Computers in Human Behavior 133 (2022) 107296

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Medical AI and human dignity: Contrasting perceptions of human and artificially intelligent (AI) decision making in diagnostic and medical resource allocation contexts

Paul Formosa [a,*], Wendy Rogers [a], Yannick Griep [b], Sarah Bankins [c], Deborah Richards [d]

[a] Department of Philosophy, Macquarie University, Australia
[b] Behavioural Science Institute, Radboud University, the Netherlands
[c] Department of Management, Macquarie University, Australia
[d] Department of Computing, Macquarie University, Australia

ARTICLE INFO

Keywords:
Artificial intelligence (AI)
Dignity
Respect
Interactional justice
Medical AI
Healthcare

ABSTRACT

Forms of Artificial Intelligence (AI) are already being deployed into clinical settings and research into its future healthcare uses is accelerating. Despite this trajectory, more research is needed regarding the impacts on patients of increasing AI decision making. In particular, the impersonal nature of AI means that its deployment in highly sensitive contexts-of-use, such as in healthcare, raises issues associated with patients' perceptions of (un) dignified treatment. We explore this issue through an experimental vignette study comparing individuals' perceptions of being treated in a dignified and respectful way in various healthcare decision contexts. Participants were subject to a 2 (human or AI decision maker) x 2 (positive or negative decision outcome) x 2 (diagnostic or resource allocation healthcare scenario) factorial design. We found evidence of a "human bias" (i.e., a preference for human over AI decision makers) and an "outcome bias" (i.e., a preference for positive over negative outcomes). However, we found that for perceptions of respectful and dignified interpersonal treatment, it matters more who makes the decisions in diagnostic cases and it matters more what the outcomes are for resource allocation cases. We also found that humans were consistently viewed as appropriate decision makers and AI was viewed as dehumanizing, and that participants perceived they were treated better when subject to diagnostic as opposed to resource allocation decisions. Thematic coding of open-ended text responses supported these results. We also outline the theoretical and practical implications of these findings.

1. Introduction

Despite concerns about "hype" and over-promising, various forms of Artificial Intelligence (AI) are already being used in clinical contexts (Rogers et al., 2021; Yin et al., 2021), while research into future uses of AI in healthcare is accelerating (Gerke et al., 2020; Lysaght et al., 2019). The forms of AI that are currently being used include both "assistive" AI, where a human clinician makes decisions with assistance from an AI (Aoki, 2021), and "autonomous" AI, where an AI makes a healthcare decision without human input (Lyell, Coiera, Chen, Shah, & Magrabi, 2021). While these advances raise various ethical concerns, including potential deskilling of healthcare workers (Rogers et al., 2021; Ross & Spates, 2020), further research is required to "assess the benefits and challenges associated with clinical AI applications" (Yin et al., 2021, p. 1), particularly from patients' perspectives (Lennartz et al., 2021). This perspective is important given research in other domains suggests that individuals subject to AI decision making dislike "being reduced to a percentage" by an algorithm and seek a "human touch" (Binns et al., 2018, p. 1; Bankins et al., 2022). While individuals generally want human interaction and to be treated with dignity and respect in decision making that impacts them, these preferences may be threatened by increasing use of AI in medicine (Shaikh, 2020). However, so far little is known about how AI decision making in healthcare impacts patients' perceptions of dignified interpersonal treatment. This leads to our study's driving research question: *How does AI decision making, compared to human decision making, impact individuals' perceptions of being treated in a dignified way in a healthcare context?*

To answer our research question, we developed an experimental

* Corresponding author.
E-mail addresses: Paul.Formosa@mq.edu.au (P. Formosa), Wendy.Rogers@mq.edu.au (W. Rogers), Yannick.Griep@ru.nl (Y. Griep), Sarah.Bankins@mq.edu.au (S. Bankins), Deborah.Richards@mq.edu.au (D. Richards).

https://doi.org/10.1016/j.chb.2022.107296
Received 10 November 2021; Received in revised form 18 March 2022; Accepted 2 April 2022
Available online 7 April 2022
0747-5632/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

survey study and constructed five healthcare decision-making scenarios that reflect the ways that AI is being (or is proposed to be) used in both diagnostic (e.g., diagnosing skin cancer) and resource allocation (e.g., distributing scarce ventilators) contexts. In each scenario we varied the decision maker (AI or human) and the decision outcome (positive or negative for the patient) and explored the impacts on a range of relevant measures, including perceptions of interactional justice (Bies, 2001, 2015), dehumanization (Binns et al., 2018), outcome satisfaction (Colquitt, 2001), and the appropriateness and trustworthiness of the decision maker (Chong, Zhang, Goucher-Lambert, Kotovsky, & Cagan, 2022). Our data show that whatever the outcomes, humans are consistently seen as appropriate decision makers and AI decisions are seen as dehumanizing. We also identify important differences between diagnostic and resource allocation decisions by showing that, in terms of respectful interpersonal treatment, it matters more who the decision maker is for diagnostic cases, and it matters more what the outcome is for resource allocation cases. Finally, we outline key differences in the ways that respect is understood by participants in these two decision contexts, before concluding with limitations and future research directions.

2. Literature review

AI has various current and potential uses in healthcare, including diagnosis, triage, screening, resource allocation, risk analysis, and surgical operations, in both assistive and autonomous modes. As AI moves into clinical practice, there is emerging research documenting types of AI applications and their clinical outcomes (Yin et al., 2021). However, this research has not yet focused on how patients view care being delivered by AI decision makers and how it impacts their relationship with their healthcare professionals. But the impact of such technologies on the patient-healthcare professional relationship is potentially significant, and the lack of information about the impact of AI on these relationships is problematic (Rogers et al., 2021). One key element to investigate is how a shift to AI decision making affects patients' perceptions about whether they are treated in dignified and respectful ways when subject to AI medical decision making.

Demands for respectful and dignified treatment are central to many ethical theories (Düwell et al., 2014; Formosa, 2017), and are particularly important in healthcare (Barclay, 2018). To explore our research question of how AI decision making impacts perceptions of dignified treatment in healthcare, we employ a range of related constructs. First, we use the construct of "interactional justice", which contrasts with procedural (were the procedures fair?) and distributive (were the outcomes fair?) understandings of justice (Cropanzano et al., 2001). Interactional justice refers to whether the "interpersonal treatment" of individuals in a decision-making process is experienced as respectful and dignified (Bies & Moag, 1986; Dai & Xie, 2016; Erdogan, 2002). Although interactional justice has been examined in relation to the experiences of healthcare workforces (Ghasi et al., 2020; H. Lee & Chui, 2019; Özer et al., 2017; Schlicker et al., 2021), it is less explored in the healthcare provider-patient relationship, despite its practicality and clear links to the types of treatment people expect from such relationships (e.g., Zhang et al., 2019). Perceptions of dignified and respectful treatment will also be influenced by several other factors. Who or what the decision maker is, and the individual's attitudes towards them, will shape their experiences of decision making. For example, people can develop expectations about the role appropriateness of a decision maker in certain contexts, such as expecting a trained healthcare professional rather than a bureaucrat or an autonomous AI to decide (e.g., Bigman & Gray, 2018). Such expectations can derive from the nature of the task, such as whether it is perceived as more or less objective (Lee, 2018). AI's role in decision making also raises concerns about reduced opportunities for human interaction, which is important for perceptions of dignified and respectful treatment. For example, the algorithmic and mathematical nature of AI can make people feel as if they have been dehumanized through "being reduced to a percentage" or number (Binns et al., 2018, p. 1; Lee, 2018; Lee, Jain, Cha, Ojha, & Kusbit, 2019). Feelings of dehumanization are thus the inverse of dignified treatment. Further, it has been argued that people will experience lower levels of trust when a decision is made by an AI rather than a human (Karunakaran, 2018), and being subject to untrustworthy decisions will likely be related to perceptions of interactionally unjust treatment. People's satisfaction with the outcome of a decision will also influence these perceptions of dignified treatment (Bankins et al., 2022), as individuals can form judgements of their treatment based on the specifics of the decision rendered and its impact upon them (Lipshitz, 1989). Taken together, incorporating perceptions of interactional justice, decision maker role appropriateness, dehumanization, trust, and outcome satisfaction should provide a broad perspective on the extent to which people subject to decisions made by others feel treated with dignity and respect. We outline our measures of these variables in our methods discussion.

To examine the impact of AI decision making on perceptions of dignified and respectful interpersonal treatment, we focus on two key factors likely to influence these perceptions: (1) who the decision maker is; and (2) what the decision outcome is. In a study focusing on AI use in human resource management (e.g., in recruitment and selection, training, and firing decisions), both factors were found to be significant in determining perceptions of dignified treatment (Bankins et al., 2022). However, further research is needed to assess whether similar effects hold in the context of healthcare and the very different patient-healthcare provider relationship.

Regarding the impact of who makes the decision, studies are increasingly contrasting perceptions of human compared to AI decision making. Lee (2018) shows that people perceive certain tasks to require skills that are either human-centered or mechanically-focused. That is, people are less likely to trust and have positive responses towards algorithmic systems undertaking the former types of tasks, but are more comfortable with these systems undertaking the latter types of tasks. There is also evidence that people perceive humans to have unique capacities for judgement, accounting for and expressing emotions, and contextualizing decision making which, for sensitive decisions such as those in medical contexts, are viewed as necessary but unavailable from AI (Binns et al., 2018; Lee et al., 2015). In reviewing the effects of automated and augmented decision making, Langer and Landers (2021) identify varied, and sometimes inconsistent, responses from potential healthcare consumers to forms of automated decision making in medical contexts. They found that individuals vary in how accurate they perceive diagnoses from automated systems to be and that they experience poorer trust in these systems compared to human decision makers (Langer & Landers, 2021). However, this appears contingent on which decisions the system is undertaking. Consistent with Lee's (2018) findings, decisions that are perceived to be morally significant could be viewed less favorably when undertaken by an automated system (Bigman & Gray, 2018), whereas more "mechanical" tasks performed by such systems may be seen as more acceptable (Palmisciano, Jamjoom, Taylor, Stoyanov, & Marcus, 2020).

