How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making?

Saumik Narayanan
Washington University in St. Louis
St. Louis, Missouri, USA
saumik@wustl.edu

Guanghui Yu
Washington University in St. Louis
St. Louis, Missouri, USA
guanghuiyu@wustl.edu

Chien-Ju Ho
Washington University in St. Louis
St. Louis, Missouri, USA
chienju.ho@wustl.edu

Ming Yin
Purdue University
West Lafayette, Indiana, USA
mingyin@purdue.edu

ABSTRACT

This paper explores the impact of value similarity between humans and AI on human reliance in the context of AI-assisted ethical decision-making. Using kidney allocation as a case study, we conducted a randomized human-subject experiment where workers were presented with ethical dilemmas in various conditions, including no AI recommendations, recommendations from a similar AI, and recommendations from a dissimilar AI. We found that recommendations provided by a dissimilar AI had a higher overall effect on human decisions than recommendations from a similar AI. However, when humans and AI disagreed, participants were more likely to change their decisions when provided with recommendations from a similar AI. The effect was not due to humans' perceptions of the AI being similar, but rather due to the AI displaying similar ethical values through its recommendations. We also conduct a preliminary analysis on the relationship between value similarity and trust, and potential shifts in ethical preferences at the population-level.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Computer supported cooperative work**.

KEYWORDS

ethical preference, AI ethics, human reliance on AI

ACM Reference Format:
Saumik Narayanan, Guanghui Yu, Chien-Ju Ho, and Ming Yin. 2023. How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making?. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 08–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3600211.3604709

This work is licensed under a Creative Commons Attribution International 4.0 License.

AIES '23, August 08–10, 2023, Montréal, QC, Canada
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0231-0/23/08.
https://doi.org/10.1145/3600211.3604709

1 INTRODUCTION

Making ethical decisions is challenging, because they often lack clear right or wrong answers. For example, during the early stage of a pandemic, local governments must decide who to vaccinate first when there are not enough vaccines available for everyone. In emergency rooms, doctors must decide how to prioritize patients who need treatments with limited amount of time and medical resources. Social workers also encounter tough choices when allocating limited resources to prevent homelessness. In these situations, decision-makers must weigh various ethical values and principles, making it difficult to find universally acceptable solutions.

In the meantime, artificial intelligence (AI) has gained significant progress in the past decade, and naturally, has been increasingly involved in decision making in our daily life, including decisions in ethically-sensitive domains. While some may fight against the implementation of AI systems being involved in real-world ethical decisions, proponents argue that AI could potentially lead to more equitable outcomes for marginalized communities by minimizing human biases [27]. In addition, the automated nature of AI can substantially speed up decision making to a level that is much faster than what humans can achieve. To leverage the benefits of AI in decision making while alleviating the concerns of having AI making ethical decisions autonomously, one approach which is getting increasing attention is to adopt the paradigm of AI-assisted decision making, where human decision makers receive recommendations from AI, which assist humans to form their final decisions.

While AI-assisted ethical decision making holds promise, incorporating AI recommendations in decision making could also lead to unintended consequences. In particular, AI algorithms exhibit their own ethical values, realized through recommendations they provide to human decision makers. Furthermore, the ethical values exhibited by AI could propagate to final decisions in different ways, depending on whether and when human decision makers decide to adopt AI recommendations. Without more research on the impacts of AI recommendations to humans in ethical decision making, we run the risk of real-world systems outpacing our understanding of these systems, potentially causing real-world harm. For example, if human decision makers always tend to accept recommendations from AI exhibiting similar ethical values and reject recommendations from AI exhibiting different ethical values, we run the risk of creating a more *polarized* decision making environment where human tend to make more extreme decisions.

In this paper, we aim to advance our understanding of incorporating AI recommendations in ethical decision making. Specifically, we investigate the research question of how value similarity between humans and AI affects the human decision makers' reliance on AI recommendations in the context of AI-assisted ethical decision making. Additionally, given the *value* exhibited by an AI is not directly observable, we are also interested in understanding whether the effect of value similarity to human reliance is influenced more by the value the AI *claims* to exhibit or the value that is demonstrated by the recommendations the AI provides.

To answer the above research questions, we have conducted a randomized human-subject experiment on Amazon Mechanical Turk (MTurk). Using the domain of kidney transplants as a case study, we first ask recruited workers to solve a series of ethical dilemmas without AI recommendation to measure their own ethical preference, which is our operationalization of the participant's "value". We then randomly assign workers into two treatments which differ on whether the AI model used in the treatment is similar or dissimilar from the participant's own ethical preference. We compare participants' decision alignment with the AI recommendation across the two treatments to understand how human-AI value similarity impacts human reliance on AI.

We find that recommendations provided by a dissimilar AI has a larger effect on human decisions than recommendations from a similar AI. However, this result is generally due to the high levels of agreement between the similar AI and user, creating less opportunities to "change their mind". If we limit our analysis to the subset of scenarios where humans and AI disagree, humans are more likely to change their decision when provided with recommendations from a similar AI than recommendations from a dissimilar AI. In addition, we find no evidence that this effect is due to humans' perceptions of the AI being similar. Instead, we find that this effect is largely due to the AI actually displaying similar ethical values through recommendations. Finally, we perform an explorative analysis that investigates potential shifts in polarization at the population-level, and find preliminary evidence that personalized AI assistants could lead to a more radicalized decision-making population.

2 RELATED WORK

There has been extensive recent work in understanding how humans rely on their AI teammates in AI-assisted decision making, and this has been studied both in domains with objectively correct decisions to be made, and domains where decisions are made according to subjective ethical practices. Our paper draws from prior work in three categories: how humans rely on AI advice, how humans trust AI advice, and how humans perceive AI values.

