Original Research Article

Big Data & Society
January–June: 1–15
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517221092956
journals.sagepub.com/home/bds

Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artificial intelligence

Kimon Kieslich[1], Birte Keller[1] and Christopher Starke[2]

Abstract

Despite the immense societal importance of ethically designing artificial intelligence, little research on the public perceptions of ethical artificial intelligence principles exists. This becomes even more striking when considering that ethical artificial intelligence development has the aim to be human-centric and of benefit for the whole society. In this study, we investigate how ethical principles (explainability, fairness, security, accountability, accuracy, privacy, and machine autonomy) are weighted in comparison to each other. This is especially important, since simultaneously considering ethical principles is not only costly, but sometimes even impossible, as developers must make specific trade-off decisions. In this paper, we give first answers on the relative importance of ethical principles given a specific use case—the use of artificial intelligence in tax fraud detection. The results of a large conjoint survey ($n = 1099$) suggest that, by and large, German respondents evaluate the ethical principles as equally important. However, subsequent cluster analysis shows that different preference models for ethically designed systems exist among the German population. These clusters substantially differ not only in the preferred ethical principles but also in the importance levels of the principles themselves. We further describe how these groups are constituted in terms of sociodemographics as well as opinions on artificial intelligence. Societal implications, as well as design challenges, are discussed.

Keywords

Ethical principles, artificial intelligence, public perception, design preferences, trade-offs

Introduction

Artificial intelligence (AI) has enormous potential to change society. While the widespread implementation of AI systems can certainly generate economic profits, policymakers and scientists alike also highlight the ethical challenges accompanied by AI. Most scholars, politicians, and developers agree that AI needs to be developed in a human-centric and trustworthy fashion, for AI that benefits the common good (Berendt, 2019; Jobin et al., 2019). Trustworthy and beneficial AI requires that ethical challenges be considered during all stages of the development and implementation process. While plenty of work addresses ethical AI development, there is surprisingly little research investigating public perceptions of those ethical challenges. This lack of citizen involvement is striking because developing ethical AI, in a normative sense, aims to be human-centric and of benefit for the whole society. Moreover, insights into citizens' perceptions of ethical principles will inform developers tasked with designing ethical AI systems and decision-makers entrusted with implementing such systems in social contexts (Berendt, 2019). We, therefore, set out to shed light on public perceptions of ethical principles outlined in ethical guidelines. Particularly, we investigate how people prioritize different ethical principles. Accounting for the trade-offs between the different ethical principles is especially important because maximizing them simultaneously often proves challenging or even impossible when designing and implementing AI systems. For instance, aiming for a

[1]Department of Social Sciences, Heinrich Heine University Düsseldorf, Düsseldorf, Germany;
[2]Amsterdam School of Communication Research, University of Amsterdam, Amsterdam, The Netherlands

Corresponding author:
Kimon Kieslich, Department of Social Sciences, Heinrich Heine University Düsseldorf, Düsseldorf, Germany.
Email: kimon.kieslich@hhu.de

high degree of explainability of AI systems can conflict with the ethical principle of accuracy, since a high degree of accuracy tends to require complex AI models that cannot be fully understood by humans, especially laypersons. Thus, taking the goal of ethical AI development seriously requires decision-makers to take the opinions of the (affected) public into account.

This paper gives first answers to the relative importance of ethical principles. We use an AI-based tax fraud detection system as a case in point. Such systems, already in use in several European countries, detect patterns in large amounts of tax data and flag suspicious cases which are then analysed in depth by a human. In a large ($n = 1099$) online survey with a conjoint design, we asked participants to rate different configurations of an AI-based tax fraud detection system; the proposed systems varied in how they comply with the seven ethical principles that are most prominent in global AI ethics guidelines (Jobin et al., 2019). As we aim for a high external information value of our results, we decided not to rely on one specific ethical guideline, but on the ethical principles that are most prominent on a global scale.

Ethical guidelines of AI development

AI increasingly permeates most areas of peoples' daily lives, whether in the form of virtual intelligent assistants, as a recommendation algorithm for movie selection, or in hiring processes. Such areas of application are only made possible by the accumulation of huge amounts of data, so-called Big Data, that people constantly leave behind in their digital lives. Although these AI-based technologies aim to take tasks off peoples' hands and make their lives easier, collecting and processing personal data is also associated with major concerns. boyd and Crawford (2012) emphasize the importance of ethically responsibly handling Big Data. Scandals such as Snowden's revelations about mass surveillance by US intelligence agencies (Steiger et al., 2017) or the collection of data from millions of Facebook users for the purpose of personalized advertising and election interference in the 2016 US presidential election by Cambridge Analytica (Hinds et al., 2020) have recently caused great public outcry. The public attention was, therefore, drawn more strongly to the issue of privacy. Policymakers are increasingly reacting to these concerns. For example, shortly after the Cambridge Analytica scandal became public, the European Union's General Data Protection Regulation came into force. This is considered an important step forward in the field of global political convergence processes and helps to create a global understanding of how to handle personal data (Bennett, 2018).

However, Big Data not only leads to privacy concerns, but also can undermine transparency for users of online services. This lack of transparency is further exacerbated by the fact that algorithms are sometimes too complex for laypersons to understand, which is often referred to as a black box (Shin and Park, 2019). Questions regarding comprehensibility and explainability are therefore at the core of algorithmic decision-making and its outcome (Ananny and Crawford, 2018). These questions become particularly relevant when algorithms make biased decisions and systematically discriminate against individual groups of people. For example, the COMPAS algorithm used by some US courts systematically disadvantaged black defendants by giving them a higher risk score for the probability of recidivism than white defendants (Angwin et al., 2016). In contrast, a hiring algorithm that Amazon developed and ultimately decided not to use systematically discriminated against female candidates (Köchling and Wehner, 2020). Algorithmic discrimination can be caused by flawed or biased input data or by the mathematical architecture of the algorithm (Shin and Park, 2019). Thus, AI systems run the risk of reproducing or even exacerbating existing social inequalities with detrimental effects for minorities. Such algorithmic unfairness then leads to the question of who is accountable for possibly biased decisions by an AI system (Busuioc, 2020; Diakopoulos, 2016). All of these questions have been extensively discussed in the fairness, accountability, and transparency in machine learning (FATML) literature (Shin and Park, 2019). The different concepts are closely intertwined. For example, Diakopoulos (2016) points out: 'Transparency can be a mechanism that facilitates accountability' (p. 58).

To address these concerns and to define common ground for (self-)regulation, governments, private sector companies, and civil society organizations have established ethics guidelines for developing and using AI. The goal is to address the challenges outlined by the scientific community and thus to ensure so-called 'human-centered AI' (Lee et al., 2017), or 'human-centric' AI (European Commission, 2019). For example, the High-Level Expert Group on AI (AI HLEG) set up by the European Commission calls for seven requirements of trustworthy AI: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) societal and environmental well-being, and finally (7) accountability (European Commission, 2019). Along similar lines, the OECD recommends a distinction between five ethical principles, namely (1) inclusive growth, sustainable development, and well-being; (2) human-centred values and fairness; (3) transparency and explainability; (4) robustness, security, and safety; and (5) accountability (OECD, 2021).

Some researchers have taken a comparative look at the numerous guidelines published in recent years and have highlighted which ethical principles are emphasized across the board (e.g. Hagendorff, 2020; Jobin et al., 2019). There is a widespread agreement on the need for ethical AI, but not on what it should look like in concrete terms.

Hagendorff (2020) highlights that the requirements for accountability, privacy, and fairness can be found in 80% of the 22 guidelines he analysed. Thus, to a large extent, the guidelines mirror the primary challenges for human-centric AI discussed in the FATML literature. At the same time, however, Hagendorff (2020) points out that it is precisely these principles that can be most easily mathematically operationalized and thus implemented in the technical development of new algorithms.

Jobin et al. (2019) conducted a systematic review of a total of 84 ethics guidelines from around the globe, although the majority of the documents originate from Western democracies. In total, the authors identify 11 overarching ethical principles, five of which (transparency, justice and fairness, non-maleficence, responsibility, and privacy) can be found in more than half of the guidelines analysed. Also, the attributes of beneficence and of freedom and autonomy can still be found in 41 and 34 of the 84 guidelines, respectively. The ethical principles of trust, sustainability, dignity, and solidarity, on the other hand, are only mentioned in less than a third of the documents (Jobin et al., 2019).