Evidence about the general preference for human over AI decision making in healthcare is increasing, although the story is complex. In a series of studies Longoni et al. (2019) found that consumers are reluctant to utilize healthcare provided by AI, and they derive negative utility if the healthcare provider is automated rather than human. Further, they found that a key concern with AI was its perceived inferior ability to account for the "uniqueness" of peoples' characteristics and circumstances. These findings are supported by subsequent research. For example, Yokoi et al. (2021), in their online scenario study, found that people prefer human to AI decision makers for diagnostic and prescription decisions. Their participants were reluctant to trust the AI even if it performed at the same level as a human doctor. However, this study focused only on diagnostic decisions and did not consider impacts on perceptions of dignified interpersonal treatment. In a study examining the evaluative attitudes of patients and their relatives to the use of AI in

neurosurgery, it was found that most participants thought it was appropriate to use AI for various functions, including imaging interpretation, but that AI should not be fully autonomous (Palmisciano, Jamjoom, Taylor, Stoyanov, & Marcus, 2020). In a literature survey by Bhandari, Purchuri, Sharma, Ibrahim, & Prior, 2021, other concerns identified by patients include the lack of human interaction from AI for tasks such as radiology, and the potential lack of liability and accountability if an AI rather than a human makes an incorrect radiological diagnosis. However, only three of the fourteen studies examined in this review focused on patient populations, suggesting more research is needed on this key stakeholder group.

There is growing evidence regarding varying perceptions of AI when used for different decisions. Lennartz et al. (2021) surveyed patients scheduled for tomography or magnetic resonance imaging regarding the use of AI for diagnosing diseases. They found very strong preferences for physicians' opinions over an AI's opinion for most clinical tasks, and for diagnostic AI to be used under physician supervision rather than to be used autonomously. Likewise, Ongena et al. (2021) found that a representative sample of the Dutch population does not support the fully autonomous use of AI for diagnostic interpretation of screening mammograms, although there was some support for AI used as a secondary backup reader to humans. Attitudes to the use of AI in healthcare also vary by demographic. Yakar et al. (2021) surveyed attitudes of the general population toward AI use in radiology, robotic surgery, and dermatology. They found that trust in AI varied for different demographics, with more highly educated, employed, or student males of Western backgrounds not recently admitted to a hospital having more trust in AI than those in other groups. However, they concluded that the general population is more distrustful of AI use in medicine than the media might suggest.

Most of these studies focus on diagnostic uses of AI. Studies that focus on AI's role in healthcare resource allocation decisions, one of its current uses, are less common. One relevant study found that consumers believe that AIs used to allocate resources efficiently follow a maximizing or "consequentialist decision strategy" that involves "morally relevant tradeoffs". People saw these trade-offs as morally "objectionable", even if the overall outcome was optimal (Dietvorst & Bartels, 2021). Shaikh (2020) argues that the increasing deployment of AI-enabled decision support systems for healthcare resource allocation reflects a focus in healthcare on concrete outcomes, such as economic benefits, rather than on less tangible outcomes, such as equity or the quality of the decision-making process. This concrete focus leads to neglect of the human decision-making process, which requires empathy and intuition; Shaikh recommends the development of AI-based resource allocation systems that are human-centered. However, this research did not include empirical evidence on the attitudes of patients towards the use of AI for resource allocation in healthcare.

This literature demonstrates that various features of AI decision making are likely to impact perceptions of dignified treatment, with people more likely to regard interpersonal treatment as respectful when it involves human rather than AI decision makers. We can therefore expect there to be a "human bias" in favor of human over AI decision makers, all else being equal. From this, we develop our first hypothesis:

H1: When compared to a human decision maker, those subject to decisions by an AI in a healthcare context will perceive lower levels of interactional justice, lower levels of outcome satisfaction, lower perceptions of decision-maker role appropriateness, lower levels of trust, and higher levels of dehumanization.

Regarding the impact of the decision outcome, research shows that people evaluate the quality of decisions based on the valence (either positive or negative) of the outcome of those decisions, sometimes regardless of the process taken to reach them (Fischhoff, 1975). This constitutes an outcome rather than a process focus (Shaikh, 2020). Labelled as "outcome bias", it reflects that evaluations of decisions often occur after the fact, thereby incorporating outcome information, despite this not necessarily reflecting the quality of the decision (Lipshitz, 1989). Outcome bias exists when individuals view more favorably the quality or fairness of decisions that have positive outcomes for them, compared to decisions with negative outcomes, even when the decision-making process is identical. This suggests that positive decisions will be perceived as being more interactionally just. This bias tends to be amplified when individuals have little information to judge decision quality (Baron & Hershey, 1988), which is relevant to vignette studies where information is limited. Little research exists showing how decision valence impacts perceptions of dignified treatment by AI in healthcare, although evidence from one study of AI use in human resource management decision making shows that the outcome bias applies to AI decisions (Bankins et al., 2022). We can thus expect there to be an "outcome bias" in favor of positive over negative decision outcomes, all else being equal. This leads to our second hypothesis:

H2: When compared to decisions with positive outcomes, those subject to decisions with negative outcomes in a healthcare context will perceive lower levels of interactional justice, lower levels of outcome satisfaction, lower perceptions of decision-maker role appropriateness, lower levels of trust, and higher levels of dehumanization.

Our two hypotheses lead us to suppose that people will feel treated with more dignity and respect when they are subject to a human rather than an AI decision maker (H1) and when they receive a positive rather than negative outcome (H2). However, H1 and H2 conflict in the case of a human making a negative decision and an AI making a positive decision. In that case our two factors compete, which leads to our exploratory question: *In conflicted cases, will the decision maker or the decision outcome be more important for determining perceptions of dignified and respectful treatment?*

A further complication, so far unexplored in the literature, is how the decision maker and valence impacts perceptions about both diagnostic and resource allocation healthcare decisions. The latter decisions have a clear valence (i.e., getting the resource is positive and not getting it is negative), but the former decisions do not (i.e., getting a diagnosis may be experienced as a positive or negative outcome). Given that our vignettes were constructed with the patient suffering various symptoms, we suggest that participants will experience being diagnosed with a specific condition as a positively valenced outcome, and not being diagnosed as a negatively valenced outcome, since without a diagnosis their symptoms remain unaccounted for. This assumption was confirmed by the data, as shown below. Exploring both diagnostic and resource allocation cases is also important as the latter raises issues around moral trade-offs (i.e., do I give the scarce resource to this person or that person?) that are not raised as directly by the former (Dietvorst & Bartels, 2021). Further, patients may be more willing to accept AI use for more "mechanical" tasks, such as allocating scarce resources efficiently, than tasks that need a more human-touch and involve judgment (Lee, 2018), such as diagnosis, but there are no empirical studies to confirm this. This leads to our final exploratory question: *Will perceptions of dignified and respectful treatment be higher in medical diagnostic or resource allocation decision contexts?* We pose this as an exploratory question since, although the background literature gives us reasons to suppose that diagnostic and resource allocation contexts are sufficiently different that they may lead to different perceptions of dignified and respectful treatment, there are no strong grounds to suppose a priori which decision problem type will lead to higher perceptions of these variables than the other.

3. Methods

3.1. Research design

Experimental surveys afford examinations of how individuals respond to particular situations through the construction of hypothetical scenarios that alter factors theorized to influence their responses (Wallander, 2009). We subjected our respondents to a 2 (human or AI decision maker) X 2 (positive or negative decision outcome) X 2 (diagnostic

or resource allocation healthcare scenario) factorial design. The diagnostic or resource allocation healthcare scenarios were created based on existing literature about the medical uses (current and near term) of AI for diagnostic (e.g., Grote & Berens, 2020; Yin et al., 2021) and resource allocation (e.g., Shaikh, 2020) cases. Specifically, we created three diagnostic cases with respect to (1) an anxiety disorder (see e.g., Nemesure et al., 2021), (2) macular disease (see e.g., Abràmoff, Lavin, Birch, Shah, & Folk, 2018), and (3) skin cancer (see e.g., Reiter, Rotemberg, Kose, & Halpern, 2019). Moreover, we created two resource allocation cases that involved distributing: (4) a kidney to a person on a transplantation waiting list (e.g., Freedman, Borg, Sinnott-Armstrong, Dickerson, & Conitzer, 2020; Schwantes & Axelrod, 2021), and (5) a ventilator to a patient when the number of patients requiring ventilation exceeded the available number of ventilators (see e.g., George et al., 2021; Yu et al., 2021). Because diagnosis and resource allocation are different types of scenarios—diagnostic cases raise issues around accuracy (i.e., was the AI's diagnosis accurate?) (e.g., Tschandl et al., 2020) whereas resource allocation cases raise issues around fairness (i.e., was AI allocating resources in that way fair?) (e.g.., Shaikh, 2020) and because our literature review tells us that these cases may be perceived differently by patients as one could be seen as more "mechanical" than the other (i.e., requiring analysis of quantitative data using objective measures, rather than subjective or intuitive assessments - Lee, 2018, p. 4) —we retained them as distinct in our analysis. We examined our 2 (human or AI decision maker) X 2 (positive or negative decision outcome) X 2 (diagnostic or resource allocation healthcare scenario) cases as summarized in Tables 1 and 2. As discussed above, for diagnostic cases, receiving a diagnosis for a set of symptoms was counted as a positive outcome (labelled as AI_D + or H_D+) and not getting diagnosed as a negative outcome (labelled as AI_D- or H_D-). For resource allocation cases (referred to as F – fairness), receiving the resource was counted as a positive outcome (labelled as AI_F + or H_F+) and not getting it as a negative outcome (labelled as AI_F- or H_F-).

3.2. Materials

The structure and length of each vignette was consistent and reflected the following information: (1) the vignette focus (i.e., the healthcare context and decision maker); (2) the importance of the situation to the individual/patient; (3) the data the healthcare decision was based upon—each vignette identified several pieces of data relevant to the decision including a mix of objectively derived data (e.g., heart rate or number of dependents) and, where appropriate, a subjectively derived piece of data (e.g., past lifestyle choices or facial expressions)—to ensure the information used in the decision was clear and consistent across human and AI versions; and (4) the decision outcome. To support external validity and best practice vignette development (Aguinis & Bradley, 2014), each vignette was reviewed by several academics, including one with expertise in both bioethics and medicine. The full text of all vignettes is accessible in the online supplementary materials.

Table 1
Vignette groupings by decision maker and outcome valence for the twelve diagnostic vignettes.

Groups	Specific vignettes
AI– Diagnostic Group [AI_D-]: AI is the decision maker and the decision is negative (i.e., no diagnosis is made)	• AI decision maker/no diagnosis of anxiety disorder • AI decision maker/no diagnosis of macular deterioration • AI decision maker/no diagnosis of skin cancer
AI+ Diagnosis Group [AI_D+]: AI is the decision maker and the decision is positive (i.e., a diagnosis is made)	• AI decision maker/diagnosis of anxiety disorder • AI decision maker/diagnosis of macular deterioration • AI decision maker/diagnosis of skin cancer
H– Diagnosis Group [H_D-]: A human is the decision maker and the decision is negative (i.e., no diagnosis is made)	• Human decision maker/no diagnosis of anxiety disorder • Human decision maker/no diagnosis of macular deterioration • Human decision maker/no diagnosis of skin cancer
H+ Diagnosis Group [H_D+]: A human is the decision maker and the decision is positive (i.e., a diagnosis is made)	• Human decision maker/diagnosis of anxiety disorder • Human decision maker/diagnosis of macular deterioration • Human decision maker/diagnosis of skin cancer

Table 2
Vignette groupings by decision maker and outcome for the eight resource allocation vignettes.

