International Journal of Digital Health

Original Article Open Access

How Neutral are Algorithms? Users' Perspectives on Bias in Medical Artificial Intelligence

Andrea Weber* **and Tanja Henking**

Institute for Applied Social Sciences IFAS, University of Applied Sciences Würzburg-Schweinfurt, Würzburg, Germany

Abstract

Background: Concerns about bias in AI mitigate high expectations regarding its benefits. Biases can emerge at various stages throughout the lifespan of an AI-based system, from development to use. The impact of biased AI partly depends on its users.

Methods: We investigated users' perspectives and expectations regarding medical AI and its implications for fair healthcare delivery in Germany. A convenience sample of 164 participants (M_{age} = 34.6 years, *SD* = 13.6; 67.1% female) answered an online questionnaire with closed and open-ended answer format. General associations with AI-based systems and perceptions of fairness of AI in a given clinical case vignette in radiology was assessed.

Results: Participants associated with AI-based systems characteristics linked to perceived intelligence, with greater disagreement characteristics linked to anthropomorphism, and were overall indifferent for characteristics linked to safety. The majority of respondents expect AI to help to reduce and detect discrimination in the given case vignette. AI was perceived significantly more accurate and diligent than humans in evaluating diverse patients. 92.1% of participants prefer a collaboration between AI and physicians in treatment over AI only (3.0%) and physician only (4.9%).

Conclusion: Results suggest an overall openness to AI use in clinical practice. While AI is expected to have a positive effect on fair healthcare delivery, participants prefer that humans have control over final decisions, at least for now. Users' varying informational needs must be met in order to ensure acceptance and appropriate use of medical AI.

Publication History:

Received: August 19, 2022
Accepted: September 15, 2022
Published: September 17, 2022

Keywords:

Medical AI, Fairness, Human control, Public opinion

Introduction

Vulnerable groups are especially prone to unfair treatment in medicine [1]. While bias in medical contexts can pose a threat to the health-related outcomes of these groups, simply neglecting differences between subpopulations can itself lead to unfair treatment. Considering differences between subpopulations with regard to certain diseases and contexts might be necessary to deliver appropriate healthcare [2,3].

While human decision-makers, such as physicians, are prone to undesirable biases e.g. [4], the emergence of technology-based applications in medicine sparked hopes for a fairer healthcare. Artificial Intelligence and Deep Learning have been discussed as tools to enable a deeply individualized and precise medicine [5] as well as decrease discrimination by overcoming human shortcomings and detecting unfair practices in medical care [6]. While information processing in AI is not always explainable (black box), it is theoretically possible to make decisions transparent. Although mathematical and cognitive challenges exist in doing so, research on explainable AI (XAI) is growing (see [7]). Kleinberg et al. [6] pointed out that AI as an bias detector is mostly "an aspiration, not a prediction" (p. 30096) and highly dependent on regulation.

Artificial intelligence (AI) applications have been developed and tested in a wide range of clinical settings, however their implementation in clinical practice is still rare [8]. Experiences from other settings however suggest that AI-based tool might be biased themselves. A well-known example is the analysis conducted by Obermeyer et al. [9], who showed that an algorithm widely used to assign health benefits to people in the United States exhibited racial bias. Panch et al. [8] defined AI bias in healthcare as "the instances when the application of an algorithm compounds existing inequities in socioeconomic status, race, ethnic background, religion, gender, disability or sexual orientation to amplify them and adversely impact inequities in health systems" (p.1).

In their groundbreaking work, Friedman and Nissenbaum [10] proposed the first categorization of computer bias, in which they differentiated among three different types of biases. Preexisting biases reflect biases that are already present in society "through the explicit and conscious efforts of individuals or institutions, or implicitly and unconsciously, even in spite of the best of intentions" [10]. Technical bias occurs due to technical constraints and considerations, whereas emergent bias occurs during the real usage of algorithmic systems due to changing societal knowledge and/or contexts of use. While their work mainly focused on simple algorithms used in airline reservation systems, it still holds explanatory power today. However, technical developments that have taken place since the first publication of their model have led to more complex and sophisticated algorithms that have partly outgrown this categorization. In particular, the emergence of self-learning and self-developing algorithms has brought about new challenges for non-biased AI that emerge in every stage of the AI lifecycle. Additionally, the medical context holds specific challenges.