In this paper, we focus on the seven most prominent ethical principles, which are discussed in most of the existing guidelines, analysed by Jobin et al. (2019). In addition to the aforementioned principles of transparency (or explainability), fairness, responsibility (accountability), and privacy Jobin et al. (2019) list non-maleficence, freedom and autonomy, and beneficence. They conceive 'general calls for safety and security' (p. 394) as non-maleficence. At the core of the principles lies the requirement for the technical security of the system, for example, in the form of protection against hacker attacks. In this way, unintended harm from AI, in particular, should be prevented (European Commission, 2019). According to Jobin et al. (2019), the freedom and autonomy issue addresses, among other things, the risk of manipulation and monitoring of the process and decisions, as also addressed by the AI HLEG. In light of this challenge, implementing human oversight in the decision-making process can ensure that human autonomy is not undermined and unwanted side effects are thus avoided (European Commission, 2019). However, decision-making procedures are perceived as fair when the procedure guarantees a maximum degree of consistency on the one hand and is free from personal bias on the other hand (Leventhal, 1980). Therefore, the neutrality of an algorithmic decision – without human bias – might explain why algorithmic decision-making is perceived as fairer than human decisions (Helberger et al., 2020; Marcinkowski et al., 2020). Even though these perceptions are context-dependent (Starke et al., 2021), it can be assumed that in some use cases no human control is desired. For example, this is especially important to consider when personal bias or, at worst, corruptibility of human decision-makers could be suspected, that is, in tax fraud detection (Köbis et al., 2021). In this sense, the use of AI can lead to less biased decisions (Miller, 2018). Finally, beneficence refers to the common good and the benefit to society as a whole. However, reaping this benefit requires algorithms that do not make any mistakes. The accuracy of AI is therefore decisive for societal benefit. This is because only a high level of predictive accuracy or correct decisions made by an AI can generate maximum benefit (Beil et al., 2019). Accordingly, AI systems used in medical diagnosis, for instance, can only improve personal and public health if they operate as accurately as possible (Graham et al., 2019).

While all ethical principles highlighted in the AI ethics guidelines seem desirable in principle, they can cause considerable challenges in practice. The reason is that when designing an AI system, it is often infeasible to maximize the different ethical aspects simultaneously. Thus, multiple complex trade-off matrices emerge (Binns and Gallo, 2019; Köbis et al., 2021). Two examples help to illustrate this point. First, the more information available about a user's wants, needs, and actions, the more helpful and accurate recommendation algorithms can make on social media platforms. This information includes not only private data about a user, such as the browsing history, but also sensitive data, such as gender. Collecting this data and simultaneously improving the recommendation can result in accuracy–privacy (Machanavajjhala et al., 2011) or accuracy-fairness trade-offs (Binns and Gallo, 2019). Second, for a company to assess if its hiring algorithm discriminates against social minorities, it needs to collect sensitive information from its applicants, such as ethnicity, which may violate fundamental privacy rights, leading to a fairness-privacy trade-off (Binns and Gallo, 2019). By adding more variables such as transparency, security, autonomy, and accountability to the mix, highly complex trade-offs between the various ethical principles emerge.

Public preferences for AI ethics guidelines

Human-centric AI is an essential, yet fuzzy concept used in ethical AI research. For instance, some authors argue that AI can only be human-centric if, on the input side, it considers the sociocultural complexity of humans and, on the output side, it provides explanations that are easy to understand for laypeople (Riedl, 2019). Another concept of human-centric AI focuses on the overarching objective of AI, namely that it is used 'in the service of humanity and the common good, with the goal of improving human welfare and freedom' (European Commission, 2019, p. 4). A common denominator of those different concepts is that accounting for perceptions of those most affected by decisions made by algorithmic systems is a key strategy to achieve human-centric AI. A recent example from the UK illustrates that violating ethical principles when designing and implementing AI—in this case, an automated system that graded students in schools—can lead to substantial public outrage (Kelly, 2021). Empirical research further suggests that perceiving AI as unethical has detrimental implications for an organization in terms of a lower reputation (Acikgoz et al., 2020) as well as a higher likelihood for protests (Marcinkowski et al., 2020) and for pursuing litigation (Acikgoz et al., 2020). Thus, to address the fundamental question of which kind of AI we want as a society, detailed knowledge about public preferences for AI ethics principles is key. A surging strand of empirical research addresses this question and finds that public preferences for AI are highly dependent on the context, as well as on individual characteristics (Pew Research Center, 2018; Starke et al., 2021). While some people perceive algorithms to be acceptable in some domains (e.g. social media recommendation), they reject them in others (e.g. predicting finance scores). Also, judgments about AI hinge considerably on sociodemographic features, such as age or ethnicity. In the US, a study by the Pew Research Center (2018) identifies four major concerns voiced by respondents: (1) privacy violation, (2) unfair outcomes, (3) removing the human element from crucial decisions, that is the belief that some tasks can be better evaluated by humans, who does not solely rely on measurable characteristics (e.g. the inclusion of empathy in a decision), and (4) inability of AI systems to capture human complexity, which refers to the notion that algorithms cannot take into account individual aspects of humans in their decisions and therefore make inaccurate decisions.

The empirical literature further shows that people largely desire to incorporate ethical principles advocated for in the legal guidelines. First, people base their assessment of an AI system on its accuracy. The seminal study by Dietvorst et al. (2015) finds that people avoid algorithms after seeing them making a mistake, even if the algorithm still outperforms human decision-makers. Along similar lines, people lose trust in faulty AI systems (Robinette et al., 2017). However, studies have found that people still follow algorithmic instructions even after seeing them err (e.g. Salem et al., 2015). Second, fairness is a crucial indicator for evaluating AI systems (Starke et al., 2021). When an AI system is perceived as unfair, it can lead to detrimental consequences for the institution implementing such a system (Acikgoz et al., 2020; Marcinkowski et al., 2020). Third, empirical evidence suggests that keeping humans in the loop of algorithmic decisions, that is, ensuring human oversight at least at some points of the decision-making process, is perceived as fairer (Nagtegaal, 2021) and more legitimate (Starke and Lünich, 2020) compared to leaving decisions to algorithms. Fourth, in terms of transparency, the literature yields mixed results. On the one hand, more openness about the algorithm is essential for building trust in AI systems (Neuhaus et al., 2019), involving the users (Kizilcec, 2016), reducing anxiety (Jhaver et al., 2018), and increasing user experience (Vitale et al., 2018). On the other hand, studies show that too much transparency can impair user experience (Lim and Dey, 2011) and confuse users, complicating the interaction between humans and AI systems (Eslami et al., 2018). Further, the results of a conjoint survey conducted by König et al. (2022) show that while the transparency of AI systems is valued, the cost consumers may have to pay is seen as a more important feature. Fifth, privacy protection can be an essential factor for evaluating AI systems, leading people to reject algorithmic recommendations based on personal data (Burbach et al., 2018). However, other studies suggest that users are often unaware of privacy risks and rarely use privacy control settings on AI-based devices (e.g. Zheng et al., 2018). Sixth, empirical research suggests that people perceive unclear responsibility and liability for algorithmic decisions as one of the most crucial risks of AI (Kieslich et al., 2020). Furthermore, accountability and clear regulations are viewed as highly effective countermeasures to algorithmic discrimination (Kieslich et al., 2020). Along similar lines, other studies found that perceptions of accountability increase people's satisfaction with algorithms (Shin and Park, 2019) as well as their trust in algorithms (Shin et al., 2020). Lastly, in terms of security, people consider a loss of control over algorithms as a crucial risk of AI systems (Kieslich et al., 2020).

Only a few studies, however, compare the influence of different ethical principles on people's preferences. In several studies, Shin and colleagues tested the effects of three crucial aspects of ethical AI: fairness, transparency, and accountability. The results, however, are mixed. While fairness had the most substantial impact on people's satisfaction with algorithms (followed by transparency and accountability) (Shin and Park, 2019), transparency was the strongest predictor for people's trust in algorithms (followed by fairness and accountability) (Shin et al., 2020). Another study found that explainability had the most decisive influence on algorithmic trust (Shin, 2020). However, to the best of our knowledge, no empirical study has looked at different trade-off matrices between the various ethical principles and investigated people's preferences for single principles at the expense of others. Therefore, we propose the following research question:

RQ1: How do varying degrees of consideration of ethical principles in the design of an AI-based tax fraud detection system influence the public's preference for prioritization among them?

