Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust

Min Kyung Lee
School of Information
University of Texas at Austin
Austin, TX, United States
minkyung.lee@austin.utexas.edu

Kate Rich*
School of Information
University of Texas at Austin
Austin, TX, United States
k.rich@utexas.edu

ABSTRACT

Emerging research suggests that people trust algorithmic decisions less than human decisions. However, different populations, particularly in marginalized communities, may have different levels of trust in human decision-makers. Do people who mistrust human decision-makers perceive human decisions to be more trustworthy and fairer than algorithmic decisions? Or do they trust algorithmic decisions as much as or more than human decisions? We examine the role of mistrust in human systems in people's perceptions of algorithmic decisions. We focus on healthcare Artificial Intelligence (AI), group-based medical mistrust, and Black people in the United States. We conducted a between-subjects online experiment to examine people's perceptions of skin cancer screening decisions made by an AI versus a human physician depending on their medical mistrust, and we conducted interviews to understand how to cultivate trust in healthcare AI. Our findings highlight that research around human experiences of AI should consider critical differences in social groups.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

Perceptions of Algorithmic Decisions, Trust, Fairness, Healthcare AI, Group-Based Medical Mistrust Scale (GBMMS), Black Perspectives

ACM Reference Format:
Min Kyung Lee and Kate Rich. 2021. Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3411764.3445570

*The second author is a Ph.D. student in the Department of Communication, University of Washington. The research was conducted at The University of Texas at Austin.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445570

1 INTRODUCTION

Artificial Intelligence (AI) is increasingly automating decision-making in diverse industry sectors. In many high-stakes domains such as healthcare, education, criminal justice systems, organizational management and public assistance [8, 9, 25, 40], AI systems are automating or augmenting decisions that human experts used to make. In order to better understand people's reactions to this change, many scholars have investigated how people perceive algorithmic decisions as compared to human decisions [6, 10, 20, 23, 24, 26]. Their findings suggest that people tend to perceive algorithmic decisions as inferior to human decisions and are resistant to following them. In several studies, people trusted algorithmic decisions less than human decisions and were less likely to adopt them, particularly when tasks were deemed to require a human's unique capabilities [23], be subjective [6] or require attention to individual uniqueness [27].

However, not everyone has an equal level of trust in human decision-makers. Those who experience marginalization from other humans may have less faith in human decisions. Specifically, anti-Black racism across multiple human-led institutions has created a climate of mistrust among many Black and African American people in the United States [31]. Do people who mistrust human decision-makers also perceive them to be more trustworthy and fairer than algorithmic decisions? Or do they trust algorithmic decisions as much as or more than human decisions?

In this paper, we examine the role of cultural mistrust related to human systems in people's perceptions of algorithmic decisions. We focus on a healthcare context, group-based medical mistrust, and Black people living in the U.S., who have been shown to have higher medical mistrust than other populations. We conducted a between-subjects online experiment to examine people's perceptions of skin cancer screening decisions made by an AI versus a human physician depending on their medical mistrust. We recruited a balanced pool of Black and white participants. The result suggests that the previous literature's finding is replicated with participants with low mistrust in human systems, but not with those with high mistrust. Participants with low mistrust trusted human decisions more than algorithmic decisions and regarded them as fairer. However, participants with high mistrust in human systems perceived algorithmic and human decisions to be equally trustworthy and fair. We conducted interviews with 21 participants to understand what contributes to mistrust in healthcare AI and what information might cultivate their trust in healthcare AI.

In this work, we make a contribution to research around human experiences of AI in the fields of Human-Computer Interaction (HCI), psychology and communication. Our work offers new insight on perceptions of healthcare AI among those with high medical

mistrust, particularly Black populations. By doing so, our work surfaces the importance of mistrust in human systems in understanding human perceptions of AI; it highlights a gap in prior work that predominantly focused on populations who are less likely to experience cultural mistrust, and calls for more research that examines perceptions of algorithmic fairness and trust among different social groups, particularly those who have reason to distrust the systems they dwell in.

2 PERCEPTIONS OF ALGORITHMIC VERSUS HUMAN DECISIONS

2.1 Resistance to algorithmic decisions

Many scholars have examined how people would perceive algorithmic decisions as compared to human decisions across a range of contexts. Lee compared how people trust, find as fair, and emotionally respond to managerial decisions made by algorithms versus human managers [23]. Castelo et al. also compared perceptions of algorithmic versus human decisions varying a subjectivity of tasks [6]. Both studies find that people are less likely to trust and adopt algorithmic decisions over human decisions when tasks are subjective. Longoni et al. compared algorithmic and human physicians' decisions for medical diagnosis and treatment, and found that people are more resistant to algorithmic decisions, potentially due to the concern that it cannot account for an individual's uniqueness [27]. The researchers also suggest that personalizing AI and having a human-in-the-loop will reduce resistance to the medical AI systems. In the hiring context, Langer et al. examined job applicants' reactions to an algorithm that evaluates candidates' interviews videos, finding that people will trust an algorithmic assessment less than a human assessment [20]. Complementing previous studies above were based online experiments. Lee and Baykal investigated people's real experience with actual algorithmic outcomes versus human allocation outcomes, finding that people perceived algorithmic allocation to be less fair [24]. Taken together, there is emerging evidence for profound resistance to algorithmic decisions and perceived superiority in human decisions.

Individual and/or group-level differences have received relatively less attention in this stream of research on comparative perceptions of human versus algorithmic decisions. In our study, we seek to examine individual level differences, with a particular focus on mistrust in human systems and Black participants. While there is abundance evidence for algorithmic bias against Black communities [5, 30], perspectives from Black individuals are relatively less understood. As a notable exception, Woodruff et al examined African American and LatinX communities' perceptions of discriminatory information tailoring such as advertisements [42]. Online information tailoring does not involve equivalent human decision-maker, so comparative evaluation of algorithmic versus human decisions was not a focus of their study.

2.2 Mistrust in human systems

Trust can be understood as a belief that an entity will help someone reach their desired outcome in a difficult or uncertain context [19]. Mistrust, however, refers to a lack of trust and is commonly experienced among Black Americans due to historical and current systemic racism [31]. Cultural mistrust specifically refers to mistrust that forms as a result of experiencing discrimination for being part of a particular social group and is often used to describe the mistrust Black Americans have in human systems [36]. This understandably large pattern of mistrust within the Black and African American community appears across several facets of interpersonal and institutional contexts. For instance, one qualitative study found negative experiences with white teachers and being one of the few Black families within a predominately white school district influenced the level of trust middle-class Black mothers had in their children's teachers [4]. This cultural mistrust also manifests in racial gaps of trust in scientific experts, police officers, and other figures or authority or expertise [13, 29].

Research about medical mistrust, which is particularly relevant to this work, dates back decades ago when nonwhite patients consistently identified more negative experiences with healthcare affecting trust than white patients [37]. Since then, the correlation between perceived racial bias in healthcare and distrust in doctors became an item of interest for scholars across many disciplines. In one study, African Americans, Latin Americans, and Asians all reported higher perceived provider discrimination and poor health compared to white Americans. Importantly, this perceived provider discrimination made racial minorities less likely to seek healthcare [22]. African American and Black communities are one of the most heavily researched racial groups that experience doctor mistrust due to perceived bias. Lack of quality care, a shortage of minority healthcare providers, and a lack of cultural competence among healthcare providers are one set of explanations for medical mistrust among Black patients[18]. This mistrust is understandable given the substantial history and contemporary issues with medical racism towards the Black community [16, 17, 33]. Moreover, the high levels of mistrust among Black patients can also result in different healthcare decisions. One study found medical mistrust led Black Americans to utilize emergency care more than primary care in comparison to whites [3].