The operationalization of medical constructs is a challenging task, as they are frequently qualitative in nature and causal links are often not fully understood in the medical field. Selecting relevant

***Corresponding Author:** Andrea Webera, Institute for Applied Social Sciences IFAS, University of Applied Sciences Würzburg-Schweinfurt, Würzburg, Germany, Tel: +49 931 3511 8175; E-mail: andrea.weber@fhws.de

Citation: Weber A, Henking T (2022) How Neutral are Algorithms? Users' Perspectives on Bias in Medical Artificial Intelligence. Int J Digt Hlthc 2: 110. doi: https://doi.org/10.15344/ijdh/2022/110

in- and output variables as well as attempting to quantify factors and translate them into code can lead to systematic errors. Additionally, inappropriate training data can lead to biased AI. Existing inequalities in healthcare, such as poorer healthcare outcomes for underserved populations, are reflected in the medical data available to algorithms. If existing medical records are used to train a medical AI tool, inequalities in the medical system will therefore be transferred to the AI system [3]. However, even unbiased training data can lead to biased AI decisions if the training data are unbalanced. The use of so-called WEIRD (Western, educated, industrialized, rich, democratic countries) samples in scientific settings neglects population diversity [11]. Panch et al. [8] highlighted the importance of the fit between the contexts of development and deployment. An AI system might be appropriate in one context but not another. Zou and Schiebinger [12] suggested an AI labeling system that informs potential users about the specifics of the training data used and how they are annotated. In addition, the training process itself can lead to biases if they are designed to strive for maximizing overall algorithm's accuracy. Doing so targets accuracy for majority groups but tends to neglect accuracy for minority groups, since the latter is less important for overall accuracy rates [12]. Alternatively the inclusion of clinically relevant outcomes such as medical improvements for subpopulations as well as therapeutic usefulness have been suggested [2,13]. Self-learning system continue developing and learning after the initial training and validation phase, leading to specific challenges. Potential developments are difficult to predict beforehand or to identify and correct later. Regular evaluation of AI systems is thus necessary while they are in use [8].

To date, most research on AI bias has focused primarily on systems themselves and less on the people affected by those systems (i.e., potential users of medical AI). Their expectations of AI will not only determine whether AI will be accepted in medical practice but also how AI results will be interpreted and used—for example, when AI is used as a medical decision-making aid. This will influence the impact AI-based systems will have on healthcare delivery. Earlier research in human–computer interaction showed contradicting results of people's perception of algorithms. On one side phenomena such as automation bias, algorithmic appreciation and an overall tendency to over-trust automated systems have been reported [14–16], on the other hand there are opposing empirical studies showing an greater distrust in algorithmic decision making compared to humans and a less forgiving attitude of algorithmic versus human failure [17]. In a representative qualitative survey of Dutch citizens the majority of participants expected decisions made using AI to be fairer than those made by humans [18]. In another qualitative study focusing on marginalized groups the majority of participants were unfamiliar with the term "algorithmic discrimination" and, when explained, perceived the scope, impact, and complexity of AI bias as minimal, indicating a lack of awareness of potential AI bias [19]. Both of these empirical studies focused on AI-based systems in a variety of non-medical application settings. Whether these results also apply to AI-based medical systems has not yet been investigated. The fit between expectations and capabilities of AI-based systems will guide informational needs of potential medical AI users (both physicians and patients) to ensure beneficial use of AI-based systems in medicine.

In this paper, we explore perceptions and expectations of medical AI among users exposed to a potentially imperfect and biased AI system. We focus on three main research questions: (1) Which associations do potential users have with AI-based systems? (2) Do potential users expect AI to be neutral? (3) What impact on medical care do potential users expect from the emerging use of medical AI? We explore general associations with AI as well as opinions about a use case of an AI-based tool in radiology.

Methods

We conducted a quantitative online survey, using stratified random sampling to select five German regions based on a ranking that considers economic, demographic, and social indicators [20]. Eligible to participate were all participants aged 18 and above, currently living in Germany. Study information and invitations to participate, including (short) URLs and QR codes linking to the online survey, were posted on social media and published in local newspapers in the selected regions. Participants marked their informed consent before participation. The online questionnaire was accessible from May 2021 until October 2021.

