Minds and Machines (2024) 34:24
https://doi.org/10.1007/s11023-024-09684-y

"The Human Must Remain the Central Focus": Subjective Fairness Perceptions in Automated Decision-Making

Daria Szafran[1,2] · **Ruben L. Bach**[2]

Received: 31 May 2023 / Accepted: 11 June 2024

Abstract
The increasing use of algorithms in allocating resources and services in both private industry and public administration has sparked discussions about their consequences for inequality and fairness in contemporary societies. Previous research has shown that the use of automated decision-making (ADM) tools in high-stakes scenarios like the legal justice system might lead to adverse societal outcomes, such as systematic discrimination. Scholars have since proposed a variety of metrics to counteract and mitigate biases in ADM processes. While these metrics focus on technical fairness notions, they do not consider how members of the public, as most affected subjects by algorithmic decisions, perceive fairness in ADM. To shed light on subjective fairness perceptions of individuals, this study analyzes individuals' answers to open-ended fairness questions about hypothetical ADM scenarios that were embedded in the German Internet Panel (Wave 54, July 2021), a probability-based longitudinal online survey. Respondents evaluated the fairness of vignettes describing the use of ADM tools across different contexts. Subsequently, they explained their fairness evaluation providing a textual answer. Using qualitative content analysis, we inductively coded those answers (N=3697). Based on their individual understanding of fairness, respondents addressed a wide range of aspects related to fairness in ADM which is reflected in the 23 codes we identified. We subsumed those codes under four overarching themes: *Human elements in decision-making*, *Shortcomings of the data*, *Social impact of AI*, and *Properties of AI*. Our codes and themes provide a valuable resource for understanding which factors influence public fairness perceptions about ADM.

Keywords Fairness · Subjective fairness perceptions · AI · Automated decision-making · Qualitative content analysis

Extended author information available on the last page of the article

Published online: 19 June 2024

1 Introduction

The rapid advances in AI and machine learning (ML) technology and the progressing digitization of modern societies have facilitated the automation of many tasks previously performed by humans (AlgorithmWatch, 2019). Both private companies and government institutions are increasingly implementing automated decision-making (ADM) tools to support or even completely replace human deciders in consequential decision-making processes in contexts such as welfare benefits transfers (Jørgensen, 2023), healthcare (Grote & Berens, 2020), and human resources (Leicht-Deobald et al., 2019).

The deployment of such tools is almost always motivated by their objectives and benefits – to make decision-making faster, more efficient, and less prone to human error (Jørgensen, 2023; Rinta-Kahila et al., 2022). However, incidents involving ADM reveal their potential to cause adverse social outcomes. Prominent examples include the Dutch government's algorithm wrongly accusing citizens of welfare fraud and denying them access to daycare benefits (Peeters & Widlak, 2023), or the Australian *Robodebt* grossly over-assessing citizens' debts causing them severe psychological distress (Rinta-Kahila et al., 2022). These instances, among others, have sparked a public debate about justice and fairness in the context of ADM, as some societal groups were disproportionately affected by the ADM systems than others. FairML scholars have since proposed a variety of metrics that aim to capture and understand bias in algorithm-driven decision processes and developed techniques to mitigate biases (Mehrabi et al., 2021; Verma & Rubin, 2018).

While formal fairness metrics and mitigation techniques are much needed, they do not necessarily include the perspective of those who are directly affected by automated decision-making, namely the citizens. Ultimately, most of a decision's impact is carried by citizens with ADM potentially having life-altering consequences. Therefore, it is imperative to explore how these stakeholders perceive fairness in the context of ADM to extend the focus of current research beyond technical fairness discussions and definitions as presented in the fairML literature.

This study sheds light on subjective fairness perceptions of citizens surrounding the topic of ADM. That is, we contribute to existing research that takes a human-centric stance towards fairness in ADM (Helberger et al., 2020; Yurrita et al., 2023). We do so by answering the following research question: *Which subjective aspects do individuals consider when assessing the fairness of a decision-making process guided by AI and ML?*

For this purpose, we explore 3697 textual responses to an open-ended survey question implemented in a 2021 German probability-based online panel using qualitative content analysis (Hsieh & Shannon, 2005). Responses come from participants who were asked to evaluate the fairness of several vignette scenarios describing the use of an AI-based decision-making process and to explain their evaluation in their own words.

Our qualitative approach allows us to immerse in subjective fairness perceptions of individuals, resulting in 23 inductively derived codes. The codes highlight a variety of factors influencing fairness in ADM that can be subsumed under four themes:

the *human aspect* of decision-making, the *data* used in the process, the *social impact* of AI, and the *properties* of algorithmic tools.

Future research may rely on our codes and themes to construct a framework of public fairness perceptions in the context of ADM. Such a framework may then become a useful resource to guide stakeholders involved in the topic of ADM about individuals' potential real-world perceptions of fairness. Our findings may help them consider individuals' perspectives not only during or after the implementation of algorithmic tools but also prompt important questions that need to be answered well *before* such tools are conceptualized and developed.

The remainder of this paper is structured as follows: first, we provide a brief background for our study; next, we describe the data, the analytical approach, and each step of our analysis; we then present the inductively derived codes before turning to the discussion of our findings.

2 Background and Related Literature

We begin this section with an introduction of the social construction of technology approach, which holds that technology is always shaped by the social contexts it is embedded in (Bijker, 2010). It motivates our perspective of focusing on individuals' perceptions of fairness in ADM systems as considering non-technical factors is crucial to understand how technology is made and how it works. We then contextualize our paper in related work by briefly reviewing relevant empirical and theoretical contributions. Specifically, we introduce algorithmic fairness as discussed in the fairML literature. This literature's focus lies on technical aspects of measuring fairness in ML-based prediction models as well as on developing solutions to mitigate biases in ML models to prevent unfair predictions and decisions resulting from the models. We then review selected findings on fairness perceptions by stakeholders of ADM systems, such as users and affected individuals.

2.1 Social Construction of Technology

The perspective of the social construction of technology (SCOT), a theory from science and technology studies, holds that technology "does not have its own intrinsic logic but is socially shaped" (Bijker, 2010:66). By taking a critical stance towards technological determinism, this perspective approaches technical artifacts and systems as constructs that are inevitably intertwined with the social. Following this notion, considering non-technical factors is crucial to understand how technology is made and how it works (Bijker, 2010). For example, human-robot interactions will largely depend on how designers of such technologies perceive and evaluate their end users (Burema, 2022). Furthermore, the understanding of technology is flexible and likely to vary across social groups (Bijker, 2010). Taking the use of AI tools in education as an example, Eynon and Young (2021) show that there is little similarity in how stakeholders from government, industry, and academia conceptualize and perceive such tools. From both, the social embeddedness of technological artifacts and their "interpretative flexibility", arise important implications for the design of vari-

ous technologies (Williams & Edge, 1996). Recent research on ADM in the context of public employment services demonstrates that conceptions of algorithmic tools as imagined by jobseekers (Scott et al., 2022) and members of non-profit organizations supporting jobseekers (Wang et al., 2023) go far beyond statistical predictions of specific outcomes. For these groups, the *practical* aspect of such tools proves the most helpful.

Discussing fairness in ADM from the constructionist perspective thus inevitably leads to questioning the status quo of such technologies. McCarthy elaborates critically on technological determinism without rejecting it (McCarthy, 2013). He argues that especially complex technologies have a "biased but ambivalent" character meaning that there exist determining forces behind them without denying technologies' "ever-present potential for change" (McCarthy, 2013:477). This ambivalence manifests itself in the asymmetrical power relations between social groups. While some human agents indeed have "the ability to control technological design and development" (McCarthy, 2013:476), others, that is, less powerful, are merely exposed to those developments without having agency. ADM systems are by no means an exception. The intention behind their deployment as well as their functioning are likely to reflect the interests of agents who are significantly engaged in their development. The examples of heavily biased ADM tools mentioned previously seem to support this assumption. Regarding fairness, the question arises to what extent ADM tools can be evaluated as fair by members of the public when the fairness criteria the tools operate on are defined by a group of selected people.

Next, we turn our focus to the fairML literature that takes a step towards mitigating the adverse outcomes and thus towards increasing fairness in algorithmic decision-making by embedding established fairness notions in the underlying functionality of such tools.

2.2 Fairness in the FairML Literature

The so-called fairML literature has been concerned with understanding, measuring, and mitigating biases in ML-based prediction tasks that may eventually result in unfair decisions (Mehrabi et al., 2021). In this literature, fairness is primarily treated as a technical feature of a machine learning model. As a result, unfair models can also be corrected, i.e., made fair, through technical approaches (Mehrabi et al., 2021).

A popular example illustrating such algorithmic fairness is the COMPAS algorithm used in the U.S. criminal justice system. Designed to help judges decide whether a defendant should be detained or released on bail while awaiting trial, the system has been shown to discriminate against Black defendants (Angwin et al., 2016). COMPAS is based on a statistical model that predicts a defendant's likelihood of being rearrested for a new crime while awaiting trial for the first crime. Briefly speaking, defendants with a low score are recommended for bail while those with a high score are to be detained. The algorithm underlying the COMPAS system must not include race as a predictor due to US anti-discrimination legislation. Still, Angwin et al. (2016) showed that Black defendants are more likely than White defendants to be incorrectly assessed as high risk, while White defendants were more likely than Black defendants to be incorrectly labeled as low risk. The reason why the tool

assigns Black individuals higher scores than White individuals is likely the disproportionately high number of Black individuals in jail, often for minor crimes. Since the system is trained with historical data, such historical discrimination against Black people in the U.S. justice system is picked up by the profiling tool.