This legacy of mistrust and negative medical experiences among Black communities leaves healthcare AI with a number of important challenges because general perceptions of innovation in medical technologies can differ by race. Previous research suggests that Black Americans were more likely than white Americans to be hesitant towards medical innovation [14]. The authors considered the introduction of new prescription drugs and medical implants to be instances of medical innovation in this case. While this study offers a useful precedent, perceptions of AI and healthcare AI among racial minorities is severely overlooked in existing literature.

2.3 Research Question and Study Overview

We are a research team of non-Black women with intersecting identities. With our positionality in mind, we used a mixed methods approach. Our experimental study prioritized measurements of experience over general demographic categories and our interview study aimed to give our participants a voice in this work by highlighting their perspectives.

Our research question is: How do people who mistrust human decision-makers perceive algorithmic decisions compared to human

decisions, particularly in a setting when algorithmic decisions automate human decisions? We explore this research question through two studies. In the online experiment, we examine whether people who mistrust human decision-makers would perceive human decisions to be more trustworthy and fairer than algorithmic decisions. Based on the literature that we reviewed above, we make the following hypotheses.

Hypothesis 1. People with low mistrust in human systems will perceive algorithmic decisions to be less trustworthy and less fair than human decisions.

Hypothesis 2. People with high mistrust in human systems will not perceive algorithmic decisions to be less trustworthy and less fair than human decisions. They will perceive algorithmic decisions to be as trustworthy and fair as or more trustworthy and fairer than human decisions.

In the interview study, we qualitatively investigate how people with high mistrust in human systems perceive healthcare AI and which information could cultivate their trust in healthcare AI.

3 STUDY 1: EXPERIMENT

We conducted a between-subjects online experiment to examine how people perceive algorithmic versus human decisions depending on their mistrust in human systems. We adopted a skin cancer screening scenario used in a previous study [27] that demonstrated that people are less likely to adopt algorithmic decisions than physicians' decisions. We recruited Black and white participants using Amazon Cloud Research mTurk Toolkit. The experiment consisted of two surveys administered one week apart. The first survey asked their levels of medical mistrust using the Group-Based Medical Mistrust (GBMMS) scale [38]. The second survey ask them to evaluate a skin cancer screening scenario in which a decision was made by either AI or a physician [27].

3.1 Method

3.1.1 Participants. We recruited participants on Cloud Research, a service that Amazon provides to enable targeted participant recruiting and use respondents from Amazon Mechanical Turk (MTurk). We chose Cloud Research as it allowed race-based participant recruiting and was the only online survey platform that allowed longitudinal research to our knowledge. Mturk has predominately white participants [15] so directly recruiting different racial groups with contrasting levels of mistrust was necessary for this research. In order to qualify for the study, participants had to reside in the US, be at least 18 years old, have completed at least 100 Human Intelligence Tasks (HITs), and have at least a 95 percent HIT approval rate. We used the Cloud Research Toolkit's functions to specifically recruit Black and white participants and administer the longitudinal, two-part study. Cloud Research collects demographics from mTurkers when they sign up and periodically validates it. Our study information was only shown to those who meet our race criteria, and did not state that we looked for Black and white participants. Thus there was no incentive for participants to misrepresent race. We also asked about the participants' race in our survey to ensure their self-reports were consistent. Overall, this method is in line with a self-report-based method of collecting race information that the US Census Bureau as well as other survey institutions use.

We created a HIT for the first two minute survey for $0.40, and set the recruiting parameters so that half of the requested participants would identify as Black or African American and half would identify as white. After one week from the first HIT's completion, we invited the participants who completed the first survey to the second 4 minute survey for $0.70. 280 participants responded to the first survey. The 278 participants who did not take the survey multiple times, passed the qualification and passed the attention check were invited to take the second part of the study a week later. 228 people responded to the second survey. We eliminated participants who completed the survey multiple times (N=19), indicated they did not reside in the US or were younger than 18 (N=10), did not pass the attention check (N=2), did not pass the qualification (N=8), or identified as a race that was not white or Black (N=2). After these eliminations, 187 participants were left. 40.6 percent of this sample identified themselves as female, and 45 percent identified themselves as Black. The participants were 18 to over 70 years old and had an average age range of 30-39. The mean education score indicated that participants held a four-year degree on average. On a four point scale, the average participant rated their health at a 3 or "Good." The average participant indicated their last doctor visit was six months ago on average. The full demographic information is reported in Table 1 and Table 2 in the Appendix C.

3.1.2 Materials. We adopted the skin cancer screening scenario in Longoni et al.'s work [27] that suggested people were less likely to adopt AI decisions compared to physicians' decisions. The scenario was informed by research on real AI that identifies skin cancer [11] and AI-based skin cancer diagnosis systems.[1] In the scenario, participants were given a definition of skin cancer with facts about the condition [2]. Then, they were asked to imagine the scenario as vividly as possible before randomly being assigned to an AI versus human scenario. Both conditions involved the same wording about seeking out a skin cancer screening and sending photos of the skin to a provider to be examined. The AI versus human physician was the only difference between these scenarios. Both providers were described as "trained to distinguish between cancerous and non-cancerous skin conditions. Training was made possible through learning the differences between cancerous and non-cancerous skin conditions using an extensive dataset of images." The scenario ended with the assurance that the provider would give them results in a week. The full scenarios are available in Appendix A.

3.1.3 Procedure. The surveys were implemented in Qualtrics. In the first survey, after assuring that respondents were over the age of 18 and a US resident as well as consenting to participate, they were given an attention check question and immediately disqualified if they answered it incorrectly. Those who passed the attention check were directed to take the GBMMS questions. Then, they rated their perception of their own health and the last time they visited a doctor. All of the participants who finished this survey and passed the attention check were invited to partake in the second part of the study a week later.

The second survey used a similar process. After assuring that they were over the age of 18 and a US resident as well as consenting

[1] https://www.fotofinder-systems.com/technology/skin-cancer-screening/artificial-intelligence/

to participate, respondents were given an attention check question and immediately disqualified if they answered it incorrectly. Those who passed the check were randomly assigned to an AI or human healthcare provider. After reading a scenario and answering questions about it, participants were asked a manipulation check question and demographics questions.

3.1.4 Measures. We measured people's mistrust in healthcare using the Group-Based Medical Mistrust Scale (GBMMS) developed by Thompson et al. to extend the measurement of cultural mistrust to medical contexts [38]. The scale has been validated across ethnoracial identities and different genders, particularly with the Black population [41], and has been shown to be negatively correlated with how likely people are to seek out healthcare [34]. This twelve item scale asks participants about their perceptions and experiences involving healthcare. Example questions are "People of my ethnic group receive the same medical care from doctors and health care workers as people from other groups." and "Doctors have the best interests of people of my ethnic group in mind." The scale was reliable (Cronbach's α=.92). Black participants had a significantly higher GBMMS score on average (Mean=3.93, SE=.12) than white participants (Mean=2.55, SE=.1, F(1,186)=78.45, p<.0001) (Table 1 in Appendix C). According to the IP addresses recorded on Qualtrics, participants were from 39 different states; there was no significant difference in mistrust level depending on states. Throughout this paper, we use use the term high mistrust to refer to participants with high GBMMS scores and low mistrust to refer to participants with low GBMMS scores.