Groups	Specific vignettes
AI– Resource Group [AI_F-]: AI is the decision maker and the decision is negative (i.e., the resource is withheld)	• AI decision maker/kidney not allocated • AI decision maker/ventilator not allocated
AI+ Resource Group [AI_F+]: AI is the decision maker and the decision is positive (i.e., the resource is given)	• AI decision maker/kidney allocated • AI decision maker/ventilator allocated
H– Resource Group [H_F-]: A human is the decision maker and the decision is negative (i.e., the resource is withheld)	• Human decision maker/kidney not allocated • Human decision maker/ventilator not allocated
H+ Resource Group [H_F+]: A human is the decision maker and the decision is positive (i.e., the resource is given)	• Human decision maker/kidney allocated • Human decision maker/ventilator allocated

Individuals who received vignettes with AI as the decision maker, termed "an AI healthcare app", received the following explanation: "The term 'AI healthcare app' refers to an app used in a healthcare context that relies to some extent on the use of artificial intelligence (AI). An AI system may, for example, make inferences and predictions based on data". Individuals who received vignettes with a human as the decision maker, termed "human healthcare professional", received the following explanation: "The term 'healthcare professional' is used to refer to a relevant human healthcare provider, such as a General Practitioner (GP), medical specialist, or nurse, depending on what is most relevant for the scenario". An example diagnostic vignette (for macular deterioration) is given below with the manipulated components identified in square brackets:

> You are worried about your eyes. Recently you have had trouble reading and you need a very bright light to see properly. Sometimes you see dark patches or things look distorted. [A human healthcare professional/An AI healthcare app] looks at your eyes, takes and examines a scan that looks at the back of your eyes (your retinas), and completes a checklist about your symptoms. Using these data, the [healthcare professional/AI healthcare app] diagnoses you as [having/not having] a serious eye condition called macular deterioration.

Resource allocation vignettes followed the same format, but instead of using the stated data to make a diagnosis, the decision maker (AI or human) uses it to allocate (or withhold) a resource.

3.3. Procedures

Participants were recruited through CloudResearch. This is a data services provider that draws on the working adult North American population. Such online panels are established as reliable sources for accessing diverse samples (Landers & Behrend, 2015), with the quality of data comparable to that from a non-paid random sample (Behrend et al., 2011) when researchers embed (as we did) attention checks in the

survey. We utilized a within-person design, with each participant invited to complete up to three randomly assigned vignettes from our pool of 20 vignettes (including both diagnostic and resource allocation scenarios). To minimize spill-over effects, or response tendencies, from one vignette to another vignette, our survey had restrictions in place to ensure participants did not receive combinations of vignettes that were confusing or contradictory. For example, participants would not have received scenarios where they were and were not diagnosed with the same condition, or scenarios where they did and did not receive the same resource. Further, the order of the vignettes that participants read was randomized, which minimized the potential for order effects. Participants were also instructed to read each vignette independently of the others and we used page breaks between vignettes to encourage this. After reading each vignette, participants were invited to complete survey measures (detailed below). We received ethics approval from our University's Human Research Ethics Committee (ref no. 52020938822719) to undertake our study and participants gave informed consent.

3.4. Sample

We recruited 743 North American participants to take part in the 20-min study for which we paid US$5.50 per participant. Upon reviewing the attention checks embedded in the survey, we removed 265 participants who failed to correctly answer two or more of our four attention checks (our most stringent cleaning process), resulting in a final sample of 478 individuals who completed the study. While participants were invited to complete three randomly assigned vignettes from our pool of 20 vignettes, some only completed two vignettes (235 respondents completed three vignettes and 243 respondents completed two vignettes). Participants were, on average, 43.57 years old (SD = 15.98), 60.67% were female and 39.12% were male. Most of our sample had University degrees (34.73% with undergraduate qualifications and 20.50% with graduate or postgraduate degrees) and 77.82% identified as Caucasian, with 9.21% identifying as Black or African American as the next most common category. In terms of marital status, 47.28% were married, 7.95% were in a de facto relationship, and 44.77% were single. Our respondents came from a wide range of sectors (top three listed here): health services (14.02%); education (10.46%); and financial services (8.79%).

3.5. Measures

We utilized the following measures for our variables of interest (collected after each vignette) with all demonstrating good reliability. We also collected several control variables (collected once at the end of the survey). In keeping with best practice recommendations (e.g., Becker et al., 2016), we compared a model in which we included our control variables with a model in which we did not include them. We found no significant effects of our control variables on the variables of interest and thus removed our control variables from the results reported below (per Bernerth & Aguinis, 2016). Participants were also invited to provide open-ended qualitative responses following the survey questions regarding interactional justice (e.g., "Can you provide further details as to why you thought the [healthcare professional's/AI healthcare app's] decision was fair or unfair?"). The measures for our variables of interest are now outlined.

Interactional Justice was measured by Colquitt's (2001) 4-item scale ranging from 1 "to a small extent" to 5 "to a large extent". An example item is: "Has the [healthcare professional/the AI healthcare app] treated you with dignity?" ($\alpha = 0.85$).

Outcome Satisfaction was measured by Colquitt's (2001) 2-item scale ranging from 1 "strongly disagree" to 5 "strongly agree". An example item is: "The outcome I received is acceptable" ($\alpha = 0.88$).

Dehumanization was measured by Bastian and Haslam's (2011) 5-item scale ranging from 1 "strongly disagree" to 5 "strongly agree". An example item is: "The [healthcare professional/the AI healthcare app] is treating me as if I were an object" ($\alpha = 0.89$).

Decision-Maker Role Appropriateness was measured by a single item constructed by the authors ranging from 1 "very inappropriate" to 7 "very appropriate". The item is: "In this scenario, how appropriate is it to have the [healthcare professional/the AI healthcare app] make this decision?"

Trust was measured by Körber's (2019) 2-item scale ranging from 1 "strongly disagree" to 5 "strongly agree". An example item is: "I trust the [healthcare professional/the AI healthcare app]" ($\alpha = 0.87$).

3.6. Analytical strategies

The data were entered into the R software package (R Core Team, 2021) for MANOVA analysis. The qualitative open-ended responses were entered into NVivo for thematic coding. Since participants were invited to complete three randomly assigned vignettes, the unit of analysis is the number of completed vignettes rather than the number of respondents. This resulted in 1191 observations overall, as 235 respondents completed three vignettes and 243 respondents completed two vignettes. These 1191 observations were spread across our three diagnostic scenarios (707 observations) and our two resource allocation scenarios (484 observations), noting again that individual respondents could have received a mix of both allocation and diagnostic scenarios. The number of observations per vignette ranged from 47 to 75, with an average of 60 observations for each of our 20 vignettes. Finally, due to the structure of our data (i.e., respondents completing multiple vignettes), we first calculated the percentage of variance in our variables of interest that is due to either within-person or between-person differences. This calculation indicated that the largest percentage of the variance in our variables of interest was attributable to between-person differences (ICC values were all below 0.05), indicating that we should not account for the nested structure of our data (Maas & Hox, 2005). Thus, any variance due to respondents completing multiple vignettes is close to zero, suggesting a negligible effect (that can be ignored) from the same person completing multiple vignettes.

MANOVA analyses. We conducted a MANOVA with our eight [2 (human or AI decision maker) X 2 (positive or negative decision outcome) X 2 (diagnostic or resource allocation healthcare scenario)] groups for our measures of outcome satisfaction, interactional justice, dehumanization, decision-maker role appropriateness, and trust as our dependent variables. We conducted Bonferroni corrected post-hoc contrast analyses.

Qualitative data. The qualitative data from the free text responses were thematically analyzed. Adopting a bottom-up "inductive analysis" allowed the organic emergence of themes from the data (Braun & Clarke, 2006, p. 83). Themes were identified at a "latent or interpretative level" (Braun & Clarke, 2006, p. 84) by coding whole passages with mentioned themes. We used investigator triangulation to ensure that different perspectives informed the thematic coding and to achieve intercoder consistency. Our process involved two of the researchers jointly coding the data from two vignettes to develop a coding scheme that captured the range of themes in the data. One of the researchers then coded the entire dataset with the coding scheme, with a second researcher checking some coding for consistency and reliability.

4. Results

4.1. Descriptive statistics

Table 3 provides an overview of the means, standard deviations, and correlations among the study's five dependent variables across all the observations in our study.

Table 3
Means, standard deviations, and correlations with confidence intervals.

Variable	M	SD	1	2	3	4
1. Outcome satisfaction	3.22	1.10				
2. Dehumanization	3.00	0.95	-.36** [-.40, −.30]			
3. Interactional Justice	3.34	1.04	.64** [.60, .67]	-.45** [-.49, −.40]		
4. Role appropriateness	4.61	1.81	.56** [.52, .60]	-.36** [-.41, −.31]	.55** [.50, .59]	
5. Trust	3.27	1.05	.77** [.75, .79]	-.42** [-.47, −.38]	.66** [.63, .69]	.65** [.62, .68]

Note. M and SD are used to represent means and standard deviations, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

4.2. Preliminary checks

Prior to presenting our MANOVA results, we assessed homogeneity and normality of the residuals for each of our five outcome variables across diagnostic and resource allocation healthcare scenarios. These preliminary checks indicate that (a) equality of variances can be assumed ($p > .05$ for Levene's test) for all but two of our cases (for outcome satisfaction resource allocation scenarios and for dehumanization in diagnostic scenarios where $p < .05$ for Levene's test) and; (b) for all outcomes normality of residuals cannot be assumed ($p < .001$). However, these violations are unproblematic as shown by Lumley et al. (2002); using a simulation study with extreme non-normal data, these scholars have demonstrated that normality of residuals is a widely but incorrectly held belief with regards to the assumptions underlying (M) ANOVA and regression.

4.3. MANOVA results

Results from our MANOVA indicated a significant difference between the eight [2 (human or AI decision maker) X 2 (positive or negative decision outcome) X 2 (diagnostic or resource allocation healthcare scenario)] groups for all dependent variables: outcome satisfaction [$F(7, 1183) = 15.66$, $p < .001$, $\eta2 = 0.085$]; interactional justice perceptions ([$F(7, 1106) = 12.27$, $p < .001$, $\eta2 = 0.072$]; mechanistic dehumanization perceptions [$F(7, 1183) = 24.03$, $p < .001$, $\eta2 = 0.124$]; decision-maker role appropriateness perceptions [$F(7, 1183) = 22.42$, $p < .001$, $\eta2 = 0.117$] and trust perceptions [$F(7, 1183) = 17.65$, $p < .001$, $\eta2 = 0.095$]. See Table 4 for an overview.