However, considering the diversity of social settings and beliefs in society, it is probable that there are trade-off differences among the public concerning the prioritization of ethical principles, respectively, ethical preference patterns of AI systems. Hence, we ask the following research question:

RQ2: Which preference patterns of ethical principles in the design of an AI-based tax fraud detection system exist in the German public?

The literature suggests that human-related factors influence the perception of AI systems. For example, empirical studies have found that age (Grgić-Hlača et al., 2020; Helberger et al., 2020), educational level (Helberger et al., 2020), self-interest (Grgić-Hlača et al., 2020; Wang et al., 2020), familiarity with algorithms (Saha et al., 2020), and concerns about data collection (Araujo et al., 2020) have effects on the perception of algorithmic fairness. Hancock et al. (2011) performed a meta-analysis of factors influencing trust in human–robot interaction and identified, among others, demographics and attitudes towards robots as possible predictors. Subsequently, we elaborate on this literature and test for differences among human-related factors concerning the emerging ethical design patterns. Hence, we ask RQ3.

RQ3: Which characteristics do people who favour a specific ethical design of an AI-based tax fraud detection system share?

Method

Sample

The data was collected via the online access panel (OAP) of the market research institute *forsa* between 16 March and 25 March 2021. The OAP is representative of the German population above 18 years of age, which at least occasionally uses the Internet. Respondents from the panel were randomly invited to participate in the survey, with each panellist having the same chance to be part of the sample. Altogether, 1204 people participated in the survey.

We cleaned the data according to three criteria: (1) low response time for the entire questionnaire (1 *SD* under average time), (2) high number of missing data in the entire questionnaire (2 *SD* above-average number of missing data), and (3) low reading time of the introduction text for the conjoint analysis (under 20 seconds reading time identified through a pre-test). Participants were excluded if *all* criteria were met. Consequently, one participant was excluded. Additionally, we excluded all respondents who rated all proposed systems in the conjoint part of the survey equally ($n = 104$). This data cleaning step was crucial, since those respondents showed no preferences for any configuration and, methodologically speaking, for those respondents, no variance can be explained in the conjoint analysis.

After data cleaning, 1099 cases remained. In total, our sample consisted of 593 (54.0%) women and 503 (45.8%) men, while 3 (0.3%) indicated nonbinary. The average age of the respondents was 47.1 years ($SD = 16.7$). Furthermore, regarding education level, 192 (17.5%) hold a low, 362 (32.9%) hold a middle, and 540 (49.1%) hold a high educational degree.[1]

Procedure

Initially, respondents were asked to answer several questions concerning their perception and opinion on AI. To evaluate the preference for ethical principles in the design of AI systems, we integrated a conjoint survey with seven attributes in the survey. The use case was an AI-based tax fraud detection system. Such systems are already in use in many European countries, for example, France, the Netherlands, Poland, and Slovenia (Algorithm Watch, 2020) and also in the state of Hesse in Germany (Institut für den öffentlichen Sektor, 2019). In our study, respondents were presented with a short text (179 words) describing the use case of AI in tax fraud detection. The text stated that these systems can be designed differently. We then described the seven key principles of AI ethics guidelines, which we derived from the review article by Jobin et al. (2019): explainability (as a measurement for the dimension 'transparency'), fairness, security (as a measurement for the dimension 'non-maleficence'), accountability (as measurement for the dimension 'responsibility'), accuracy (as a measurement for the dimension 'societal well-being'), privacy, and limited machine autonomy (for exact wording of the attributes, see Table 1). Notably, we chose to include machine autonomy as we assumed that in the special case of tax fraud detection, no human oversight might be preferred due to possible bias reduction. In the following, the ethical principles are also called 'attributes'.[2]

After reading the short introductory text, respondents were told that an AI system can have different configurations of the ethical principles. If the system satisfied a principle, it was indicated with a green tick; if the property was not met, it was marked with a red cross. Respondents were presented with a total of eight cards showing different compositions of AI systems in randomized order. The configurations varied only in the different fulfilment of the ethical principles. For each card, we asked respondents to indicate how much they preferred the configuration of ethical principles shown on the card. At the end of the questionnaire, respondents had to indicate some sociodemographic information.

Measurement

Conjoint design. The strength of conjoint surveys lies in the ability to analyse a variety of possible attributes simultaneously (Green et al., 2001). This is particularly relevant for attributes that can potentially offset each other in reality, as argued in the trade-offs of the ethical principles. While asking for the approval of the principles separately is likely to yield high scores across the board, conjoint surveys

Table 1. Desciption of the attributes.

Ethical principle	Description
Explainability	*Explanation of the decision*: Each/any person concerned is explained in a generally understandable way why the system has classified him/her as a potential tax fraudster.
Fairness	*No systematic discrimination*: No persons (groups) are systematically disadvantaged by the automated tax investigation.
Security	*State-of-the-art security technology*: The protection of the computer system against hacker attacks is always kept up to date with the latest security technology.
Accountability	*Full responsibility with the tax authority*: Should the automated tax investigation system lead to false accusations, the responsible tax authority bears full responsibility for any damage incurred.
Accuracy	*Virtually no errors in decision-making*: The automated identification of tax fraud by the computer system works almost without errors.
Privacy	*Exclusively earmarked use of data*: Only the necessary data is used by the automated tax investigation system. Any other use of the considered data is excluded.
Machine autonomy	*No human supervision*: The identification of suspicious cases remains the sole responsibility of the automated tax investigation system.

force respondents to make a choice between the imperfect configurations of the principles. Furthermore, conjoint surveys can also be conducted with a partial factorial design. Thus, it is possible to predict respondents' preferences for all possible combinations, even if they only rate a small fraction of them. As described above, we treated the seven most prominent ethical design principles outlined by Jobin et al. (2019) as attributes (transparency, fairness, non-maleficence, responsibility, beneficence, privacy, and machine autonomy). We chose sub-codes for some ethical principles to tailor the broad concepts to our use case of tax fraud detection. As attribute levels, we simply marked if an ethical principle was complied with or not.

To determine the different compositions of the cards used in our study, we calculated a fractional factorial design using a standard 'order' allocation method and random seed. This method produces an orthoplan solution in which combinations of attributes are well balanced (see Table 2).

Measures

System approval. The approval of each system configuration was measured using a single item on a seven-point Likert scale (1='do not like the presented system at all'; 7='really like the presented system').

Interest in AI. To gauge people's interest in AI, respondents were asked to rate four items on a five-point Likert scale (1='not true at all'; 5='very true'); for instance, 'In general, I am very interested in artificial intelligence' (see Supplemental material for exact wording). We used the four items to compute a highly reliable mean index ($M = 2.79$; $SD = 1.07$; $\alpha = 0.94$). Scale and wording were adopted from the *Opinion Monitor Artificial Intelligence* (Meinungsmonitor Künstliche Intelligenz, 2021).

Acceptance of AI in domains. Respondents were asked whether they support the use of AI in 14 different domains on a five-point Likert scale (1='no support at all'; 5='totally support'). For every domain, we grouped the support values (4 and 5) as acceptance for AI in the specific domain. Afterwards, we calculated a sum index of acceptance; thus, the sum index ranges from 0='support in none application domain' to 14='support in all application domains, $M = 3.96$ ($SD = 2.96$). The measurement was adapted from the *Opinion Monitor Artificial Intelligence* (Meinungsmonitor Künstliche Intelligenz, 2021).

Risk awareness of AI. We measured risk awareness of AI with three items by asking respondents: 'You can associate both advantages and disadvantages with artificial intelligence. Completely independent of how big you think a possible benefit is: How great do you think the risk posed by artificial intelligence is?' Respondents gave their opinion on their risk perception for the whole society and themselves, as well as for family and friends. The items were measured on a 10-point Likert scale (1='no risk at all' to 10='very high risk'). We adapted the measurement by Liu and Priest (2009) and calculated a highly reliable mean index ($M = 4.90$; $SD = 1.92$; $\alpha = 0.91$).

Opportunity awareness of AI. Along similar lines, respondents were asked to rate three items on a 10-point Likert scale to the question: 'Completely independent of the risk, how great do you think is the benefit to be gained from artificial intelligence'. Again, they had to rate the benefit perception for the whole society, themselves, and friends and family. We adapted the measurement by Liu and Priest (2009) and computed a highly reliable mean index ($M = 5.82$; $SD = 1.76$; $\alpha = 0.87$).