We measured fairness as a scale of three questions about perceived fairness in decision outcomes, providers' interaction such as considering patient concerns and providing the same level of care adopting [12]. Questions about perceived fairness ranged on a scale from 1 being "very unfair" to 7 being "very fair." The scale was reliable (Cronbach's α=.81). To measure trust, we asked two questions about trust in the outcome and the decision-making process. Questions about trust used a scale with 1 being "completely distrust" and 7 being "completely trust." The scale was reliable (Cronbach's α=.95). We also included demographic questions and questions about their perceived health ratings and last doctors appointments. All survey questions used a 7-point scale with the exception of demographic questions, health ratings, and last doctor's appointments. As a recall-based manipulation check question, we asked whether participants recalled an AI or human physician made the decision in the scenario that they read. Two participants answered these questions incorrectly, and we excluded them following [35].

3.1.5 Analysis. We divided participants into low versus high GBMMS groups using the median. 79.8% of the low GBMMS group was white, and 72.73% of the high GBMMS group was Black (Table 2, Appendix C). For each of the GBMMS group, we conducted the one-way ANOVA to test the effect of decision-maker (AI versus physician) on perceived fairness and trust of the decisions.

3.2 Results

The analysis revealed a significant decision-maker effect on perceived fairness and trust in the low GBMMS group, replicating the result in [27] (Figure 1a). The algorithmic decision was perceived significantly as less fair (Mean=5.18 (SE=.14)) than the human physician's decision ((Mean=5.99 (SE=.14), F(1,97)=17.68, p<.0001). Additionally, the algorithmic decision was trusted significantly less (Mean=4.98 (SE=.16)) than the human physician's decision (Mean=5.78 (SE=.16), F(1,97)=12.60, p=.0006). On the contrary, the decision-maker did not have a significant effect on perceived fairness and trust in the high GBMMS group (Figure 1b). The algorithmic decision was perceived as fair (Mean=4.81, (SE=.18)) as the human decision (Mean=5.17 (SE=.19), n.s.) and as trustworthy (Mean=4.66 (SE=.19)) as the human decision (Mean=5.01 (SE=.21), n.s.). These results support our Hypotheses 1 and 2.

In order to examine whether Black non-trusting participants are different than the white non-trusting participants, we conducted additional ANOVAs comparing white vs Black participants within both low GBMMS and high GBMMS groups.[2] In the low GBMMS group, perceived fairness and trust did not differ between white and Black participants. On the contrary, in the high GBMMS group, Black participants reported significantly lower perceived fairness (Mean=4.82 (SE=.15) and trust (Mean=4.6 (SE=.16)) than white participants (Fairness Mean=5.4 (SE=.25), F(1,87)=4.06, p=.05; Trust Mean=5.46 (SE=.27), F(1, 87)=7.96, p=.006). We then analyzed further differences between Black versus white participants in the high GBMMS group. The Black participants had significantly higher mistrust (Mean=4.47 (SE=.09)) than the white participants (Mean=3.94, SE=.14) (F(1,87)=9.83, p=.002); they did not differ in terms of their gender, age, education, income, self-health rating, and last doctor's visit. These results suggest that Black non-trusting participants have higher mistrust and find healthcare decisions as less fair and trustworthy than white non-trusting participants.

We also ran all analyses reported above controlling for gender. In the high GBMMS group, gender was insignificant; in the low GBMMS group, gender had a marginal effect (p=.06), females trusting decisions less (Mean=5.09 (SE=.19)) than males (Mean=5.54 (SE=.14)). This effect did not change the significant main effect of the decision-maker (AI vs human). Compared with Mallari et al. [28] that found a significant gender effect in the evaluation of Black vs white defendants, we believe different tasks contributed to the difference in findings. In their study, participants assessed others, whereas in our study, participants answered about themselves.

4 STUDY 2: INTERVIEW

4.1 Method

We conducted 30 minute semi-structured interviews with 21 participants in order to get qualitative insights on how participants perceive AI decisions in healthcare and learn their ideas on what information could make AI more trustworthy. We first began by understanding participants' general perceptions of healthcare AI and what makes it trustworthy; we then gave them examples of AI descriptions in order to probe what information could help AI be more trustworthy.

4.1.1 Participants. We recruited participants by posting a HIT asking people to fill out a survey with an optional field to write down

[2]We note that the interaction effects of race and decision-maker, gender and decision-maker, and race and gender were insignificant.

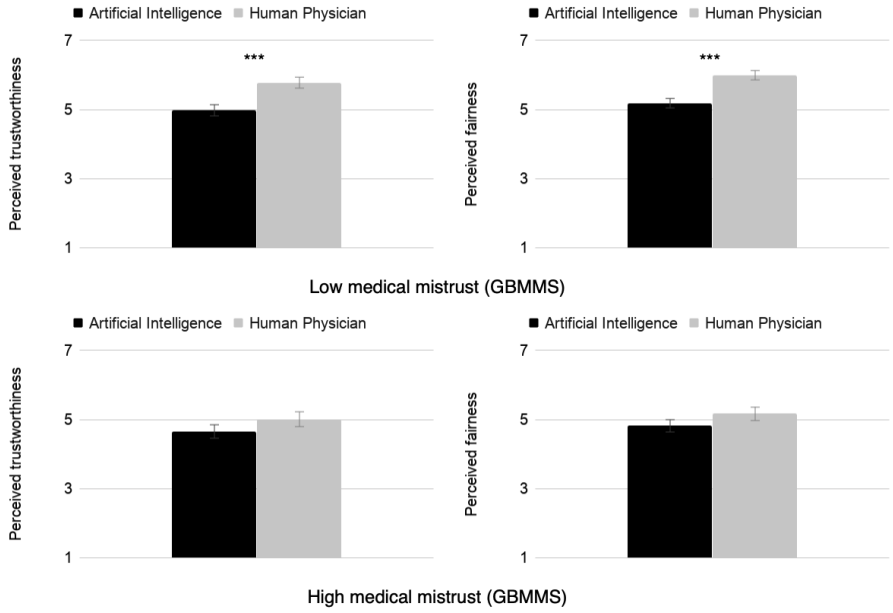

Figure 1: Perceived fairness and trustworthiness depending on the decision-maker in the low versus high mistrust (GBMMS) groups. a) Participants with low medical mistrust perceive the algorithmic decision as less fair and less trustworthy than the the human decision. b) Participants with high medical mistrust perceive the algorithmic decision as fair and trustworthy as the human decision.

their email address if they are interested in a 30 minute remote interview for a $20 Amazon gift card. We began by inviting our experiment participants, then recruited new participants due to a low response rate. For new participants, the recruiting survey included the GBMMS scale as well as demographic questions. Based on the responses, we sampled participants mainly with high GBMMS scores but included a few participants with low GBMMS scores. Seventeen out of twenty-one total participants had relatively high GBMMS scores. Our participants had scores ranging from 2 to 6.75 and had an average GBMMS score of 4.126. Six identified as men, fourteen identified as women and one identified as a transgender man. Participants ranged from 21 to 79 years old with an average age range of 30-39 years old. Fifteen of our participants identified as Black and six identified as white. One of the participants was a nursing student and one mentioned they had extensive experience receiving care from dermatologists for skin cancer. The full participant information is reported in Table 3 in Appendix C.