To formalize such fairness violations, various fairness metrics have been proposed, which can be broadly categorized into three groups (Gajane & Pechenizkiy, 2018). The first set of fairness notions focuses on the equality of treatment, under which an algorithm may be considered fair if protected attributes are not explicitly used in the model-building process (*Fairness through Unawareness*; e.g., Grgić-Hlača et al., 2016). Note, however, that this (naive) approach fails to mitigate discrimination when legitimate factors are correlated with illegitimate factors, such as location and ethnicity in the context of racial segregation (Pedreshi et al., 2008). A second set of fairness notions focuses on the equality of outcomes, i.e., the fairness of predictions that are obtained from an algorithm. Examples in this context include parity-based definitions, which state that an algorithm can be considered fair if members of protected groups have the same probability of being assigned to a positive outcome as members of unprotected groups (*Demographic Parity;* Dwork et al., 2012), given a set of legitimate factors (*Conditional Statistical Parity;* Corbett-Davies et al., 2017). The third group of fairness notions can be characterized as focusing on the equality of errors that are made by the prediction model. These fairness notions are largely centered around different types of errors that can arise in classification settings when comparing the number of false negatives (individuals who are falsely predicted to fall into the negative class) and false positives (individuals who are falsely predicted to fall into the positive class) to different baselines (Rodolfa et al., 2019). According to this notion, 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;* Chouldechova, 2016) between members of protected and unprotected groups. This principle can be applied to various error metrics and their combinations (e.g., false discovery rates, false omission rates, accuracy, ROC-AUC). This research has resulted in a plethora of criteria: In a review of the literature, Verma and Rubin (2018) list 20 definitions of (algorithmic) fairness.

Given the number of different fairness notions that may be used to evaluate a prediction algorithm, it is unclear which definition is appropriate for a given type of ADM application. It is particularly worth noting that different fairness definitions have been shown to conflict with each other, i.e., many fairness objectives cannot be achieved at the same time (Berk et al., 2018; Friedler et al., 2016). Likewise, it is unclear whether these algorithmic fairness approaches match individuals' evaluations of fairness in ADM systems.

2.3 Fairness Perceptions

A second stream of literature focuses on *perceptions of fairness by stakeholders*, such as individuals affected by ADM decisions and those using ADM tools in their decision-making. This literature acknowledges that there may be a mismatch between (technical) perception of algorithmic fairness in the fairML literature and *perceptions* of fairness by stakeholders. So far, however, it is much less developed than

the fairML literature. In a recent review, Starke et al. (2022), for example, identified only 58 papers that provide empirical insights on perceptions of algorithmic fairness. According to this review, current literature can be organized along four key dimensions that shape perceptions of fairness. The first dimension are papers that deal with algorithmic fairness as introduced in subsection 2.2. That is, several studies investigate whether the ways that an ADM system is designed affects people's perceptions of fairness. In these studies, algorithmic fairness is a crucial factor when evaluating algorithms, as, for example, systems with low algorithmic fairness tend to be perceived as less fair. The second dimension is concerned with individual predictors of perceptions of fairness, such as sociodemographic variables, individuals' self-interest, and familiarity with data algorithms. The third dimension focuses on comparative effects, that is, differences between human decision-making and algorithmic decision-making. However, no clear pattern across studies can be identified. While some studies report that ADM systems are perceived to be fairer than solely human judgment, others find the opposite. The fourth dimension focuses on the consequences of perceived fairness of algorithmic decision-making systems. For example, perceived fairness seems to impact trust in and satisfaction with algorithms and systems implemented perceived to be unfair in the workplace environment may negatively impact outcomes such as organizational commitment.

A few papers use empirical approaches like ours, that is, closed-ended survey questions to measure fairness evaluations in combination with open-ended survey questions to allow for a more nuanced understanding of users' fairness perceptions. These previous papers contextualize ADM across various settings related to employment, medicine, or banking.

Bankins et al. (2022) presented vignette scenarios to 446 North American respondents describing a fictitious company that offers an extensive training course to its employees. Due to the limited number of participants, either an algorithm or a human manager had to screen the eligible employees. The survey respondents were asked to place themselves as the company's employees. Subsequently, they were asked to provide a textual response on why they felt treated (dis-)respected in this situation as a measure for interactional justice. The study's findings are nuanced, but the authors conclude that, overall, respondents felt more respected when a human manager was involved in the decision-making process. While some respondents highlighted the allegedly objective character of algorithms, the textual responses reveal that others described it as inappropriate or lacking the necessary emotional intelligence. Next, Bedemariam and Wessel (2023) surveyed 282 North American respondents via Amazon MTurk. The survey was designed to simulate a typical job application process for an online task for which respondents would be paid money. After having applied, the respondents were either accepted for the job or rejected, a decision that was made either by an employee in the human resources department or an AI system. Respondents were then asked to evaluate the decision based on measures of procedural and interactional justice and provide more detail on their evaluation via a textual response. Overall, the human decision-maker was favored over the AI system. Respondents described the human as having the ability to recognize characteristics that are not detectable to AI. Further, the AI system was seen as only executing instructions according to its programming. Interestingly, despite only small differ-

ences between Black and White respondents, both aspects regarding the human and algorithmic decision-maker were mentioned more frequently by Black respondents. Juijn et al. (2023) studied 225 responses by individuals who were presented with hypothetical scenarios involving recruitment algorithms. The respondents were then asked how fair they think the algorithm is and they were prompted to provide a textual response about the most relevant reasons behind their evaluation. To ensure that respondents had a constant reference point, the authors provided them with a fairness definition according to Mehrabi et al. (2021). Comparing the textual responses to mathematical fairness definitions, Juijn et al. (2023), found that the respondents' reasoning corresponded mostly to distributive (focused on the outcome) and procedural fairness (focused on the decision-making process).

Two studies focused on decisions regarding loans. Schoeffer et al. (2021) conducted a survey among 196 respondents who were presented with hypothetical scenarios in which a person was denied access to a loan. The decision was made either by an ADM tool or a human decider. While respondents highlighted the fairness of the algorithmic tool due to its ability to treat individuals equally and objectively, they also mentioned the fact that algorithms are being programmed by humans who have an impact on how a computational tool operates. Overall, the authors conclude that "automated decisions are perceived as more informationally fair than human-made decisions" (Schoeffer et al., 2021:6). Similarly, Yurrita et al. (2023), using an online convenience sample of 267 English-speaking individuals, show that explanations, human oversight, and contestability affect loan approval decisions involving AI systems in high- and low-stakes contexts and users' fairness perceptions of these systems. A qualitative analysis of responses to open-ended survey questions revealed three main areas of tension, namely perceptions of informational fairness and two aspects related to perceptions of procedural fairness.

In a study by Formosa et al. (2022), 478 North American respondents were asked to assess the fairness of hypothetical scenarios describing either a human or an algorithmic tool making decisions about resource allocation in the medical sector (e.g., organ transplantation) or disease diagnosis. In all scenarios, respondents were asked to place themselves as the affected patient. After having evaluated the fairness of the decision, they were prompted to provide more detail on why they thought it was fair or not. In general, the findings indicate that humans were seen as more appropriate decision-makers across all vignettes. Moreover, the algorithmic tool was reported to cause experiences of dehumanization, which highlights the importance of interpersonal contact in medical settings.

Finally, two studies did not contextualize the algorithmic tool but rather asked respondents more broadly about their fairness perceptions. Helberger et al. (2020) asked 958 respondents in a Dutch adult population survey to rate whether an AI or a human would make a fairer decision. Overall, they find that AI systems are perceived to be fairer than human decision-makers. In addition, the analysis of the textual responses where respondents elaborate on their fairness judgments showed that emotions, expectations about the data, calculations that are part of an AI system, and the role of human AI designers are important aspects of (un)fairness perceptions towards AI or humans. Van Nuenen et al. (2022) surveyed 663 respondents with a disadvantaged or marginalized background (in terms of e.g., gender, ethnicity, or

disability status) about their experiences of being treated unfairly by an automated computational system. The respondents identified more than 20 types of unfair treatment by various systems. Among the most prominent types was the experience of being denied access to resources, such as jobs and being forced into an inaccurate category by the computational system, which might have serious consequences for decision-making. Moreover, respondents frequently mentioned the systems' opacity that contributes to the experience of unfair treatment.

The reviewed literature on technical fairness metrics and subjective fairness perceptions implies that reaching the goal of designing fair and acceptable algorithms requires a comprehensive consideration of all stakeholders involved. Beyond that, from the perspective of the social construction of technology, ADM tools can be seen as what they are, namely socially embedded artifacts. By taking a step back, this perspective allows us to fundamentally rethink the properties and objectives of ADM tools so that they conform to a broader, perhaps *social*, understanding of fairness. To tackle this empirically, we describe our data in the next section.

3 Data

The data analyzed in this paper were collected in July 2021 as part of the 54th wave of the German Internet Panel (GIP; Blom et al., 2015; Blom et al., 2021), a probability-based longitudinal online survey. The GIP runs several times a year among a nationwide representative sample of the German population aged between 16 and 75 years. Regarding the distribution of sociodemographic characteristics in our data, 51.8%[1] of the respondents reported to be male and 48.2% to be female. 31.3% were 62 years old or older, 33% between 47 and 61 years old. Approximately 29.7% of the sample were between 27 and 46 years old, while 6% were 26 years old or younger. Most of the sample (51%) had a high educational level (ISCED 5–8), followed by 45% with a medium educational level (ISCED 3–4). The remaining 4% of the sample reported a low educational level (ISCED 1–2).

The 54th wave had a section dedicated to ADM. Respondents were presented with various vignettes describing hypothetical scenarios that depicted the use of a computer program developed to make decisions based on individuals' data across four different contexts: (i) banking, (ii) criminal justice, (iii) hiring, and (iv) unemployment. Moreover, the content of the vignettes varied concerning three dimensions. First, they described either assistive or punitive actions, e.g., hiring new employees or terminating employees' contracts after the probation period, respectively[2]. Second, the personal data used for decision-making could be either solely related to the social context of the decision-making process or additionally coming from the internet. Third, the vignettes varied regarding the degree of human leeway and the involvement of the computer program in making a decision. The computer program, introduced as processing individuals' data, was described as either making a decision autonomously, i.e., without human leeway, as recommending a decision to the human

[1] Percentages regarding gender, age, and education are based on complete responses.

[2] Vignettes in the criminal justice context described solely assistive decision-making.

who would then make the final decision or simply as a computer program used to process individuals' data that the human decider evaluates to make the decision. The last case is somewhat ambiguous since it does not clarify how exactly the program or its output are to be used in making a decision.