In studies of humans' reliance on AI advice, there have been mixed results on whether humans rely more on human advice or AI advice. Many papers have shown evidence of algorithmic aversion, which is the notion that humans tend to relatively distrust AI advice, and prefer to receive advice from other humans [7, 29, 34]. This aversion extends to second and third parties, who may prefer decision-makers to use no advice, rather than AI advice [36, 43]. On the other hand, despite the evidence that decision-makers tend to subjectively prefer human advice over AI advice, Logg et al. [28] found that human-decision makers tend to rely more on AI advice in practice. This finding has been validated not only in objective domains, but ethical decision-making domains where there are no correct answers [32, 41]. One potential explanation is that humans perceive AI to be more rational and unbiased [8]. Human decision-makers may also want to shift the cognitive burden of ethical decision making off of them [25], as society tends to hold humans to higher standards of being unbiased than AI [4].

One aspect which affects human reliance on AI is trust, or more generally, the level of confidence that humans have in AI outputs. Bansal et al. [3] investigated the mental models that humans have in AI behavior, and found that when model outputs are more understandable, humans are better able to incorporate these outputs into their own decision-making strategies, leading to better team performance. Yin et al. [45] looked at the relationship between model accuracy and trust, and found that humans tend to both trust and rely on advice with a higher stated accuracy more than advice with a lower stated accuracy. Schmitt et al. [35] found that when humans are exposed to AI advice and later shown that the prior advice was incorrect, their trust in the AI actually increases. Zhang et al. [47] looked at methods for calibrating human trust in AI, and found that confidence scores improve trust calibration, though this doesn't necessarily improve overall decision making performance.

Our work focuses on the effects of value similarity to human reliance in AI-assisted ethical decision making. There is a rich body of sociological work understanding the effects of value similarity on humans. For example, Sitkin and Roth [38] found that improving reliability is insufficient for restoring trust in interpersonal relationships or inter-organizational mechanisms, and a better method for improving trust is to show value similarity. Siegrist et al. [37] analyzed the effects of value similarity in risk management, and found that increased value similarity leads to increased trust and is a significant predictive factor in the outcome of risk-benefit analysis for new technology.

In the last few years, more work has begun on understanding how value similarity affects interactions between humans and AI assistants. One of the closest work to ours is by Grgić-Hlača et al. [18]. They focused on objective (non-ethical) domains and measured AI similarity by comparing model output with human decisions. Similar to our observations, they found that advice from similar AIs is more likely to change the mind of a human decision maker, but dissimilar AIs have more opportunities to change minds, giving them a bigger overall impact. Mehrotra et al. [31] and Yokoi and Nakayachi [46] both analyzed the effects of value similarity on AI trust in various ethical decision-making domains, and found that AI assistants with a higher value similarity lead to higher levels of trust in the AI assistant. However, the latter two papers only look at subjective measures of trust in these ethical decision-making domains, without empirically validating changes in user reliance. We have already seen paradoxical results when looking at reliance on human and AI advice, where decision-makers prefer and trust human advice more, but rely on AI advice more. As such, we aim to fill this research gap in AI-assisted ethical decision-making, by showing that value similarity in AI recommendations leads to both increased reliance and increased trust.

In this work, we perform experiments in the area of medical resource allocation, specifically, kidney transplant allocation, as a case study. There has been a rich body of literature which has looked

at the ethics of medical resource allocation [12, 13, 17, 33]. Taking from this literature, there have been a few algorithmic experiments understanding human values for kidney allocation. Freedman et al. [16] created a methodology for estimating human values for kidney allocation, and proposed kidney exchange algorithmic improvements which better take into account human values. Narayanan et al. [32] expanded on this research by incorporating both verifiable information and predictive information into the solicitation of human ethical preferences. Research on ethics on scarce allocation actually informs real-world kidney exchange algorithms. For example, the United Network for Organ Sharing published a report detailing changes they made to their kidney algorithm in the last year, and showed that outcomes are now more equitable for racial minorities and other vulnerable groups [15].

3 EXPERIMENT

The aim of our experiment is to investigate the influence of value similarity between humans and artificial intelligence (AI) on human reliance on AI for ethical decision-making. In pursuit of this objective, we present scenarios involving ethical dilemmas to recruited participants and measure their ethical preferences in varying conditions. These conditions include instances where participants are provided with no AI recommendations, recommendations from AI with similar ethical preferences (similar AI), and recommendations from AI with dissimilar ethical preferences (dissimilar AI). We pose two main research questions, and design our experiment to validate the following hypotheses.

Research Question 1: How does value similarity affect reliance on AI recommendations?

- **H1:** Recommendations made by a dissimilar AI will create a greater change in alignment than recommendations made by a similar AI.
- **H2:** When considering scenarios where humans originally disagreed with the AI, recommendations made by a similar AI will cause a greater change in alignment than recommendations made by a dissimilar AI.

Research Question 2: Are the effects of value similarity on reliance caused by claims of value similarity or because the recommendations actually align with human values?

- **H3:** The effect of value similarity is primarily due to humans relying on AI recommendations which claim to share similar values, and it is less important for humans reliance that AI actually follows its claimed values.

3.1 Experiment Task

To test the aforementioned hypotheses, we conduct a case study in which we recruit participants to make a series of ethical decisions pertaining to the allocation of kidneys. Each decision presents participants with a hypothetical scenario where two patient candidates are in need of a kidney transplant, but only one kidney is available. Participants are required to evaluate the information provided about both candidates and express their preference for which candidate should receive the kidney first.

To align our task design with well-established ethical preference frameworks, we follow the extensive literature on the ethical principles in allocating scarce medical interventions [12, 13, 17, 32, 33]. In particular, we adopt the ethical preference framework proposed by Persad et al. [33], which describes four categories of ethical values: Treating People Equally, Favoring the Worst-Off, Promoting Social Usefulness, and Maximizing Total Benefits. Narayanan et al. [32] differentiated between the first three categories and the last, denoting the first three as *verifiable* and the last as *predictive*. They found that this predictive category can have an out-sized effect on the verifiable categories, especially when the prediction is considered to be AI-determined. To avoid these effects, we only display the three verifiable categories in our experiments, and select the following factors to represent these categories.