4.1.2 Procedure and Interview Questions. The interviews were conducted through video chat or telephone and recorded. We first asked participants to provide their personal definitions of fairness and trust and their experiences with human doctors. We then asked about their perceptions of healthcare AI and what they wanted to know about it. After this first part of the interview, participants were shown four descriptions of the same AI designed to detect skin cancer to understand how people respond to different descriptions of AI. The first description used the description from our experiment (Baseline AI Description). The second description included more detailed information on types of algorithms, training data sets and accuracy using language from real dermatology AI [11] (Data-Driven AI Description). The third and fourth descriptions added different statements about the AI fairness to the Data-Driven AI Description. These statements about fairness were drawn from Google AI's Responsible AI Practices [1]. The third description assured the user that a broad definition of fairness was being applied to the AI (Fairness-Driven AI Description) while the fourth description specifically committed to eliminating racial bias, gender discrimination, and all other forms of discrimination (Anti-Discrimination AI Description). These descriptions were meant to give participants potential examples and measure their reaction to different communication strategies. The full descriptions shown to participants are available in Appendix B. With each scenario, participants were asked to read the description out loud and share their thoughts about each one. They also provided their ratings (on a seven-point scale) of how fair and trustworthy they believed the description made the AI sound. At the end, they were asked to compare the descriptions and make suggestions about what measures might improve their trust or perception of fairness.

4.1.3 Analysis. The interview recordings were first transcribed using otter.ai, a natural language processing based transcription service. Two researchers went over the transcripts to verify quality and manually fix any automatic transcription errors. We followed a qualitative data analysis method [7, 32]. Using Dedoose[3], one researcher coded the transcripts at the sentence or paragraph level. The codes were discussed in meetings with another researcher, and were further grouped based on emerging themes. We grouped

[3]https://www.dedoose.com

participants in low, medium, and high mistrust groups: There were four participants with low GBMMS scores ranging from 2 to 2.75, nine participants with medium GBMMS scores ranging from 3.92 to 4.67, and eight participants with high GBMMS scores from 4.75 to 6.75. When synthesizing the codes, participants' GBMMS scores, race, and gender are considered together to produce a more nuanced understanding of how cultural mistrust and personal experiences with difference might impact perceptions of healthcare AI. We report the number of participants' responses that belong to each theme to show their distribution in our sample. This is a small sample, qualitative study, so these numbers do not imply different theme's relative importance or generalizability in a bigger sample.

4.2 Findings

Our findings describe participants' experiences with human doctors and their role in medical mistrust, belief on how biased AI could or could not be, desired information about healthcare AI, and responses to our AI descriptions. Throughout this section, high, medium, low mistrust will be used to describe participants with corresponding GBMMS scores.

4.2.1 Medical Mistrust and Experiences with Human Doctors. Overall, there was a noticeable pattern between participants' mistrust levels and how they described their previous experiences with doctors. Participants with low mistrust tended to regard their previous experiences with doctors as positive. Conversely, participants with high mistrust in our study were more likely to describe their experiences as unpleasant, mixed, or show general skepticism towards medical systems. Some participants also gave responses that indicate the identity of their doctor made a difference in how they perceived their treatment.

Fourteen of the participants, including all four with low mistrust, described their previous experiences with human doctors as relatively positive or did not elaborate about their healthcare experiences extensively. All of the participants reported that getting access to healthcare was relatively easy for them, but some complained about the availability of their doctors.

There were important differences among participants with different mistrust. Seven participants out of seventeen with high and medium mistrust, disclosed negative experiences with healthcare providers. Six of these seven were Black. Although most did not explicitly describe their encounters as racist or otherwise discriminatory, many had concerns about not being listened to or talked down to by their providers. Participant 9 was among those with medium mistrust who described her experiences in medicine as positive and negative depending on the context. She explained that in the emergency room, doctors often assumed she had diabetes because of her weight and she had to clarify that "every bigger person doesn't have diabetes." She also noted that her tendency to seek out nonwhite doctors may explain why her medical experiences outside of emergency rooms were mostly positive. Although she did not express direct criticisms of white doctors, Participant 9 felt her excellent experiences with previous doctors might be explained by the lack of white doctors she had seen throughout her life. She also explained that the woman of color doctor she used to see was one of the best experiences she could remember. Participant 19 was another person with a high mistrust who disliked their experiences with doctors, specifically in the United States. She was originally from another country and said that when it came to seeking healthcare in the United States, "I just don't feel safe or feel comfortable going as often as I normally would."

There were participants with high and medium mistrust who felt they had positive experiences with healthcare professionals. Some elaborated on this being a byproduct of their environment or their relationship with a particular provider. For instance, Participant 14, a Black woman with high mistrust, shared "I live in a primarily Black community, and I am Black. So I've mainly been treated by Black doctors or minority doctors. So I feel like for the most part I've been treated fairly." This statement suggests that, in particular contexts, high mistrust might be an indicator of how participants mistrust future medical interactions or the larger medical system, but have a different attitude towards individual doctors in their communities. In other words, the location and the identity of medical professionals may have played a role in how participants previously experienced medical systems.

4.2.2 AI as Biased versus Unbiased. Participants' mistrust and racial backgrounds had a noticeable association with their preexisting notion of how biased or unbiased AI is. Participants with high mistrust in our study worried about bias in AI whereas participants with varying levels of mistrust were more likely to believe AI could not be biased. These notions were expressed when we asked about how fair they thought the AI could be before they were exposed to our descriptions about healthcare AI.

Six participants with all levels of mistrust indicated that AI were unbiased and could not be unfair. Half of these participants were white and half were Black. Three of these respondents, two of whom were Black with medium to high mistrust and one who was white with low mistrust, indicated that doctors were biased towards the Black and African American community and that AI may circumvent this issue as they thought AI was unbiased. For example, when comparing human doctors to AI, Participant 5 remarked if human doctors have something against maybe a certain race of people or a certain gender people, they may bring that with them when they're practicing medicine. "I mean, you might not know it, but it's there. And so that is something that you wouldn't experience with AI." Other participants who shared the belief that AI could not be biased did not mention potential biases in human doctors but felt, as Participant 2 shared, "if the AI is purely scientific, if it can be purely based on facts, I don't see how it could be unfair." All six of these participants expressed being surprised by the AI description that included fairness and anti-discrimination policy in the later session of the interview.

Black participants with higher mistrust were more likely to express concern about healthcare AI's potential to be biased than white participants with lower mistrust. Six participants with medium and high mistrust raised concerns about if the AI would be biased towards certain people before being shown our AI descriptions. All of these participants identified as Black and most cited racial bias as a concern. As Participant 21 shared, "what was the demographic of the people tested on because for a variety of reasons, generally speaking, any kind of AI or computer systems tend to be biased towards Black people."

The mixture of participants who believed AI would be unbiased versus those who suspected it would be biased towards Black people is a notable finding. All of the participants who suspected the AI might be biased were Black and had moderate to high mistrust. There were three Black participants who thought AI could not be biased, and none of the white participants expressed a concern about bias prior to their exposure to our AI descriptions. Participant 3, a white male participant with a low mistrust even indicated at the end of the interview that he regretted not considering the potential for algorithmic bias before the final description prompted him to do so. As Participant 3 put it, "so reading this makes me realize my own bias. In this whole thing, because I wasn't even thinking about the potential, I was thinking more about privacy. To be honest with you, I wasn't even thinking about this in terms of race in terms of, I guess, you know, I'm guilty cuz I kind of assumed that that was baked in but that's that's my white privilege for you." This pattern suggests that even some participants who are open to the idea that AI can be biased, may not consider the possibility of bias without being prompted to do so or being personally affected by that bias in some way.