The following text illustrates an exemplary vignette in the context of hiring describing assistive decision-making by a human being based on the program's recommendation using additional data available on the internet:

> A company has developed a computer program to hire new employees. This program uses data from the person's resume as well as publicly available information about the person from the internet. The program compares this information with that of other individuals already working in the company. The program gives a recommendation to an employee in the human resources department whether to hire the person. The employee determines whether the person will get hired."[3]

Every respondent was consecutively shown four vignettes, one randomly chosen from each context[4]. The order of the four contexts was also randomized within respondents. After having read a vignette, respondents were asked to evaluate the fairness of the hypothetical scenario on a four-point scale ranging from "Not at all fair" to "Very fair" (exact wording of the question (translated from German): "How fair do you find it is to make a decision in this way?"). Additionally, they were asked to provide an open-ended answer explaining their fairness evaluation on the first of the presented vignettes (exact wording of the question (translated from German): "Why do you find this way of making a decision [not at all fair/not very fair/somewhat fair/very fair]?". Here, the respondent's respective answer from the previous question regarding the fairness evaluation was included into the question wording)[5]. The authors of the survey deliberately refrained from referencing an established definition of fairness, allowing respondents to answer the questions based on their subjective understanding of fairness (Kern et al., 2022). In total, 4090 respondents participated in the survey. After discarding invalid responses (open-ended answers shorter than five characters were found to be meaningless upon manual inspection, including missing data), our final dataset consists of 3697 open-ended responses[6].

Note that the original survey collected additional information, such as sociodemographic information and attitudes and behaviors, which we do not consider further in our study (see the complete questionnaire of the survey at https://search.gesis.org/research_data/ZA7762). Given our study's qualitative approach and its focus on the

[3] See Online Resource 1 in the Supplementary Information for all 42 vignettes (originally in German, translated to English by the authors).

[4] For a distribution of the 3404 respondents across scenarios and vignette dimensions see Online Resource 2.

[5] The vignette and the question regarding the fairness evaluation were presented to the respondents simultaneously. The open-ended question about the evaluation's explanation was shown on a subsequent page of the survey. Here, the vignette was no longer displayed, however, respondents had the option to jump to the previous page by clicking on a "Back" button.

[6] 960 responses in the banking scenario, 953 in hiring, 906 in criminal justice, and 878 in unemployment.

textual responses, we do not consider any other information collected in the survey except for the closed-ended fairness rating (see subsection 6.2). Future research may extend our approach to consider information from the additional survey items using, e.g., a mixed-methods approach that combines our qualitative analysis with statistical analysis of the remaining survey data.

4 Analytical Approach

We use conventional content analysis as described by Hsieh and Shannon (2005) to analyze the open-ended responses. This approach is especially useful when the literature about the phenomenon under study is scarce. Rather than using predetermined categories, it allows for an inductive development of categories that emerge directly from the available data. There are two main reasons why we chose this approach: first, we followed the notion applied in the survey of abstaining from exposing the respondents to specific fairness conceptions to obtain unprimed and unlimited real-world evaluations of ADM (Kern et al., 2022). Second, we aim to provide an understanding of subjective fairness perceptions that is grounded in our actual data. The conventional content analysis (Hsieh & Shannon, 2005) is widely used in social science and public health research to analyze qualitative data such as open-ended survey responses (Foulkes et al., 2021; Liem, 2019; Munro et al., 2022), interviews (Berg et al., 2023; Guenna Holmgren et al., 2022; Spreckley et al., 2022), or newspaper articles (Cengiz & Eklund Karlsson, 2021).

5 Analysis

Data analysis was conducted between February 10 and May 24, 2023, under the lead of the first author using Microsoft Excel (Version 16.67) for data management and coding. In the first step, all responses were read in their entirety to gain a comprehensive picture of the data. During this process, the first patterns in the data started to emerge. Then, the responses were re-read and the majority were already coded. A part of the responses was initially left uncoded due to uncertainty. Simultaneously, we developed the coding scheme which was gradually expanded by adding new codes that emerged from the data. Next, we read through all the responses assigned to each code to ensure that they were consistent with the code. During this process, the coding scheme was finalized by refining the code definitions and selecting example quotes which were translated from German to English. Lastly, using the coding scheme, we coded the remaining open-ended responses. All steps of the analysis were discussed between both authors during regular meetings throughout the analysis process.

In summary, 3404 of the 3697 responses were manually coded into one or more codes (up to five codes per response). 293 responses could not be coded as they were either not interpretable (e.g., a single, vague word, such as “Feeling.”) or unrelated to the topic (e.g., “The employment agency is completely unsuitable for the placement of employees. In my opinion, it should only manage the statistics and let a private company handle the job placement.”). These responses were discarded from

the descriptive analysis shown in the next section. The descriptive analysis was conducted using R (Version 4.0.4, R Core Team, 2023).

6 Results

Individual responses varied in length, ranging from a minimum of one word to a maximum of 154 words and with a minimum of six to a maximum of 1024 characters (see Online Resource 3 for the distribution of response length). The average length of a response was 15.7 words, the median was 12; the mean number of characters was 113.7, and the median was 83 characters. Most responses were short with only a few cases being longer than 70 words. Previous research refers to the length and the interpretability of textual responses as an indicator of their quality (Grauenhorst et al., 2016; Mavletova, 2013; Schmidt et al., 2020). The response length in our study corresponds roughly to previous studies investigating response quality through response length (Meitinger et al., 2021; Schmidt et al., 2020). The relatively low number of both item-nonresponse (out of 3930 respondents who provided a valid fairness evaluation, 3697 answered the open-ended question) and not interpretable responses (293 of the 3697 textual responses were not interpretable) as well as the richness of identified aspects across all responses indicate overall high response quality.

6.1 Summary of Results

This study aims to explore respondents' subjective perceptions of fairness in the context of ADM. In total, we identified 23 individual codes, which were then grouped into four overarching themes. Respondents addressed a variety of aspects related to the role of *humans in decision-making*, the characteristics of *data* being used in the process of decision-making, the *social impact* that algorithmic decision-making might have, and the *properties* of algorithm-operated decision tools. The identified themes, codes as well as code descriptions, and example quotes are presented in Table 1. All code descriptions in this section concisely represent the content of the textual responses. We describe them in more detail in the next subsections.

6.1.1 Human elements in decision-making

Seven codes were grouped under the theme *Human elements in decision-making*. This theme was the most frequently represented one in the data with a total of 1886 assigned codes, which makes up approximately 43.2% of the total number of assigned codes.

6.1.1.1 Human individuality The first code refers to the individuality of a person. Respondents highlighted that each human, i.e., their personality, behavior, goals, needs, and desires, is unique. Therefore, individuals cannot be compared to each other or put into universal patterns. Further, no deductions about the behavior of one person can be drawn from the behavior of other people. Such calculated predictions (in the context of ADM) increase the risk of undeservedly punishing or rewarding

Table 1 Identified themes and codes, code descriptions, example quotes, and absolute and relative frequencies

Theme	Code	Short description	Example quote	*N* (%)
Human elements in decision-making	Human individuality	Each individual is unique, making it impossible to compare them or generalize their behavior.	"You can NOT compare apples with oranges. Every person is different and reacts differently." (R1499)	745 (17.1)
	Human decider	Only human beings are equipped with the necessary skills to make decisions; therefore, they should decide about other humans.	"Such a serious decision should not be dictated by an algorithm. Such cases cannot always be compared on the basis of pure numbers." (R3339)	513 (11.7)
	Interpersonal contact	Decision-making should involve an interpersonal dialogue in which all parties have the opportunity to speak.	"I think you also have to talk to these people personally, take into account their motivations, their view of things." (R1202)	388 (8.9)
	Human prejudice	Decisions made by humans may be influenced by their personal preferences.	"It cannot be ruled out that the deciding person would discriminate against me for "personal reasons"!" (R450)	96 (2.2)
	More than one decider	At least two individuals should be involved in the decision-making process to ensure objectivity.	"This decision should be made by a board of experts." (R1614)	68 (1.6)
	Autonomy of the affected person	ADM tools diminish individuals' perceived autonomy in shaping their own future.	"Since the customer has no influence on this, they are being externally controlled." (R820)	62 (1.4)
	Lack of trust towards AI	Algorithms and their recommendations or decisions are not trusted.	"I don't trust a computer program to make decisions about people. People, or more precisely experts, should make these decisions." (R1286)	14 (0.3)

Table 1 (continued)

Theme	Code	Short description	Example quote	*N* (%)
Shortcomings of the data	Limited data	The complexity of a human life might not be captured in data.	"A resume and data found on the Internet can represent only a small part of a person's complex personality." (R1596)	600 (13.7)
	Distorted data	Data might not truly reflect reality.	"The sources of internet data are not transparent. […] Unverifiable is fake information and information that maliciously gets posted online due to bullying." (R2315)	252 (5.8)
	Outdated data	Data might not reflect the current circumstances of a person.	"Also, it doesn't take into account that people can change their opinions and behaviors." (R898)	225 (5.2)
	Historical bias in AI	The reliance of algorithms on training data introduces the risk of perpetuating historical biases that discriminate against specific social groups.	"Prisons are institutions based on structural oppression, such as racism and classism. And this becomes visible, among other things, in the time people spend in prisons (who, why, how long, etc.). This decision would reinforce discriminatory structures." (R1067)	11 (0.3)

Table 1 (continued)

Theme	Code	Short description	Example quote	*N* (%)
Social impact of AI	Violation of privacy	The processing of personal data without individuals' explicit consent violates their privacy.	"I consider the involvement of the internet to be critical. Is that covered by data protection? Besides, the internet doesn't forget anything (e.g., youth sins that have nothing to do with work performance)." (R2307)	222 (5.1)
	Purpose of AI	As long as ADM tools serve a specific purpose, their deployment seems to be justified.	"The company looking for new employees has clear interests and formulated goals. In the search for "suitable" candidates a good program can help and support the selection." (R2281)	149 (3.4)
	Unequal access to resources	The algorithm's classification of individuals may systematically exclude certain individuals from the access to resources.	"People's life courses are ultimately very different and can only be compared to a limited extent. People who are not considered by the program may undeservedly fall through the cracks." (R1721)	134 (3.1)
	Meritocracy	Decisions about individuals should be based on their merits as this ensures that everyone receives what they deserve.	"If someone works poorly and slow, they can be terminated. It doesn't matter if the person is not up to the job and can't perform or doesn't try hard enough and falls below their capabilities." (R1949)	126 (2.9)
	Risk of being influenced by AI	Human deciders may unquestioningly adopt algorithm's recommendations.	"It is always said that the employee decides in the end, but in practice, recommendations from such programs are adopted without criticism due to time pressure." (R2906)	32 (0.7)
	Surveillance by AI	ADM tools might be used to monitor and control individuals.	"The basic attitude is distrust and control of a social life issue, the person is degraded and reduced to the "obedience principle". The person loses their individuality by a perverse equalization […]." (R3076)	25 (0.6)
	Loss of diversity	ADM tools tend to favor individuals with similar characteristics to the training data which compromises social diversity of groups.	"And a new employee, who perhaps doesn't quite fit into the "mold" of the previous employees, may be just the right person and bring new energy and new ideas to the company." (R2636)	16 (0.4)