- *Kidney Donor Status (Promoting Social Usefulness)*: If the candidate has donated a kidney of their own in their past. This is a binary feature, with possible values of {Not prior donor, Prior Donor}.
- *Wait Time (Treating People Equally)*: How long the candidate has been waiting to receive a kidney. This feature has possible values of {Less than 1 year, 1 year, 2 years, 3 years, 4 years, 5 years}.
- *Kidney Disease Stage (Favoring the Worst-Off)*: How severe the candidate's kidney disease is. This is a binary feature, with possible values of {Stage 4 (Severe kidney damage), Stage 5 (Kidney failure or near-failure)}.

It is worth noting that in the ethical principle framework proposed by Persad et al. [33], each factor has a default preference ordering in cases where all other factors are equal. If one candidate is a prior donor, and the other isn't, then the default ordering prioritizes the prior donor. If one candidate has been waiting for a longer period than the other, the default ordering prioritizes this candidate. If one candidate's kidney disease is at a higher stage than the other, the default ordering prioritizes this candidate. In our study, we presented various scenarios to online workers to investigate how individuals make trade-offs between these three factors, which correspond to the stated ethical principles.

3.1.1 Scenario construction. In our experiment, workers are asked to make a series of ethical decisions. Specifically, we generate scenarios with two candidates, and workers are asked to express their ethical preference on which candidate should receive a kidney transplant first. When eliciting workers' ethical preferences, these scenarios can be split into three categories.

The first category includes scenarios where the two candidates differ in only one factor, and share the same values for the other two factors. For example, in one scenario, Candidate A may be a prior donor, while Candidate B is not; both candidates have been waiting for 3 years and have Stage 4 Kidney Disease. The primary objective of this category is to elicit workers' baseline preferences for each of the factors individually (in this case, *Donor Status*). The second category consists of scenarios to understand workers trade-offs between two factors. In this category, the two candidates share the same value for one factor, one factor should prioritize the first candidate, and the remaining factor should prioritize the second candidate (according to the default preference ordering). For example, Candidate A may be a prior donor, while Candidate B is not, Candidate A may have been waiting for 2 years, while

Candidate B has been waiting for 4 years, and both candidates have Stage 5 Kidney Disease. This category enables us to isolate the trade-offs between pairs of factors (in this case, *Donor Status* and *Wait Time*). The third category involves scenarios where the two candidates have different values in all three factors. One candidate is prioritized in one factor, while the other candidate is prioritized by the other two factors. For example, Candidate A may be a prior donor, while Candidate B is not, Candidate A may have been waiting for 2 years, while Candidate B has been waiting for 4 years, and Candidate A may have Stage 4 Kidney Disease, while Candidate B has Stage 5 Kidney Disease. This category enables us to represent more complex interactions between the factors.

In each of these categories, there are three unique scenarios, giving us a total of nine scenarios. For each user, we realize each scenario with random values that preserve the preference order. For instance, if the disease stage needs to be equal, we may display both patients as "Stage 4" or "Stage 5". We also limit wait time differences between candidates to be no more than 2 years.

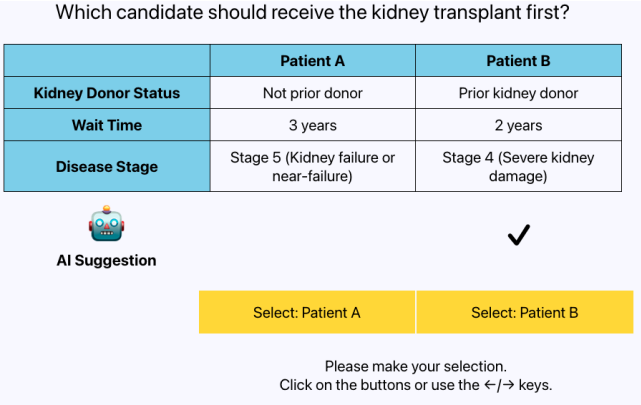

Figure 1: An example of the task interface for our experiment. This interface corresponds to the task of Stage 2 in our experiment design as described in Section 3.2.

3.1.2 Creating AI with Similar/Dissimilar Ethical Preferences. The goal of this work is to investigate the influence of value similarity between humans and AI on human reliance for ethical decision-making. Given our domain application, we use the similarity of ethical preferences to represent the value similarity. We now describe how we create AI with similar or dissimilar ethical preferences with a given worker.

For a worker's ethical preference, we can measure their answers on a set of given scenarios, i.e., their choices on who to receive a kidney first among several pairs of candidates, when they are not provided AI recommendations. Using their answers, we can compute their (prior) ethical preferences without seeing AI recommendations. A worker's ethical preference is represented by three values, each indicating how often workers' answers align with the default ethical ordering of each factor. This alignment is measured separately for each factor, and indicates how often the worker chooses the preferred factor value (e.g. "Prior Donor" over "Not Prior Donor" for the "Donor Status" factor), across all scenarios. For example, in the scenario presented in Figure 1, if the worker selected Patient A, then their answer aligns with the preferred factor for the "Wait Time" and "Disease Stage" factors, but not the "Donor Status" factor. We would then average the number of times the worker aligns with each preferred factor across all scenarios to generate the alignment values for each factor.

Using these values, we use the $A > B > C$ notation to denote a worker's value ordering in their ethical preferences over factors A, B, and C. For example, if a worker aligns with the "Donor Status" factor in 30% of scenarios, with the "Wait Time" factor 80% of the time, and the "Disease Stage" factor in 50% of scenarios, then their prior ethical preference ordering would be "Wait Time">"Disease Stage">"Donor Status".

Based on a worker's value ordering in the prior ethical preference, we can design a similar AI and a dissimilar AI that share similar and dissimilar ethical preferences with the worker. In particular, if a worker's value ordering is $A > B > C$, the ethical preferences for the similar/dissimilar AI for that worker are specified below:

- Similar AI: The ethical preference order for a similar AI is chosen uniformly at random to be either $A > B > C$ or $A > C > B$, i.e., the top factor of the similar AI is the same as the top factor of the worker.
- Dissimilar AI: The ethical preference order for a dissimilar AI is chosen uniformly at random to be either $C > A > B$ or $C > B > A$, i.e., the top factor of the dissimilar AI is the same as the bottom factor of the worker.