4.2.3 Desired Information about Healthcare AI. Participants mentioned, prior to seeing our descriptions of healthcare AI, several information types that they would like to see to trust healthcare AI: length of use, privacy, accuracy, clarity about the AI design and decision process such as information about the data set and the involvement of human oversights, and official certification. The desired information type itself had minimal patterns related to GBMMS scores, race, or gender, but participants with higher mistrust sought out more details for each information type, particularly privacy, length of use, and accuracy specific to Black people or different skin colors.

Four high and medium mistrust participants were curious about how long the AI had been in use. They indicated that they would find the technology more trustworthy and reliable if they knew it had been used for a considerable amount of time. As Participant 15 explained, "I would want to know how long the system has been in use. If it was like purely experimental or like tried and true." Three participants with medium to high mistrust indicated they had concerns about the AI's privacy standards and who would have access to their data. All three of these participants were 21-29 year old Black women with medium or high mistrust. Their individual concerns about privacy varied. Participant 9 wondered if the images in the AI's data set were given voluntarily. Participant 15 worried that "if I am sending images, especially if it's in like a more concealed part of my body, I'm concerned with them just possibly be leaked or hacked somehow."

Accuracy was one of the most mentioned with thirteen participants out of twenty-one total. These responses often involved inquiries about percentages and numerical details about how consistent the AI was. The participants did not indicate what specific percentage or numerical assurance of accuracy they preferred. However, Black participants with medium and high mistrust were the only ones who mentioned how accurate AI might be on certain skin colors and wanted to know more details about how accurate it was with darker skin tones. The same six participants who expressed concerns about bias in AI included potential racial disparities as a part of their worries about accuracy. For example, Participant 21 shared, "I really need to know who they tested this stuff on [...] sensors in general for things originally, they had a very hard time picking up dark skinned people." This particular finding suggests that Black people with high mistrust may be more likely than white people with low mistrust to be concerned about accuracy on darker skin tones. In short, participants with high mistrust outlined a more specific accuracy concerned with skin color than low mistrust participants who were concerned with accuracy at large.

Four participants across all mistrust levels mentioned wanting some kind of official entity or certification to assure them of the AI's reliability. For instance, Participant 18 remarked "I would want to get some stamp of approval like the refrigeration is the cars that we drive." Five participants across all mistrust levels wanted to understand how the AI analyzed pictures and what process it went through when it learned to diagnose skin cancer. Two Black participants with high mistrust specifically wanted to know about the data set the AI learned from and where those images were obtained from. Two of these five participants with varying mistrust levels specifically wanted to know if a human doctor would be part of the process or double-check the AI's results. These results indicate that information about the processes of AI development and usage could improve people's trust in AI.

4.2.4 Responses to Different AI Descriptions. In this section, we report participants' responses to our Data-Driven, Fairness-Driven, and Anti-Discrimination AI Descriptions. The Data-Driven AI Description was generally regarded as stronger than the Baseline AI Description that lacked specific details about the AI. The Fairness-Driven AI Description received mixed responses from participants with low mistrust who did not believe the AI was capable of being unfair. Participants with high mistrust sometimes appreciated the statement, but felt it did not go far enough or explicitly name the potential for racial bias in AI. Most participants across all levels of mistrust preferred the Anti-Discrimination AI Description, even if they were not concerned with algorithmic bias at the beginning of the interview. The participants with the highest mistrust in this sample thought it was good that equity was considered in the Anti-Discrimination AI Description, but were ultimately skeptical of how genuine and effective it was in preventing bias.

Participants generally agreed the Data-Driven AI Description offered much stronger content that made them more likely to trust the AI compared to the Baseline AI Description. All of the participants expressed that the specific numerical data in this description offered them some kind of reassurance. For instance, Participant 11 said, "I didn't think that it would be as good as it was. Okay, over 129,000 and it tested about the same as a human dermatologist. My thoughts are like woah."

Participants with high mistrust tended to have more questions about the Data-Driven AI Description than those with low mistrust. More precisely, ten of the seventeen participants with medium and high mistrust had further questions about which human dermatologists were being compared to AI and what their accuracy rate was. More information, such as a specific percentage of the AI's accuracy or an explanation of its process, was also requested by these respondents. The six participants who had concerns about bias, particularly the AI's ability to diagnose different skin colors,

were impressed by the extensive number of images in the data base but three of them wanted to know details about how many of those images were representative of different skin colors.

The Fairness-Driven AI Description elicited a mixed set of responses that differed across GBMMS scores in this sample. All of the low mistrust participants, particularly those who believed AI were unbiased, said they did not need an assurance about fairness but did not mind it. As Participant 4 put it, "it's going to be fair anyway. So for me, this doesn't really (like) add anything." This group of people considered the AI in the Fairness-Driven AI Description to be equally trustworthy and fair compared to the Data-Driven AI Description.

The six participants who previously expressed concerns about AI bias described the Fairness-Driven AI Description as hopeful but ultimately vague. Participant 20, a Black transgender person with one of the highest GBMMS scores in our study, responded negatively to the description. They said: "in my personal experience of people who use the phrase fairness is that it's been a really white based culture." Three other participants with medium and high mistrust questioned why this assurance about fairness was even needed in the first place. Participant 5 was particularly clear that this description made her worry more about the AI because she had previously not considered fairness. On a different note, Participant 16, a white man with high mistrust, disliked this description citing fairness as a political concept that he felt did not need to be part of science. In summary, participants with high mistrust had stronger reactions and opinions involving the Fairness-Driven AI Description than those with low mistrust.

When faced with the Anti-Discrimination AI Description, fifteen of the twenty-one overall participants indicated it was their preferred description. Three of the four low mistrust participants indicated it made no difference for them personally but they could see why other people might like it. Participant 3, a white man with low mistrust, speculated that the anti-discriminatory language might be comforting for some but wanted to know more about tangible steps the creators of the AI were taking to ensure equity. Nevertheless, all four participants with low mistrust indicated that the Anti-Discrimination AI Description was the best of the options.

Seventeen participants with medium and high mistrust had less converging responses. Three of them participants preferred the Data-Driven AI Description because they believed the AI was unbiased in the first place. Another three preferred the Fairness-Driven AI Description because it was broader. As Participant 18, a Black woman with high mistrust explained, "It doesn't fit the discriminatory. I don't…I think fairness sounds sounds much more appropriate than discriminatory and fairness sounds a lot better than the breakdown of race, gender biases." The remaining eleven participants who did prefer the Anti-Discrimination AI Description either responded with praise, cautious optimism, or skepticism. Participants with higher mistrust were more likely to share their criticisms or worries about this description than those with medium mistrust. Participant 15, a Black woman with high mistrust stated, "I think it's better than the last statement (Fairness-Driven AI Description). Just because we know how they're trying to promote equity in this city, the anti-discrimination measures that they're targeting. Um, but again, how?" Some were even more hesitant to say that they preferred the Anti-Discrimination AI Description because they felt it was not sincere. Participant 21, a Black man with the highest GBMMS score in our study, felt the description was "lip service" and that this could not be a true commitment to anti-discrimination "unless they're going out a way to give this stuff to Black people and other people of color for free. Or at least run them through clinical trials and recruiting in areas like that." Participant 20 had similar concerns about the authenticity of this description and wondered if the anti-racism aims of the hypothetical programmers were reflected in their hiring practices. Our findings indicate that Black participants with medium or low group-based medical mistrust tended to respond positively to the Anti-Discrimination AI Description while Black participants with high mistrust were more likely to be critical of the statement as an indicator of actual equity.