Sample

A total of 167 participants completed the online survey (convenience sample), of whom 164 were included in the analysis. Three were excluded due to a lack of plausibility in response patterns. Of the included participants, 110 (67.1%) identified as women, 53 (32.2%) as men, and 1 (0.6%) as non-binary. Participants' average age was 34.6 years (SD = 13.6, range = 21–73), and 53.7% held a degree from a higher education institution. With regard to perceived expertise in AI, 63 (38.4%) stated that they had good or expert knowledge of AI systems, whereas 101 (61.6%) had only minimal knowledge. Participants who had never heard the term "artificial intelligence" were excluded from the analysis. A total of 105 (64.0%) participants indicated that they had experienced discrimination in the health sector before, of whom 18 (10.9%) had experienced it often or very often.

Measures

The questionnaire consisted of two parts: (1) general associations with AI and humans and (2) a use case of medical AI (clinical case vignette). All items were presented in German.

Associations with AI: Participants were asked for their associations with AI. To receive answers about overall associations with the term Artificial Intelligence, no definition of AI was given at this point of the questionnaire. 20 characteristics were presented in opposing pairs and rated on an eight-point Likert scale (1 to 8) to force choice. The 20 characteristics consisted of items from the Godspeed questionnaire and its subscales (perceived intelligence, anthropomorphism, and safety) [21]; items adapted from the Robotic Social Attributes Scale RoSAS [22]; and deducted items from Helberger et al. [18]. All items were additionally presented and rated for humans. Items were chosen based on relevance to the current study's focus on fairness perceptions. All translations and adaptations were discussed in a multidisciplinary team.

Clinical case vignette: A definition of discrimination by the German Federal Anti-Discrimination Agency was presented [1]. Four items on the perceived degree of discrimination in the medical field as well as the respondent's own prior experiences of discrimination followed. We used a clinical case vignette of a radiology department wherein computer tomography (CT) was in use. An AI-based tool had been introduced to help physicians evaluate CT scans by marking striking areas to guide their attention. An extension was used to later on in the questionnaire to describe a setting in which the AI-based tool operated on its own under physicians' supervision. The clinical case vignette and its extension are displayed in Figure 1.

Imagine a radiology practice that offers examinations using CT (computed tomography, "tube"). Up to now, the doctors have evaluated the images of the CT and made the diagnosis (situation 1).

For some time now, artificial intelligence has been used to support the doctors' work. The images produced by the CT are not only evaluated by the doctors in charge, but also by an artificial intelligence. The artificial intelligent tool automatically analyses the images and compares them with reference values. Abnormalities are marked in colour. The doctors can look at these areas again in more detail. The final diagnosis is still made by the doctors and communicated to the patients (situation 2).

Extension: In the radiology practice, it is now possible for artificial intelligence to evaluate images and make a diagnosis on its own. The doctors monitor the decisions of the artificial intelligence and continue to have personal conversations with the patients (situation 3).

Figure 1: Clinical Case Vignette.

Following the presentation of the original case vignette, participants were asked to rate whether they considered discrimination to be a problem in the given case and if the use of AI would reduce, increase, or help detect discrimination. Participants were then asked to indicate whether the AI-based system treated people with diverse backgrounds with equal accuracy and diligence. The same was asked for physicians. Answers were given on a four-point Likert scale ranging from "disagree" (1) to "agree" (4). Afterwards the extension of the clinical case vignette was presented to all participants. The final items of the questionnaire asked participants about their treatment preferences (solely AI, solely physician, or physician and AI) and allowed them to provide open-ended explanations.

All quantitative analyses were conducted using IBM SPSS version 26. Open-ended questions were coded independently by two raters, and discrepancies were discussed to reach consensus.

Results

General Associations with AI

The vast majority of participants expected AI-based systems to be reliable (91.5%), competent (87.8%), knowledgeable (90.2%), capable (90.2%), objective (81.1%), fact-driven (97.6%), considering (73.8%), reasonable (81.7%), unnatural (88.4%), unsocial (78.7%), strange (79.9%), peaceful (82.3%), insensitive (98.8%), unfeeling (95.7%), and without consciousness (86.0%). Results were more inconsistent for ease of manipulation (56.1%), trustworthiness (65.9%), trust (52.4%), transparency (69.5%), and danger (51.8%), on which participants split almost evenly. In terms of Bartneck et al.'s [21] classification, participants associated with AI-based systems strongly and with broad consensus characteristics linked to perceived intelligence, rather low and with greater disagreement characteristics linked anthropomorphism, and were overall indifferent for items linked to safety.