Our second research question aims to understand how the claim of similarity affects human reliance on AI recommendations. We therefore need to be able to distinguish between cases where AI is truly following its value preferences, and cases where the AI is only claiming to follow its value preferences, but no more. To create this distinction, we instruct our AI to act as follows:

- Deterministic AI: The AI will deterministically follow its ethical preference ordering. If the AI's top ethical preference has different values for the two candidates, then the AI will pick the candidate whose factor value aligns with its preference. If the values are tied, then the AI will move to the second preference, and then the third if necessary.
- Random AI: The AI chooses the recommendation entirely randomly, without any regard for the candidate attributes.

Using this design, we can distinguish between cases where user reliance is affected by both the value similarity claim and similar recommendations (Deterministic), and cases where user reliance is affected only by the value similarity claim (Random). When we describe the AI to workers in our experiment, we explicitly inform workers of the AI's ethical preferences and that the AI makes stochastic recommendations.

3.2 Experiment Design

To understand the effect of AI similarity on the usage of AI recommendations in ethical decision making, we conducted a two-stage, two-treatment randomized behavioral experiment. A general schematic of our experiment design can be found in Figure 2.

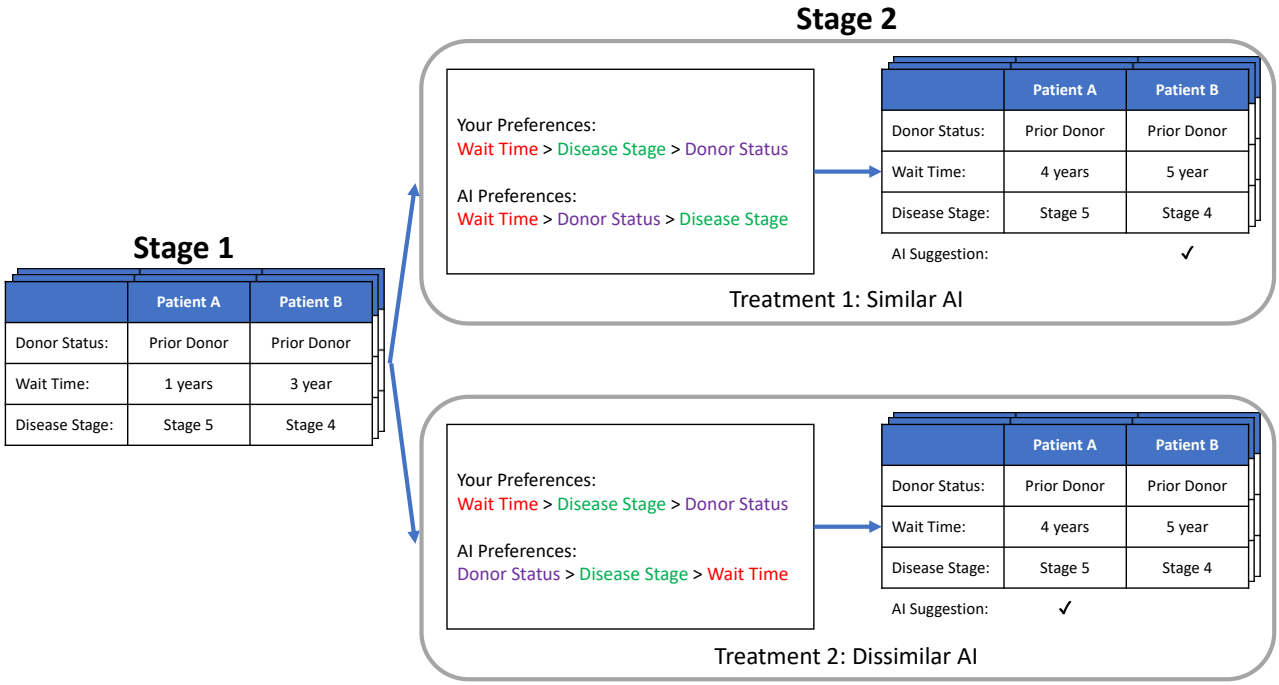

Figure 2: A general illustration of our experiment design. In the first phase, we present the user with a series of scenarios, and use this data to understand the user's ethical preferences. Using this, we create similar and dissimilar AI assistants in the second phase, and display them to the user. We then present the user additional scenarios, with the AI recommendation visible.

In our experiment, each recruited worker begins with the first stage, where they are asked to express their ethical preferences in 9 scenarios, generated using the approach described in Section 3.1.1. After eliciting workers' prior ethical preferences, we then randomly assign workers to two treatments:

- **Treatment 1 (Similar AI)**: In the second stage, each worker in this treatment group are shown recommendations from AI with similar ethical preferences to their own ethical preferences.
- **Treatment 2 (Dissimilar AI)**: In the second stage, each worker in this treatment group are shown recommendations from AI with dissimilar ethical preferences to their own ethical preferences.

After the first stage, workers are presented with a summary of their own ethical preferences and the ethical preference of the AI that will make recommendations during their decision-making during the second stage. Workers are also asked three survey questions — how confident they are in their own answers, if they think our estimation of their preferences is accurate, and how much trust they would have in an AI which behaves according to the displayed preferences. Each of these is graded on a 5-point Likert scale.

In the second stage, workers are presented with 18 additional scenarios where they make their decisions with the assistance of the provided AI. An illustration of our experiment scenario layout in the second stage is shown in Figure 1. The scenarios are generated the same way as in the first stage, but the number of scenarios are doubled and the realizations of the factor values might not be the same. In both treatments, workers will encounter a deterministic AI in 9 scenarios, and a random AI in the other 9 scenarios. These are shuffled so workers don't know whether recommendations are deterministic or random. Because the Random AI could still pick the patient according to its original value preference ordering by chance, the combined AI (Deterministic+Random) follows its stated value preference ordering stochastically, about 75% of the time.