Trust in AI. We measured trust in AI with four items to the question 'How much do you trust systems of artificial intelligence already today…' on a 10-point Likert scale (1='do not trust at all' to 10='trust completely'). An example item read as follows: '…recognize patterns in large data sets'. We calculated a reliable mean index ($M = 5.81$; $SD = 1.73$; $\alpha = 0.76$). The question wording was adapted from Lee (2018). The items are based on the dimensions proposed by Kieslich et al. (2021).

Table 2. Orthoplan.

Card-ID	Explainability	Fairness	Security	Accountability	Accuracy	Privacy	Machine Autonomy
A	Yes	Yes	Yes	No	No	No	No
B	Yes	No	No	Yes	No	No	Yes
C	No	No	Yes	No	Yes	No	Yes
D	Yes	Yes	Yes	Yes	Yes	Yes	Yes
E	No	Yes	No	No	No	Yes	Yes
F	Yes	No	No	No	Yes	Yes	No
G	No	No	Yes	Yes	No	Yes	No
H	No	Yes	No	Yes	Yes	No	No

Results

All calculations were performed in R (V4.0.3). The analysis code, including R packages used, is available upon request.

Relative importance of ethical principles

To answer RQ1, we calculated a conjoint analysis in R. In particular, we computed linear regressions for every respondent with the attributes as independent variables (dummy coded) and the ratings of the cards as the dependent variable. Thus, 1099 regression models were calculated to show the preferences of every respondent; the regression coefficients are called the part-worth values (Härdle and Simar, 2015). We then calculated the average value of the regression coefficients to determine the preferences of the German population for an ethical design of AI.

Part-worth of attributes

Predictably, all regression coefficients (part-worths) were positive ($b_{Accountability} = 0.80$; $b_{Accuracy} = 0.64$; $b_{Explainability} = 0.57$; $b_{Fairness} = 0.66$; $b_{Autonomy} = 0.32$; $b_{Privacy} = 0.66$; $b_{Security} = 0.66$; SE for all $b = .032$). Hence, the compliance with every ethical principle, except for limited machine autonomy, positively influences the approval rating of an AI system. As mentioned earlier, machine autonomy can be seen as conducive to objectivity in some cases. Hence, it can be preferred to human oversight. In the given case, the respondents aim at an average for a solution where tax fraud is arguably detected unbiasedly.

We looked more closely at the differences in the importance of fulfilling the ethical principles, or, in other words, people's preferences for certain ethical principles over others. For that, we calculated the importance weights for each attribute (see Figure 1). Importance weights can be obtained by dividing each mean attribute part-worth by the total sum of the mean part-worths.

The results suggest that accountability is, on average, perceived as the most important ethical principle. Fairness, security, privacy, and accuracy are on average equally important to the respondents. The explainability of AI systems is slightly less important. Lastly, machine autonomy is the least important for the respondents. Though, the importance weights of the attributes are relatively balanced in aggregate.

Preference patterns among the public

To answer RQ2 and RQ3, we conducted k-means clustering. K-means clustering is a method used to split observations into k mutually exclusive groups, called clusters, whereby group members within a group are as similar as possible and as dissimilar as possible from other groups (Boehmke and Greenwell, 2020). Thus, k-means clustering provides solutions for a differentiation of respondents based on a given set of properties.

We used respondents' regression coefficients as cluster-forming variables. The number of clusters was determined using the within-cluster sum of square ('elbow') method with Euclidean distance measure. Euclidean distance measure was chosen since the cluster variables follow a Gaussian distribution and have few outliers. The results suggest a solution of $k = 5$ or $k = 11$ clusters. [3] Since we aim for a comprehensible cluster solution and k is commonly determined on convenience (Boehmke and Greenwell, 2020), we decided to choose the five-cluster solution in our analysis, as it is clearer to interpret and allows for further description and comparisons of the groups. Afterwards, we computed the k-mean clusters using the algorithm of Hartigan and Wong (1979) using 20 different starting points.

Figure 2 shows the preference profiles of the five cluster groups. The yellow group includes people who do not seem to care much about the ethical design of systems. The purple group values fairness, accuracy, and accountability. The green group demands privacy, security, and accountability. The blue group considers all ethical principles equally important. Finally, the main characteristic of people belonging to the red group is described through high disapproval of machine autonomy.

In the next step, we labelled the clusters and calculated the cluster sizes. Cluster 1 (red) was labelled as 'Human in the Loop' cluster 2 (blue) as 'Ethically Concerned' cluster 3 (green) as 'Safety Concerned', cluster 4 (purple) as 'Fairness Concerned', and cluster 5 (yellow) as

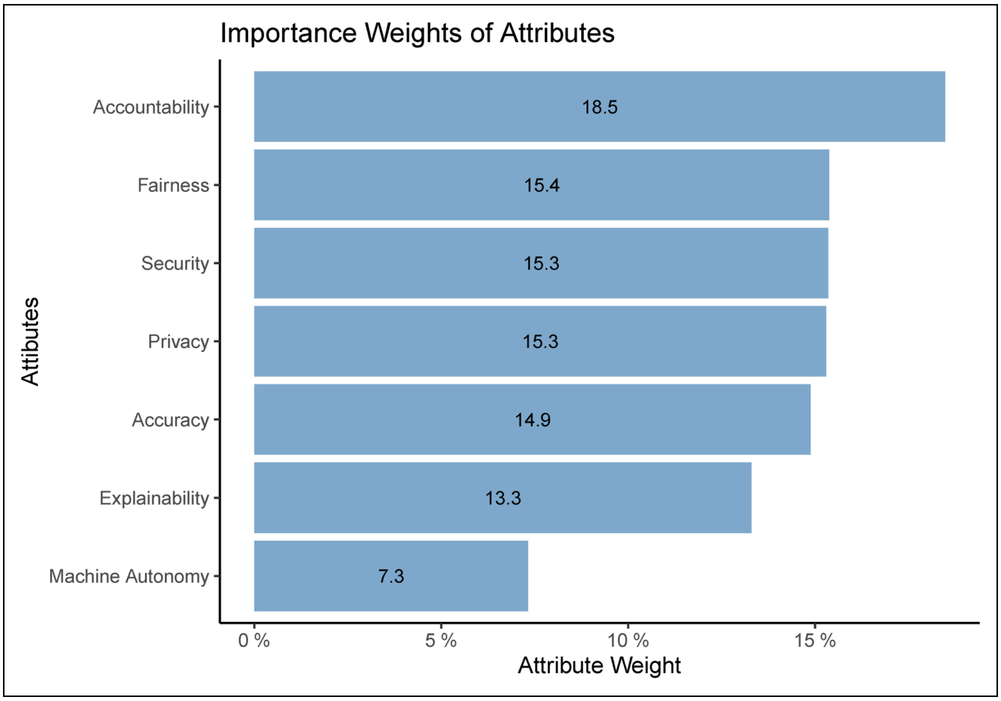

Figure 1. Overall importance weights of attributes.

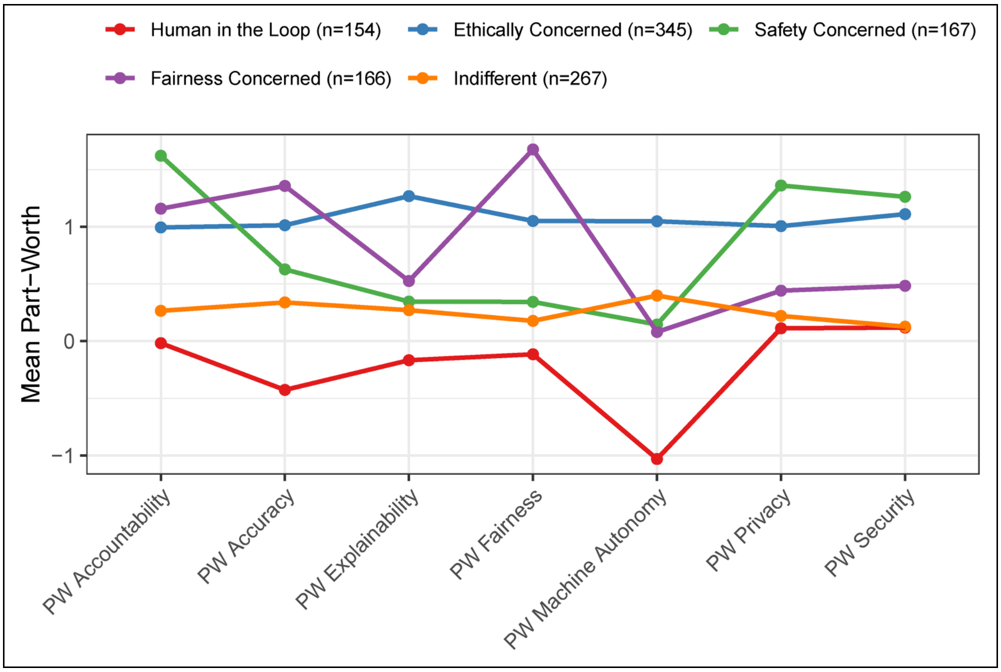

Figure 2. Preference profiles.