Diligence and accuracy of medical AI

Based on the clinical case vignette, participants were asked to evaluate the accuracy and diligence of AI-based systems when confronted with patients with diverse backgrounds. Answers were given on a four-point Likert scale (1 to 4). Respondents reported that they expected AI-based systems to generally treat different patients with the same diligence (M = 3.54, SD = .75) and accuracy (M = 3.34, SD = .79). Participants were also asked about their evaluations of humans regarding accuracy and diligence. On average, physicians were expected to be less accurate (M = 2.57, SD = .87) and to act with less diligence (M = 2.35, SD = .88) than AI-based systems when confronted with diverse patients. These differences were significant for both accuracy ($t(163) = -9.87$, $p < .01$, $d = -.77$) and diligence ($t(163) = -13.80$, $p < .01$, $d = -1.08$).

Preference for treatment

When asked for their preferred treatment delivery (only physician, only AI, or AI and physician), most participants indicated that they would prefer a collaboration between physicians and AI (92.1%; 3.0% AI only, 4.9% physician only).

A reoccurring theme in the open-ended section was the complementarity of human and AI capabilities. While physicians were assigned characteristics such as empathetic, experienced, creative, and intuitive, AI was perceived as objective and precise. Physicians' exhaustion and distraction were mentioned in favor of additional AI use. Approximately one-third of participants mentioned the complementary benefits of a collaboration between physicians and AI. One participant stated: "AI might provide a more thorough analysis that might be missed by the human eye. However, physicians are able to assess how severe and relevant the detected anomaly is" (Participant 82). Some participants also mentioned the necessity of empathy, personal contact, and human interaction in the given case vignette: "It is about my health, I trust a person more than AI because humans are empathetic and can put themselves in my position" (Participant 165). A few participants indicated that they expected a period of transition, in which medical AI will gradually take a more central role in healthcare in the future. Human involvement and supervision are currently still needed but might be less important, if not redundant, in the future. As one participant stated, "AI is always a work in progress—a physician possibly identifies nuances and irregularities that AI cannot detect yet" (Participant 18). Another respondent added: "There is no long-term experience with image analysis by AI yet. If it works well and is safe for a few years, I might reconsider my choice" (Participant 103). A lack of trust in current medical AI, which made human involvement indispensable, was also mentioned by some other participants. By way of explanation, one participant stated: "I have no trust in a diagnosis solely conducted by AI, simply because I don't know it and am not used to it" (Participant 51).

Discussion

Our results suggest that AI's perceived strengths include its ability to deliver correct clinical diagnoses. A large majority (over 90%) of respondents associated with AI characteristics such as reliability, knowledge, capability, and fact-drivenness. These characteristics are part of the intelligence scale to measure perceptions of robots developed by Bartneck et al. [21]. On the other hand, emotions are not associated with AI-based systems, with more than 95% of respondents viewing AI as both "unfeeling" and "insensitive." The majority of respondents in our study saw these characteristics as unique to humans. These results are in line with those of Helberger et al. [18], who also observed associations between AI and objectivity, on the one hand, and humans and emotions, on the other. Interestingly, some

Table 1: Diagnostic model performed well in both training and testing.

