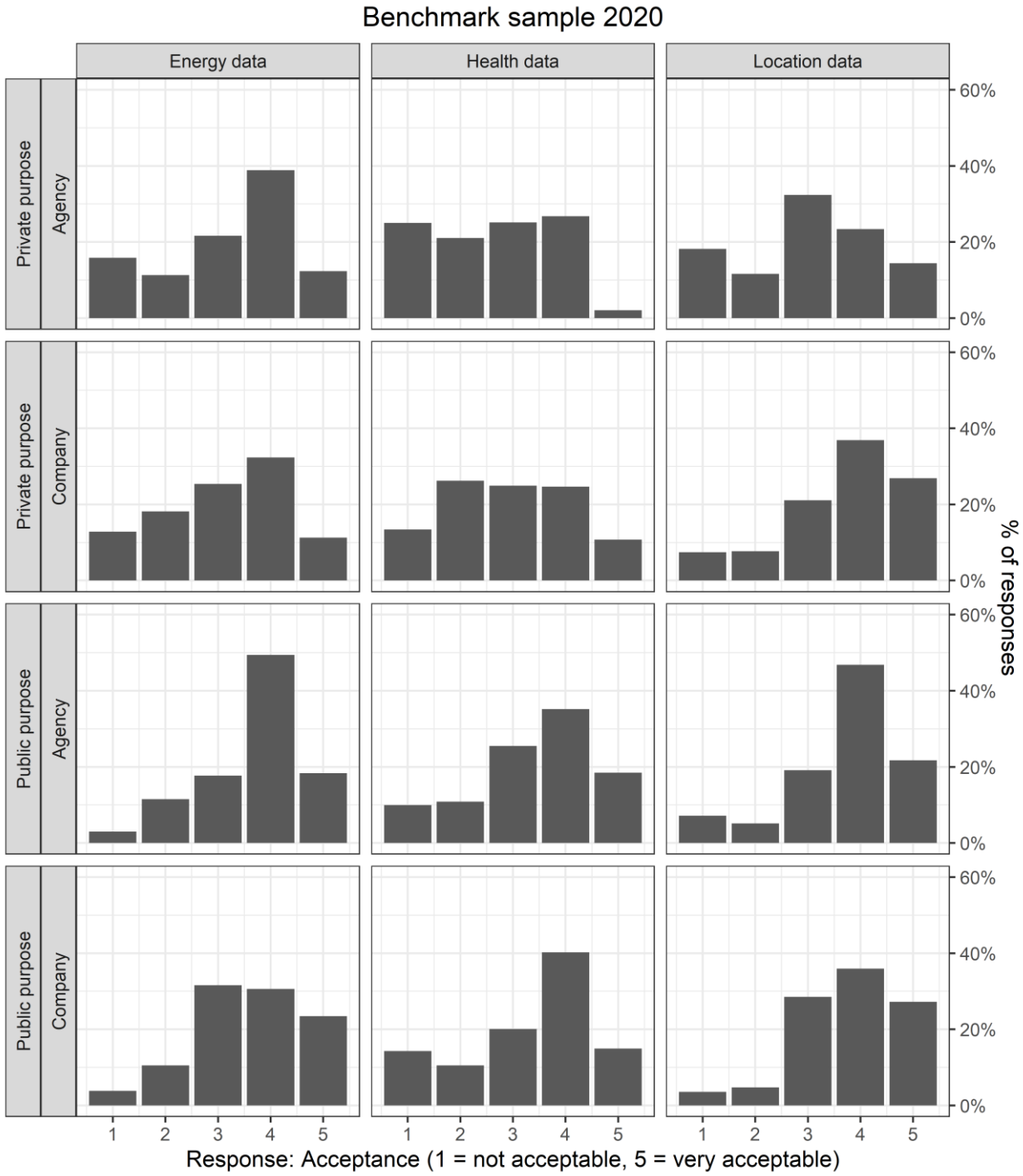


Figure A4.2e. Distribution of responses in the benchmark sample 2020. Weighted analysis.



7.4. Appendix for Chapter 5

Section M

Details on methods, questionnaire, and sample

In the following, I provide further information on the questionnaire, pre-testing, and sample.