'Indifferent'. Table 3 depicts the average approval ratings for each cluster group per card and in total across all cards.

The results shown in Figure 2 suggest that the largest group consists of people who treat all ethical principles equally and highly important, $n = 345$ (31.4%). Hence, people in the *Ethically Concerned* group appreciate systems that satisfy all ethical principles. Otherwise, the approval ratings are quite low.

In contrast, the second largest group consists of people whose system approval ratings are only slightly affected

Table 3. Mean card ratings per group.

Group name	Card A	Card B	Card C	Card D	Card E	Card F	Card G	Card H	Average rating
Human in the loop	4.19	3.14	3.01	2.82	3.32	3.87	4.56	3.79	3.59
Ethically concerned	2.28	2.17	2.03	6.34	1.96	2.14	1.97	1.91	2.60
Safety concerned	2.38	2.54	2.47	6.14	2.28	2.77	4.68	3.02	3.29
Fairness concerned	3.02	2.10	2.25	6.05	2.53	2.66	2.42	4.52	3.19
Indifferent	3.18	3.54	3.46	4.40	3.40	3.43	3.21	3.38	3.50

by an ethical design of an AI system, $n = 267$ (24.3%). We label them as *Indifferent*. Those people do not seem to care much about the ethical design of the system. However, persons in this cluster group show a medium acceptance for all presented systems.

A total of 167 (15.2%) respondents were considered as *Safety Concerned*. For those, AI systems must be safe, privacy has to be protected, and the responsibility of a specific entity has to be ensured. These ethical principles are far more important than fairness, accuracy, or explainability. Across all presented systems, the approval ratings are on a low to medium level.

The group of *Fairness Concerned* consists of 166 (15.1%) respondents who especially consider fairness and accuracy to be the important principles, whereas privacy and security hardly affected positive ratings. The *Fairness Concerned* are quite sceptical towards the presented systems if they do not follow their demanded ethical principles.

In the fifth cluster, people oppose machine autonomy and accordingly strive for human control, $n = 154$ (14.0%). We term this group of respondents *Human in the Loop* as limited machine autonomy is the only factor that highly affects the ratings of the AI systems. Hence, for this group, it is relevant to build systems that are under human control. However, approval of the presented systems is on average at a medium level.

Cluster description

We address RQ3 by describing the five cluster groups based on several characteristics, which we group into two categories: socio-demography (*age*, *educational level*) and AI opinions (*interest*, *acceptance*, *risk awareness*, *opportunity awareness*, *trust*). We only included those respondents ($n = 913$), who answered all included variables (no missing values). In the first step, we calculated the mean values for each explanatory variable for each cluster group. To test for mean differences between the cluster groups, we ran a multivariate analysis of variance (MANOVA) with the cluster group as the independent grouping variable and the seven characteristics outlined above as dependent variables. We checked the assumptions and found homogeneity of variance–covariance matrices using Box's M test, $M = 137.62$, $p = 0.05$. As Box's M test is very sensitive, values lower than .001 are considered to be not trusted (Field, 2011). Further, we tested for normal distribution of the dependent variables with visual inspection and multivariate Shapiro–Wilk test. The Shapiro–Wilk test showed a significant deviance from normality, $W(913) = 0.98$, $p < 0.01$. Moreover, visual inspection revealed that the data was non-normal distributed. However, MANOVA is rather robust to a violation of normal distribution (Field, 2011). We used Pillais' trace test statistic, as it is the most robust test for violations of the underlying assumptions (Field, 2011).

As the MANOVA shows statistical significance, $V = 0.11, F(4{,}908) = 3.74$, $p < 0.01$, we performed subsequent analysis of variance (ANOVA) analyses for each of the dependent variables. Further, post-hoc tests with Tukey-honestly significant difference correction were used to test for mean differences between the groups for every dependent variable (see Table 4).

The ANOVA results show that the clusters significantly deviate from each other in all characteristics analysed. In the following, we will describe the profile of each cluster group in detail. All mean scores of the variables are displayed in Table 4.

Human in the loop. This group overwhelmingly demands human control and, thus, is strongly opposed to machine autonomy. Persons belonging to this group tend to be older and less educated. Regarding AI opinions, they are rather uninterested in AI and have a low acceptance of AI. Moreover, they are comparatively more aware of risks, have quite low opportunity perceptions, and low levels of trust in AI.

Ethically concerned. People who demand high standards on all ethical principles are comparatively young and well educated. They also have a high interest and trust in AI. Furthermore, they tend to accept AI in various domains and have a relatively high opportunity perception and a relatively low risk perception.

Safety concerned. Respondents belonging to the *Safety Concerned* group are located in between the other groups regarding the sociodemographic variables and AI opinions. They are, of average age and education. Furthermore, they are somewhat interested in AI, accept AI in some

Table 4. Cluster descritpion.

	Human in the loop	Ethically concerned	Safety concerned	Fairness concerned	Indifferent	*F*	*p*
Sociodemographics							
Age	49.52 (15.63)^{a}	44.87 (15.87)^{ab}	46.28 (16.48)^{a}	40.54 (17.66)^{b}	49.46 (16.66)^{a}	8.18	0.00
Educational level	2.24 (0.76)^{c}	2.49 (0.68)^{ab}	2.34 (0.76)^{bc}	2.60 (0.65)^{a}	2.13 (0.74)^{c}	12.72	0.00
AI opinions							
Interest	2.77 (1.11)^{ab}	3.00 (1.10)^{a}	2.97 (0.95)^{ab}	3.01 (0.95)^{a}	2.71 (1.00)^{ab}	3.44	0.01
Acceptance	3.58 (2.79)^{c}	4.52 (3.03)^{ab}	4.22 (2.77)^{bc}	5.15 (2.79)^{a}	3.71 (2.78)^{c}	7.83	0.00
Risk awareness	5.28 (1.93)^{ab}	4.82 (1.94)^{ab}	4.71 (2.04)^{bc}	4.18 (1.68)^{c}	5.29 (1.78)^{a}	9.27	0.00
Opportunity awareness	5.55 (1.69)^{bc}	6.04 (1.79)^{ab}	5.98 (1.77)^{abc}	6.18 (1.65)^{a}	5.54 (1.80)^{c}	4.75	0.00
Trust in AI	5.50 (1.75)^{b}	6.05 (1.60)^{a}	5.91 (1.58)^{ab}	6.38 (1.48)^{a}	5.57 (1.95)^{b}	7.39	0.00

Note: MANOVA significant using Pillais' trace test statistic, $p < .05$. Cells show mean values and standard deviation (in brackets) for the cluster groups. *F* and *p* show the significance of the subsequent ANOVAs performed. Means in a row without a common superscript (a–c) letter differ ($p < .05$) as analysed by the ANOVA and the TUKEY post-hoc test.

application domains, are medium risk and opportunity aware, and trust AI systems to an average extent.

Fairness concerned. The *Fairness Concerned* group is comparatively young and well educated. Out of all clusters, the *Fairness Concerned* also perceive the lowest risks and the greatest opportunities of AI. They further have the highest trust in AI systems, are most accepting of AI, and are one of the groups with the highest interest in AI.

Indifferent. The *Indifferent* can be described—together with the *Human in the loop*—as the group with the most negative opinions on AI. People who do not demand ethically designed systems have relatively low acceptance, low opportunity perceptions, and little trust in AI. Further, they have a high risk awareness and are comparatively uninterested in AI.

Discussion

This study sheds light on a crucial area of ethical AI, namely public perceptions of ethical challenges that come along with developing algorithms. Empirical insights into citizens' preferences for fundamental principles of ethical AI and the trade-offs between them are essential to advance the notion of human-centric AI. Our results can further have practical implications: They can inform developers about prioritizing certain ethical principles when designing AI systems. The findings also provide vital information for decision-makers tasked with implementing AI systems into society according to fundamental societal values.