The results for each dependent variable are presented in Table 5 for resource allocation and Table 6 for diagnostic cases. Tables 5 and 6 are ordered according to decision maker (showing bias for human decision makers), decision outcome (showing bias for positive outcomes [i.e., allocation of the resource or diagnosis made]), and conflicting cases where the pro-human and pro-outcome biases conflict or accord. A higher score is considered "better" for all variables, except for the dehumanization variable where lower is considered "better". We now narrate the results by reporting statistically significant differences ($p < .05$) between our groups from the MANOVA analysis for our five dependent variables, while omitting mention here of non-significant differences (see full details in Tables 5 and 6).

In terms of decision makers, while holding the decision outcome constant, we found, reflecting a pro-human bias, participants report statistically significant better ratings in resource allocation cases when:

Table 4
Overview of MANOVA Results for Significant Difference Between the Eight [2 (human or AI decision maker) X 2 (positive or negative decision outcome) X 2 (diagnostic or resource allocation healthcare scenario)] Groups for all Five Dependent Variables.

	Human X positive X resource allocation	AI X positive X resource allocation	Human X negative X resource allocation	AI X negative X resource allocation	Human X negative X diagnostic	AI X negative X diagnostic	Human X positive X diagnostic	AI X positive X diagnostic			
	M (SE)	M (SE)	M (SE)	M (SE)	M (SE)	M (SE)	M (SE)	M (SE)	F	p	W^2
Outcome satisfaction	3.57 (.10)	3.55 (.10)	2.77 (.09)	2.69 (.10)	3.34 (.08)	3.00 (.08)	3.53 (.08)	3.34 (.08)	15.66	<.001	.09
Interactional Justice	3.55 (.10)	3.37 (.11)	3.06 (.09)	2.80 (.10)	3.55 (.07)	3.14 (.07)	3.69 (.07)	3.39 (.08)	12.27	<.001	.07
Dehumanization	2.62 (.09)	3.31 (.08)	3.01 9.08)	3.52 (.08)	2.78 (.07)	3.26 (.07)	2.50 (.07)	3.17 (.07)	24.03	<.001	.12
Role appropriateness	4.92 (.16)	4.28 (.16)	4.61 (.14)	3.66 (.15)	5.28 (.13)	3.84 (.13)	5.45 (.12)	4.55 (.13)	22.42	<.001	.12
Trust	3.63 (.10)	3.14 (.09)	3.23 (.08)	2.701 (.09)	3.55 (.08)	2.93 (.07)	3.69 (.07)	3.23 (.08)	17.65	<.001	.10

Note for Table 4: M = mean value and SE = standard error.

Table 5
Overview of MANOVA results for resource allocation cases for the five dependent variables.

	H_F+ vs. AI_F+	H_F- vs. AI_F-	H_F+ vs. H_F-	AI_F+ vs. A_F-	H_F- vs. AI_F+	H_F+ vs. AI_F-
Outcome satisfaction	3.57 (H+) = 3.55 (AI+)	2.77 (H-) = 2.69 (AI-)	3.57 (H+) > 2.77 (H-)	3.55 (AI+) > 2.69 (AI-)	2.77 (H-) < 3.55 (AI+)	3.57 (H+) > 2.69 (AI-)
Interactional justice	3.55 (H+) = 3.37 (AI+)	3.06 (H-) = 2.80 (AI-)	3.55 (H+) > 3.06 (H-)	3.37 (AI+) > 2.80 (AI-)	3.06 (H-) = 3.37 (AI+)	3.55 (H+) > 2.80 (AI-)
Dehumanization	2.62 (H+) < 3.31 (AI+)	3.01 (H-) < 3.52 (AI-)	2.62 (H+) < 3.01 (H-)	3.31 (AI+) = 3.52 (AI-)	3.01 (H-) < 3.31 (AI+)	2.62 (H+) < 3.52 (AI-)
Role appropriateness	4.92 (H+) > 4.28 (AI+)	4.61 (H-) > 3.66 (AI-)	4.92 (H+) = 4.61 (H-)	4.28 (AI+) > 3.66 (AI-)	4.61 (H-) = 4.28 (AI+)	4.92 (H+) > 3.66 (AI-)
Trust	3.63 (H+) > 3.14 (AI+)	3.23 (H-) > 2.70 (AI-)	3.63 (H+) > 3.23 (H-)	3.14 (AI+) > 2.70 (AI-)	3.23 (H-) = 3.14 (AI+)	3.63 (H+) > 2.70 (AI-)

Table 6
Overview of MANOVA results for diagnostic cases for the five dependent variables.

	H_D+ vs. AI_D+	H_D- vs. AI_D-	H_D+ vs. H_D-	AI_D+ vs. A_D-	H_D- vs. AI_D+	H_D+ vs. AI_D-
Outcome satisfaction	3.53 (H+) = 3.34 (AI+)	3.34 (H-) > 3.00 (AI-)	3.53 (H+) = 3.34 (H-)	3.34 (AI+) > 3.00 (AI-)	3.34 (H-) = 3.34 (AI+)	3.53 (H+) > 3.00 (AI-)
Interactional justice	3.69 (H+) > 3.39 (AI+)	3.55 (H-) > 3.14 (AI-)	3.69 (H+) = 3.55 (H-)	3.39 (AI+) = 3.34 (AI-)	3.55 (H-) = 3.39 (AI+)	3.69 (H+) > 3.14 (AI-)
Dehumanization	2.50 (H+) < 3.17 (AI+)	2.78 (H-) < 3.26 (AI-)	2.50 (H+) < 2.78 (H-)	3.17 (AI+) = 3.26 (AI-)	2.78 (H-) < 3.17 (AI+)	2.50 (H+) < 3.26 (AI-)
Role appropriateness	5.45 (H+) > 4.55 (AI+)	5.28 (H-) > 3.84 (AI-)	5.45 (H+) = 5.28 (H-)	4.55 (AI+) > 3.84 (AI-)	5.28 (H-) > 4.55 (AI+)	5.45 (H+) > 3.84 (AI-)
Trust	3.69 (H+) > 3.23 (AI+)	3.55 (H-) > 2.93 (AI-)	3.69 (H+) = 3.55 (H-)	3.23 (AI+) > 2.93 (AI-)	3.55 (H-) > 3.23 (AI+)	3.69 (H+) > 2.93 (AI-)

Note for Table 5–6: Presented values are mean values of the five listed dependent variables. The symbol ">" refers to significantly higher ($p < .05$) mean scores on the dependent variable; the symbol "<" refers to significantly lower ($p < .05$) mean scores on the dependent variable; and the symbol " = " refers to no significant difference in mean scores on the dependent variable. A higher score is considered "better" for all variables except for dehumanization.

(1) A human makes a positive decision compared to an AI making a positive decision for the dehumanization, appropriateness, and trust variables, and (2) A human makes a negative decision compared to an AI making a negative decision for the dehumanization, appropriateness, and trust variables. In diagnostic cases participants report statistically significant better ratings when: (1) A human makes a positive decision compared to an AI making a positive decision for the interactional justice, dehumanization, appropriateness, and trust variables, and (2) A human makes a negative decision compared to an AI making a negative decision for all five dependent variables.

In terms of decision outcomes, while holding the decision maker constant, we found, reflecting a bias for positive outcomes, participants report statistically significant better ratings in resource allocation cases when: (1) A human makes a positive decision compared to a human making a negative decision for the outcome satisfaction, interactional justice, dehumanization, and trust variables, and (2) An AI makes a positive decision compared to an AI making a negative decision for the outcome satisfaction, interactional justice, appropriateness, and trust variables. In diagnostic cases participants report statistically significant better ratings when: (1) A human makes a positive decision compared to a human making a negative decision for the dehumanization variable, and (2) An AI makes a positive decision compared to an AI making a negative decision for the outcome satisfaction, appropriateness, and trust variables.

In terms of the two conflicting cases (positive AI decision vs. negative human decision, and negative AI decision vs. positive human decision), we found that participants report significantly better ratings in resource allocation cases when: (1) A human makes a positive decision compared to an AI making a negative decision for all five dependent variables, and (2) A human makes a negative decision compared to an AI making a positive decision for the dehumanization variable AND when an AI makes a positive decision compared to a human making a negative decision for the outcome satisfaction variable. In diagnostic cases participants report significantly better ratings when: (1) A human makes a positive decision compared to an AI making a negative decision for all five dependent variables, and (2) A human makes a negative decision compared to an AI making a positive decision for the dehumanization, appropriateness, and trust variables.

Finally, in terms of our final exploratory question, "Will perceptions of dignified and respectful treatment be higher in medical diagnostic or resource allocation decision contexts?", we compared respondents' mean scores on our dependent variables when comparing diagnostic and resource allocation healthcare scenarios. Our results indicated that respondents in the diagnostic scenarios ($M = 3.31$, $SD = 1.06$) compared to respondents in the resource allocation scenarios ($M = 3.11$, $SD = 1.15$) reported significantly higher outcome satisfaction, $t(483) = 3.80$, $p < .001$. Moreover, we found that respondents in the diagnostic scenarios ($M = 3.45$, $SD = 0.98$) compared to respondents in the resource allocation scenarios ($M = 3.17$, $SD = 1.10$) reported significantly higher perceptions of interactional justice, $t(458) = 5.53$, $p < .001$. Next, we found that respondents in the diagnostic scenarios ($M = 2.92$, $SD = 0.91$) compared to respondents in the resource allocation scenarios ($M = 3.12$, $SD = 1.00$) reported significantly lower perceptions of dehumanization, $t(484) = 4.49$, $p < .001$. Next, we found that respondents in the diagnostic scenarios ($M = 4.78$, $SD = 1.76$) compared to respondents in the resource allocation scenarios ($M = 4.35$, $SD = 1.85$) reported significantly higher perceptions of role appropriateness, $t(483) = 5.07$, $p < .001$. Finally, we also found that respondents in the diagnostic scenarios ($M = 3.35$, $SD = 1.02$) compared to respondents in the resource allocation scenarios ($M = 3.16$, $SD = 1.08$) reported significantly higher levels of trust, $t(483) = 3.90$, $p < .001$.

Returning to our hypotheses, our results support H1 regarding a bias in favor of human decision making, with some exceptions and differences between allocation and diagnostic cases (see Tables 5 and 6). In comparing the same decision made by a human or an AI, for diagnostic cases, there was evidence of human bias for all five dependent variables except for outcome satisfaction with positive decisions, and for allocation cases there was evidence of human bias for the dehumanization, appropriateness, and trust variables.