On the introduction page of the survey, respondents were asked for consent. After informed consent, respondents were asked to agree to carefully read and answer the questions in order to avoid speeding (Conrad et al., 2017). However, respondents could proceed without agreeing to carefully reading the questionnaire. Respondents were then asked about gender, age, and whether the respondent currently resided in the respective country. These three questions were mandatory, and respondents who indicated an age lower than 18 or who indicated that they did not currently reside in the respective country were screened out. After respective screenouts and quota checks, respondents were presented with the vignettes and subsequently items on, e.g., privacy concerns and trust. All items, including information on randomization of items within pages, are available in the questionnaires in the Online Appendix (*Section Q*).

Different versions of several vignettes, the vignette question, and the answer categories were pre-tested via several cognitive interviews for the initial German questionnaire. The participants came from the social circles of the author and of a student assistant. All interviews focused on selected elements of the questionnaire and especially relied on the "think aloud" technique and sometimes on comprehension probing (see Lenzner et al., 2016). The feedback was used particularly to improve the understandability of the vignettes and the usefulness of the question and answer scale.

After finalizing the German questionnaire, automated translations into English and Spanish were produced. The Spanish questionnaire was carefully controlled and corrected by a professional translator and the author also consulted a Spanish-speaking survey expert. After the English version was checked by the author, an academic working at a UK university was consulted to check for potential problems. Quantitative pilot surveys were run in Germany (n = 237), Spain (n = 232), and UK (n = 235) (after removing two German and four UK answer sets from respondents that likely participated in the survey twice). The pilot included closed and open-ended probing questions (see Behr et al., 2017). One probing question asked whether respondents found that they could well express their personal opinion with the provided response options for the vignette-related question, to which about 94 percent of respondents

agreed. The pilots further included open-ended questions on the understandability of vignettes and asking for further comments. The responses did not reveal any frequent concerning issues.

The pilot experimentally varied whether respondents received vignettes always with or without the following statement: "This information does not contain the person's name or address." Respondents overall tended to rate vignettes with this statement more negatively than vignettes without this statement. To reduce the respondents' leeway in assumptions about the anonymity of the data, this statement was shown in all vignettes in the main studies.

In the following, I provide more detailed sample sizes and exclusion criteria for the main surveys in December 2022 and May 2023. In wave 1, 2,109 respondents (Germany: 702; Spain: 704; United Kingdom: 703) completed the survey. In wave 2, 2,225 respondents (Germany: 746; Spain: 742; United Kingdom: 737) completed the survey. Among the respondents in wave 2, 1,574 (Germany: 556; Spain: 541; United Kingdom: 477) already participated in wave 1.

Before running the analysis, I first exclude respondents that appear to have participated in the same survey wave twice (Wave 1: n = 6 in each Germany and UK, n = 4 in Spain; Wave 2: n = 10 in Germany and UK, n = 2 in Spain). Then, I exclude all respondents who did not indicate to commit to carefully reading and answering the questionnaire (Wave 1: n = 12; Wave 2: n = 7). Among the remaining respondents, I furthermore exclude all respondents that were defined as speeders (Wave 1: n = 360; Wave 2: n = 365) or for whom no duration information is available (Wave 1: n = 36; Wave 2: n = 34). As the speeding threshold I use a relative measure of 60% of the median response time (Roßmann, 2010) of those respondents for whom response times were available, while I define the medians and speeders separately for each country. Speeding was defined after removing respondents who appear to have participated more than once, but before the other exclusion steps. After these steps, I finally exclude five respondents who answered the question about gender with "Other" (Wave 1: n = 3; Wave 2: n = 2), as one cannot make valid inferences based on this small sample size.

The final sample sizes for analysis are as follows:

- Wave 1: 1,682 respondents (Germany: 562; Spain: 564; United Kingdom: 556)
- Wave 2: 1,795 respondents (Germany: 594; Spain: 603; United Kingdom: 598)

Note that I exclude further cases for the regression analyses (see Section *Results*).