In this paper, we investigated opinions about the ethical design of AI systems by jointly considering different essential ethical principles and shedding light on their relative importance (RQ1). We further explored different preference patterns (RQ2) and how these patterns can be described by sociodemographics as well as AI-related opinions (RQ3).

From ethical guidelines to legal frameworks?

Our results show that there are no major differences within the German population with regard to the relative importance of ethical principles. However, we find a slight accentuation of *accountability* as the most important ethical principle; moreover, the respondents consider *limited machine autonomy* slightly less important than the other ethical principles. Initially, these aggregate results indicate a balanced view on ethical AI. None of the ethical principles are strongly preferred over the other, leading to the conclusion that German citizens seem to have no critical blind spots. For a good rating of an AI system, all ethical principles are more or less equally important. Hence, developers and organizations should not neglect some ethical principles, while emphasizing others. Based on these results, it seems that compliance with multiple ethical principles is important for an AI system to receive a positive rating.

Thus, ethical guidelines are not only present in a vacuum, but also address the needs of the public. In the case of German citizens, accountability is foremost demanded. In the context of our study, accountability is equal to liability; hence, there is a need for a clear presentation of an actor, who can be accounted for losses and who—in the end—can be regulated. This is in line with empirical evidence showing that legal regulations are perceived not only as effective, but also as demanded countermeasures against discriminatory AI (Kieslich et al., 2020). As AI technology is considered a potential risk or even threat—at least among a share of the public (Kieslich et al., 2021; Liang and Lee, 2017)—setting up a clear legal framework for regulation might be a way to further enhance trust and acceptance toward AI. In this respect, the EU has already taken on a pioneering role, as the EU commission recently proposed a legal framework for the handling of AI technology (European Commission, 2021). With this, they set up a classification framework for high-risk technology and even list specific applications that should be closely controlled or even banned. Considering the results of our study, this might be a fruitful way to include citizen perceptions in this process and, for example, specifically make clear, who takes responsibility for poor decisions made by AI systems. Besides, it is the articulated will of the European Commission to put humans at the centre of AI development. Our empirical results suggest that ethical design matters and—if the EU takes their goals seriously—ethical challenges should play a major role in the future. Strictly speaking, ethical AI primarily requires regulatory political or legal actions. Hence, the implementation of ethical AI is a political task, which must not necessarily include computer scientists. However, from our results, we can also draw conclusions for the ethical design of AI systems in a technological sense.

Ethical design and demands of potential stakeholder groups

Our results also suggest that citizens value ethical principles differently. After clustering the respondents' preferences, we found five different groups that differ considerably in their preferences for ethical principles. This suggests that there might not be a universal understanding and balance of the importance of ethical principles in the German population. People have different demands and expectations regarding the ethical design of AI systems. Thus, these different preference patterns have implications for the (technical) design and implementation of AI systems. For example, the *Fairness Concerned* group should be addressed in different ways than the *Safety Concerned* or the *Human in the Loop* group. Several studies on the inclusion of stakeholders in the design process have already been conducted, especially for fairness (e.g. Webb et al., 2018).

Concerning the results of our study, for example, given the case of an algorithmic admission system in universities (Dietvorst et al., 2015), system requirements articulated by the affected public (in this case, students) might widely differ from those of a job seeker categorization system (e.g. the algorithmic categorization system used by the Austrian job service) (Allhutter et al., 2020). While students supposedly are younger, well educated, and more interested in AI, those affected by a job seeker categorization system are supposedly older and less positive about AI. Our results suggest that operators of AI systems should address the needs of the stakeholders differently if aiming for greater acceptance. For the admission system, it might be useful to highlight that such systems deliver precise results and treat students equally, since students—based on their group characteristics—primarily belong to the group of the *Fairness Concerned*. For the job seeker categorization system, on the other hand, it might be more promising to focus on safety issues or the presence of human responsibility, as most affected stakeholders may be assigned to the group of *Safety Concerned* or *Human in the Loop*. It should be noted that we explicitly highlight that we believe that every ethical design principle is of great importance and that developers should address all issues accordingly. We simply outline that communication about such systems could differ concerning the affected public.

Notably, there is also a group of people of substantial size (the *Indifferent*), who are only slightly concerned with the ethical design of AI systems. This group does not oppose AI systems in general (in fact, they have on average the second highest approval scores of all cluster groups for the presented systems), but they are not affected by compliance with ethical principles. This might be problematic, since this group arguably will not set high expectations for companies that develop AI systems. For example, Elzayn and Fish (2020) showed that achieving *fairness* in AI systems is very costly and that the market does not reward putting a massive amount of money into collecting data of marginalized groups, whether for monopolists or under competition. This becomes more alarming when considering the share of the *Indifferent* in society (24%). One might assume that the combination of lack of reward for ethical principles by the market and a potential lack of public outcry – at least in some parts of society – might lead to a sloppy implementation of ethical principles in practice. This is especially important to consider because ethical considerations are often left out of software development (McNamara et al., 2018). Again, Elzayn and Fish (2020) propose policy solutions to tackle this issue. Besides policy actions, organizations that are concerned with the ethical design of AI (e.g. *Algorithm Watch*) could actively reach out to the *Indifferent* and try to create awareness of the consequences of non-compliance with ethical principles. As it is part of the strategy of these organizations to fuel public awareness and discussion about AI across all parts of society, it could be beneficial to reach out to people who are currently unconcerned about ethical issues. According to our results, generating at least some interest as well as trust in the capacities of AI could lead to greater engagement with ethical design challenges.

The largest share of the German population equally values all ethical principles and, thus, sets very high standards for ethical AI development. In fact, this leads to the observation that the bar for approval of AI systems is very high for this group. However, if the principles are complied with, ethical AI can lead to high acceptance of AI. Common characteristics of this group are a high level of education, young age, and high interest in AI as well as high acceptance of AI. This group is especially demanding in terms of AI design. This may, in consequence, lead to a serious problem for AI design. As outlined, some trade-off decisions must be made eventually, as the simultaneous maximizing of all ethical principles is very challenging. However, our results suggest that it will not be easy to satisfy the demands of ethically concerned people. If some ethical trade-offs are taken, it may very well lead to reservations against AI.

However, considering only the public perspective in AI development and implementation might also have serious ramifications. Srivastava et al. (2019) show, regarding algorithmic fairness, that the broad public prefers simple and easy to comprehend algorithms to more complex ones, even if the complex ones achieved higher factual fairness scores. As AI technology is complex in its nature, it is possible that many people will not understand some design settings. In the end, this might lead to a public demand for systems that are easier to understand. However, it might very well be that a more thorough design of those systems would follow ethical principles to an even higher extent. Thus, we highlight that the public perspective on AI development definitely needs more attention in science as well as in technology development and implementation. We emphasize that the public perspective should rather complement and not dominate other perspectives on AI development and implementation.

Limitations

This study has some limitations that need to be recognized. We used an algorithmic tax fraud identification system as a use case in our study. Hence, our results are only valid for the specific context. However, as we wanted to describe preference profiles and cluster characteristics, we decided to present only one use case. This approach is similar to studies in the field of fairness perceptions (Grgić-Hlača et al., 2018; Shin and Park, 2019; Shin, 2021). However, public perceptions of AI are highly context-dependent. It might be that importance weights and cluster profiles differ concerning the particular use case. Therefore, further studies should test for various use cases simultaneously and compare the results regarding those contexts. Context-comparing studies have already been performed for public perception of trust in AI (Araujo et al., 2020) and threat perceptions (Kieslich et al., 2021).

Additionally, the survey was conducted in Germany, and the findings are thus only valid for the German population. We encourage further studies that replicate and enhance our study in other countries. Cross-national studies could detect specific nation patterns regarding the importance weights and preference profiles of ethical principles. The comparison to the US, Chinese and UK population would be especially interesting, since those countries follow a different national strategy for the development of AI than Germany.

Conclusion

Ethical AI is a major societal challenge. We showed that compliance with ethical requirements matters for most German citizens. To gain wide acceptance of AI, these ethical principles have to be taken seriously. However, we also showed that a notable portion of the German population does not demand ethical AI implementation. This is critical, as compliance with ethical AI design is, at least to some level, dependent on the broad public. If ethical requirements are not explicitly demanded, one might fear that implementation of those principles might not be on the highest standard, especially because ethical AI development is expensive. However, we showed that people who demand high-quality standards are interested in AI as well as aware of the risks. Thus, to raise demands for ethical AI, it would be a promising way to raise public interest in the technology.