In terms of our second hypotheses, our results clearly support H2 regarding a bias in favor of positive outcomes for allocation cases, however the support for H2 in diagnostic cases applies mainly to AI decision making only (see Tables 5 and 6). In comparing the different decision outcomes made by the same decision maker, for allocation cases there was evidence of outcome bias for all five dependent variables, except for role appropriateness with human decision making and dehumanization with AI decision making. In contrast, for diagnostic cases, there was no evidence of an outcome bias for human decisions except for the dehumanization variable, whereas there was evidence of an outcome bias for AI decisions with significant differences for three out of the five dependent variables (outcome satisfaction, appropriateness, and trust).

For our first exploratory question about the conflict between human and outcome bias, in comparing human decisions with negative outcomes to AI decisions with positive outcomes, for diagnostic cases we found that the human bias outweighed the outcome bias for three of our five variables (dehumanization, appropriateness, and trust). However, for resource allocation cases, most variables showed no significant

differences between the two decisions, except for the human bias (for dehumanization) and the outcome bias (for outcome satisfaction) each trumping in one case.

For our final exploratory research question with regards to the differences between medical diagnostic or resource allocation decision contexts, we found that respondents' perceptions were more favorable (i.e., lower dehumanization and higher outcome satisfaction, interactional justice, role appropriateness, and trust) in the diagnostic scenarios compared to the resource allocation scenarios.

4.4. Post-hoc sensitivity results

We conducted two sets of post-hoc sensitivity analyses to explore any differences between our three diagnosis scenarios and between our two allocation scenarios. First, we compared respondents' mean scores on our dependent variables across our three different diagnostic healthcare scenarios (i.e., macular disease, anxiety disorder, and skin cancer). Our results indicated that respondents in the macular disease scenario ($M = 3.47$, $SD = 0.96$) compared to respondents in both the anxiety disorder ($M = 3.24$, $SD = 1.04$) and skin cancer scenarios ($M = 3.20$, $SD = 1.14$) reported significantly higher outcome satisfaction, $t(240) = 3.26$, $p = .001$ and $t(237) = 3.58$, $p < .001$, respectively. We found no significant differences in outcomes satisfaction between respondents in the anxiety disorder and skin cancer scenarios, $t(237) = 0.53$, $p = .594$. Second, we found no significant differences in interactional justice perceptions between respondents in the macular disease scenario ($M = 3.52$, $SD = 0.91$) compared to respondents in both the anxiety disorder ($M = 3.44$, $SD = 1.00$) and skin cancer scenarios ($M = 3.40$, $SD = 1.04$), $t(216) = 1.10$, $p = .268$ and $t(210) = 1.67$, $p = .097$, respectively. Moreover, we found no significant differences in perceptions of interactional justice between respondents in the anxiety disorder and skin cancer scenarios, $t(210) = 0.61$, $p = .543$. Third, we found no significant differences in dehumanization perceptions between respondents in the macular disease scenario ($M = 2.86$, $SD = 0.89$) compared to respondents in both the anxiety disorder ($M = 2.93$, $SD = 0.91$) and skin cancer scenarios ($M = 2.98$, $SD = 0.93$), $t(227) = -1.18$, $p = .239$ and $t(237) = -2.00$, $p = .056$, respectively. Moreover, we found no significant differences in perceptions of dehumanization between respondents in the anxiety disorder and skin cancer scenarios, $t(237) = -.83$, $p = .407$. Fourth, we found that respondents in the macular disease scenario ($M = 5.02$, $SD = 1.67$) compared to respondents in both the anxiety disorder ($M = 4.69$, $SD = 1.72$) and skin cancer scenarios ($M = 4.62$, $SD = 1.86$) reported significantly higher perceptions of role appropriateness, $t(227) = 2.91$, $p = .004$ and $t(237) = 3.35$, $p < .001$, respectively. Moreover, we found no significant differences in perceptions of role appropriateness between respondents in the anxiety disorder and skin cancer scenarios, $t(237) = 0.60$, $p = .548$. Finally, we found that respondents in the macular disease scenario ($M = 3.48$, $SD = 0.95$) compared to both those in the anxiety disorder ($M = 3.29$, $SD = 1.02$) and skin cancer scenarios ($M = 3.28$, $SD = 1.08$) reported significantly higher trust, $t(227) = 2.76$, $p = .006$ and $t(237) = 2.79$, $p = .006$, respectively. Moreover, we found no significant differences in trust between respondents in the anxiety disorder and skin cancer scenarios, $t(237) = 0.14$, $p = .886$. Overall, these results show that respondents' reactions were more favorable for all but two (i.e., interactional justice and dehumanization) of our five variables in the macular disease scenario compared to the anxiety disorder and skin cancer scenarios, whereas there were no differences across all variables when comparing the anxiety disorder and skin cancer scenarios.

Second, we compared respondents' mean scores on our dependent variables across our two resource allocation scenarios (i.e., kidney transplant and ventilator). We found no significant differences in outcome satisfaction between respondents in the kidney transplant ($M = 3.08$, $SD = 0.96$) and ventilator allocation scenarios ($M = 3.13$, $SD = 1.07$), $t(237) = -.65$, $p = .518$. We also found no significant differences in interactional justice perceptions between respondents in the kidney transplant ($M = 3.16$, $SD = 1.11$) and ventilator allocation scenarios ($M = 3.19$, $SD = 1.09$), $t(223) = -.40$, $p = .691$. We found no significant differences in dehumanization perceptions between respondents in the kidney transplant ($M = 3.10$, $SD = 1.02$) and ventilator allocation scenarios ($M = 3.15$, $SD = 0.97$), $t(237) = -.83$, $p = .409$. We also found no significant differences in role appropriateness perceptions between respondents in the kidney transplant ($M = 4.30$, $SD = 1.89$) and in the ventilator allocation scenarios ($M = 4.41$, $SD = 1.81$), $t(237) = -.95$, $p = .343$. Finally, we also found no significant differences in trust between respondents in the kidney transplant ($M = 3.12$, $SD = 1.11$) and in the ventilator allocation scenario ($M = 3.19$, $SD = 1.05$), $t(237) = 1.02$, $p = .310$. Overall, these results show that respondents' reactions do not differ across all variables when comparing the kidney transplant and ventilator allocation scenarios.

4.5. Qualitative results

All data were coded under one of the three major themes of being positive, negative, or neutral for respectful and dignified interpersonal treatment, and then under one of several minor themes that emerged organically. Descriptions, illustrative quotes, and relative frequency of each of our major and minor themes are presented in Table 7. The group from which each quote is sourced is indicated in square brackets after the quote and in the format of: Decision maker [**AI** or **H**uman]_Decision Type [**D**iagnostic or **F**airness (i.e. resource allocation)]_Decision outcome [+ positive/**–** negative].

By analyzing the data by frequency across our three major themes for each of the four groups, we can assess the extent to which the qualitative data support the presence of a bias toward human decision makers (H1) and positive outcomes (H2) regarding experiences of dignified interpersonal treatment. Fig. 1 contrasts the percentages of positive, neutral, and negative interpersonal justice themes for different decision makers and Fig. 2 for different decision outcomes. Fig. 1 demonstrates a clear human bias, with human decisions (on the left) leading to a large majority of positive (in blue) over negative (in orange) themes, whereas AI decisions (on the right) lead to a majority of negative over positive themes, across both diagnostic and resource allocation cases. Fig. 2 demonstrates a clear outcome bias, with decisions with positive outcomes (on the left) leading to a large majority of positive over negative themes, whereas decisions with negative outcomes (on the right) lead to a large majority of negative over positive themes, across both diagnostic and resource allocation cases. This broadly supports our quantitative results. However, we can also surface more nuanced insights.

We can better understand the outcome bias by contrasting different outcomes. For human resource allocation decisions with positive outcomes, the most common themes were a good, accurate or fair outcome ("I felt treated with dignity and respect because I got a good result" [H_F+]), that is data driven, which comes from a respectful decision maker, and involves a fair process. In contrast, for human resource allocation decisions with negative outcomes, the most common themes were a bad or unfair outcome ("He was unfair in deciding I wasn't good enough to get the ventilator" [H_F-]) from an inappropriate decision maker ("They should not be deciding who gets one [kidney] next" [H_F-]). Concerns around disrespect were also much more common for negative (as opposed to positive) decisions ("No one feels respected if they are basically told they have no hope and are a number to die so others can live" [H_F-]). We saw a similar pattern with AI resource allocation decisions. For AI allocation decisions with positive outcomes, the most common themes were a good or fair outcome ("They have a screening tool and it was used and they followed it. That is fair" [AI_F+]), although themes about AI being an inappropriate decision maker remained comparatively high even for positive decisions. In contrast, for negative AI resource allocation decisions, the most common themes were, once again, a concern with AI as an inappropriate decision maker, followed by a bad or unfair outcome ("I should have been able to receive [the] ventilator and if they refused me, I do not think I was treated with dignity" [AI_F-]). This demonstrates how using the same

Table 7
Summary of themes with illustrative quotes.