Wave 2 of the survey comprises respondents who already participated in wave 1, as well as newly recruited respondents to compensate for drop-outs. An additional longitudinal data set comprises only those respondents who participated in both waves. I apply the same exclusion criteria to the longitudinal data set as above: I first exclude 21 respondents who appeared to have participated twice within any of the single waves, and a further 163 respondents who fell

under at least one same exclusion criterion in both waves. Then, 280 respondents were removed from the longitudinal data set completely if they fell under any of the exclusion criteria in any of the waves. This results in a longitudinal data set with 8,880 responses from 1,110 respondents (Germany: n = 382; Spain: n = 389; UK = 339).

References

Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2017). Web probing – implementing probing techniques from cognitive, interviewing in web surveys with the goal to assess the validity of survey questions. *GESIS Survey Guidelines*. https://doi.org/10.15465/GESIS-SG_EN_023

Conrad, F., Tourangeau, R., Couper, M., & Zhang, C. (2017). Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods*, *11*(1), 45–61. https://doi.org/10.18148/SRM/2017.V11I1.6304

Lenzner, T., Neuert, C., & Otto, W. (2016). Cognitive Pretesting. *GESIS Survey Guidelines*. https://doi.org/10.15465/GESIS-SG_EN_010

Roßmann, J. (2010). *Data quality in web surveys of the German longitudinal election study 2009*. Presentation at the 3rd ECPR Graduate Conference, Dublin, Ireland.

Section F

Additional figures

Figure A5.1a. Distribution of responses to vignettes in wave 1.