5 DISCUSSION

Our online experiment and interview results suggest that participants who mistrust human medical providers such as doctors and nurses perceive healthcare AI as equally untrustworthy and as unfair as human medical providers. The use of the mistrust scale allows us to account for varying Black experiences rather than using race as a single demographic category. Not all Black participants reported experiencing high mistrust, but those who did had significantly higher mistrust than white participants. The interviews shed light on important intersections and factors that solely focusing on demographic categories could not account for, such as personal experiences with medical discrimination.

Our studies have several implications for research around human perceptions of AI decisions. Many studies use human decision-making as a benchmark to measure people's perceptions of AI, and their findings indicate that people trust AI decisions significantly less than human decisions, particularly for decisions that are subjective and/or deemed to require human capabilities [6, 23, 26, 27]. Our work points out a critical gap in prior work: it does not reflect important differences in social groups. While most of the existing studies do not report demographic information about their participants, all studies were conducted on mTurk, whose users are predominantly white [15]. In the one study that reported participant demographic information, 79% of participants were white while 3.8% were Black [23]. It is likely that the prior work predominantly reflects the perspectives of those who trust human systems. Our work suggests that it is important to acknowledge that not everyone perceives human decision-making the same way, especially if they experience marginalization at the hands of other humans. This calls for future research that purposely recruits and studies different social groups and different dimensions that can account for individual differences in experiences with AI. The practice of reporting ethnoracial information about participants should be further encouraged so that the research community can collectively examine whose perspectives are included and whose perspectives are left out.

Additionally, our findings suggest that previous work's approach of increasing the role of human decision-makers in AI systems to increase trust may be ineffective for Black people with high mistrust. For example, in emerging studies, human augmentation, or human decision-makers using AI, has been proposed as a way to improve people's trust in AI [27]. This work raises an important question

as to whether human augmentation will improve trust in AI even when people have higher mistrust in human systems. Future studies would need to investigate whether one approach will universally improve people's trust in AI and/or whether a different approach would be needed. This is a critical step to avoid unintentionally reinforcing existing health inequity by making healthcare AI more trustworthy only for a certain group.

Our interview findings also highlight that healthcare AI should be designed with an explicit consideration for people who currently experience more negative medical encounters. Some interview participants highlighted how their medical experiences were influenced by sizeism, transphobia, ableism, sexism, and other factors. While a recent paper starts to recognize that healthcare AI may diminish existing trust between physicians and healthcare consumers [21], it does not recognize certain social groups' high medical mistrust in healthcare providers. Previous work in HCI considers how communities of color seek safe spaces and healing from microaggressions on digital platforms [39]. Healthcare AI, however, presents a different sort of challenge, as interpersonal and systematic forces are intertwined in medicine. As our interview results suggest, Black patients who experience mistrust are not only seeking refuge from the verbal microaggressions they may face from human doctors; the quality and accuracy of the care they receive was a particularly noticeable concern among these participants. While some were assured by commitments to anti-discrimination in theory, practical evidence and data of anti-racist aims were required for participants with higher levels of mistrust. Individuals with these concerns often asked directly about how representative the data set and training materials were for people with darker skin tones. They also suggested providing a link to a website on the anti-discrimination clause with a more extensive explanation of how the programmers were committed to anti-racism in their technology and their workplace. This finding suggests that building trust with Black communities with high cultural mistrust requires more than simply articulating intentions. Rather than aspiring to be equitable, practitioners in healthcare AI may benefit from explicitly demonstrating how equitable the process of building their product is. This reveals the need for a scaffolding interface for healthcare AI information, as well as research on healthcare AI-specific guidelines for the regulation of training data sets that can gain people's trust.

Finally, as indicated by participants' responses before being exposed to our example AI descriptions, this study also raises questions about how many people are concerned about AI's potential to be biased and how many people believe AI cannot be biased. It is important to understand what factors might lead someone to hold each of these beliefs.

We also would like to highlight that studying underrepresented groups with a crowdsourcing platform poses a challenge. A recent report suggests that the mTurk population is not representative of U.S. working adults and that people of color are underrepresented [15]. For example, only 7% of the mTurk population in 2016 was Black. This means that Black people, Indigenous people and other people of color will be underrepresented in mTurk studies unless researchers target specific populations. In our study, we learned that seeking Black participants requires additional fees, and one might need to use higher incentives for people to participate. This financial factor could serve as a barrier in encouraging researchers to study underrepresented populations. Future research and community efforts should explore how we can enable more accessible routes to studying the experiences of people of color.

6 LIMITATIONS

Like any study, our study has many limitations that readers should take into consideration. Our participants were recruited through Amazon Cloud Research's mTurk Toolkit and shared opinions on one hypothetical medical service through online experiments and interviews. The findings would need to be further evaluated with real services across diverse contexts, for example, other medical services or decisions in government or management systems. We believe our findings will not be applicable to a setting where human systems are universally mistrusted. In our experiment, we did not vary the descriptions of human physician and AI, which should be further explored. Additionally, both the interview and experiment sample had a very small percentage of transgender participants. Future studies could benefit from seeking to understand how transgender and non-binary individuals experience medical mistrust in relation to healthcare AI. Finally, we focused on Black and white populations recruited via mTurk; future studies would need to investigate perceptions among other people of color such as the Latinx and Asian populations and non-mTurk users.

7 CONCLUSION

As algorithmic systems continue to play a more substantial role in various institutions and processes, the perceptions of social groups with high levels mistrust in human systems offer insights for the development and adoption of these technologies. Our findings from the online experiment and qualitative interviews evidence the complexity of addressing different perspectives within the Black community and among people with differing levels of mistrust. The implications for HCI research are discussed in detail and future directions for this research are also provided. Ultimately, our aim is to make this study a starting point for considering perceptions of algorithmic fairness and trust among different social groups and people with intersecting identities who struggle to trust the systems they dwell in.

ACKNOWLEDGMENTS

We thank our participants who provided valuable insights and Aashna Lal who helped us with the literature review, study design, and paper writing. Our research was supported by National Science Foundation Award IIS-1939606, UT Austin Humanities Institute Grant for Healthcare Technology, Communication, and Privacy, and Good Systems[4], a UT Austin Grand Challenge to develop responsible AI technologies.

REFERENCES

[1] [n.d.]. Responsible AI Practices – Google AI. https://ai.google/responsibilities/responsible-ai-practices/?category=fairness. (Accessed on 09/13/2020).
[2] [n.d.]. Skin Cancer Facts & Statistics - The Skin Cancer Foundation. https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/. (Accessed on 09/13/2020).
[3] MJ Arnett, Roland J Thorpe, DJ Gaskin, JV Bowie, and TA LaVeist. 2016. Race, medical mistrust, and segregation in primary care as usual source of care: findings