Theme	Illustrative Quotations	Code freq.
Negative for justice		44%
Inappropriate decision maker	"A real doctor should determine that, not an app" [AI_F-] "AI will not make life or death decisions for me" [AI_D+] "No one alone should be allowed to make that decision" [H_F-]	10%
Bad, inaccurate, or unfair outcome	"I don't believe it would be accurate" [AI_F+] "Unfair because my life is as important as anybody else" [AI_F-]	8%
Decision based on wrong, irrelevant, or missing data	"They should have asked more questions, ran more tests" [H_D-] "They didn't get enough information to make that diagnosis" [H_D+] "It only considered data and not circumstances" [AI_F+]	7%
Decision maker was disrespectful	"Machine will treat anybody as a machine" [AI_D-] "It is not respectful or dignifying to have an AI make the decision for whether or not a sick person should receive a transplant" [AI_F-] "Because I was treated as a number in an equation" [H_F-]	7%
Decision maker lacks emotional intelligence or emotions	"It's a machine with no feelings" [AI_D-] "There's no feelings shown" [H_D+] "There just was no empathy" [AI_F-]	3%
Decision maker is untrustworthy or unreliable	"I would not trust an app to diagnose such a serious problem" [AI_D+] "They are not reliable" [H_D-] "Distrust AI" [AI_D-]	2%
Lack of explanation for the decision	"They didn't explain why they made this decision" [H_D+] "I want a human to explain" [AI_F-]	2%
Lack of human interaction	"There's no human interaction" [AI_D-] "There is little interaction and questioning" [H_D-]	2%
Decision maker can't make that decision	"I don't think AI are capable of properly diagnosing these issues" [AI_D-] "A healthcare professional cannot diagnose an anxiety disorder in this manor" [H_D+]	1%
Human needs are not met	"[I]t's [not] okay to deny people … something they may need" [H_F-] "The kidney may be needed as an emergency" [AI_F-]	1%
Decision maker might be biased	"They are biased" [H_F+] "Doctors are human and often financially motivated" [H_D-]	1%
Neutral for justice		**11%**
Neither respected nor disrespected; neither fair nor unfair	"I think the AI app can't really treat any [one] with or without dignity and respect" [AI_D+] "AI has no feelings so there is no dignity or respect involved" [AI_D-] "I don't think this is fair or unfair … it is based on facts" [H_F+]	8%
Lack information to answer	"Really not enough info in the scenario to say" [H_D+]	3%
Positive for justice		**45%**
Decision maker is respectful	"I think I was treated with respect because the AI app asked me appropriate questions about my symptoms" [AI_D+] "I was treated with dignity and respect" [H_D-]	15%
Good, accurate, or fair outcome	"Got an accurate diagnosis" [H_D+] "I thought [it was] fair because it benefited me" [AI_F+]	13%
Decision based on relevant data	"If the data says its cancer then it is probably cancer" [H_D+] "All the factors were weighed, and the decision was made based upon facts" [AI_F-]	6%
Appropriate decision maker	"I think the healthcare provider is doing the job they were trained to do" [H_D+] "A Healthcare professional should be the one to decide" [H_F+]	4%
Fair and accurate process	"I think it is a fair way to decide" [H_F+]	2%
Human needs are met	"It was fair … to give it to me [since] … I need it more than others" [H_F+] "There are other people who need it more than me" [AI_F-]	2%
Decision maker is unbiased	"It used data in an unbiased manner" [AI_F+] "AI makes decisions based on an algorithm. There is no bias in its decision making" [AI_F+]	1%
Decision maker is trustworthy or reliable	"I trust the health care professional and his expertise" [H_D-] "I find it credible and trustworthy" [H-D-]	1%
Human interaction and feeling	"I feel the human-to-human connection allows for the professional to present more emotion and socially correct feelings" [H_D+] "Showed compassion" [H_F+]	1%

Note for Table 7: Bold frequencies are cumulative percentages for each major theme.

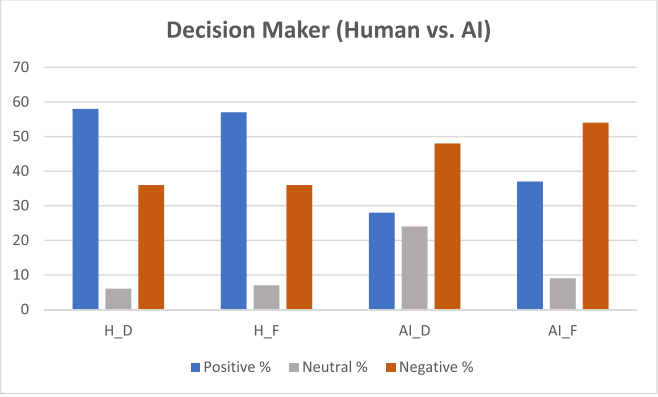

Fig. 1. Percent frequency of themes for different decision makers. Notes: human (H) vs. AI for diagnostic (D) and fairness/resource allocation (F) cases.

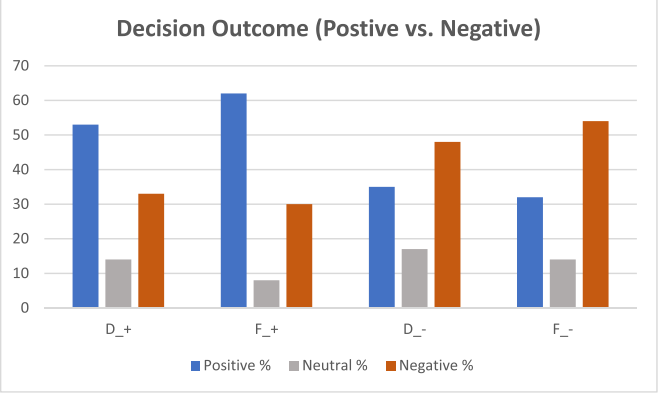

Fig. 2. Percent frequency of themes for different decision outcomes. Notes: positive (+) vs. negative (−) for diagnostic (D) and fairness/resource allocation (F) cases.

data and the same process can be viewed positively when it leads to a good outcome (e.g., the decision was based on appropriate data and the outcome was fair) and negatively when it doesn't (e.g., the decision maker didn't consider important data and the outcome was unfair). However, while this outcome bias was common (as shown in Fig. 2), it was not universal, with some participants recognizing that even though they didn't get a good outcome, the decision could still be fair ("I, like all the other patients being cared for, was judged solely on criteria that affects all of us" [H_F-]).

This outcome bias was also seen in the contrast between positive and negative outcomes for diagnostic decisions, but with two key differences. First, there were more themes around respectful decision makers, as we explore further below. Second, the data suggest that the outcome bias was less pronounced for diagnostic cases. This appears due to resource allocation cases having a very clear positive valence (i.e., getting a resource is good) and a very clear negative valence (i.e., not getting a resource is bad), whereas getting a diagnosis is a double-edged sword. This is because the patient has a diagnosis for their symptoms, which is positive ("I feel it's taking my symptoms very seriously and giving me accurate diagnosis") [AI_D+], but it also confirms that they have a serious medical condition, which is negative ("I think I was treated fair because they said I didn't have that disease" [H_D-]).

We can explore the human bias by contrasting human and AI decision makers. In resource allocation cases, what seems to explain the large disparity of positive vs. negative themes for humans compared to AI decision makers (see Fig. 1) is that AI is seen as an inappropriate decision maker that should not make important healthcare decisions about who gets access to resources ("A real doctor should determine that, not an app" [AI_F-]). Other important, but less frequent, themes raised were a concern about a lack of human interaction with AI decision makers ("There is no human … [contact], just vitals … [inputted into] an app" [AI_F-]) and AI's lack of emotions and emotional intelligence ("AI has no feelings it deals in data" [AI_F-]). Shifting to diagnostic decisions, the two most common positive themes were identical and in the same order for both groups; the decision maker was respectful ("The professional saw the patient as a person and took the time to examine the patient" [H_D+]), and there was a good or fair outcome ("It seems like a helpful app for health" [AI_D-]). However, the negative themes were differently ordered, with the themes of the decision being based on wrong, irrelevant, or missing data being the most common for the human groups ("They just took a look they didn't do any proper testing" [H_D-]), whereas the concern that AI is an inappropriate decision maker was (again) the most common theme for the AI group ("should be done by another human being and not AI" [AI_D+]). Further, both AI groups received a high number of neutral themes for AI neither respecting nor disrespecting them in diagnostic decisions ("AI does not have the ability to distinguish dignity vs not, and it is impossible for it to be anything other than neutral" [AI_D+]), whereas this issue was largely absent for human decision making in both cases and AI decision making in resource allocation cases.

This points to two key differences between diagnostic and resource allocation decisions. First, the themes of respectful treatment were mentioned much more frequently for diagnostic decisions (mentioned 145 times or 24% of total themes across all diagnostic cases) compared to allocation decisions (mentioned 74 times or 8% of total themes across all resource allocation cases) for both human and AI decision makers. The data suggest the reason for this is that participants saw getting tests and scans, whether ordered by a human or an AI, as itself indicating respectful treatment since this action signaled taking their concerns, and thus themselves as persons, seriously ("I think I was treated with respect because the AI app asked me appropriate questions about my symptoms" [AI_D+]). In contrast, decisions about the allocation of resources less clearly indicated to participants that they were regarded as persons of worth ("Did not see me as a person" [AI_F-]; "Because I was treated as a number in an equation" [H_F-]), leading to a lower frequency of themes about respectful treatment for both decision makers. Second, the high levels of the neutral theme of the decision maker being neither respectful nor disrespectful were only common in the context of AI diagnostic decisions. The data suggest the reason for this is that when an AI made a diagnostic decision, some participants felt that questions of respect or disrespect were not relevant since the AI was simply trying to diagnose a condition ("This app was only responding to the picture of the mole that I submitted to the app. It doesn't react to me personally in any other way" [AI_D-]), whereas they responded differently to resource allocation decisions made by an AI since there were implicitly moral questions about fairness, and thus respectful treatment, at play ("I felt I was being treated as just a thing based on its computations" [AI_F+]).

5. Discussion

AI systems are increasingly taking on decision making roles in healthcare. However, the continuing uptake, effectiveness, and ethicality of these systems will depend, in part, on whether healthcare patients perceive that they are being treated with dignity and respect when subject to their decision making. Our study provides important insights into these perceptions. In terms of our two hypotheses, we found evidence of both a human bias (i.e., preferring human over AI decisions) and an outcome bias (i.e., preferring positively over negatively valenced decisions). People felt they were treated with more dignity and respect when subject to a human decision maker or a decision with a positive outcome, which broadly aligns with existing research. However, our data reveal several novel theoretical and practical implications that we now discuss.

5.1. Theoretical and practical implications

First, our results suggest that whatever the outcomes, humans are consistently seen as appropriate decision makers and AIs are consistently seen as causing experiences of dehumanization. For resource allocation decisions, the only cases where there was no outcome bias were for dehumanization for AI decision makers (i.e., participants felt dehumanized whether or not they received the resource) and for appropriateness of human decision makers (i.e., participants felt human decision makers were appropriative even if the resource was denied). This finding was supported by the qualitative data where appropriate and respectful decision maker themes were more frequent for human (compared to AI) decision makers, and inappropriate and disrespectful decision maker themes were more frequent for AI (compared to human) decision makers. Further, many participants mentioned feeling dehumanized by AI decisions regardless of the decision outcome. This finding fits with existing research about the dehumanizing impacts of AI decision making (Binns et al., 2018) and perceptions about the inappropriateness of AI making some morally significant decisions (Formosa & Ryan, 2021). It also meshes with research showing that people are concerned that AI, compared to humans, cannot account for their "uniqueness" (Longoni et al., 2019); that AI, by reducing individuals to a number which is interchangeable with any other equivalent number, misses something unique about them. This fits with our qualitative data which referenced AI as being unable to capture certain information that was irreducible to a number. The concern with uniqueness and being reduced to a number also points to a core feature of the recognition of human dignity, which is that each person is uniquely valuable and cannot be replaced without loss by some other person. This is central to the Kantian contrast between things with a "price" that can "be replaced by something else as its equivalent", and persons whose dignity raises them above all "price" and therefore cannot be replaced by others without loss (Kant, 1996, p. 4:434). A concern with AI decision making and its "consequentialist decision strategy" (Dietvorst & Bartels, 2021) is that it reduces everything to a common currency that allows persons with dignity to be weighed up as if they were things with a price.