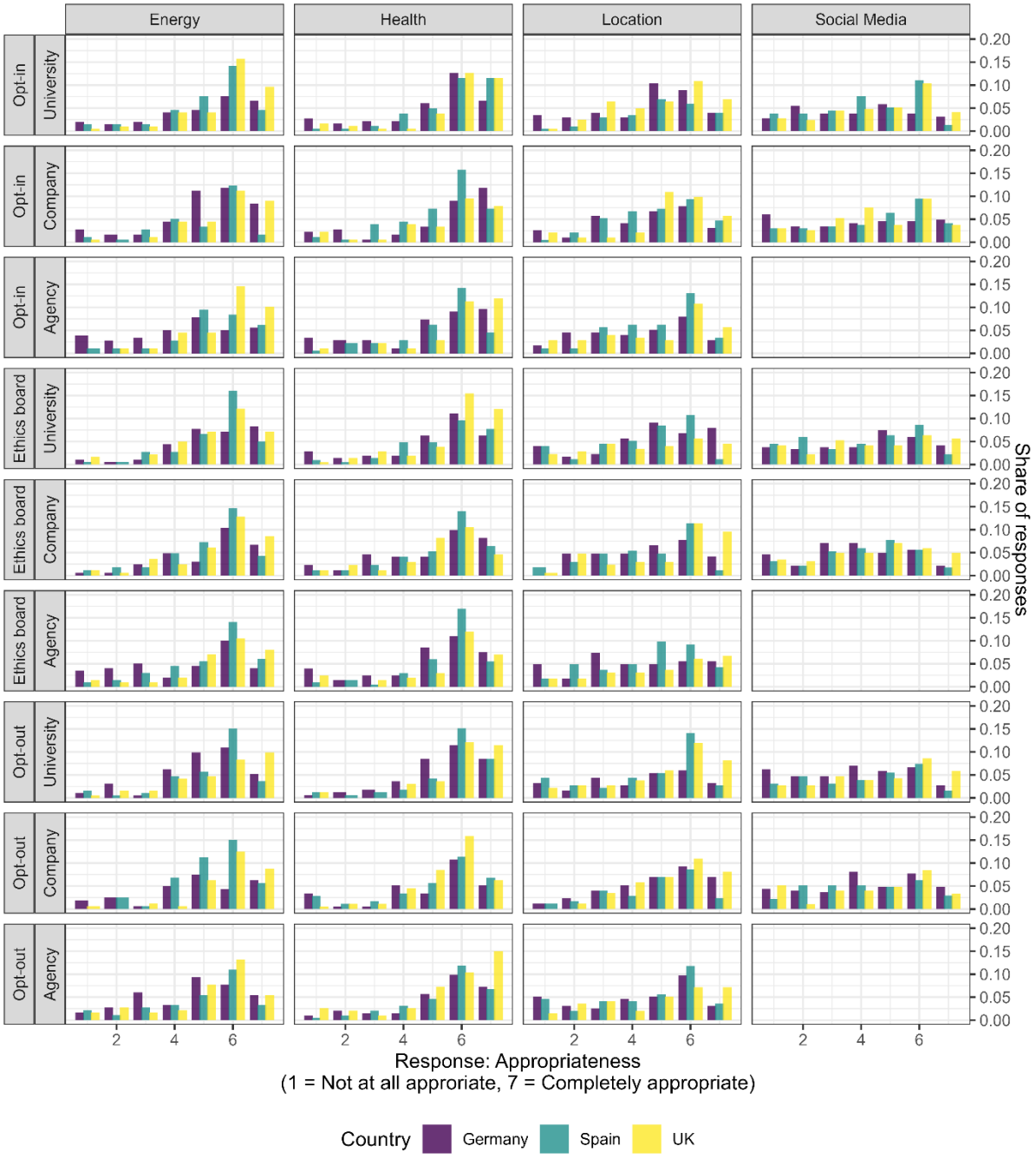


Figure A5.1b. Distribution of responses to vignettes in wave 2.