[4] https://goodsystems.utexas.edu

from the exploring health disparities in integrated communities study. *Journal of Urban Health* 93, 3 (2016), 456–467.
[4] Karen Stansberry Beard and Kathleen M Brown. 2008. 'Trusting'schools to meet the academic needs of African-American students? Suburban mothers' perspectives. *International Journal of Qualitative Studies in Education* 21, 5 (2008), 471–485.
[5] Ruha Benjamin. 2019. *Race after technology : abolitionist tools for the New Jim Code.* Polity, Cambridge, UK Medford, MA.
[6] Noah Castelo, Maarten W Bos, and Donald R Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825.
[7] Juliet Corbin, Anselm Strauss, and Anselm L Strauss. 2014. *Basics of Qualitative Research.* Sage.
[8] John Danaher. 2016. The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology* 29, 3 (2016), 245–268.
[9] John Danaher, Michael J Hogan, Chris Noone, Rónán Kennedy, Anthony Behan, Aisling De Paor, Heike Felzmann, Muki Haklay, Su-Ming Khoo, John Morison, et al. 2017. Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society* 4, 2 (2017).
[10] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.
[11] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.
[12] Mark Fondacaro, Bianca Frogner, and Rudolf Moos. 2005. Justice in health care decision-making: Patients' appraisals of health care providers and health plan representatives. *Social justice research* 18, 1 (2005), 63–81.
[13] Cary Funk, Meg Hefferon, Brian Kennedy, and Courtney Johnson. 2019. Trust and mistrust in Americans' views of scientific experts. *Pew Research Center* (2019).
[14] Peter W Groeneveld, Seema S Sonnad, Anee K Lee, David A Asch, and Judy E Shea. 2006. Racial differences in attitudes toward innovative medical technology. *Journal of general internal medicine* 21, 6 (2006), 559–563.
[15] Paul Hitlin. 2016. Research in the crowdsourcing age, a case study. *Pew Research Center* 11 (2016).
[16] John Hoberman. 2012. *Black and blue: The origins and consequences of medical racism.* Univ of California Press.
[17] Rana A Hogarth. 2017. *Medicalizing Blackness: making racial difference in the Atlantic world, 1780-1840.* UNC Press Books.
[18] Bernice Roberts Kennedy, Christopher Clomus Mathis, and Angela K Woods. 2007. African Americans and their distrust of the health care system: healthcare for diverse populations. *Journal of cultural diversity* 14, 2 (2007).
[19] Roderick M Kramer. 1999. Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual review of psychology* 50, 1 (1999), 569–598.
[20] Markus Langer, Cornelius J König, and Maria Papathanasiou. 2019. Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment* 27, 3 (2019), 217–234.
[21] Emily LaRosa and David Danks. 2018. Impacts on trust of healthcare AI. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* 210–215.
[22] Chioun Lee, Stephanie L Ayers, and Jennie Jacobs Kronenfeld. 2009. The association between perceived provider discrimination, health care utilization, and health status in racial and ethnic minorities. *Ethnicity & disease* 19, 3 (2009), 330.
[23] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 1–16.
[24] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.* ACM, 1035–1048.
[25] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems.* 1603–1612.
[26] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.
[27] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. 2019. Resistance to medical artificial intelligence. *Journal of Consumer Research* 46, 4 (2019), 629–650.
[28] Keri Mallari, Kori Inkpen, Paul Johns, Sarah Tan, Divya Ramesh, and Ece Kamar. 2020. Do I Look Like a Criminal? Examining how Race Presentation Impacts Human Judgement of Recidivism. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–13.
[29] Rich Morin and Renee Stepler. 2016. The racial confidence gap in police performance. *Pew Research Center* 29 (2016).
[30] Safiya Noble. 2018. *Algorithms of oppression : how search engines reinforce racism.* New York University Press, New York.
[31] Shayla C Nunnally. 2012. *Trust in Black America: Race, discrimination, and politics.* NYU Press.
[32] Michael Q Patton. 1980. *Qualitative Research and Evaluation Methods.* Sage.
[33] Dorothy E Roberts. 1999. *Killing the black body: Race, reproduction, and the meaning of liberty.* Vintage.
[34] Rachel C Shelton, Gary Winkel, Stacy N Davis, Nicole Roberts, Heiddis Valdimarsdottir, Simon J Hall, and Hayley S Thompson. 2010. Validation of the group-based medical mistrust scale among urban black men. *Journal of General Internal Medicine* 25, 6 (2010), 549–555.
[35] Harold Sigall and Judson Mills. 1998. Measures of independent variables and mediators are useful in social psychology experiments: But are they necessary? *Personality and Social Psychology Review* 2, 3 (1998), 218–226.
[36] Francis Terrell and Sandra L Terrell. 1981. An inventory to measure cultural mistrust among Blacks. *Western Journal of Black Studies* 5, 3 (1981), 180–184.
[37] David H Thom and Bruce Campbell. 1997. foTlG INALRESEARCH Patient-Physician Trust: An Exploratory Study. *The Journal of family practice* 44, 2 (1997), 169.
[38] Hayley S Thompson, Heiddis B Valdimarsdottir, Gary Winkel, Lina Jandorf, and William Redd. 2004. The Group-Based Medical Mistrust Scale: psychometric properties and association with breast cancer screening. *Preventive medicine* 38, 2 (2004), 209–218.
[39] Alexandra To, Wenxia Sweeney, Jessica Hammer, and Geoff Kaufman. 2020. " They Just Don't Get It": Towards Social Technologies for Coping with Interpersonal Racism. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–29.
[40] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, 440.
[41] Christopher W Wheldon, Stephanie K Kolar, Natalie D Hernandez, and Ellen M Daley. 2017. Factorial invariance and convergent validity of the group-based medical mistrust scale across gender and ethnoracial identity. *Journal of Health Care for the Poor and Underserved* 28, 1 (2017), 88–99.
[42] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 chi conference on human factors in computing systems.* 1–14.

A EXPERIMENT SCENARIO DESCRIPTIONS

Please read the following scenario carefully and imagine the situation as vividly as possible, as if you were going through it in real time.

Skin cancer is the most common cancer in the United States. 1 in 5 Americans will develop skin cancer in their lifetime. People of all colors and races can get skin cancer. There are many different types of skin cancer, including carcinoma and melanoma. Carcinoma is the most common form of skin cancer; melanoma is the deadliest. With early detection and proper treatment, both carcinoma and melanoma have a high cure rate.

A.1 AI Condition

Imagine that the American Academy of Dermatology is promoting a new type of preventive screening for skin cancer. This initiative is meant to foster detection of skin cancer in its earliest, most treatable stage, to reduce the incidence of the skin cancer and raise awareness of effective skin cancer prevention techniques.

Imagine that you decide to get this particular screening. The screening service is provided through a website. The service asks you to take photos of the skin around your scalp, face, mouth, hands, feet, trunk and extremities, eyes and eyelids, ears, fingers, toes and toenails, but you can ask for more (or fewer) areas to be checked. You will submit the photos to a website to be processed.

Then, the photos of your body will be sent to an Artificial Intelligence (AI) dermatologist trained to develop pattern recognition skills. This dermatologist is an algorithm trained to distinguish between cancerous and non-cancerous skin conditions. Training was made possible through learning the differences between cancerous and non-cancerous skin conditions using an extensive dataset of images. This AI dermatologist will analyze the photos of your body,

and email the results and recommendations back to you within one week from the screening.

A.2 Physician Condition

Imagine that the American Academy of Dermatology is promoting a new type of preventive screening for skin cancer. This initiative is meant to foster detection of skin cancer in its earliest, most treatable stage, to reduce the incidence of the skin cancer and raise awareness of effective skin cancer prevention techniques.

Imagine that you decide to get this particular screening. The screening service is provided through a website. The service asks you to take photos of the skin around your scalp, face, mouth, hands, feet, trunk and extremities, eyes and eyelids, ears, fingers, toes and toenails, but you can ask for more (or fewer) areas to be checked. You will submit the photos to a website to be processed.

Then, the photos of your body will be sent to a dermatologist trained to develop pattern recognition skills. This person is a dermatologist trained to distinguish between cancerous and non-cancerous skin conditions. Training was made possible through learning the differences between cancerous and non-cancerous skin conditions using an extensive dataset of images. This dermatologist will analyze the photos of your body, and email the results and recommendations back to you within one week from the screening.