Table 1 (continued)

Theme	Code	Short description	Example quote	*N* (%)
Properties of AI	AI as an assisting tool	Advantages of algorithmic tools should be harnessed within decision-making as long as they are being controlled by humans, who make final decisions.	"Decision makers don't have the time to check backgrounds on all possible early dismissals, so machine support is very helpful." (R1770)	404 (9.3)
	Objectivity of AI	Algorithms can make objective and rational decisions because, unlike humans, they do not have emotions and prejudices.	"The program compares data, decides objectively not according to its own feelings." (R1079)	169 (3.9)
	The person behind AI	ADM tools are developed by humans, whose worldviews are likely to be embedded in the design of such tools.	"Such computer programs may not be free from the views and decisions of its creators." (R271)	50 (1.1)
	AI fallibility	Algorithmic tools are prone to errors which may lead to potentially severe consequences for those affected.	"Computers make mistakes!" (R3434)	40 (0.9)
	Lack of transparency	Algorithmic decisions cannot be fully understood.	"Because how the computer program finds the information, what the program finds, and what it doesn't find is not transparent to the unemployed person." (R2861)	25 (0.6)
			Total:	4366 (100%)

certain individuals based on the deeds of others. Human beings deserve to be seen and evaluated independently from each other as illustrated by the following example quotes:

> "It's implied that people with similar experiences would behave similarly. But that is not the case." (R1178)

> "Each person is unique and acts depending on their environment and socialization." (R1483)

> "This kind of decision-making does not include the individuality and the context of the human being; [it] presumes statistically calculable homogeneous entities." (R1557)

> "A computer program that targets the personal values of others and sets them in relation to each other cannot be judged as fair in my eyes." (R1715)

6.1.1.2 Human decider Other respondents argued that consequential decisions having a significant impact on the future of a person should be made by another human. Only a human being (compared to an AI) has the social competence to approach another human being with the appropriate respect and understanding. Respondents also noted that throughout their careers, individuals acquire the necessary expertise that entitles them to make decisions about other human beings. The use of ADM gives rise to the suspicion that individuals may lose this competence in the long run as expressed in the following example quotes:

> "No machine or software can decide about humans. Only a person decides on the fate of a person. For this purpose, we have laws, public prosecutors, and judges." (R1314)

> "There are always reasons why that person said or did something. Personally, I think that decisions made by a computer program are inhumane, not compassionate. A program can't deliberate, a person however can." (R3276)

Further, respondents mentioned that a human decider can be held accountable if their decision is erroneous. With the use of ADM, personal accountability becomes blurred:

> "(…) because it releases the supervisor from their responsibility." (R2486)

6.1.1.3 Interpersonal contact Another code refers to the social aspect of decision-making. Respondents noted that the decision-making process should occur in the context of a social interaction between at least two individuals. Only in a face-to-face conversation, all involved parties can communicate their concerns, ask questions, and explain their circumstances. Furthermore, interpersonal interactions offer the possibility to collect additional data through observation – something that an algorithm is unable to do. Both aspects are crucial for the actual decision-making. The following quotes reflect this aspect exemplarily:

> "The person may not even be given the opportunity for a face-to-face interview after being "eliminated" by the program, or to present individual circumstances that would justify why there were problems with earlier loans etc." (R837)

> "A computer lacks empathy, lacks eye contact, lacks conversation." (R1040)

> "Because the human aspect is missing. In a personal meeting, an employer can

get an overall impression of the applicant and the applicant can demonstrate their various qualities." (R1981)

6.1.1.4 Human prejudice However, other respondents critically recognized that human beings often make decisions that are not objectively understandable because they are motivated by their personal preferences, prejudices, or whims – an issue illustrated by the following example quotes:

"Human decisions are directly or indirectly influenced by emotions, e.g., racism, bad mood, and so on." (R1824)

"The final decision is made by the employee. If they do not like the person they are dealing with, that is certainly going to influence the decision." (R3693)

6.1.1.5 More than one decider Next, some respondents argued that decisions affecting a human life should generally be made by at least two different individuals or a whole group of accordingly qualified people should be involved in the decision-making process to ensure objectivity:

"That a single individual decides is not acceptable. The decision should be cross-checked by at least one independent person." (R3609)

"This decision should be made by a committee of experts." (R1614)

6.1.1.6 Autonomy of the affected person Another aspect mentioned in the responses subsumed under *Human elements in decision-making* was the autonomy of the person affected by the decision. Respondents argued that the use of algorithmic decision tools greatly reduces their ability to make their own decisions about their future. The fact that the person being decided upon has no say leads to a perceived control loss as reflected in the following example quotes:

"I don't really have direct control." (R306)

"It is important to have a face-to-face dialogue, to involve the jobseeker in the decision (…)." (R3109)

"(…) the person themselves must be able to make their decisions freely without being monitored." (R3632)

6.1.1.7 Lack of trust towards AI Lastly, respondents explicitly said that they do not trust the algorithmic decision tool and their decisions or recommendations especially when they can significantly impact an individual's life:

> "I do not trust a mere comparison of data collected by the computer as a decision-making process about a person's fate." (R1609)

> "No trust in computer programs that "interfere" in decisions of such magnitude." (R1982)

6.1.2 Shortcomings of the data

The next theme encompasses four codes related to various shortcomings stemming from the data used in the decision-making process. In sum, 1088 of these codes were assigned, which makes up approximately 25% of the total number of assigned codes.

6.1.2.1 Limited data First, some respondents argued that the personal data considered in the context of ADM do not reflect the individual's situation in its entirety. Rather, they are a snapshot of reality and do not capture the complexity and interplay of motives behind it. Respondents also mentioned a variety of latent characteristics that are hard or impossible to quantify. Therefore, an algorithmic tool can consider only a greatly limited picture of an individual as expressed exemplarily by the following quotes:

> "There are other factors that come into play, such as empathy, collegiality, trustworthiness, and reliability. No computer can evaluate that." (R1872)

> "The algorithm only compares data. It does not see the person behind it and does not further investigate. For example, in the case of periods of unemployment, it could have been a 3-year cancer illness or parental leave. However, it is possible that such a long period of unemployment may lead to exclusion, even though there were valid reasons for it, and it does not diminish the applicant's competence." (R2132)

6.1.2.2 Distorted data Other respondents noted that the personal data considered in the context of ADM might not truly reflect the individual's characteristics or circumstances. The data might have been either consciously or unconsciously manipulated

by the individual or a third person, regardless of the manipulation's purpose as mentioned in these example quotes:

> "Personally, I have 20-30 applicants a year – what is written in the applications usually does not correspond to reality and how a person appears. There are often worlds in between." (R2133)

> "Information about people that can be found online may or may not be true. There are also rumors and lies circulating on the internet, which could be a disadvantage for the person in case of doubt when making a decision in this way." (R1743)

6.1.2.3 Outdated data Another aspect mentioned in the responses focusing on the data was the timeliness of data. Personal data considered in the context of ADM refer only to the individual's past. Respondents argued that the individual (e.g., their behavior) or the circumstances may have changed between the time points of data collection and decision-making. Consequently, decisions affecting a person's future should be based on current data that represent their present circumstances as can be derived from the following example quotes:

> "I can change my views and actions in the course of my experiences." (R351)

> "(…) a resume or previous activities do not say anything about the person at the moment. People are constantly developing for the better as well as for the worse." (R1853)

> "Everyone makes mistakes and has done something foolish at some point, so when this is held against you, it's not fair. People can change, too." (R1907)

6.1.2.4 Historical bias in AI Finally, respondents addressed the training data aspect. The performance of algorithmic decision tools is highly dependent on the training data. These may contain a historical bias that discriminates against certain social groups. Therefore, the tool might reproduce discriminating structures which contradicts not only the current socio-political sentiment but also legal norms as this quote shows exemplarily:

> "Moreover, this selection may only reinforce structures that already exist (e.g., the selection of certain genders or demographic groups)." (R2325)

6.1.3 Social impact of AI

The third theme includes seven codes and refers to the social impact of AI. Combined, 704 of these codes were assigned, which constitutes approximately 16.2% of the total number of assigned codes.

6.1.3.1 Violation of privacy The most frequently assigned code under this theme refers to individuals' privacy. The processing of personal data for algorithmic decision-making without the knowledge and explicit consent of the individuals was evaluated as a violation of their privacy. Respondents often referred to official data protection regulations with individual cases explicitly mentioning the EU's General Data Protection Regulation (GDPR).

> "If this program searches out information about the person from the internet, this is more than questionable from a data protection point of view." (R261)

Further, respondents noted that the use of personal data that is not related to the actual decision should not be evaluated in its context as illustrated by this respondent's response:

> "Political attitudes, religious affiliation or a person's sexual orientation belong in a private sphere and must therefore not be considered by the employer just because they have been found on a social media site on the internet." (R2725)

6.1.3.2 Purpose of AI Another aspect mentioned by the respondents was the purpose of algorithmic decision tools. Their deployment was often legitimized by highlighting their purpose. Respondents evaluated ADM mostly as fair when it aimed to bring a certain advantage to a larger group of people or an institution, e.g., to secure a bank's financial transactions, to increase a company's productivity, or to (re-)integrate individuals on the job market:

> "Comparing applicants with current employees allows selecting those applicants who fit in well because of similar characteristics [which can] bring about a positive development for the company." (R2216)

> "To actively seek employment can be required of anyone who lives at the expense of the community – some kind of "benchmark" is not too unreasonable to expect." (R3593)

However, some of those respondents viewed the algorithm's purpose as independent of fairness:

> "I find the word "fair" inappropriate in this context. AI is going to monitor us increasingly in the future, whether we like it or not. Is that what you mean by fair? I think it's logical for companies to proceed in this way." (R85)

Other respondents questioned the extent to which the purpose of AI aligns with the actual purpose of institutions' decision-making processes as shown by this exemplary response:

> "Invest your energy/program for the person to find a job. The approach is missing the point. Employment agency is not a punitive institution." (R3567)