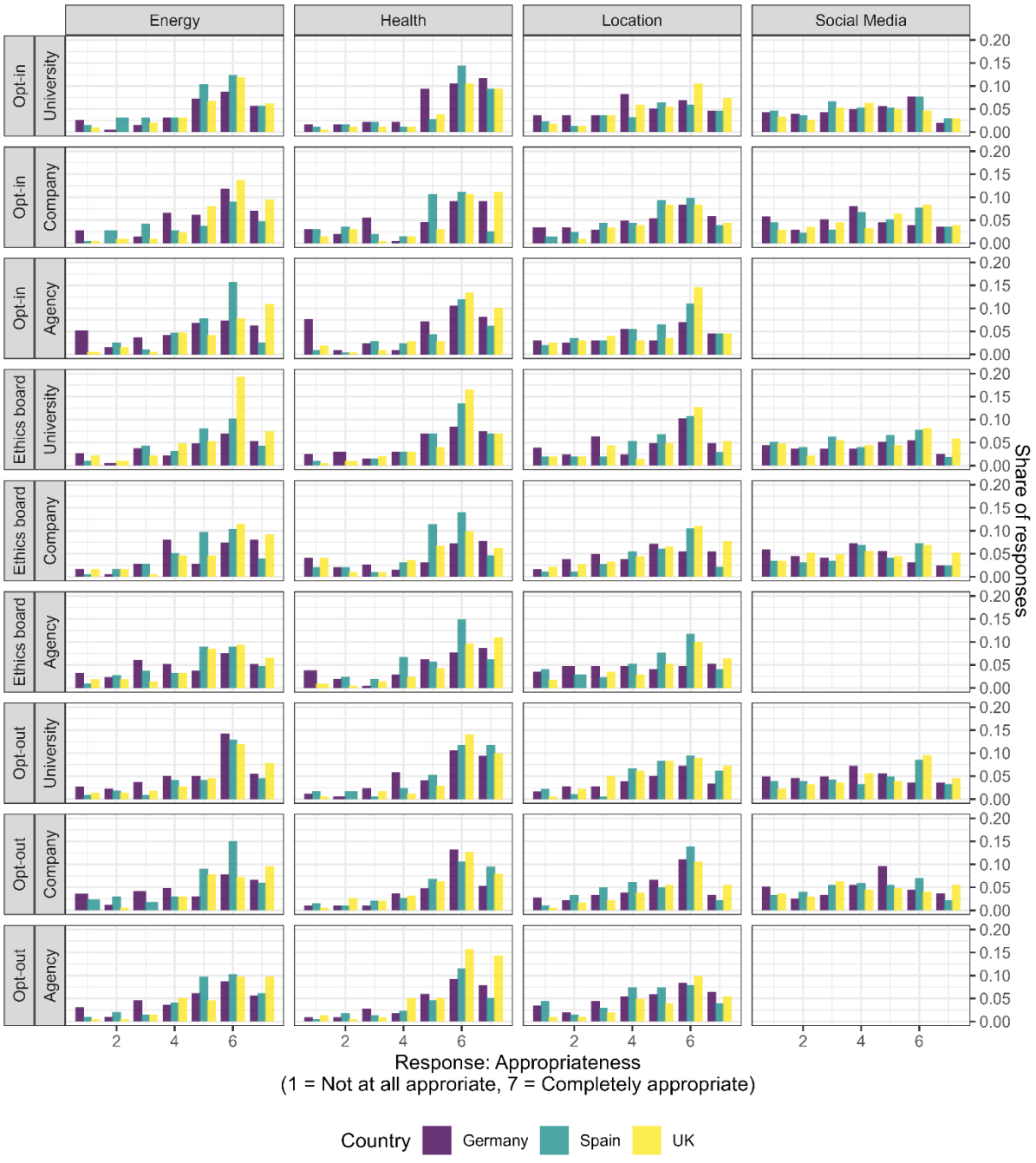


Figure A5.2. Arithmetic mean values of perceived appropriateness of all vignette scenarios in wave 2 (May 2023). Each column represents one data recipient, each row one transmission principle. Each box shows the arithmetic mean values for each data type and for each country. Number of responses per country: Germany: 2,376; Spain: 2,412; UK: 2,392.