Practically speaking, given the significance of health for human wellbeing, concerns about dehumanization must be taken seriously by

those designing and implementing AI in healthcare. An AI that makes patients feel dehumanized and is seen as an inappropriate decision maker may struggle to be accepted by patients, even if it is otherwise effective or efficient. The physical configuration of the AI and the way that patients interact with it may also affect peoples' perceptions about dignified or dehumanized care, given that an AI embodied in more humanoid robotic forms is likely to be anthropomorphized to a greater degree than one that is not (Formosa, 2021). When implementing AI in healthcare it will be important to provide information to patients about the basis for the AI decision (if known). For example, the literature on autonomous AI IDx-DR emphasizes the fact that the AI uses the same information about the person's eyes that a clinician would use to make a diagnosis of diabetic retinopathy (Abramoff et al., 2018), which may go some way to assuaging concerns about loss of individual uniqueness.

Second, we found that the bias for human decision makers is stronger in diagnostic cases, while the bias for positive outcomes is stronger in resource allocation cases. In diagnostic cases, while there was clear evidence of a human bias, we found no evidence (outside of the dehumanization variable) of an outcome bias for human diagnostic decisions. This suggests that in diagnostic cases, people prefer a human decision maker, and they feel they were treated appropriately whatever diagnostic outcome they received from a human (i.e., positive or negative). In contrast, for resource allocation decisions there is clear evidence of an outcome bias for both human and AI decision makers. This suggests that in resource allocation cases, people are comparatively more concerned with positive outcomes than with who makes the decisions. We again see support for this in the qualitative data, where themes about respectful decision makers are more common for diagnostic decisions compared to resource allocation decisions. We see further evidence of this general trend in the conflict cases. There, for diagnostic cases the human bias trumps the outcome bias for three out of the five variables, with the others showing no significant difference. In contrast, for allocation cases, both the human bias (for dehumanization) and the outcome bias (for outcome satisfaction) trump for one variable each, with the other three variables showing no significant difference. Overall, in terms of perceptions of respectful interpersonal treatment, it matters more who makes diagnostic decisions and it matters more what outcome resource allocation decisions have. Existing literature shows that people prefer human decision makers over AI in various diagnostic contexts (e.g., Lennartz et al., 2021; Yokoi et al., 2021). Our study adds nuance to this literature by showing that a similar, but distinct, preference or "human bias" also applies to resource allocation decisions, and that this preference or bias is expressed across several distinct but related variables relevant to interpersonal treatment.

Practically speaking, these results add to other studies which show that patients want humans involved when AI is used, indicating a preference for assistive over autonomous uses of medical AI (e.g., Ongena et al., 2021). However, these studies focused on diagnostic contexts, and our comparative research shows that the concern for a human to be involved is greater in diagnostic decisions than it is in resource allocation decisions. One reason for this may be a desire to ensure that human interaction remains part of the healthcare context (e.g., Bhandari, Purchuri, Sharma, Ibrahim, & Prior, 2021), although this may be achieved, perhaps more efficiently, if an AI first makes the diagnosis and then a human healthcare professional negotiates the management of the condition with the patient. However, our research did not investigate this distinction between diagnosis and management or responses to AI-human synergies. The preference for a positive outcome, irrespective of the decision maker, has potentially problematic implications for practice because both human and AI decision makers will at times make negative decisions. It might be possible to build in a layer of human oversight of AI-generated negative decisions, such that all of these are referred for human review and communication. However, this might come at the cost of any efficiency gains made by using the AI in the first place.

Third, we found that respectful treatment is understood differently by participants in diagnostic and resource allocation cases. For diagnostic decisions, themes about respectful decision makers were comparatively more common, and taking symptoms seriously seemed to constitute respectful treatment for both human and AI decision makers. In contrast, for resource allocation decisions, respectful treatment focused more on whether a person felt valued as a person, rather than as a mere input into a calculation. This helps to make sense of the significant difference we found between diagnostic and resource allocation cases, with higher perceptions of dignified and respectful treatment across all five variables for the former in comparison to the latter. Overall, our qualitative data suggests that compared to resource allocation decisions people are more likely to experience diagnostic decisions as respectful regardless of the outcome, and that allocation decisions are less likely to be experienced as respectful treatment regardless of the decision maker. Part of the reason for this difference in respectful decision maker themes between diagnostic and resource allocation cases is that participants often felt they were reduced to a number by both human and AI decision makers in resource allocation cases. This was probably due to perceptions about the mechanical and impersonal nature of such allocation decisions, compared to the face-to-face and interpersonal nature of diagnostic decisions. This finding fits with research by Lee (2018) that people are more comfortable with AI undertaking what are seen as mechanical tasks, such as resource allocation, rather than human-centered tasks, such as diagnostic interactions. Reinforcing this point, a lack of human interaction in the AI decision cases and the presence of human interaction in the human decision cases were mentioned by participants. As noted above, we found consistent perceptions about respectful treatment within problem types, with no significant difference between our two resource allocation cases and two out of our three diagnostic cases. The exception was the macular deterioration scenario, which might be explained by the apparent seriousness of the eye condition compared to the other two conditions or by the comparatively greater sophistication of the diagnostic technology (i.e., retinal photo) used in this example, although further research is needed to confirm this.

Practically speaking, this emphasizes the importance of maintaining the interpersonal relationship between health care professionals and their patients in the diagnostic context, and the importance of putting a 'human face' on resource allocation decisions regardless of the ultimate decision maker. This fits with concerns in the literature about AI affecting the healthcare professional-patient relationship (e.g., Rogers et al., 2021), the greater acceptance of diagnostic AI if it is implemented in a way that maintains the integrity of that relationship (Nelson et al., 2020), and the preference for the use of medical AI only in assistive and backup roles to humans (e.g., Ongena et al., 2021).

Fourth, concerns around algorithmic bias are common in the AI literature (e.g., Danks & London, 2017), and such concerns are repeated in the medical AI literature where, for example, algorithms can exhibit racial bias against minorities (e.g., Obermeyer et al., 2019). However, algorithmic bias was rarely raised in our qualitative data as grounds for disrespectful interpersonal treatment. Where it was mentioned, bias was raised more often as a concern with human decision making and a lack of bias raised more often as a feature of AI decision making. This indicates that our participants, while familiar with everyday human biases (such as the financial incentives of doctors), showed less understandings of the important concerns about biases in data and algorithms that can impact the fairness of AI decision making and which may reflect broader human biases.

Practically speaking, this suggests that in addition to measures to reduce bias in algorithms and datasets, further public education about the impacts of algorithmic bias in AI healthcare decisions is warranted. For example, patients could be offered specific questions to ask before agreeing to AI healthcare, such as whether the AI is accurate for "people like them" or whether it works equally well for all genders and racialized groups.

5.2. Limitations and future research directions

Experimental vignette studies come with limitations. Vignettes are limited in scope and may fail to capture important elements of a phenomenon. Further, given that the scenarios are hypothetical, it may be difficult for participants to accurately predict how they would respond to real-world versions, especially when dealing with unfamiliar situations. To counteract this limitation, we used expert review of our vignettes to build external validity (Cruz, 2021) and assigned participants randomly to different vignettes to reduce ordering and unfamiliarity effects. However, it would be helpful for other studies to employ different methods (Langer & Landers, 2021), such as interviewing or surveying patients subject to AI decision making in healthcare to confirm our results in clinical settings, or using discrete choice experiments to further investigate how patients might be willing to trade off different aspects of care (Clark et al., 2014). Further, as other research indicates that different levels of education, gender, and cultural groups can all impact attitudes toward AI use in healthcare (Yakar et al., 2021), it would be helpful to replicate our study with different samples. Finally, in focusing on varying only the decision maker and decision outcome, we did not explore how other salient dimensions of AI that are known to impact human attitudes towards its use, such as its transparency and explainability (Springer & Whittaker, 2020), could impact these results.

In terms of future research, studies could examine different types of scenarios, because other work has shown this to be important (e.g., Langer & Landers, 2021) and we saw some differences in our macular deterioration scenario. Exploring whether the preference for human decision makers holds across multiple contexts is required, as some research suggests that patients might be more comfortable talking with AIs or robots, compared to humans, when discussing stigmatizing conditions such as mental health concerns (Duan et al., 2021). Other types of scenarios could also be explored, such as surgical assistance and operation, disease screening, and medical risk analysis, since our results show there are significant differences when using AI for different problem types (i.e., diagnostic as opposed to allocation cases). It was also unclear whether our sample thought AI was better (or worse) than humans at making accurate diagnoses and efficiently allocating resources. Future research could explore how information about the relative accuracy and efficiency of AI decision making impacts perceptions of respectful treatment. Finally, future research could further unpack the constellation of the variables examined here. That is, in the current study we focused on the differences in perceptions of outcome satisfaction, dehumanization, interactional justice, role appropriateness and trust across the different vignettes. While this provides us with important and insightful findings, these also serve as steppingstones for future research examining models in which some of these outcomes serve as mediators (e.g., role appropriateness) and moderators (e.g., dehumanization) when predicting other dependent variables, such as outcome satisfaction. In order to build these models and explore them, future work could use a longitudinal design with a minimum of three measurement points to assess mediation or moderated mediation, since our current experimental survey design results in cross-sectional data which would result in a misspecified model of change and should not be used to estimate a mediation model (e.g., Mitchell & Maxwell, 2013), or a sequential design with a minimum of two measurement points to assess moderation (MacCallum & Austin, 2000).

6. Conclusion

Individuals will be increasingly subject to decision making by AI in healthcare contexts. Maintaining broad acceptance of medical AI, ensuring its ethical deployment, and actualizing the benefits it can bring to healthcare, will rely in part on patients perceiving its use as maintaining respectful and dignified interpersonal treatment. Our study advances knowledge in these areas by exploring how perceptions of interactional justice and a range of related measures are impacted by different decision makers (human or AI), decision outcomes (positive or negative), and decision types (diagnostic or resource allocation healthcare scenarios). We show the presence of a human bias and an outcome bias in these scenarios, and we demonstrate that in terms of perceptions of respectful and dignified interpersonal treatment, it matters more who makes diagnostic decisions and it matters more what outcome is generated in resource allocation decisions. We also found that participants perceived they were treated better when subject to diagnostic as opposed to resource allocation decisions. Finally, we outlined several future research directions.

Funding

We received funding from Facebook (Ethics in AI Research Initiative for the Asia Pacific). Paul Formosa also received funding from the Australian Research Council (DP190100734).