Characteristics	Group descriptives		Paired differences			
	M (SD)	M (SD)	M (SD)	*t* (*df* = 163)	*p*	Cohen's *d*
reliable–unreliable	Human	3.77 (1.34)	1.18 (2.01)	7.52	<.001	.59
	AI	2.59 (1.47)				
competent–incompetent	Human	3.70 (1.31)	.97 (1.84)	6.75	<.001	.53
	AI	2.73 (1.64)				
knowledgeable–ignorant	Human	3.54 (1.32)	1.19 (1.99)	7.65	<.001	.60
	AI	2.35 (1.46)				
capable–incapable	Human	3.28 (1.39)	1.48 (1.77)	10.72	<.001	.84
	AI	1.80 (1.13)				
easy to manipulate–hard to manipulate	Human	2.79 (1.28)	-1.34 (2.50)	-6.87	<.001	-.54
	AI	4.13 (2.12)				
subjective–objective	Human	2.66 (1.37)	-3.71 (2.65)	-17.94	<.001	-1.40
	AI	6.37 (1.85)				
fact-driven–intuitive	Human	4.89 (1.49)	3.16 (1.86)	21.79	<.001	1.70
	AI	1.73 (1.06)				
considering–hasty	Human	4.30 (1.55)	.98 (2.83)	4.41	<.001	.34
	AI	3.33 (2.13)				
reasonable–unreasonable	Human	4.14 (1.34)	.90 (2.42)	4.75	<.001	.37
	AI	3.24 (1.78)				
trustworthy–untrustworthy	Human	3.81 (1.44)	-.29 (2.20)	-1.70	.09	-.13
	AI	4.10 (1.70)				
transparent–opaque	Human	3.84 (1.53)	.16 (2.90)	.73	.47	.06
	AI	3.68 (2.21)				
natural–unnatural	Human	2.83 (1.68)	-3.80 (2.48)	-19.65	<.001	-1.54
	AI	6.63 (1.65)				
has my trust–does not have my trust	Human	3.72 (1.52)	-.81 (2.42)	-4.29	<.001	-.34
	AI	4.53 (2.02)				
social–unsocial	Human	3.05 (1.50)	-3.15 (2.53)	-15.97	<.001	-1.25
	AI	6.20 (1.78)				
compassionate–insensitive	Human	3.12 (1.42)	-3.98 (2.00)	-25.52	<.001	-1.99
	AI	7.10 (1.29)				
feeling–unfeeling	Human	2.02 (1.35)	-5.40 (2.00)	-34.63	<.001	-2.70
	AI	7.41 (1.19)				
conscious–unconscious	Human	1.84 (1.28)	-4.96 (2.27)	-27.95	<.001	-2.18
	AI	6.80 (1.76)				
strange–familiar	Human	5.47 (1.75)	2.31 (2.75)	10.74	<.001	.84
	AI	3.16 (1.75)				
dangerous–harmless	Human	4.45 (1.65)	.03 (2.23)	.17	.86	.01
	AI	4.41 (1.81)				
aggressive–peaceful	Human	4.70 (1.41)	-1.27 (2.02)	-8.07	<.001	-.63
	AI	5.98 (1.56)				

N = 164, 8-point Likert scale (1 to 8)

Citation: Weber A, Henking T (2022) How Neutral are Algorithms? Users' Perspectives on Bias in Medical Artificial Intelligence. Int J Digt Hlthc 2: 110. doi: https://doi.org/10.15344/ijdh/2022/110

participants saw emotions and empathy, which are highly associated with human decision-making, as interfering with fair decision-making. We did not specifically assess this issue in our study, but some comments in the open-ended section indicated that some participants viewed personal contact and human interaction as essential aspects of good and appropriate clinical decision-making. One reason for these deviating results might be the different settings of the two studies; personal contact might be deemed specifically important for medical decisions.

This perception is complemented by the expectation that AI will be both diligent and accurate when dealing with data on people with diverse backgrounds. On average, respondents strongly agreed that AI-based systems would provide equal and appropriate treatment in terms of accuracy and even more so in terms of diligence. The opposite was true for physicians. Respondents were overall more sceptical of physicians' provision of equal and appropriate treatment with regard to both accuracy and diligence. In particular, the belief in AI's accuracy is in contrast to current observations in AI development and training. As AI performance during training aims to maximize overall accuracy, accuracy for minority groups might not be appropriately considered and thus be lower than overall accuracy [12, 23]. The accuracy of AI-based systems is also highly dependent on the training data used [12]. As a result, AI is less capable of detecting diseases in minority groups than majority groups. Due to the quantitative nature of the questionnaire, this study was limited to an overall perception of accuracy that leaves room for interpretation. We are unable to confirm whether participants' high expectations of accuracy are due to a lack of awareness of potential differences in accuracy levels for different subpopulations or whether participants are in fact aware of this issue but assumed that an appropriate and balanced training dataset had been used in the study scenario. Both lines of interpretation might lead to different user needs as well as deviating outcomes. Lacking awareness might result in overall acceptance and, potentially, over-trust in AI-based systems, making education and awareness-raising activities regarding AI and its capabilities necessary. On the other hand, users with high expectations regarding appropriate training and evaluation of AI prior to its use might ask for information when confronted with AI-based tools in the medical field. These users might be disappointed when confronted with AI's limitations in real-life application. This might lead to mistrust and rejection of AI-based systems that do not meet these criteria. Future qualitative research could shed light on this issue.