B AI DESCRIPTIONS USED AS PART OF THE INTERVIEWS

B.1 Baseline AI Description

The American Academy of Dermatology is promoting a new type of preventive screening for skin cancer. This initiative is meant to foster detection of skin cancer in its earliest, most treatable stage, to reduce the incidence of the skin cancer and raise awareness of effective skin cancer prevention techniques. The screening service is provided through Artificial Intelligence (AI). This algorithm is trained to distinguish between cancerous and non-cancerous skin conditions. Training was made possible through learning the differences between cancerous and non-cancerous skin conditions using an extensive dataset of images.

B.2 Data-Driven AI Description

The American Academy of Dermatology is promoting a new type of preventive screening for skin cancer. This initiative is meant to foster detection of skin cancer in its earliest, most treatable stage, to reduce the incidence of the skin cancer and raise awareness of effective skin cancer prevention techniques.The screening service is provided through Artificial Intelligence (AI). This algorithm, Deep Convolutional Neural Networks (CNNs), is trained to distinguish between cancerous and non-cancerous skin conditions. Training was made possible through learning the differences between cancerous and non-cancerous skin conditions using an extensive dataset of images. More specifically, it is trained using a dataset of 129,450 clinical images consisting of different types of cancer. Its performance was tested against 21 board-certified dermatologists on biopsy-proven clinical images. It achieves performance on par with all tested experts across different tasks, demonstrating an AI capable of classifying skin cancer with a level of competence comparable to human dermatologists.

B.3 Data-Driven AI Description with the Fairness Statement

The American Academy of Dermatology is promoting a new type of preventive screening for skin cancer. This initiative is meant to foster detection of skin cancer in its earliest, most treatable stage, to reduce the incidence of the skin cancer and raise awareness of effective skin cancer prevention techniques.The screening service is provided through Artificial Intelligence (AI). This algorithm, Deep Convolutional Neural Networks (CNNs), is trained to distinguish between cancerous and non-cancerous skin conditions. Training was made possible through learning the differences between cancerous and non-cancerous skin conditions using an extensive dataset of images. More specifically, it is trained using a dataset of 129,450 clinical images consisting of different types of cancer. Its performance was tested against 21 board-certified dermatologists on biopsy-proven clinical images. It achieves performance on par with all tested experts across different tasks, demonstrating an artificial intelligence capable of classifying skin cancer with a level of competence comparable to human dermatologists. Ensuring the fairness of these tools also provides an opportunity and a challenge. The programmers of this AI are committed to building AI tools for healthcare that are fair to all patients.

B.4 Data-Driven AI Description with the Anti-Discrimination Statement

The American Academy of Dermatology is promoting a new type of preventive screening for skin cancer. This initiative is meant to foster detection of skin cancer in its earliest, most treatable stage, to reduce the incidence of the skin cancer and raise awareness of effective skin cancer prevention techniques. The screening service is provided through Artificial Intelligence (AI). This algorithm, Deep Convolutional Neural Networks (CNNs), is trained to distinguish between cancerous and non-cancerous skin conditions. Training was made possible through learning the differences between cancerous and non-cancerous skin conditions using an extensive dataset of images. More specifically, it is trained using a dataset of 129,450 clinical images consisting of different types of cancer. Its performance was tested against 21 board-certified dermatologists on biopsy-proven clinical images. It achieves performance on par with all tested experts across different tasks, demonstrating an artificial intelligence capable of classifying skin cancer with a level of competence comparable to human dermatologists. Ensuring these tools are not discriminatory also provides an opportunity and a challenge. The programmers of this AI are committed to building AI tools for healthcare with anti-discrimination measures that prevent racial bias, gendered bias, and all forms of inequality.

C PARTICIPANT DEMOGRAPHIC

We report our experiment participant demographic information depending on their race (Table 1). We also report their demographic information depending on their group-based medical mistrust scale (GBMMS) score (Table 2). Table 3 includes interview participants' demographic information.

Table 1: Demographic information about our Black and white participants

	Black	**White**
Age	Mean = 35.37 SD = 10.97	Mean = 37.07 SD = 12.44
GBMMS	Mean = 3.93 SD = 1.20	Mean = 2.55 SD = 0.93
Gender	50% Male 50% Female	66.99% Male 33.01% Female
Income	Mean = $59,910.21 SD = $44,935.65	Mean = $42,839.40 SD = $25,622.90
Education		
Less than High School	2.38%	0%
High School / GED	8.33%	17.48%
Some College or Currently in College	20.24%	20.39%
2-Year College Degree	19.05%	5.83%
4-Year College Degree	33.33%	42.72%
Masters Degree	15.48%	11.65%
Doctoral Degree	0%	0%
Professional Degree (JD, MD)	1.19%	1.94%
Health Rating	Mean = 2.54 (Good) SD = 0.80	Mean = 2.51 (Good) SD = 0.78
Last Doctor Visit	Mean = 2.65 (6 months ago) SD = 1.10	Mean = 2.36 (6 months ago) SD = 1.19

Table 2: Demographic information about our participants with low versus high Group-Based Medical Mistrust scale (GBMMS) scores. The median score was used to divide participants into the low versus high GBMMS groups.

	Low GBMMS	**High GBMMS**
Age	Mean = 36.46 SD = 12.64	Mean = 36.12 SD = 10.55
Race	20.20% Black 79.80% White	72.73% Black 27.27% White
Gender	64.65% Male 35.35% Female	53.41% Male 46.59% Female
Race x Gender	11.11% Black Female 9.09% Black Male 24.24% White Female 55.56% White Male	35.23% Black Female 37.50% Black Male 11.36% White Female 15.91% White Male
Income	Mean = $56,819.91 SD = $41,672.30	Mean = $60,118.52 SD = $39,518.70
Education		
Less than High School	1.01%	1.14%
High School / GED	18.18%	7.96%
Some College or Currently in College	21.21%	19.32%
2-Year College Degree	4.04%	20.46%
4-Year College Degree	39.39%	37.50%
Masters Degree	13.13%	13.64%
Doctoral Degree	0%	0%
Professional Degree (JD, MD)	3.03%	0%
Health Rating	Mean = 2.61 (Good) SD = 0.78	2.43 (Good) SD = 0.78
Last Doctor Visit	Mean = 2.36 (6 months ago) SD = 1.19	Mean = 2.64 (6 months ago) SD = 1.12

Table 3: Interview participant demographic information

	Age	Race	Gender	GBMMS
Participant 1	40-49	Black	Man	2.00
Participant 2	50-59	White	Man	2.00
Participant 3	30-39	White	Man	2.08
Participant 4	30-39	White	Woman	2.75
Participant 5	30-39	Black	Woman	3.92
Participant 6	30-39	White	Woman	4.08
Participant 7	70-79	White	Woman	4.17
Participant 8	30-39	Black	Woman	4.25
Participant 9	21-29	Black	Woman	4.33
Participant 10	50-59	Black	Woman	4.42
Participant 11	40-49	Black	Woman	4.50
Participant 12	21-29	Black	Woman	4.67
Participant 13	21-29	Black	Man	4.67
Participant 14	21-29	Black	Woman	4.75
Participant 15	21-29	Black	Woman	5.17
Participant 16	40-49	White	Man	5.17
Participant 17	21-29	Black	Woman	5.25
Participant 18	50-59	Black	Woman	5.33
Participant 19	30-39	Black	Woman	5.42
Participant 20	30-39	Black	Trans Man	5.67
Participant 21	30-39	Black	Man	6.75