CRediT author statement

Paul Formosa: Conceptualization, Methodology, Qualitative analysis, Writing – original draft, Writing – review & editing. **Wendy Rogers:** Methodology, Writing – review & editing. **Sarah Bankins:** Funding acquisition, Project administration, Conceptualization, Methodology, Writing – review & editing. **Yannick Griep:** Conceptualization, Methodology, Formal analysis, Writing – review & editing. **Deborah Richards:** Conceptualization, Methodology, Writing – review & editing.

Acknowledgements

We received research assistance in running the formal analysis from Omid Ghasemi.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chb.2022.107296.

References

Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine, 1*(1), 39.

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods, 17*(4), 351–371.

Aoki, N. (2021). The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence. *Computers in Human Behavior, 114*, Article 106572.

Bankins, S., Formosa, P., Griep, Y., & Richards, D. (2022). AI decision making with dignity? *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10223-8

Barclay, L. (2018). Dignitarian medical ethics. *Journal of Medical Ethics, 44*(1), 62–67.

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology, 54*(4), 569–579.

Bastian, B., & Haslam, N. (2011). Experiencing dehumanization. *Basic and Applied Social Psychology, 33*(4), 295–303.

Becker, T. E., Atinc, G., Breaugh, J. A., Carlson, K. D., Edwards, J. R., & Spector, P. E. (2016). Statistical control in correlational studies. *Journal of Organizational Behavior, 37*(2), 157–167.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods, 43*(3), 800–813.

Bernerth, J. B., & Aguinis, H. (2016). A critical review and best-practice recommendations for control variable usage. *Personnel Psychology, 69*(1), 229–283.

Bhandari, A., Purchuri, S. N., Sharma, C., Ibrahim, M., & Prior, M. (2021). Knowledge and attitudes towards artificial intelligence in imaging. *Clinical Imaging, 80*, 413–419.

Bies, R. (2001). Interactional (in)justice. In *Advances in organizational justice* (pp. 89–118). Stanford University Press.

Bies, R. J. (2015). Interactional justice. In *The Oxford Handbook of justice in the workplace*. Oxford University Press.

Bies, R., & Moag, J. (1986). Interactional justice. *Research on Negotiation in Organizations, 1*, 43–55.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21–34.
Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). It's reducing a human being to a percentage. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–14).
Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101.
Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves. *Computers in Human Behavior, 127*, Article 107018.
Clark, M., et al. (2014). Discrete choice experiments in health economics. *PharmacoEconomics, 32*(9), 883–902.
Colquitt, J. A. (2001). On the dimensionality of organizational justice. *Journal of Applied Psychology, 86*(3), 386–400.
Cropanzano, R., Rupp, D. E., Mohler, C. J., & Schminke, M. (2001). Three roads to organizational justice. In *Research in personnel and human resources management* (pp. 1–113). Emerald.
Cruz, K. S. (2021). Does anyone care about external validity? *Group & Organization Management, 46*(6), 974–983.
Cumming, G. (2014). The new statistics. *Psychological Science, 25*(1), 7–29.
Dai, L., & Xie, H. (2016). Review and prospect on interactional justice. *Open Journal of Social Sciences, 4*(1), 55–61.
Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 4691–4697.
Dietvorst, B. J., & Bartels, D. M. (2021). Consumers object to algorithms making morally relevant tradeoffs because of algorithms' consequentialist decision strategies. *Journal of Consumer Psychology*.
Duan, Y., Yoon, M., Liang, Z., & Hoorn, J. F. (2021). Self-disclosure to a robot. *Robotics, 10*(3), 98.
Düwell, M., Braarvig, J., Brownsword, R., & Mieth, D. (2014). *The Cambridge Handbook of human dignity*. Cambridge University Press.
Erdogan, B. (2002). Antecedents and consequences of justice perceptions in performance appraisals. *Human Resource Management Review, 12*(4), 555–578.
Fischhoff, B. (1975). Hindsight != foresight. *Journal of Experimental Psychology, 1*, 288–299.
Formosa, P. (2017). *Kantian ethics, dignity and perfection*. Cambridge University Press.
Formosa, P. (2021). Robot autonomy vs. Human autonomy. *Minds and Machines, 31*, 595–616.
Formosa, P., & Ryan, M. (2021). Making moral machines. *AI & Society, 36*(3), 839–851.
Freedman, R., Borg, J. S., Sinnott-Armstrong, W., Dickerson, J. P., & Conitzer, V. (2020). Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence, 283*, Article 103261.
George, N., Moseley, E., Eber, R., Siu, J., Samuel, M., Yam, J., … Lindvall, C. (2021). Deep learning to predict long-term mortality in patients requiring 7 days of mechanical ventilation. *PLoS One, 16*(6).
Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. *Artificial Intelligence in Healthcare*, 295–336.
Ghasi, N. C., Ogbuabor, D. C., & Onodugo, V. A. (2020). Perceptions and predictors of organizational justice among healthcare professionals in academic hospitals in South-Eastern Nigeria. *BMC Health Services Research, 20*(1), 301.
Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics, 46*(3), 205–211.
Kant, I. (1996). Groundwork of the metaphysics of morals. In *Practical philosophy* (pp. 37–108). Cambridge University Press.
Karunakaran, A. (2018). In cloud we trust, 2018 *Academy of Management Proceedings*, (1), Article 13700.
Körber, M. (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. *Proceedings of the 20th Congress of the International Ergonomics*, 13–30.
Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth. *Industrial and Organizational Psychology, 8*(2), 142–164.
Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work. *Computers in Human Behavior, 123*, Article 106878.
Lee, M. K. (2018). Understanding perception of algorithmic decisions. *Big Data & Society, 5*(1).
Lee, H., & Chui, J. (2019). The mediating effect of interactional justice on human resource practices and organizational support in a healthcare organization. *Journal of Organizational Effectiveness, 6*(2), 129–144.
Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural justice in algorithmic fairness. In *Proceedings of the ACM on Human-Computer Interaction, 3 (CSCW)* (pp. 1–26).
Lee, M. K., Kusbit, D., Metsky, E., & Dabbish, L. (2015). Working with machines. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1603–1612.
Lennartz, S., Dratsch, T., Zopfs, D., Persigehl, T., Maintz, D., Große Hokamp, N., & Pinto dos Santos, D. (2021). Use and control of artificial intelligence in patients across the medical workflow. *Journal of Medical Internet Research, 23*(2), Article e24221.
Lipshitz, R. (1989). The effects of success and failure on the evaluation of decision making and decision makers. *Organizational Behavior and Human Decision Processes, 44*(3), 380–395.
Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650.
Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health, 23*(1), 151–169.
Lyell, D., Coiera, E., Chen, J., Shah, P., & Magrabi, F. (2021). How machine learning is embedded to support clinician decision making. *BMJ Health & Care Informatics, 28*(1), Article e100301.
Lysaght, T., Lim, H. Y., Xafis, V., & Ngiam, K. Y. (2019). AI-assisted decision-making in healthcare. *Asian Bioethics Review, 11*(3), 299–314.
Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modelling. *Methodology, 1*(3), 85–91.
MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*(1), 201–226.
Mitchell, M. A., & Maxwell, S. E. (2013). A comparison of the cross-sectional and sequential designs when assessing longitudinal mediation. *Multivariate Behavioral Research, 48*(3), 301–339.
Nelson, C. A., Pérez-Chada, LM., Creadore, A., Li, S. J., Lo, K., Manjaly, P., … Mostaghimi, A. (2020). Patient perspectives on the use of artificial intelligence for skin cancer screening. *JAMA Dermatology, 156*(5), 501–512.
Nemesure, M. D., Heinz, M. V., Huang, R., & Jacobson, N. C. (2021). Predictive modelling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Scientific Reports, 11*(1).
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.
Ongena, Y. P., Yakar, D., Haan, M., & Kwee, T. C. (2021). Artificial intelligence in screening mammography. *Journal of the American College of Radiology, 18*(1), 79–86.
Özer, Ö., Uğurluoğlu, Ö., & Saygili, M. (2017). Effect of organizational justice on work engagement in healthcare sector of Turkey. *Journal of Health Management, 19*(1), 73–83.
Palmisciano, P., Jamjoom, A. A. B., Taylor, D., Stoyanov, D., & Marcus, H. J. (2020). Attitudes of patients and their relatives toward artificial intelligence in neurosurgery. *World Neurosurgery, 138*, e627–e633.
R Core Team. (2021). *R. R Foundation for statistical computing*. http://www.R-project.org/.
Reiter, O., Rotemberg, V., Kose, K., & Halpern, A. C. (2019). Artificial intelligence in skin cancer. *Current Dermatology Reports, 8*(3), 133–140.
Rogers, W. A., Draper, H., & Carter, S. M. (2021). Evaluation of artificial intelligence clinical applications. *Bioethics, 35*(7), 623–633.
Ross, P., & Spates, K. (2020). Considering the safety and quality of artificial intelligence in health care. *Joint Commission Journal on Quality and Patient Safety, 46*(10), 596–599.
Schlicker, N., Langer, M., Ötting, S. K., Baum, K., König, C. J., & Wallach, D. (2021). What to expect from opening up 'black boxes. *Computers in Human Behavior, 122*, Article 106837.
Schwantes, I. R., & Axelrod, D. A. (2021). Technology-enabled care and artificial intelligence in kidney transplantation. *Current Transplantation Reports, 8*(3), 235–240.
Shaikh, S. J. (2020). Artificial intelligence and resource allocation in health care. *AAAI Fall 2020 Symposium on AI for Social Good, 8*.
Springer, A., & Whittaker, S. (2020). Progressive disclosure. *ACM Transactions on Interactive Intelligent Systems, 10*(4), 1–32.
Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., … Kittler, H. (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine, 26*(8), 1229–1234.
Wallander, L. (2009). 25 years of factorial surveys in sociology. *Social Science Research, 38*(3), 505–520.
Yakar, D., Ongena, Y. P., Kwee, T. C., & Haan, M. (2021). Do people favor artificial intelligence over physicians? *Value in Health*, 0(0).
Yin, J., Ngiam, K. Y., & Teo, H. H. (2021). Role of artificial intelligence applications in real-life clinical practice. *Journal of Medical Internet Research, 23*(4), Article e25759.
Yokoi, R., Eguchi, Y., Fujita, T., & Nakayachi, K. (2021). Artificial intelligence is trusted less than a doctor in medical treatment decisions. *International Journal of Human-Computer Interaction, 37*(10), 981–990.
Yu, L., Halalau, A., Dalal, B., Abbas, A. E., Ivascu, F., Amin, M., & Nair, G. B. (2021). Machine learning methods to predict mechanical ventilation and mortality in patients with COVID-19. *PLoS One, 16*(4), Article e0249285.
Zhang, X., Guo, X., Lai, K., & Yi, W. (2019). How does online interactional unfairness matter for patient–doctor relationship quality in online health consultation? *European Journal of Information Systems, 28*(3), 336–354.