Once the worker finishes the second stage of the experiment, they fill out an additional survey where we ask workers for a general demographic description, and two more questions about their experience — which dimension (Prior Donor, Wait Time, Disease Stage) most impacted their decision making without the AI, and how much did they think they relied on the AI when making decisions in the second stage.

4 RESULTS

We recruited a total of 303 workers, with 160 workers being assigned to the first treatment, and 143 workers being assigned to the second treatment. 67% of participants were male, and 33% were female. 86% of participants were white. 81% of participants had a bachelor's or higher. Median pay for workers was approximately $10 per hour. This study was approved by our institution's IRB.

4.1 Effect of Value Similarity on AI Reliance

We start by answering our first research question, which analyzes how value similarity affects reliance on AI recommendations. We measure reliance in two different ways. First, we express reliance as the overall change in alignment between the human and AI between the first and second stages. Then, we express reliance as the change in decision-making behavior, computed only on the subset of scenarios where the human and AI differ in the first stage. We present results for both of these metrics in Figure 3. We report the statistical significance values using a t-test and the effect sizes using Cohen's d. Error bars in plots represent standard errors.

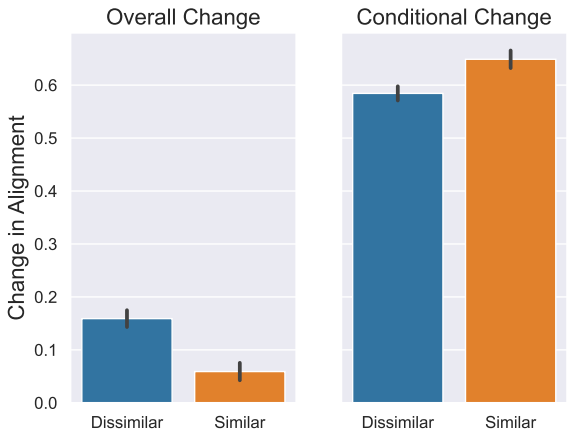

Figure 3: The effect of value similarity on alignment change between Stages 1 and 2. In the left figure, we find across all scenarios, the dissimilar AI has a significantly larger change in alignment ($p < .001$). In the right figure, we find that in scenarios where the human and AI disagree, the similar AI has a significantly larger change in alignment ($p = 0.003$).

4.1.1 Overall Change in Alignment. In order to measure the overall change in alignment, we compare the rate at which users match with the (unseen) AI in the first stage with the matching rate in the second stage. We find that adding a recommendation from a similar AI significantly increases alignment by 5.9% ($t(1286) = 3.58, p < .001, d = 0.10$), while adding a recommendation from a dissimilar AI significantly increases alignment by 15.9% ($t(1439) = 9.98, p < .001, d = 0.26$). The difference between the two increases is also significant with $t(2705) = 4.35, p < .001, d = 0.17$. **Overall, we find that dissimilar AIs have a bigger overall impact on overall alignment, confirming our first hypothesis.**

While this result may seem unintuitive, it can be explained by the fact that users tend to agree more with a similar AI than a dissimilar AI, so there is less room to increase agreement for a similar AI in the second stage.

4.1.2 Conditional Change in Alignment. As a perhaps more useful measure of reliance, we can choose to consider only scenarios where the AI gives recommendations which go against the decision that the user made in the first stage. This comparison is possible because our experiment design guarantees that each of the nine possible scenarios appear once in the first stage, and twice in the second stage.

We find that when the AI gives a recommendation which goes against the user's Stage 1 decision, alignment with a similar AI increases by 64.9%[1], while alignment with a dissimilar AI increases by 58.4%. This difference is significant with $t(1302) = -3.00, p = 0.003, d = 0.17$. **Overall, we find that similar AIs have a bigger impact on human alignment when the AI goes against human prior preferences, confirming our second hypothesis.**

4.2 Effect of Value Similarity *Claims* on Alignment Change

For our second research question, we try to understand why we see effects of value similarity on AI reliance. Specifically, we want to see if the increases in AI alignment caused by value similarity in Sections 4.1 can be explained by the workers' belief that the AI shares a similar set of values to the workers, or if the increase in AI alignment is due to the actual similarity in values exposed in AI recommendations reinforcing the workers' own preferences.

In our experiment design, half of the AI recommendations in the second stage are generated deterministically according to the claimed ethical preference, and half of the AI recommendations are generated randomly. When the AI is random, any alignment increase is only due to the perception of the AI having similar or dissimilar values. When the AI is deterministic, alignment increases are explained by both user perception of AI similarity and the effect of the AI actually acting according to its preferences. As a result, we can compare these two to find the isolated effect of AI claims.

We measure the effect of value similarity on conditional AI alignment (as in Section 4.1.2), and break this data down by AI Behavior — whether the AI is deterministic or random. These results are presented in Figure 4. In this experiment, we have two independent variables (deterministic vs random AI, and similar vs dissimilar AI). The dependent variable is the conditional alignment. To examine the significance of the results, we first conduct a two-way ANOVA test and find a significant interaction effect between the two independent variables ($F(1) = 6.86, p = 0.009$). We then conduct post-hoc Tukey's HSD tests. We find that when the AI is deterministic, there is a significant difference in the conditional AI alignment between similar and dissimilar AI ($p < 0.001$). However, when the AI is random, we see no significance in the conditional AI alignment between similar and dissimilar AI ($p = 0.58$). The results suggest that workers' reliance on AI is influenced by the realized AI recommendation instead of the value AI claims to exhibit. **With this result, we find no evidence to support our third hypothesis, as we see no effect from AI similarity claims alone on reliance.**

4.3 Exploratory Analysis

Now that we have answered our main research questions, we perform a few follow-up investigations of our data to shed further

[1]Because we are only examining scenarios where the human originally disagreed with the AI, these increases can be interpreted as total alignment in the second phase. E.g., in this subset of scenarios, workers choose to follow similar AI recommendations 0% of the time in the first stage, and 64.9% of the time in the second phase.