6.1.3.3 Unequal access to resources Respondents also noted critically that algorithmic decision tools use personal data to classify individuals. Such *social selection* is evaluated as unfair because it creates an artificial social class system that excludes individuals from access to resources (e.g., jobs, credits) based on their characteristics which are often purely ascriptive. In the long run, such practices lead to an amplification of existing inequalities by systematically favoring certain individuals while disadvantaging others – a concern illustrated exemplarily by the following quotes:

> "Because you're just making comparisons without having seen or talked to the person. So, you're excluding potential employees who might be just the right people!" (R2379)

> "Even if a judge decides in the end, some candidates may never be recommended by the program, and thus never have a chance for an early release." (R1249)

6.1.3.4 Meritocracy Meritocracy was another aspect mentioned by respondents. Here, they argued that individual accomplishments should be the foundation of the decision-making process. Only by judging people based on their merits will everyone receive what they deserve. Accordingly, employees should be fired if they perform worse than their colleagues and able-bodied individuals who stay unemployed due to laziness should not receive any social benefits:

> "Companies can find quality staff this way. Employees put more effort into performing well. No more dragging along unqualified workers. Performance must be rewarded." (R2241)

> "Those who work harder should also get more. This program seems to take this into account." (R2930)

Some respondents even went a step back in the decision-making process, saying that individuals who have committed crimes should not be granted the opportunity for a new decision, and thus earlier release, at all:

> "Why should anyone be released early at all? If they are proven guilty, the sentence must also be served." (R1827)

6.1.3.5 Risk of being influenced by AI The next code related to the social impact of AI is the risk that algorithms might influence human deciders in their decision-making. Respondents pointed out that humans working with algorithmic tools might adopt the tools' recommendations without the necessary critical assessment. Consequently, the final decision could differ from the counterfactual, i.e., the outcome of a decision-making process without the involvement of software. Moreover, respondents said that the human decider might be both conscious (e.g., to minimize their personal responsibility) or unconscious (e.g., due to cognitive bias) about this influence process as reflected by these example quotes:

> "Bank employees will predominantly follow the computer's prescription as a safeguard." (R618)

> "Because it depends on the algorithm and because people tend to get comfortable, meaning the recruiter may hastily agree." (R2499)

6.1.3.6 Surveillance by AI Another concern raised by respondents is the possibility of using AI for surveillance purposes. They suspected a fundamental lack of trust towards people on the part of the institution or company using algorithmic tools for decision-making. As illustrated by this example quote, respondents associated the scenarios with a dystopian vision of society:

> "Ethically, it is not okay to collect this information about [human] behavior. The way back to the "control state" is not far with this. Lack of trust in humanity also does not lead to the desired result." (R2900)

6.1.3.7 Loss of diversity Finally, respondents noted that ADM might jeopardize social diversity. As algorithms compare individuals with already existing data, it is thus likely that such tools will systematically recommend those individuals whose characteristics are similar to the training data. Respondents argued that this may com-

promise the social diversity of groups and, consequently, affect group synergy negatively. The following quotes reflect this concern exemplarily:

> "The employees are eventually all the same […] This thwarts new ideas and innovations." (R2594)

> "Comparisons with "own" employees are not serious. Every person has their style and experience and that makes up their personality. I'm not looking for a clone of my employee." (R2536)

6.1.4 Properties of AI

The final theme encompasses five codes and relates to the properties of AI. In total, 688 responses were assigned these codes, which amounts to approximately 15.8% of the total number of assigned codes.

6.1.4.1 AI as an assisting tool The code assigned most often under this theme refers to certain advantages of algorithmic tools, such as the ability to screen and compare enormous data sets in a short time. Respondents argued that such properties should be used in the decision-making process. However, AI assistance should be combined with human supervision. The combination of both can positively affect the efficiency and quality of the decision as reflected in the following responses:

> "It has been demonstrated that statistical evaluation of such data brings added value and that the results are often correct. In addition, there is still a human control authority involved." (R1449)

> "I think such a program can improve the efficiency of decision-making. But the ultimate decision and responsibility should be with a human being." (R2837)

6.1.4.2 Objectivity of AI Another AI property that was mostly evaluated as fair was objectivity. Respondents acknowledged that algorithmic decision tools can make objective or neutral decisions that are free of emotions, prejudices, and preferences. Because such tools decide based solely on data, they treat people equally without preferring or rejecting individuals based on their characteristics:

> "Basically, there is fairness because the same regulations would apply to everyone." (R1085)

> "A computer program cannot be manipulated and decides rationally." (R1651)

6.1.4.3 The person behind AI Other respondents highlighted the fact that algorithmic tools used for decision-making are effectively developed by humans. Therefore, such tools and the selection of features considered are likely to reflect the opinions, worldviews, and biases of those people who write and maintain them as expressed by the following example quotes:

> "Because programs are written by people like me. And people make mistakes. That's why programs are full of mistakes. It is an illusion to think that computers are smarter than humans in decision-making. The computers only execute algorithms." (R1589)

> "Whether this procedure is fair or not depends on the programming. Such programming could, for example, contain racist items." (R1943)

However, some respondents also noted the ambiguity of this aspect. The quality of the algorithmic tool may vary greatly depending on the people who participate in its development:

"If the program is made by true professionals, the calculated output can actually be helpful." (R3096).

6.1.4.4 AI fallibility Furthermore, respondents critically noted that decisions made by an algorithmic tool can simply be wrong which can lead to serious consequences for the affected individuals. This fallibility was seen as an inherent characteristic. The following example illustrates the expressed concern about the replacement by automated decision tools that are not necessarily better than their human counterparts:

> "This is too much programmed knowledge for me – the human component is completely missing. After all, a wrong decision could be made – which program is absolutely foolproof? We rely more and more on computers and lose our human ability to observe and make decisions." (R2097)

6.1.4.5 Lack of transparency The last code under this theme refers to the lack of transparency in ADM. Respondents argued that the decision or recommendation made by the algorithmic decision tool as well as the underlying reasons might not be fully comprehensible. They also doubted whether individuals working with the tool or individuals being decided upon know the exact decision-making process, e.g., the entirety of criteria used as well as their weighting. However, a decision or recommendation along with their justification should be completely transparent to be fair as can be derived from the following responses:

> "Black box algorithm. Criteria for decision-making are not comprehensible, especially when AI is used." (R1810)

"The software's parameters are developed by others. Does the employee in the employment agency see through these algorithms? Is the suggestion of the program completely comprehensible for them?" (R3597)

6.2 Identified Codes, Fairness Evaluation, and Context Dependency

As we mentioned above, our study's focus lies on the qualitative analysis of the textual fairness explanations. However, respondents' textual explanations may depend on their fairness rating on the closed-ended fairness question (see section "Data"). To investigate how the identified codes are distributed across respondents' fairness evaluations, we briefly examine this aspect quantitatively (Fig. 1).

Overall, 19 out of the identified codes were predominantly associated with negative fairness evaluations of the presented scenarios. On the contrary, *AI as an assisting tool*, *Objectivity of AI*, *Purpose of AI*, and *Meritocracy* were coded more frequently in connection with positive fairness evaluations. This uneven distribution of our codes seems to reflect Kern et al.'s (2022) findings who, using quantitative analysis of the closed-ended fairness ratings, concluded that respondents were rather skeptical about the use of automated decision tools especially in high-stakes scenarios. Interestingly, the four codes assigned with predominantly fair evaluations refer to aspects surrounding the efficiency of algorithmic tools, which is a property highlighted as one of the central advantages by supporters of such technologies. In other words, individuals who perceive ADM systems to be fair seem to do so due to sharing the notion that ADM systems and AI can reach better (i.e., more objective) decisions in a shorter time.

Respondents' textual responses may also depend on the scenario (banking, hiring, criminal justice, unemployment) and the dimension levels of the vignette preceding the open-ended fairness evaluation since the open-ended question was asked after one randomly chosen vignette per respondent only. As shown in Kern et al. (2022), fairness ratings, measured via the fairness rating scale, depend on the context of the application. To briefly investigate this possibility, we contrast our results against the hypothetical scenarios. Overall, there does not seem to be a strong tendency for respondents' textual responses on fairness, as studied in this paper, to depend on the context of the application Supplementary Material 1. Likewise, we do not find strong dependence on the individual dimension levels, with a few notable exceptions. When looking at the distribution of codes according to the first vignette dimension, namely action type, several aspects stand out (see Online Resource 5). The code *Interpersonal contact*, as well as the codes relating to algorithms' properties, mentioned previously in this subsection, namely *Objectivity of AI*, *Meritocracy*, and *Purpose of AI* were coded more frequently in connection with punitive actions. This indicates, on the one hand, that interpersonal dialogue and empathy are particularly important when a decision is likely to negatively affect an individual. On the other hand, the emphasis on algorithms' characteristics and the intention behind their use might indicate that it is precisely the belief in the alleged algorithmic objectivity and the meritocratic nature that serves as an additional factor legitimizing decisions that