Acknowledgements

The authors would like to thank Pero Došenović, Marco Lünich and Frank Marcinkowski for their feedback on the paper. Furthermore, the authors would like to thank Pascal Kieslich and Max Reichert for their feedback on the methodological implementation and evaluation of the results.

Declaration of conflicting interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was conducted as part of the project *Meinungsmonitor Künstliche Intelligenz* (opinion monitor artificial intelligence). From January 2020 to March 2021 the project was funded by the *Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen* (Ministry of Culture and Science of the State of North Rhine-Westphalia), Germany. Since April 2021 the project is funded by the *Stiftung Mercator*.

ORCID iDs

Kimon Kieslich https://orcid.org/0000-0002-6305-2997
Birte Keller https://orcid.org/0000-0002-3145-5206
Christopher Starke https://orcid.org/0000-0001-7899-6029

Supplemental material

Supplemental material for this article is available online. The analysis code is available on request.

Notes

1. Five persons (0.05%) in the sample did not indicate their educational level.
2. As the ethical principles outlined by Jobin et al. (2019) are rather broad, we consulted the guidelines of the EU commission (European Commission, 2019) for some formulations of the attributes. We take this measure as the German AI strategy is oriented on the EU guidelines and we aimed for a comprehensible design of the attributes.

3. Additionally, we searched for the optimal number of clusters using the Silhouette method as well as Gap statistic with 100 bootstrapping iterations. Results from the calculation with Silhouette method showed the best solution for k=2 or k=8 clusters. However, k=5 also represents a local maximum and can be considered as satisfying. Gap statistics reached highest values for k=2 and k=6. However, k=5 reached equally high values in the gap statistic. Taken together the results of the three search methods, we decided to chose k=5, since it depicts a good cluster solution in all methods used.

References

Acikgoz Y, Davison KH, Compagnone M, et al. (2020) Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment* 28(4): 399–416.

Algorithm Watch (2020) Automating Society Report 2020. Available at: https://automatingsociety.algorithmwatch.org/ (accessed 27 April 2021).

Allhutter D, Mager A, Cech F, et al. (2020) Der AMS-Algorithmus: Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-System (AMAS). Available at: https://epub.oeaw.ac.at/ita/ita-projektberichte/2020-02.pdf (accessed 6 May 2021).

Ananny M and Crawford K (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20(3): 973–989.

Angwin J, Larson J, Mattu S, et al. (2016) Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. May 2016. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 6 May 2021).

Araujo T, Helberger N, Kruikemeier S, et al. (2020) In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society* 35(6): 611–623.

Beil M, Proft I, van Heerden D, et al. (2019) Ethical considerations about artificial intelligence for prognostication in intensive care. *Intensive Care Medicine Experimental* 7(70): 1–13.

Bennett CJ (2018) The european general data protection regulation: An instrument for the globalization of privacy standards? *Information Polity* 23(2): 239–246.

Berendt B (2019) AI for the common good?! Pitfalls, challenges, and ethics pen-testing. *Paladyn, Journal of Behavioral Robotics* 10(1): 44–65.

Binns R and Gallo V (2019) Trade-offs. Available at: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-trade-offs/ (accessed 28 April 2021).

Boehmke BC and Greenwell B (2020) *Hands-on machine learning with R. Chapman & Hall/CRC the R series*. Boca Raton: CRC Press Taylor & Francis Group.

Boyd D and Crawford K (2012) Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5): 662–679.

Burbach L, Nakayama J, Plettenberg N, Ziefle M, Valdez AC, et al.(2018) *User preferences in recommendation algorithms*. In: Pera S, Ekstrand M, Amatriain X and O'Donovan J (eds) *Proceedings of the 12th ACM conference on recommender systems*, Vancouver, BC, Canada, 2–7 October 2018, pp. 306–310. New York, NY, USA: ACM.

Busuioc M (2020) Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review* 29(1): 4349.

Diakopoulos N (2016) Accountability in algorithmic decision making. *Communications of the ACM* 59(2): 56–62.

Dietvorst BJ, Simmons JP and Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology* 144(1): 114–126.

Elzayn H and Fish B (2020) *The effects of competition and regulation on error inequality in data-driven markets*. In: Hildebrandt M, Castillo C, Celis E, Ruggieri S, Taylor L and Zanfir-Fortuna G (eds) *Proceedings of the 2020 conference on fairness, accountability, and transparency*, Barcelona Spain, 27–30 January 2020, pp. 669–679. New York, NY, USA: ACM.

Eslami M, Krishna Kumaran SR, Sandvig C, et al. (2018) *Communicating algorithmic process in online behavioral advertising*. In: Mandryk R, Hancock M, Perry M and Cox A (eds) *Proceedings of the 2018 CHI conference on human factors in computing systems*, Montreal, QC, Canada, 21–26 April 2018, pp. 1–13. New York, NY, USA: ACM.

European Commission (2019) Ethics guidelines for trustworthy AI. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed 29 April 2021).

European Commission (2021) Proposal for a regulation of the European Parliament and of the council: Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. Available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=75788 (accessed 10 May 2021).

Field A (2011) *Discovering Statistics Using SPSS: (and Sex and Drugs and Rock 'n' Roll)*. 3rd ed., reprinted. Los Angeles, CA: Sage.

Graham S, Depp C, Lee EE, et al. (2019) Artificial intelligence for mental health and mental illnesses: An overview. *Current Psychiatry Reports* 21(11): 1–18.

Green PE, Krieger AM and Wind Y (2001) Thirty years of conjoint analysis: Reflections and prospects. *Interfaces* 31-(3_supplement): S56–S73.

Grgić-Hlača N, Redmiles EM, Gummadi KP, et al. (2018) *Human perceptions of fairness in algorithmic decision making*. In: Champin PA, Gandon F, Lalmas M and Ipeirotis PG (eds) *WWW '18: Proceedings of the 2018 world wide web conference*, Lyon, France, 23–27 April 2018, pp. 903–912. New York, NY, USA: ACM Press.

Grgić-Hlača N, Weller A and Redmiles EM (2020) Dimensions of diversity in human perceptions of algorithmic fairness, http://arxiv.org/pdf/2005.00808v1.

Hagendorff T (2020) The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines* 30(1): 99–120.

Hancock PA, Billings DR, Schaefer KE, et al. (2011) A meta-analysis of factors affecting trust in human–robot interaction. *Human Factors* 53(5): 517–527.

Härdle WK and Simar L (2015) *Conjoint measurement analysis*. In: Härdle WK and Simar L (eds) *Applied multivariate statistical analysis*, vol. 27. Berlin, Heidelberg: Springer, pp. 473–486.

Hartigan JA and Wong MA (1979) algorithm AS 136: A K-means clustering Algorithm. *Applied Statistics* 28(1): 100–108.

Helberger N, Araujo T and de Vreese CH (2020) Who is the fairest of them all? public attitudes and expectations regarding automated decision-making. *Computer Law & Security Review* 39: 105456.

Hinds J, Williams EJ and Joinson AN (2020) "It wouldn't happen to me": Privacy concerns and perspectives following the Cambridge analytica scandal. *International Journal of Human–Computer Studies* 143: 102498.

Institut für den öffentlichen Sektor (2019) Land Hessen: Künstliche Intelligenz in der Steuerfahndung. Available at: https://publicgovernance.de/html/de/8497.htm (accessed 27 April 2021).

Jhaver S, Karpfen Y and Antin J (2018) *Algorithmic anxiety and coping strategies of Airbnb hosts*. In: Mandryk M, Hancock M, Perry M and Cox A (eds) *Proceedings of the 2018 CHI conference on human factors in computing systems*, Montreal, QC, Canada, 21–26 April 2018, pp. 1–12. New York, NY, USA: ACM.

Jobin A, Ienca M and Vayena E (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1(9): 389–399.