According to Kleinberg et al. [6], AI may be able to function as a discrimination detector if proper regulation strategies are in place. Our participants shared this view, with the majority indicating that AI could help detect discrimination in the given clinical case vignette. Interestingly, some participants also indicated in the open-ended section that AI could also help reduce discrimination by putting physicians under pressure to perform better and discriminate less. This implies that, for at least some respondents, physicians' discriminatory behaviour is deemed attributable to negligence that can be avoided. This is in line with participants' lower ratings of physicians in terms of diligence when treating patients with different backgrounds, in which participants were almost evenly divided between those who believed physicians would provide diligent treatment for all patients and those who did not. Answers in the open-ended section also shed some light on the nature of this perception, especially lack of diligence. Participants mentioned internal factors such as limited cognitive resources due to exhaustion or distraction, similar to other research on biases in human decision-making (e.g. [24]). Our participants thus viewed AI as able to not only detect and reduce unfair treatment in healthcare due to its inherent characteristics but also avoid unfair treatment by physicians by acting as a sort of supervision tool.

Interestingly, over 90% of participants were still open to AI use in clinical settings, although most preferred that physicians be involved to varying degrees. One possible interpretation of these results is that participants see the potential value of AI for delivering accurate diagnoses but still lack experience with medical AI in real-life settings. The finding that a majority of respondents generally associated AI with strangeness rather than familiarity supports this hypothesis. Another interpretation is the importance of emotions and empathy—both associated with humans—for appropriate clinical decision-making. With the collaboration of physicians and AI-based systems, participants might expect to benefit from the accuracy of AI-based systems as well as the emotional characteristics of humans.

In sum, our results showed that participants are overall hopeful about the impact of AI on antidiscrimination and willing to accept AI as part of clinical practice. While AI can have a positive effect on fair healthcare delivery by being less prone to bias than humans, as well as by improving the performance of physicians and detecting existing biases, participants preferred physicians to remain involved. Physician involvement is seen as indispensable, at least for now, and to varying degrees, ranging from AI as a provider of second opinions to AI as a controlling tool for physicians' decisions. Considering issues of over-trust, automation bias, and mathwashing, our results suggest an overall tendency to overestimate AI's capabilities to deliver neutral and bias-free results. If the use of AI-based systems is intended to improve healthcare outcomes, focus on AI-based systems alone seems insufficient. Potential users of AI must be considered and educated on the capabilities of the AI-based systems with which they are confronted.

This entails questions regarding the informational needs of involved stakeholders, from patients to officials involved in regulation and decisions related to the application of medical AI. Considering the results of our study, patients might have varying degrees of prior knowledge and misconceptions about AI and its capabilities. Questions arise regarding what and how much information is needed to help without overwhelming patients. Physicians not only need to advise patients about the risks and benefits of medical AI use but might also be confronted with myths and misconceptions. This requires an appropriate level of understanding of the functioning of medical AI and issues such as the biases attached to it. Basic technical knowledge might be necessary for future physicians, especially regarding the nature and information-processing capabilities of AI. Appropriate understanding might also help physicians debunk their own potential misconceptions. What information and education is needed will also depend on improvements in the transparency, explainability, and interpretability of AI-based systems. In situations with deviating or unexpected AI suggestions, an appropriate understanding of and appropriate levels of trust in AI might help physicians differentiate between AI malfunctions that require human intervention and cases in which AI actually outperforms humans. Physicians' abilities to adequately differentiate between the two options described and the impact of their decisions linked to it can also potentially guide future trust levels of both patients and physicians.