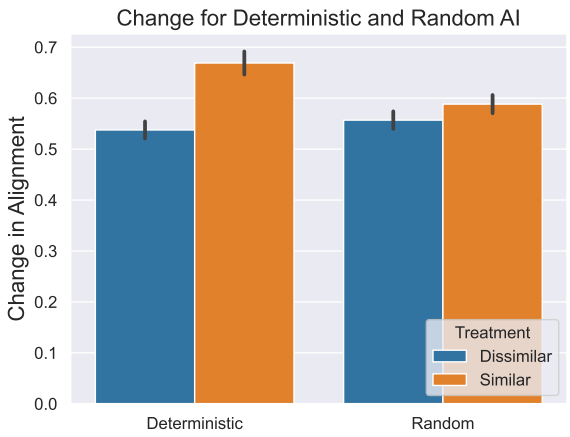

Figure 4: The effect of value similarity on alignment change between Stages 1 and 2, across combinations of Deterministic/Random and Similar/Dissimilar. When the AI is Deterministic, the Similar AI leads to a significantly larger change in conditional alignment ($p < .001$). However, when the AI is Random, there is no significant difference between Similar and Dissimilar AI ($p = .58$).

light on the effects and implications of using AI recommendations in problems of ethical decision making. We note these analysis is intended to be exploratory and hope that this additional analysis provides a starting point for future work to study these topics in more depth.

First, we look at the relationship between AI similarity and people's subjective beliefs of self-confidence, trust in AI, and perceived usage of AI with the type of AI they used (similar/dissimilar). This can be considered an extension of Mehrotra et al. [31], which investigated the relationship between people's subjective beliefs of AI trust and AI similarity. In addition, we broaden the scope of our results in Section 4.1 to understand not only the individual-level effects of AI assistance on reliance, but population-level shifts which personalized AI recommendations can create.

4.3.1 Subjective Perceptions. In our experiment, we asked users three subjective questions which relate to their perceptions of their own decisions or the AI's decisions: How confident were they in their own decisions made in the first stage (Self-Confidence), how much they trust AI to make decisions on its own (AI-Trust), and how much they believed they relied on the AI in the second stage (AI-Reliance). Each of these questions were asked on a 5-point Likert scale, where "1" represents strongly confident, strongly trust, and strongly reliant, respectively.

We compare the results of these questions across the two experiment treatments - whether they were presented with similar or dissimilar AI recommendations. It should be noted that the first two subjective questions, on Self-Confidence and AI-Trust, were asked directly after we presented them with a summary of their own values (calculated using their responses from the first stage) and the values of the AI assistant assigned to them. The third question, on AI-Reliance, was asked after the second stage.

We find that users shown a similar AI had a Self-Confidence score of 1.82, an AI-Reliance score of 2.01, and an AI-Trust score of 2.02. Users assigned to a dissimilar AI had a Self-Confidence score of 1.88, an AI-Reliance score of 2.14, and an AI-Trust score of 2.18. However, none of these differences across treatments are significant, with p-values of 0.41, 0.23, and 0.41, respectively.

We highlight this last result specifically, as it is similar to the analysis done by Mehrotra et al. [31]. However, they found a significant correlation between value similarity and trust in a smaller study (89 users), while we were not able to replicate this finding in a larger experiment (303 users). We speculate that this lack of replication is due to the choice of ethical values used for determining similarity. In Mehrotra et al. [31], they described their AI assistants to workers using a generic set of ethical values, only some of which were actually relevant to their ethical decision-making problem. This could have lead workers to have high trust in AI recommendations based on values relevant to the problem, and low trust in AI recommendations based on values irrelevant to the problem. In contrast, we exclusively present values which are relevant to our ethical problem; this may cause a smaller effect when comparing the values against each other.

4.3.2 Population-Level Shifts. In this section, we investigate potential population-level shifts in user behavior as a result of using personalized (similar or dissimilar) AI recommendations. Specifically, we aim to understand if populations become more divided in their ethical preference strengths, and potential implications on population polarization.

First, we discuss the metric ΔP, introduced by Awad et al. [2], which represents a worker's ethical preference in a single factor (e.g. Prior Donor Status). We can calculate ΔP on this factor by taking all decisions where the factor is unequal across candidates, and computing the difference in preferences across options. For example, if a worker views four scenarios where one candidate is a prior donor and the other candidate is not, and the worker selects the prior donor three times, their ΔP for the Prior Donor factor is $0.75 - 0.25 = 0.5$. For each worker, we generate a vector of ΔP values (or $\boldsymbol{\Delta P}$ for short) to represent the worker's ethical preferences across the three factors.

Using $\boldsymbol{\Delta P}$, we can then generate our population-level metric, the normalized stated preference (or stated preference). Recall that we asked workers to express the dimension they care about the most in the post-experiment survey (we call this dimension "preferred factor"). To generate the normalized stated preference, we normalize each worker's $\boldsymbol{\Delta P}$ to be length one and select the value in the dimension of the workers' preferred factor. For example, if a user's normalized $\boldsymbol{\Delta P} = (0.8, 0.6, 0)$, and the user's preferred factor is the Prior Donor factor (the first dimension of the vector), then their normalized stated preference value would be 0.8. The reason we normalize $\boldsymbol{\Delta P}$ before selecting this preferred factor is to better measure the relative preferences between a user's stated preference and the other two preferences, without giving extra weight to users with a higher overall ethical preference strength.

The intuition of using this normalized stated preference as a metric is to measure how divided a population is. For example, people

generally have varying priorities on what government should focus on (e.g. the economy, health care, climate change, security) [23]. If a population has a relatively low stated preference, then this can be interpreted as the population having relatively weak preferences towards their highest priority over the other policy options. If the population has a high stated preference, this means that people strongly believe in their top policy over the others.