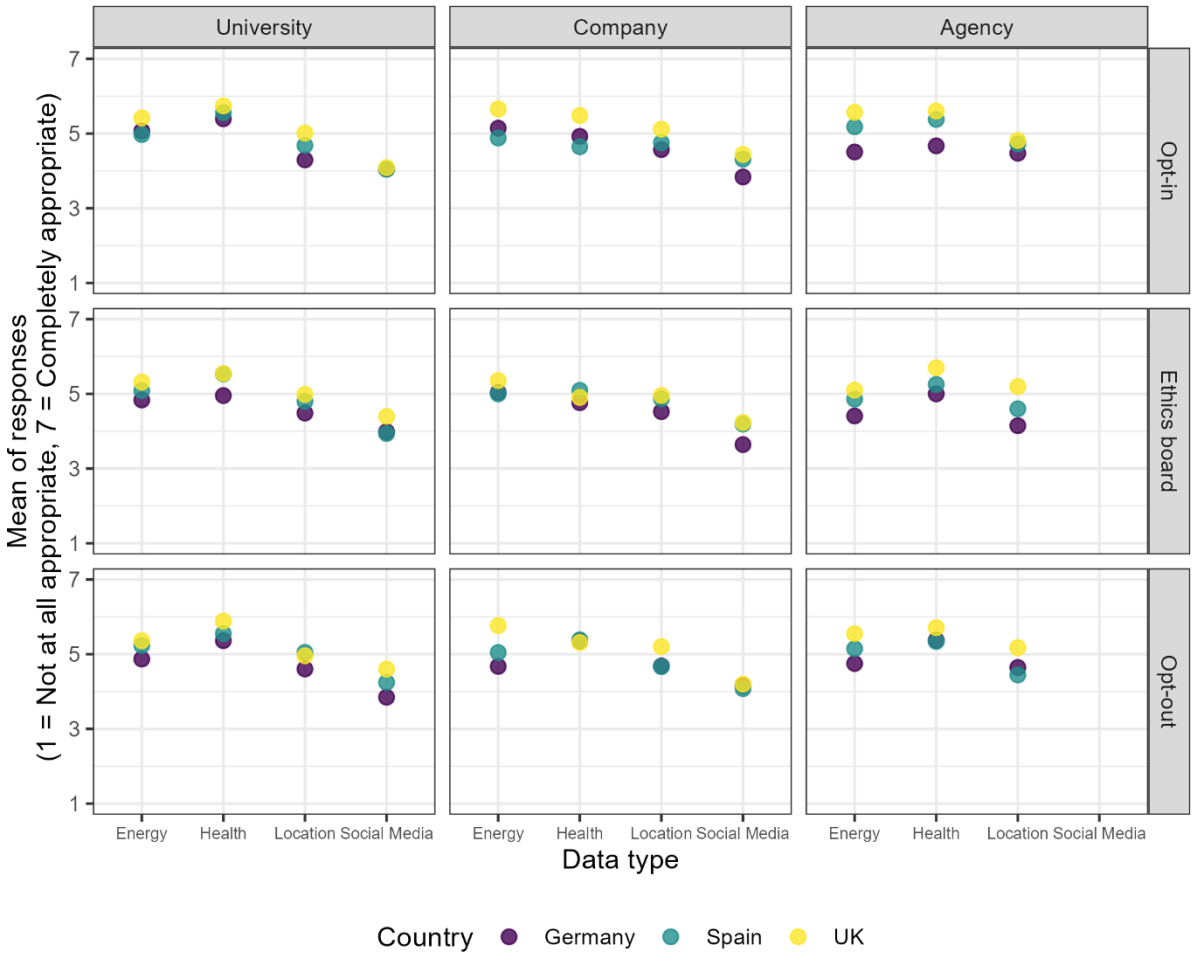


Figure A5.3. Changes in arithmetic means of responses from wave 1 (December 2022) to wave 2 (May 2023) among those respondents who participated in both waves (including speeders). Aggregated for country, data type, recipient, or transmission principle. Based on 8,880 responses from 1,110 respondents who participated in both waves.

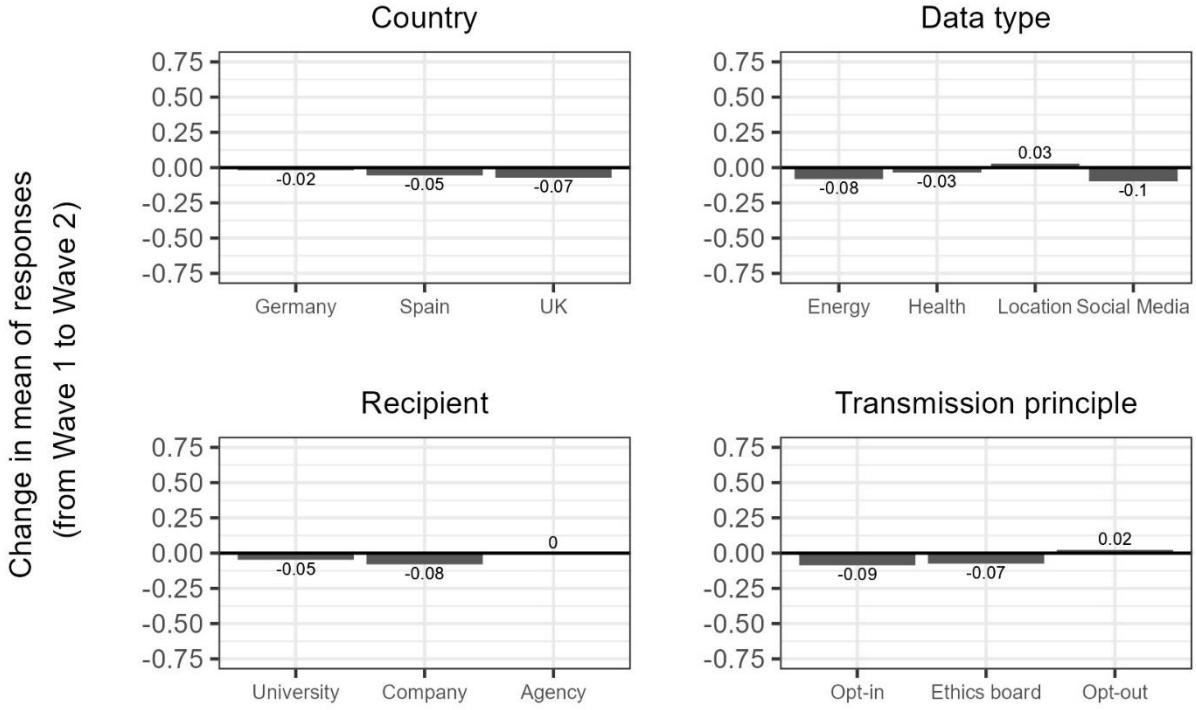


Figure A5.4. Changes in arithmetic means of responses from wave 1 to wave 2 among those respondents who participated in both waves (including speeders). Differentiated by country, data type, and data recipient. Based on 12,328 responses from 1,541 respondents who participated in both waves.