Kelly A (2021) A tale of two algorithms: The appeal and repeal of calculated grades systems in england and Ireland in 2020. *British Educational Research Journal* 47(3): 725–741. doi:10.1002/berj.3705

Kieslich K, Lünich M and Marcinkowski F (2021) The threats of artificial intelligence scale (TAI). *International Journal of Social Robotics* 13(7): 1563–1577. doi:10.1007/s12369-020-00734-w

Kieslich K, Starke C, Došenović P, et al. (2020) Artificial intelligence and discrimination. Available at: https://www.cais.nrw/en/factsheet-2-ai-discrimination/ (accessed 1 June 2021).

Kizilcec RF (2016) *How much information?* In: Kaye J, Druin A, Lampe C, Morris D and Hourcade JPSEP (eds) *Proceedings of the 2016 CHI conference on human factors in computing systems*, San Jose, CA, USA, 7–12 May 2016, pp. 2390–2395. New York, NY, USA: ACM.

Köbis N, Starke C and Rahwan I (2021) Artificial intelligence as an anti-corruption tool (AI-ACT): Potentials and pitfalls for top-down and bottom-up approaches. https://arxiv.org/abs/2102.11567.

Köchling A and Wehner MC (2020) Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research* 13(3): 795–848.

König PD, Wurster S and Siewert MB (2022) Consumers are willing to pay a price for explainable, but not for green AI. Evidence from a choice-based conjoint analysis. *Big Data & Society* 9(1): 205395172110696.

Lee MK (2018) Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5(1): 1–16.

Lee MK, Kim JT and Lizarondo L (2017) *A human-centered approach to algorithmic services: considerations for fair and motivating smart community service management that allocates donations to non-profit organizations*. In: Mark G, Fussell S, Lampe C, Schraefel M, Hourcade JP, Appert C and Wigdor D (eds) *Proceedings of the 2017 CHI conference on human factors in computing systems*, Denver, CO, USA, 6–11 May 2017, pp. 3365–3376. New York, NY, USA: ACM.

Leventhal GS (1980) *What should be done with equity theory?* In: Gergen KJ, Greenberg MS and Willis RH (eds) *Social exchange*. Boston, MA: Springer US., pp. 27–55.

Liang Y and Lee SA (2017) Fear of autonomous robots and artificial intelligence: Evidence from national representative data with probability sampling. *International Journal of Social Robotics* 9(3): 379–384.

Lim BY and Dey AK (2011) *Investigating intelligibility for uncertain context-aware applications*. In: Landay J, Shi Y, Patterson DJ, Rogers Y and Xie X (eds) *Proceedings of the 13th international conference on ubiquitous computing – UbiComp'11*, Beijing, China, 17–21 September 2011, p. 415. New York, NY, USA: ACM Press.

Liu H and Priest S (2009) Understanding public support for stem cell research: Media communication, interpersonal communication and trust in key actors. *Public Understanding of Science* 18(6): 704–718.

Machanavajjhala A, Korolova A and Sarma AD (2011) Personalized social recommendations - accurate or private?. *Proceedings of the VLDB Endowment (PVLDB)* 4(7): 440–450. http://arxiv.org/pdf/1105.4254v1

Marcinkowski F, Kieslich K, Starke C, et al. (2020) *Implications of AI (un-)fairness in higher education admissions*. In: Hildebrandt M, Castillo C, Celis E, Ruggieri S, Taylor L and Zanfir-Fortuna G (eds) *Proceedings of the 2020 conference on fairness, accountability, and transparency*, Barcelona, Spain, 27–30 January 2020, pp. 122–130. New York, NY, USA: ACM.

McNamara A, Smith J and Murphy-Hill E (2018) *Does ACM's code of ethics change ethical decision making in software development?* In: Leavens GT, Garcia A and Păsăreanu CS (eds) *Proceedings of the 2018 26th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, Lake Buena Vista, FL, USA, 4–9 November 2018, pp. 729–733. New York, NY, USA: ACM.

Meinungsmonitor Künstliche Intelligenz (2021) Ein Instrument zur kontinuierlichen Beobachtung öffentlicher und veröffentlichter Meinung zu KI. Available at: https://www.cais.nrw/memoki/ (accessed 25 April 2021).

Miller AP (2018) Want less-biased decisions? Use algorithms. *Harvard Business Review*. Available at: https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms (accessed 4 May 2021).

Nagtegaal R (2021) The impact of using algorithms for managerial decisions on public employees' procedural justice. *Government Information Quarterly* 38(1): 101536.

Neuhaus R, Laschke M, Theofanou-Fülbier D, et al. (2019) *Exploring the impact of transparency on the interaction with an in-car digital AI assistant*. In: Janssen CP, Donker SF, Chuang LL and Ju W (eds) *Proceedings of the 11th International conference on automotive user interfaces and interactive vehicular applications: Adjunct proceedings*, Utrecht, Netherlands, 21–25 September 2019, pp. 450–455. New York, NY, USA: ACM.

OECD (2021) Recommendation of the Council on Artificial Intelligence. Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed 3 May 2021).

Pew Research Center (2018) Public attitudes towards computer algorithms. Available at: https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/ (accessed 4 May 2021).

Riedl MO (2019) Human–centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies* 1(1): 33–36.

Robinette P, Howard AM and Wagner AR (2017) Effect of robot performance on human–robot trust in time-critical situations. *IEEE Transactions on Human–Machine Systems* 47(4): 425–436.

Saha D, Schumann C, McElfresh DC, et al. (2020) *Human comprehension of fairness in machine learning*. In: Markham A, Powles J, Walsh T and Washington AK (eds) *Proceedings of the AAAI/ACM conference on AI, ethics, and society*, New York, NY, USA, 7–9 February 2020, p. 152. New York, NY, USA: ACM.

Salem M, Lakatos G, Amirabdollahian F, et al. (2015) *Would you trust a (faulty) robot?* In: Markham A, Powles J, Walsh T and Washington AL (eds) *Proceedings of the tenth annual ACM/IEEE international conference on human–robot interaction*, Portland, OR, USA, 2–5 March 2015, pp. 141–148. New York, NY, USA: ACM.

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

Shin D (2021) The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human–Computer Studies* 146: 102551.

Shin D and Park YJ (2019) Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior* 98: 277–284.

Shin D, Zhong B and Biocca FA (2020) Beyond user experience: What constitutes algorithmic experiences? *International Journal of Information Management* 52(3): 102061.

Srivastava M, Heidari H and Krause A (2019) *Mathematical notions vs. human perception of fairness*. In: Teredesai A, Kumar V, Li Y, Rosales R, Terzi E and Karypis G (eds) *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, Anchorage, AK, USA, 4–8 August 2019, pp. 2459–2468. New York, NY, USA: ACM.

Starke C, Baleis J, Keller B, et al. (2021) Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. http://arxiv.org/pdf/2103.12016v1.

Starke C and Lünich M (2020) Artificial intelligence for political decision-making in the European Union: Effects on citizens' perceptions of input, throughput, and output legitimacy. *Data & Policy* 2(e16): 1–17. doi:10.1017/dap.2020.19

Steiger S, Schünemann WJ and Dimmroth K (2017) Outrage without consequences? post-snowden discourses and governmental practice in Germany. *Media and Communication* 5(1): 7–16.

Vitale J, Tonkin M, Herse S, et al. (2018) *Be more transparent and users will like you*. In: Kanda T, Ŝabanović S, Hoffman G and Tapus A (eds) *Proceedings of the 2018 ACM/IEEE international conference on human–robot interaction*, Chicago, IL, USA, 5–8 March 2018, pp. 379–387. New York, NY, USA: NY: ACM.

Wang R, Harper FM and Zhu H (2020) *Factors influencing perceived fairness in algorithmic decision-making*. In: Bernhaupt R, Mueller F, Verweij D, Andres J, McGrenere J, Cockburn A, Avellino I, Goguey A, Bjørn P, Zhao S, Samson BP and Kocielnik R (eds) *Proceedings of the 2020 CHI conference on human factors in computing systems*, Honolulu, HI, USA, 25–30 April 2020, pp. 1–14. New York, NY, USA: ACM.

Webb H, Koene A, Patel M, et al. (2018) *Multi-stakeholder dialogue for policy recommendations on algorithmic fairness*. In: Gruzd A, Jacobsen J, Mai P, et al. (eds) *Proceedings of the 9th international conference on social media and society*, Frederiksberg, Denmark, 18–20 June 2018, pp. 395–399. New York, NY: ACM.

Zheng S, Apthorpe N, Chetty M, et al. (2018) User perceptions of smart home IoT privacy. *Proceedings of the ACM on Human–Computer Interaction* 2(CSCW): 1–20.