We analyze the average stated preference of the two stages. We find an average stated preference of 0.151 in the first stage, and an average stated preference of 0.173 in the second stage. This increase is not significant ($t(895) = -0.52, p = 0.60, d = 0.04$). However, if we compare the increase in the average stated preference with similar AI and the increase with dissimilar AI, We see that a similar AI increases stated preference to 0.226, and a dissimilar AI decreases stated preference to 0.125. This difference is statistical significant, with $t(595) = -2.09, p = 0.037, d = 0.17$. Overall, the results suggest that the use of similar AI recommendations leads to higher stated preferences than using dissimilar recommendations.

5 DISCUSSION

In this section, we discuss the limitations, implications, and future work of our study.

Limitations and generalization. We discuss the limitations of this study. First, we have conducted our experiments using crowdsourcing with users recruited from Amazon Mechanical Turk. While crowdsourcing is getting increasing popularity in conducting user studies, the nature of distributed work of the platform raises questions about the engagement of workers and the quality of their responses. The common approaches to improve the quality of crowdsourced data collection include post-hoc aggregation [6, 19, 20, 42, 48], designing proper incentives [21, 22, 24, 30, 44], and improving the task design [1, 9–11, 14, 40]. However, the subjective nature of our task makes it challenging to ensure data quality as we cannot evaluate whether the workers are providing truthful answers. Moreover, the hypothetical nature of the presentation of the moral dilemma, although being a standard practice for academic studies [2, 16], may not reflect human ethical preferences in real-life scenarios. Additionally, the study surveyed ethical preferences from a general population of laypeople, who may interpret the moral dilemma differently from relevant domain stakeholders. Therefore, surveying preferences from stakeholders such as medical doctors or policymakers could provide valuable insights on how these results could inform real-world implementation of AI-assisted human decision making on kidney allocation.

Second, we have conducted a case study in the domain of kidney allocation to investigates the effects of value similarity to human reliance in the context of AI-assisted ethical decision making. Given the nature of case study, we cannot guarantee that the results and findings carry over to other domains. However, kidney allocation is an example of a general family of problem in scarce resource allocation. Therefore, we conjecture that our results could translate to other domains in this family of problems, such as vaccine distribution or homelessness resource allocation. However, it is important to carefully study applications in other domains before using these results to inform implementation in real-world systems.

Implications. In this work, we find that human reliance on AI is influenced by the value similarity between humans and AI. This result showcases the complexity of understanding the impacts of incorporating AI recommendations in ethical decision making, as the final decisions made by human-AI teams would depend on not only the ethical values exhibited by humans and AI algorithms but also the similarity between them. For example, if workers' ethical preferences are reinforced by AI with similar ethical preference, in the sense that they put more focus on the top factor in making ethical decisions, when we provide personalized assistive AI (with similar values to decision makers) in AI-assisted ethical decision making, it could create an effect similar to the *echo chamber* effect [5] that make the ethical decisions made by AI-assisted decision making more polarized, focusing on more extreme factors.

Moreover, the fact that human decisions are influenced by AI assistance also creates potential concerns of manipulation. For example, through leveraging the techniques from the literature on information design [26, 39], the advantageous party (e.g., the party that provides the AI assistance, usually the party with more power and information advantage) might strategically choose the assistance to lead human decision makers to take certain decisions. Therefore, as the growing prevalence of AI involvements in decision making in high-stake domains, having more studies on how humans reliance on AI evolves and whether it is possible to be manipulated are important to ensure the introduction of AI in decision making creates positive impacts to the society.

Future work. Our work has presented interesting findings on the effect of value similarity to human reliance in AI-assisted ethical decision making. There are still a lot of open questions that deserve future study. First, it is worth exploring the other factors that might impact human reliance on AI in the domain of ethical decision making. For example, if we provide explanations on why the AI recommendations exhibit certain ethical values, are human decision makers more likely to follow the recommendations? Moreover, as brought up by the above discussion on the limitations and implications, investigating the impact of AI assistance in different problem domains and with different stakeholder populations would help us understand the generalizability of the results. It is also important to study how the overall ethical preferences evolve when introducing AI to help humans make decisions in ethically-sensitive domains.

6 CONCLUSION

We investigate the impact of value similarity to human reliance in AI-assisted ethical decision making. We find that recommendations provided by a dissimilar AI have a higher impact on human decision-making than those given by a similar AI. However, this result is primarily due to the fact that a similar AI typically has a higher level of agreement with the human decision maker, leaving fewer opportunities for persuasion. When we focus on scenarios where humans and AI disagree, we have observed that humans are more likely to change their decision when given recommendations from a similar AI rather than a dissimilar one. We have found no evidence to suggest that this effect is a result of humans perceiving the AI as being similar. Instead, our findings indicate that this effect is mainly due to the AI's ability to display similar ethical values through its recommendations.

ACKNOWLEDGMENTS

This work is supported in part by the NSF under grant IIS-1850335, the NSF FAI program in collaboration with Amazon under grant IIS-1939677 and IIS-2040800, and the Office of Naval Research grant N00014-20-1-2240.

REFERENCES

[1] Harini Alagarai Sampath, Rajeev Rajeshuni, and Bipin Indurkhya. 2014. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3665–3674.

[2] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[4] Yochanan E Bigman, Desman Wilson, Mads N Arnestad, Adam Waytz, and Kurt Gray. 2022. Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General* (2022).

[5] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118.

[6] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* 39 (1977), 1–38.

[7] 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.

[8] Jaap J Dijkstra, Wim BG Liebrand, and Ellen Timminga. 1998. Persuasiveness of expert systems. *Behaviour & Information Technology* 17, 3 (1998), 155–163.

[9] Ryan Drapeau, Lydia Chilton, Jonathan Bragg, and Daniel Weld. 2016. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4.

[10] Xiaoni Duan, Chien-Ju Ho, and Ming Yin. 2020. Does Exposure to Diverse Perspectives Mitigate Biases in Crowdwork? An Explorative Study. In *Proceedings of the 8th AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 155–158.