Citation: Heironimus J, Stromberg A, Steiner RWP (2022) Detecting Covid-19 Respiratory Markers with Ordinary Mobile Phones. Int J Digt Hlthc 2: 109. doi: https://doi.org/10.15344/ijdh/2022/109

Limitations

Several limitations should be considered when interpreting the results of our study. First, our research sample was rather small and not representative of the German population, even less for other countries and contexts. Participants were overall young and well educated. Additionally, the survey was conducted solely online, with no paper-and-pencil option, which may have excluded people without digital skills or access and lead to a selection bias in recruitment. Our results can provide a first glimpse into perceptions of AI bias, but future research should address this issue at a larger scale and in other countries. Similar to the non-representativeness of our sample, it seems worth mentioning that the research team solely consisted of white, highly educated women. Second, the quantitative and descriptive nature of the study led to results that provide a general overview of the topic. Our results indicate future areas of research that need to be addressed with more in-depth study designs. Third, we used a fictitious case vignette in our study that was taken from a realistic scenario in radiology. Our participants based their answers on this specific application of AI use and a fictitious case. Future research should assess people's perceptions in other areas of medical AI use as well as (future) real-life scenarios.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

Andrea Weber: Conceptualization, methodology, formal analysis, investigation, resources, writing original draft, visualization

Tanja Henking: Conceptualization, supervision, writing- review& editing

References

1. Beigang S, Fetz K, Kalkum D, Otto M (2017) Diskriminierungserfahrungen in Deutschland. Ergebnisse einer Repräsentativ- und einer Betroffenenbefragung. Baden-Baden: Nomos.
2. Starke G, De Clercq E, Elger BS (2021) Towards a pragmatist dealing with algorithmic bias in medical machine learning. Med Health Care Philos 24: 341-349.
3. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, et al. (2020) Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit Med 3: 1-11.
4. Saposnik G, Redelmeier D, Ruff CC, Tobler PN (2016) Cognitive biases associated with medical decisions: a systematic review. BMC Medical Inform Decis Mak 16: 1-14.
5. Fogel AL, Kvedar JC (2018) Artificial intelligence powers digital medicine. NPJ Digit Med 1: 1-4.
6. Kleinberg J, Ludwig J, Mullainathan S, Sunstein CR (2020) Algorithms as discrimination detectors. Proceedings of the National Academy of Sciences 117: 30096-30100.
7. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, et al. (2018) A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51: 1-42.
8. Panch T, Mattie H, Atun R (2019) Artificial intelligence and algorithmic bias: implications for health systems. J Glob Health 9: 020318.
9. Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366: 447-453.
10. Friedman B, Nissenbaum H (1996) Bias in computer systems. ACM Transactions on Information Systems (TOIS) 14: 330-347.
11. Nugent SE, Jackson P, Scott-Parker S, Partridge J, Raper R, et al. (2020) Recruitment AI has a Disability Problem: questions employers should be asking to ensure fairness in recruitment. Institute for Ethical Artificial Intelligence.
12. Zou J, Schiebinger L (2018) AI can be sexist and racist - it's time to make it fair. Nature 559: 324-332.
13. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G (2018) Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 178: 1544-1547.
14. Benenson F (2016) Mathwashing. Facebook and the zeitgeist of data worship. Technical. Ly Brooklyn.
15. Goddard K, Roudsari A, Wyatt JC (2012) Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 19: 121-127.
16. Logg J, Minson J, Moore D (2019) Algorithm appreciation : People prefer algorithmic to human judgement, Organizational Behavior and Human Decision Processes, Bd. 151, S. 90-103.
17. Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General 144: 114–126
18. Helberger N, Araujo T, 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.
19. Woodruff A, Fox SE, Rousso-Schindler S, Warshaw J (2018) A qualitative exploration of perceptions of algorithmic fairness. Proceedings of the 2018 chi conference on human factors in computing systems, pp. 1-14.
20. Slupina M, Dähner S, Klingholz R, Reibstein L, Amberger J, et al. (2019) Die demografische Lage der Nation. Wie zukunftsfähig Deutschlands Regionen sind. Berlin: Berlin Institut für Bevölkerung und Entwicklung.
21. Bartneck C, Kulić D, Croft E, Zoghbi S (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International Journal of Social Robotics 1: 71-81.
22. Carpinella C M, Wyman A B, Perez MA, Stroessner SJ (2017) The robotic social attributes scale (rosas) development and validation. In Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction, pp. 254-262.
23. London A J (2019) Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent Rep 49: 15-21.
24. Bornstein BH, Emler AC (2001) Rationality in medical decision making: a review of the literature on doctors' decision-making biases. J Eval Clin Pract 7: 97-107.