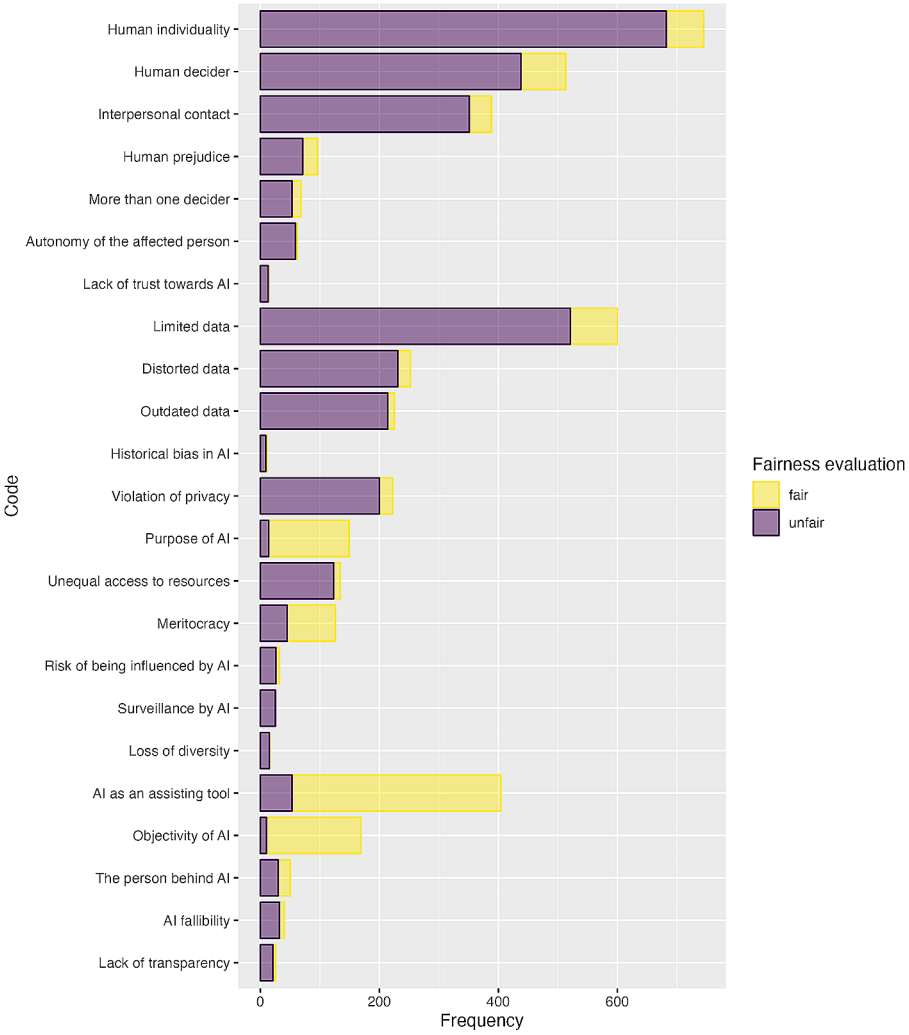

Fig. 1 Number of textual responses, by code and fairness evaluation. Note: "Fair" refers to respondents who chose *very fair* or *somewhat fair* on the closed-ended fairness question. "Unfair" refers to respondents who chose *not at all fair* or *not very fair*

have severe consequences for individuals, even if the decisions might be unfair. In contrast, the codes *Autonomy of the affected person* and *Unequal access to resources* were assigned more frequently in connection with assistive actions. This seems plausible, as punitive actions usually have a top-down character and are more likely to be accepted as being a consequence of a certain behavior, which limits the scope of individual autonomy. In contrast, this autonomy becomes even more important when it comes to access to resources as it empowers individuals to advocate for themselves. When looking at the distribution of codes according to the next vignette dimension – internet data being either used or not – two codes stand out: *Distorted data* and *Violation of privacy* (see Online Resource 6). Both were assigned more

frequently in connection with scenarios in which additional internet data was used for decision-making. This seems plausible as the risk of manipulation or incorrect representation of reality is higher for internet data compared to e.g., administrative data. At the same time, *how* a person chooses to present themselves on the internet is a private matter that should not underlie a decision-making process. When looking at the distribution of codes across the levels of the last vignette dimension (type of decision-making about the degree of human leeway, see Online Resource 7), three codes stand out: *Limited data*, *Human decider*, and *AI as an assisting tool*. The first two codes were assigned more frequently in connection with scenarios describing an algorithm making the decision autonomously, i.e., without human involvement. In the case of *Limited data*, the code might reflect the contrast between a human and an algorithm regarding the request of additional data. While an algorithm makes the decision based on the (likely limited) data material that is available at the time the decision is made, the human decider can obtain more in-depth information if they believe it is necessary to make a well-considered decision. The code *Human decider* may also have been assigned more often in connection with this dimension level due to the contrast between a human and an algorithm.

7 Discussion

In the following, we address our most striking findings as well as the resulting implications for the use of ADM and directions for further research. The aim of our study was to explore which subjective aspects do individuals consider when assessing the fairness of ADM. Respondents addressed a wide range of topics which is reflected in the 23 codes and four themes that we identified: *Human elements in decision-making*, *Shortcomings of the data*, *Social impact of AI*, and *Properties of AI*.

7.1 Aspects Affecting Fairness

Regarding our research question, we found that individuals take a strong human-centric stance when elaborating on fairness in ADM. By far the most frequently assigned codes relate to the human aspects of decision-making processes. On the side of those affected by a decision, the recognition of their individuality is of great importance. Accordingly, a person's characteristics and needs should be considered independently so that the decision can be tailored towards them. Despite respondents' awareness that human deciders might be biased, it is them and not the algorithms that can recognize the humanness in another individual. One response pinpoints this ambivalence: "Humans decide not only rationally, but especially emotionally. Such decisions are certainly not always good, but they make us human." (R80). Even aspects that we have summarized under the seemingly technical theme *Shortcomings of the data* were often framed from a human standpoint. Thus, the complexity and dynamic nature of a human identity is not easy to capture in numbers that could be processed by an algorithm. Such shortcomings might be balanced out by human deciders, as an individual's current and detailed circumstances can be further explored in a face-to-

face conversation: "A one-hour interview can possibly result in a better assessment than just reviewing the resume." (R1852).

Next, we identified certain algorithmic properties that might positively affect a decision-making process. First, respondents evaluated the human-in-the-loop scenarios mostly as fair based on the algorithm's ability to process large amounts of data within a short time (recall the three codes mostly associated with a fair rating in Fig. 1). Second, because algorithmic tools process only data, they are believed to be objective. This neutrality was frequently associated with equal treatment. One respondent said explicitly that a fully automated system "could be an opportunity for those who could not find work due to their appearance." (R3507). This notion contradicts the view that algorithms systematically exclude people from access to resources, which we also found in the data. This contrast can be interpreted twofold. First, concerns about the impact of AI might arise due to AI awareness as evidence shows that individuals who possess more AI knowledge tend to perceive it as riskier (Yigitcanlar et al., 2022). Second, individuals, especially those who have experienced discrimination in decision-making processes guided solely by humans, might see a potential in AI for objective and fair decisions.

Another intriguing finding is that the deployment of algorithms seems to be legitimized and therefore mostly evaluated as fair when the purpose for the use is highlighted. While such a goal-oriented attitude is not necessarily harmful, it raises some issues and should thus be critically examined. For example, one respondent said: "In our welfare state, too many people hide behind social networks, transparency must be established, a high-performing society cannot afford parasites in the long run." (R3678). This response seems to reflect a neoliberal notion of a meritocratic society. This might be investigated further by examining how societal beliefs regarding the redistribution of resources affect fairness evaluations of ADM.

Although we are cautious about the generalizability of our qualitative findings, there are obvious parallels to studies with similar designs that need to be addressed. As in the studies by Helberger et al. (2020), Bankins et al. (2022), and Formosa et al. (2022), we were able to show that individuals attach great importance to the human aspect of decision-making processes. Humans, as emotional and empathic beings, are believed to understand and consider the whole context of another person's circumstances and thus make fair decisions (Bedemariam & Wessel, 2023). At the same time, our study also shows that an algorithm's objectivity seems to be among the main reasons why automated decisions were evaluated as fair (Bankins et al., 2022; Schoeffer et al., 2021). Similar to Yurrita et al.'s (2023) findings, our responses indicate a preference for keeping a human as the final decider even if an algorithm is involved in the decision-making process.

Ultimately, the textual responses address specific values such as privacy, autonomy, or transparency. These are not only closely related to an individual's rights, but also have relevance within the jurisdiction of the EU, which regulates the use of information technologies, e.g., with the EU AI Act. Although the presented scenarios in our study were purely hypothetical, similar tools are already used, e.g., in HR and recruiting, and discussed in, e.g., public employment services (Bach et al., 2023). Similar profiling tools that focus on providing a risk score and that leave the discussion with a human, are likewise widespread, as evidenced by, e.g., SCHUFA,

a German credit scoring system that is often required when applying for a credit card or signing a rental agreement (Elmer, 2021).

From a public law perspective, the increasing use of ADM tools will certainly affect some personal rights such as non-discrimination or information rights (Hofmann, 2023). A possible remedy to the negative outcomes due to ADM might be the "right to an explanation" of an algorithmic decision as guaranteed by the EU's GDPR. Besides debates around the interpretation of said right, research shows that the practical realization of this right is not necessarily straightforward: on the one hand, the explanations provided to the individual upon their request might vary in the breadth and depth of their content (Dexe et al., 2020), which is likely to affect the individual's understanding of the decision-making process. On the other hand, individuals might prioritize different aspects of an explanation depending on their personal background or domain knowledge (Hamon et al., 2021; Kern et al., 2023).

7.2 Limits to Technical Fairness Metrics

In addition to the strong human-centric stance emerging from the textual responses that we described in the last subsection, respondents often addressed certain characteristics of the data being used in the process of decision-making. Those aspects are an important intersection between our findings and existing technical fairness metrics. Indeed, these metrics aim at increasing fairness by focusing on the data processed in the context of ADM. However, they often provide targeted solutions and cannot account for fundamental shortcomings due to poor data quality. Since algorithmic systems are increasingly used by public authorities, there is an urgent need for standardized data quality requirements and regulations at, e.g., the EU level. The development of such standards is one of the goals of the European Commission to make AI "inclusive, non-biased, and trustworthy" (Balahur et al., 2022). Standards would thus target the source of potentially adverse social outcomes caused by ADM, including those we identified based on the textual responses. Ideally, such data quality requirements should accommodate both proposed metrics of fairML scholars' and individuals' fairness perceptions regarding personal data.

Furthermore, the responses reveal a nuanced understanding of fairness that goes beyond proposed fairML metrics. The bottom-up approach to fairness highlights a variety of individual views that are not free of contradictions as illustrated by the examples in subsection 7.1. These are largely determined by intersections of various characteristics. Kern et al. (2022) who explore the same dataset quantitatively by focusing on respondents' fairness ratings using the closed-ended scale mentioned above, acknowledge that varying fairness evaluations might indicate the presence of self-interest or social identity effects. While our findings point to the limitations of fairML metrics, they do not allow for a concise fairness definition that would consider all identified aspects. We provide a broad insight to individual fairness perceptions; however, more research is needed that should shed light on the interplay between individual characteristics and contextual factors in ADM that affect those subjective fairness perceptions. A sound theoretical basis could be a first step towards analyzing this interplay in the future. In the context of this paper, we have grounded our theoretical considerations on fairness in ADM in the social construction of tech-

nology (Bijker, 2010), which we discuss against the backdrop of our findings in the next subsection.