Figure A5.5a. Changes in arithmetic means of responses from wave 1 to wave 2 among those respondents who participated in both waves (only non-speeders). Differentiated by country, data type, and data recipient, and transmission principle. Based on 8,880 responses from 1,110 respondents who participated in both waves.

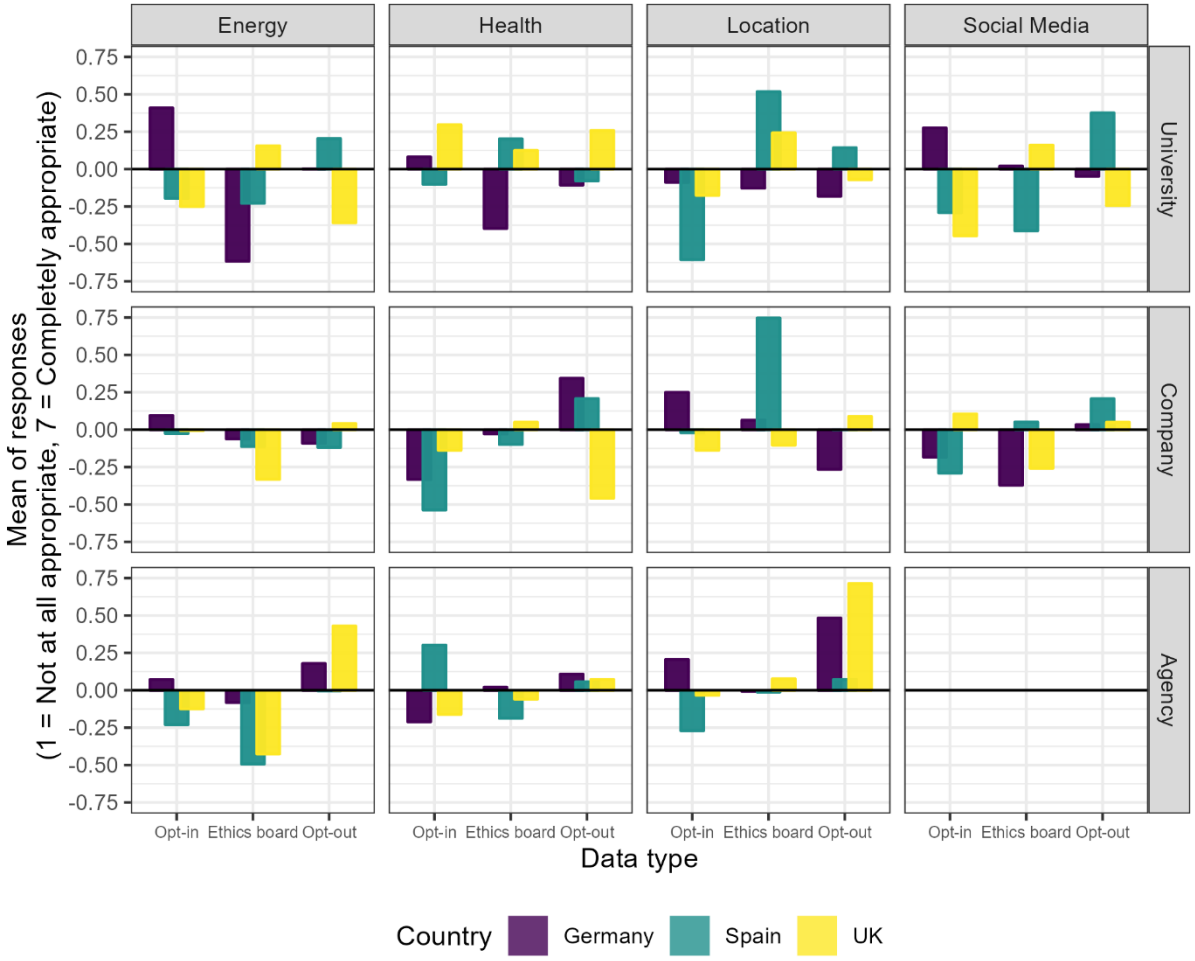


Figure A5.5b. Changes in arithmetic means of responses from wave 1 to wave 2 among those respondents who participated in both waves (including speeders). Differentiated by country, data type, and data recipient, and transmission principle. Based on 12,328 responses from 1,541 respondents who participated in both waves.