[11] Xiaoni Duan, Chien-Ju Ho, and Ming Yin. 2022. The influences of task design on crowdsourced judgement: A case study of recidivism risk evaluation. In *Proceedings of the ACM Web Conference 2022*. 1685–1696.

[12] Ezekiel J Emanuel, Govind Persad, Ross Upshur, Beatriz Thome, Michael Parker, Aaron Glickman, Cathy Zhang, Connor Boyle, Maxwell Smith, and James P Phillips. 2020. Fair allocation of scarce medical resources in the time of Covid-19. , 2049–2055 pages.

[13] Ezekiel J Emanuel and Alan Wertheimer. 2006. Who should get influenza vaccine when not all can? *Science* 312, 5775 (2006), 854–855.

[14] Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. 2013. Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*. 1–4.

[15] United Network for Organ Sharing. 2022. *One-year monitoring report shows increases in kidney transplants for Black, Hispanic, Asian and pediatric patients following policy changes.* https://unos.org/news/1-yr-kidney-data-report-transplant-increases/

[16] Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P Dickerson, and Vincent Conitzer. 2020. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence* 283 (2020), 103261.

[17] Adrian Furnham. 1996. Factors relating to the allocation of medical resources. *Journal of Social Behavior and Personality* 11, 3 (1996), 615–624.

[18] Nina Grgić-Hlača, Claude Castelluccia, and Krishna P Gummadi. 2022. Taking Advice from (Dis) Similar Machines: The Impact of Human-Machine Similarity on Machine-Assisted Decision-Making. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 74–88.

[19] Chien-Ju Ho, Rafael Frongillo, and Yiling Chen. 2016. Eliciting categorical data for optimal aggregation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2450–2458.

[20] Chien-Ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. 2013. Adaptive task assignment for crowdsourced classification. In *Proceedings of the 30th International Conference on Machine Learning*. 534–542.

[21] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing High Quality Crowdwork. In *Proceedings of the 24th International Conference on World Wide Web (WWW)*.

[22] Chien-Ju Ho, Yu Zhang, Jennifer Wortman Vaughan, and Mihaela Van Der Schaar. 2012. Towards social norm design for crowdsourcing markets. In *Proceedings of the 4th Human Computation Workshop*.

[23] Juliana M Horowitz, Ruth Igielnik, and Rakesh Kochhar. 2020. Most Americans say there is too much economic inequality in the US, but fewer than half call it a top priority. *Pew Research Center* 9 (2020).

[24] John Joseph Horton and Lydia B. Chilton. 2010. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce (EC)*.

[25] Antoine Hudon, Théophile Demazure, Alexander Karran, Pierre-Majorique Léger, and Sylvain Sénécal. 2021. Explainable artificial intelligence (XAI): how the visualization of AI predictions affects user cognitive load and confidence. In *Information Systems and Neuroscience: NeuroIS Retreat 2021*. Springer, 237–246.

[26] Emir Kamenica and Matthew Gentzkow. 2011. Bayesian persuasion. *American Economic Review* 101, 6 (2011), 2590–2615.

[27] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[28] 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.

[29] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. 2019. Resistance to medical artificial intelligence. *Journal of Consumer Research* 46, 4 (2019), 629–650.

[30] Winter Mason and Duncane Watts. 2009. Financial Incentives and the "Performance of Crowds". In *Proceedings of the 1st Human Computation Workshop (HCOMP)*.

[31] Siddharth Mehrotra, Catholijn M Jonker, and Myrthe L Tielman. 2021. More similar values, more trust?-the effect of value similarity on trust in human-agent interaction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 777–783.

[32] Saumik Narayanan, Guanghui Yu, Wei Tang, Chien-Ju Ho, and Ming Yin. 2022. How Does Predictive Information Affect Human Ethical Preferences?. In *ACM Conference on AI, Ethics, and Society*.

[33] Govind Persad, Alan Wertheimer, and Ezekiel J Emanuel. 2009. Principles for allocation of scarce medical interventions. *The lancet* 373, 9661 (2009), 423–431.

[34] Marianne Promberger and Jonathan Baron. 2006. Do patients trust computers? *Journal of Behavioral Decision Making* 19, 5 (2006), 455–468.

[35] Anuschka Schmitt, Thiemo Wambsganss, Matthias Söllner, and Andreas Janson. 2021. Towards a Trust Reliance Paradox? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice. In *International Conference on Information Systems (ICIS)*.

[36] Victoria A Shaffer, C Adam Probst, Edgar C Merkle, Hal R Arkes, and Mitchell A Medow. 2013. Why do patients derogate physicians who use a computer-based diagnostic support system? *Medical Decision Making* 33, 1 (2013), 108–118.

[37] Michael Siegrist, George Cvetkovich, and Claudia Roth. 2000. Salient value similarity, social trust, and risk/benefit perception. *Risk analysis* 20, 3 (2000), 353–362.

[38] Sim B Sitkin and Nancy L Roth. 1993. Explaining the limited effectiveness of legalistic "remedies" for trust/distrust. *Organization science* 4, 3 (1993), 367–392.

[39] Wei Tang and Chien-Ju Ho. 2021. On the Bayesian Rational Assumption in Information Design. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. 120–130.

[40] Wei Tang, Ming Yin, and Chien-Ju Ho. 2019. Leveraging peer communication to enhance crowdsourcing. In *The World Wide Web Conference*. 1794–1805.

[41] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[42] Jacob Whitehill, Ting fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems (NIPS)*.

[43] James R Wolf. 2014. Do IT students prefer doctors who use IT? *Computers in Human Behavior* 35 (2014), 287–294.

[44] Ming Yin and Yiling Chen. 2016. Predicting crowd work quality under monetary interventions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4. 259–268.

[45] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[46] Ryosuke Yokoi and Kazuya Nakayachi. 2021. The effect of value similarity on trust in the automation systems: A case of transportation and medical care. *International Journal of Human–Computer Interaction* 37, 13 (2021), 1269–1282.

[47] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

[48] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment* 10, 5 (2017), 541–552.