7.3 Re-Imagining Algorithmic Tools

The perspective of social construction of technology (Bijker, 2010) assumes that technologies are socially shaped artifacts. This critical stance towards technological determinism offers some leeway for re-imagining technology. Recent evidence from workshops (Scott et al., 2022) and interviews combined with design fiction exercises (Wang et al., 2023) conducted to explore alternative technologies in the context of public employment agencies show that AI technologies as imagined by jobseekers and those who support them have little in common with currently deployed algorithms for statistical prediction of outcomes. According to these stakeholders, the intention behind such technologies should primarily focus on *empowering* people and seeing them as whole individuals. We observed some parallel findings in our data. For example, one respondent criticized the punitive character of an algorithmic tool in the *unemployment* scenario: "…because one can see the intention to cut payments without a genuine recognition of the affected individual's situation…" (R3045). Another respondent took the opportunity to creatively describe the concept of an algorithmic system that could prove helpful for jobseekers without restricting their autonomy: "I see this as a kind of job fair, where the unemployed can look at what others with a similar resume have done to re-enter the job market. They can use this as an orientation to consider whether, e.g., a retraining could be useful and beneficial for themselves." (R3372). These instances show that fairness seems to be dependent on the underlying intention to use the algorithm for a purpose. In turn, the purpose of such tools is likely to drive their development in a specific direction. Prominent examples of unfair social outcomes caused by algorithms might be an occasion to rethink the purpose of such systems or as Wang et al. expressed it: "it is clear that a shift away from the focus on predicting social outcomes of individuals is required" (Wang et al., 2023:2). Including the ideas and perspectives of the most affected stakeholders by consulting them during both the early conceptualization and the critical stages throughout the design process of algorithmic technologies might positively contribute to the perceived fairness of such systems. At the same time, given the rapid advances in the automation of various decision-making processes during the last decades, it is imperative to examine how much the members of the general population know about this automation development, whether and how they assess the associated risks that such systems might pose to their e.g., autonomy, and what opportunities they have to express their opinions and concerns. In the following subsection, we refer to recent research on this topic and discuss it in light of our findings.

7.4 Public Awareness Regarding AI

The last point we highlight is AI awareness by the public. A recent paper by Kieslich et al. (2023) investigated Germans' public opinion towards AI. Their results reveal a rather indifferent attitude of German citizens regarding AI with educational level

and interest in the topic being predictors of concerns with AI. Issues such as fairness, accountability, or transparency, although frequently addressed in political and academic discourse, do not seem to resonate with the general population (Kieslich et al., 2023). Against this backdrop, our data shows a puzzling variety of ethical concerns regarding social inequality, transparency, or autonomy in the context of ADM. One explanation might be the circumstance that the textual responses we examined were provided as direct responses to the vignette scenarios presented, which were a steady point of reference. Nevertheless, isolated responses on semi- or fully automated decision-making indicate a low level of AI awareness: "I cannot imagine that there is a computer program for this." (R343), "I have heard little of it and do not really know." (R1214), and "[B]ecause I am not that informed." (R3119). Those responses were left uncoded due to the lack of thematic relatedness to fairness. We do not claim that they confirm Kieslich et al.'s (2023) findings, but the discrepancies in our results provide a starting point for further research on AI awareness in the general population.

8 Strengths and Limitations

The major strength of our study lies in the combination of a qualitative approach that allows for data immersion with a probability-based sample that is representative for the German population. Previous qualitative research on ADM was mostly conducted on small samples, sometimes recruited by convenience sampling approaches that leave it open whose opinion is represented. Our design, combining a large probability sample with a qualitative approach to uncover the nuances in respondents' opinions and views, aims to balance sample breadth and generalizability of the findings versus the depth of the phenomena studied.

However, this study also has some limitations. First, the open-ended responses were collected in an online survey without the possibility of further inquiry into respondents' initial thoughts. Therefore, the individual responses, though allowing for more nuanced analysis than closed-ended survey responses, are inevitably superficial.

Second, respondents evaluated the fairness of hypothetical scenarios. Therefore, the content of the vignettes may have influenced not only their fairness evaluation but also the explanation they provided with the textual responses. However, our brief contrast of the fairness codes versus the ADM context revealed only minor context and vignette dependency of fairness perceptions (see subsection 6.2).

Third, we acknowledge some ambiguity in the vignette content regarding the third dimension (see section "Data", the degree of human leeway concerning the computer program). Regardless of the level of this dimension, all vignettes mentioned a program that was developed specifically to make a particular decision. This means that even in the scenario with maximum human leeway, the human decision-maker still decides based on data used by the computer program. Due to this ambiguity, we cannot rule out that respondents who evaluated those scenarios nevertheless thought of an algorithm or its recommendation assisting the human decision-maker. This is supported by the fact that AI-specific codes such as *AI as an assisting tool*, *Purpose of*

AI, or *Objectivity of AI* were assigned in connection with scenarios in which a human decider compared data processed by the computer program.

Fourth, although we derived our codes directly from the data and discussed every step of the analysis to reach a consensus, we cannot entirely rule out subjectivity which could potentially have biased our findings.

Fifth, we focused our attention on getting a nuanced understanding of participants' textual responses while mostly ignoring additional data collected in the survey through closed-ended survey questions. Future research may want to revisit our data and findings, while taking additional respondent characteristics into account, e.g., by using a combined qualitative-quantitative analytical strategy. Regarding the gender variable in the data, we furthermore point out that it was collected in a binary form. An answer other than "female" or "male" resulted in item-nonresponse (this applies to two respondents). Consequently, our data provides only limited insight into potential differences in fairness perceptions between genders. Future research should therefore use inclusive measurement instruments that allow representation of all genders in the data.

Ultimately, respondents were not provided with a fairness definition when answering the survey questions. This means that they assessed the fairness of the scenarios according to their subjective understanding of fairness. The seemingly conflicting codes discussed in the previous section seem to reflect the differences in individual fairness perceptions. However, similar frictions also exist between established fairness definitions used in the fairML literature as highlighted in recent publications (Alves et al., 2023; Garg et al., 2020). Future research may want to examine the similarities between these frictions.

9 Conclusion

The increasing deployment of ADM tools in both private and public sectors along with prominent examples revealing their discriminating potentials have sparked discussions about the consequences of algorithms on fairness and inequality. Existing fairness evaluations in ADM focus on technical notions of fairness without considering how fairness in the context of ADM is understood by potentially affected individuals. To shed light on those subjective fairness perceptions, we analyzed 3697 textual responses to hypothetical ADM scenarios embedded in the German Internet Panel (Wave 54, July 2021), a probability-based longitudinal online survey. The inductive content analysis yielded 23 individual codes reflecting various aspects relating to four overarching themes in the context of ADM: *Human elements in decision-making*, *Shortcomings of the data*, *Social impact of AI*, and *Properties of AI*. Our findings show that individual understandings of fairness are nuanced, exceeding the scope of hitherto proposed mathematical fairness concepts in ADM. Our study contributes to previous research on subjective fairness perceptions in ADM by highlighting the limited focus of technical fairness metrics and provides valuable insights regarding the factors that might affect fairness in ADM from the perspective of affected stakeholders.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11023-024-09684-y.

Acknowledgements We thank Marike Andreas, members of the Kreuter-Keusch research group, and the anonymous reviewers for their constructive comments on an earlier version of this paper.

Authors Contribution Conceptualization: D.S. and R.B., Formal analysis: D.S., Funding acquisition: R.B., Methodology: D.S., Validation: R.B., Visualization: D.S., Writing - original draft: D.S., Writing - review & editing: D.S. and R.B.

Funding 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".
Open Access funding enabled and organized by Projekt DEAL.

Data 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.

Declarations

Ethical Approval Not applicable.

Consent of Publication Not applicable.

Competing Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

AlgorithmWatch (2019). Automating society 2019. In AlgorithmWatch. https://algorithmwatch.org/en/automating-society-2019/.

Alves, G., Bernier, F., Couceiro, M., Makhlouf, K., Palamidessi, C., & Zhioua, S. (2023). Survey on fairness notions and related tensions. *EURO Journal on Decision Processes*, *11*, 100033. https://doi.org/10.1016/j.ejdp.2023.100033.

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

Bach, R. L., Kern, C., Mautner, H., & Kreuter, F. (2023). The impact of modeling decisions in statistical profiling. *Data & Policy*, *5*, e32. https://doi.org/10.1017/dap.2023.29.

Balahur, A., Jenet, A., Hupont, I. T., Charisi, V., Ganesh, A., Griesinger, C. B., Maurer, P., Mian, L., Salvi, M., Scalzo, S., Soler, J. G., Taucer, F., & Tolan, S. (2022). Data quality requirements for inclusive, non-biased and trustworthy ai: putting-science-into-standards. https://doi.org/10.2760/365479.

Bankins, S., Formosa, P., Griep, Y., & Richards, D. (2022). AI decision making with dignity? Contrasting workers' justice perceptions of human and ai decision making in a human resource management context. *Information Systems Frontiers*, *24*(3), 857–875. https://doi.org/10.1007/s10796-021-10223-8.

Bedemariam, R., & Wessel, J. L. (2023). The roles of outcome and race on applicant reactions to AI systems. *Computers in Human Behavior*, *148*, 107869. https://doi.org/10.1016/j.chb.2023.107869.

Berg, J., Lipponen, E., Sailas, E., Soininen, P., Varpula, J., Välimäki, M., & Lahti, M. (2023). Nurses' perceptions of nurse–patient communication in seclusion rooms in psychiatric inpatient care: A focus group study. *Journal of Psychiatric and Mental Health Nursing*, 781–794. https://doi.org/10.1111/jpm.12907.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: T state of the art. *Sociological Methods & Research*. https://doi.org/10.1177/0049124118782533.

Bijker, W. E. (2010). How is technology made?—That is the question! *Cambridge Journal of Economics*, *34*(1), 63–76.

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.

Blom, A. G., Fikel, M., Gonzalez Ocanto, M., Krieger, U., Rettig, T., & SFB 884 ´Political economy of reforms´, university of mannheim. (2021). German internet panel, Wave 54 (July 2021). *GESIS Data Archive Cologne*, *ZA7762 Data file Version 1.0.0*. https://doi.org/10.4232/1.13835.

Burema, D. (2022). A critical analysis of the representations of older adults in the field of human–robot interaction. *AI & Society*, *37*(2), 455–465. https://doi.org/10.1007/s00146-021-01205-0.

Cengiz, P. M., & Eklund Karlsson, L. (2021). Portrayal of immigrants in Danish media—a qualitative content analysis. *Societies*, *11*(2), 45. https://doi.org/10.3390/soc11020045.

Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments (issue arXiv:1610.07524). arXiv. https://doi.org/10.48550/arXiv.1610.07524.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A. Algorithmic decision making and the cost of fairness, & Mining (2017). 797–806. https://doi.org/10.1145/3097983.3098095.

R Core Team (2023). R: A language and environment for statistical computing.

Dexe, J., Ledendal, J., & Franke, U. (2020). An empirical investigation of the right to explanation under gdpr in insurance. In S. Gritzalis, E. R. Weippl, G. Kotsis, A. M. Tjoa, & I. Khalil (Eds.), *Trust, privacy and security in digital business* (pp. 125–139). Springer International Publishing. https://doi.org/10.1007/978-3-030-58986-8_9.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through Awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, *214-226*. https://doi.org/10.1145/2090236.2090255.

Elmer, C. (2021). Algorithms in the spotlight: Collaborative investigations at Der Spiegel. *The Data Journalism Handbook: Towards a critical data practice* (pp. 257–264). Amsterdam University. https://doi.org/10.1515/9789048542079.

Eynon, R., & Young, E. (2021). Methodology, legend, and rhetoric: The constructions of ai by academia, industry, and policy groups for lifelong learning. *Science Technology & Human Values*, *46*(1), 166–191. https://doi.org/10.1177/0162243920906475.

Formosa, P., Rogers, W., Griep, Y., Bankins, S., & Richards, D. (2022). Medical AI and human dignity: Contrasting perceptions of human and artificially intelligent (AI) decision making in diagnostic and medical resource allocation contexts. *Computers in Human Behavior*, *133*, 107296. https://doi.org/10.1016/j.chb.2022.107296.

Foulkes, L., Reddy, A., Westbrook, J., Newbronner, E., & McMillan, D. (2021). Social relationships within university undergraduate accommodation: A qualitative study. *Journal of Further and Higher Education*, *45*(10), 1469–1482. https://doi.org/10.1080/0309877X.2021.1879745.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). *On the (Im)Possibility of Fairness* (Issue arXiv:1609.07236). arXiv. https://doi.org/10.48550/arXiv.1609.07236.

Gajane, P., & Pechenizkiy, M. (2018). On formalizing fairness in prediction with machine learning (issue arXiv:1710.03184). arXiv. https://doi.org/10.48550/arXiv.1710.03184.

Garg, P., Villasenor, J., & Foggo, V. (2020). Fairness metrics: A comparative analysis. 2020 IEEE International Conference on Big Data (Big Data), 3662–3666. https://doi.org/10.1109/BigData50022.2020.9378025.

Grauenhorst, T., Blohm, M., & Koch, A. (2016). Respondent incentives in a national face-to-face survey: Do they affect response quality? *Field Methods*, *28*(3), 266–283. https://doi.org/10.1177/1525822X15612710.

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. NIPS Symposium on Machine Learning and the Law, 1(2), 1–11.

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, *46*(3), 205–211. https://doi.org/10.1136/medethics-2019-105586.

Guenna Holmgren, A., Juth, N., Lindblad, A., & von Vogelsang, A. C. (2022). Nurses' experiences of using restraint in neurosurgical care – a qualitative interview study. *Journal of Clinical Nursing*, *31*(15–16), 2259–2270. https://doi.org/10.1111/jocn.16044.

Hamon, R., Junklewitz, H., Malgieri, G., HertP. D., Beslay, L., & Sanchez, I. (2021). Impossible explanations? Beyond explainable AI in the GDPR from a COVID-19 use case scenario. *Proceedings of the 2021 ACM Conference on Fairness Accountability and Transparency*, *549–559*. https://doi.org/10.1145/3442188.3445917.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in supervised learning (issue arXiv:1610.02413). arXiv. https://doi.org/10.48550/arXiv.1610.02413.

Helberger, N., Araujo, T., & de Vreese, C. H. (2020). Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Computer Law & Security Review*, *39*, 105456. https://doi.org/10.1016/j.clsr.2020.105456.

Hofmann, H. C. H. (2023). Automated decision-making (ADM) in EU public law. *SSRN Scholarly Paper 4561116*. https://doi.org/10.2139/ssrn.4561116.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687.

Jørgensen, R. F. (2023). Data and rights in the digital welfare state: The case of Denmark. *Information Communication & Society*, *26*(1), 123–138. https://doi.org/10.1080/1369118X.2021.1934069.

Juijn, G., Stoimenova, N., Reis, J., & Nguyen, D. (2023). Perceived algorithmic fairness using organizational justice theory: An empirical case study on algorithmic hiring. *Proceedings of the 2023 AAAI/ACM Conference on AI Ethics and Society*, *775-785*. https://doi.org/10.1145/3600211.3604677.

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). https://doi.org/10.1016/j.patter.2022.100591.

Kern, D. R., Stevens, G., Dethier, E., Naveed, S., Alizadeh, F., Du, D., & Shajalal, M. (2023). Peeking inside the schufa blackbox: Explaining the German housing scoring system. *arXiv*. https://doi.org/10.48550/arXiv.2311.11655. arXiv:2311.11655.

Kieslich, K., Lünich, M., & Došenović, P. (2023). Ever heard of ethical AI? Investigating the salience of ethical ai issues among the German population. *International Journal of Human–Computer Interaction*, *0*(0), 1–14. https://doi.org/10.1080/10447318.2023.2178612.

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. (2019). The challenges of algorithm-based hr decision-making for personal integrity. *Journal of Business Ethics*, *160*(2), 377–392. https://doi.org/10.1007/s10551-019-04204-w.

Liem, A. (2019). Indonesian clinical psychologists' perceptions of complementary and alternative medicine research and knowledge: A content analysis study. *The Journal of Mental Health Training Education and Practice*, *14*(3), 164–173. https://doi.org/10.1108/JMHTEP-03-2018-0018.

Mavletova, A. (2013). Data quality in PC and mobile web surveys. *Social Science Computer Review*, *31*(6), 725–743. https://doi.org/10.1177/0894439313485201.

McCarthy, D. R. (2013). Technology and `the International' or: How I learned to stop worrying and love determinism. *Millennium*, *41*(3), 470–490. https://doi.org/10.1177/0305829813484636.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 115:1–115. https://doi.org/10.1145/3457607.

Meitinger, K., Behr, D., & Braun, M. (2021). Using apples and oranges to judge quality? Selection of appropriate cross-national indicators of response quality in open-ended questions. *Social Science Computer Review*, *39*(3), 434–455. https://doi.org/10.1177/0894439319859848.

Munro, M., Cook, A. M., & Bogart, K. R. (2022). An inductive qualitative content analysis of stigma experienced by people with rare diseases. *Psychology & Health*, *37*(8), 948–963. https://doi.org/10.1080/08870446.2021.1912344.

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *560*, 568. https://doi.org/10.1145/1401890.1401959.

Peeters, R., & Widlak, A. C. (2023). Administrative exclusion in the infrastructure-level bureaucracy: The case of the Dutch daycare benefit scandal. *Public Administration Review*, 1–15. https://doi.org/10.1111/puar.13615.

Rinta-Kahila, T., Someh, I., Gillespie, N., Indulska, M., & Gregor, S. (2022). Algorithmic decision-making and system destructiveness: A case of automatic debt recovery. *European Journal of Information Systems*, *31*(3), 313–338. https://doi.org/10.1080/0960085X.2021.1960905.

Rodolfa, K. T., Saleiro, P., & Ghani, R. (2019). Bias and fairness. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane (Eds.), Big Data and Social Science: A Practical Guide to Methods and Tools (2nd ed.).

Schmidt, K., Gummer, T., & Roßmann, J. (2020). Effects of respondent and survey characteristics on the response quality of an open-ended attitude question in web surveys. *Methods Data Analyses*, *14*(1). https://doi.org/10.12758/mda.2019.05.

Schoeffer, J., Machowski, Y., & Kuehl, N. (2021). Perceptions of fairness and trustworthiness based on explanations in human vs. automated decision-making (arXiv:2109.05792). arXiv. https://doi.org/10.48550/arXiv.2109.05792.

Scott, K. M., Wang, S. M., Miceli, M., Delobelle, P., Sztandar-Sztanderska, K., & Berendt, B. (2022). Algorithmic tools in public employment services: Towards a jobseeker-centric perspective. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2138–2148. https://doi.org/10.1145/3531146.3534631.

Spreckley, M., de Lange, J., Seidell, J. C., & Halberstadt, J. (2022). Patient insights into the experience of trying to achieve weight-loss and future expectations upon commencement of a primary care-led weight management intervention: A qualitative, baseline exploration. *PLOS ONE*, *17*(6), e0270426. https://doi.org/10.1371/journal.pone.0270426.

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), 20539517221115189. https://doi.org/10.1177/20539517221115189.

van Nuenen, T., Such, J., & Cote, M. (2022). Intersectional experiences of unfair treatment caused by automated computational systems. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 4451–44530. https://doi.org/10.1145/3555546.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. Proceedings of the International Workshop on Software Fairness, 1–7. https://doi.org/10.1145/3194770.3194776.

Wang, S. M., Scott, K. M., Artemenko, M., Miceli, M., & Berendt, B. (2023). We try to empower them—exploring future technologies to support migrant jobseekers. *2023 ACM Conference on Fairness, Accountability, and Transparency. (Forthcoming)*.

Williams, R., & Edge, D. (1996). The social shaping of technology. *Research Policy*, *25*(6), 865–899. https://doi.org/10.1016/0048-7333(96)00885-2.

Yigitcanlar, T., Degirmenci, K., & Inkinen, T. (2022). *Drivers behind the public perception of artificial intelligence: Insights from major Australian cities*. AI & SOCIETY. https://doi.org/10.1007/s00146-022-01566-0.

Yurrita, M., Draws, T., Balayn, A., Murray-Rust, D., Tintarev, N., & Bozzon, A. (2023). disentangling fairness perceptions in algorithmic decision-making: The effects of explanations, human oversight, and contestability. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–21. https://doi.org/10.1145/3544548.3581161.

Authors and Affiliations

Daria Szafran[1,2] · Ruben L. Bach[2]

✉ Daria Szafran
daria.szafran@uni-mannheim.de

[1] School of Social Sciences, University of Mannheim, A5, 6, 68159 Mannheim, Germany

[2] Mannheim Centre for European Social Research (MZES), University of Mannheim, A5, 6, 68159 Mannheim, Germany