How Does Predictive Information Affect Human Ethical Preferences?

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

Wei Tang
Washington University in St. Louis
St. Louis, Missouri, USA
w.tang@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

Artificial intelligence (AI) has been increasingly involved in decision making in high-stakes domains. Meanwhile, involving AI in these high-stake decisions has created ethical concerns on how to balance different trade-offs to respect human values. One approach for aligning AIs with human values is to elicit human ethical preferences and incorporate this information in the design of computer systems. In this work, we explore how human ethical preferences are impacted by the information shown to humans during elicitation. In particular, we aim to provide a contrast between verifiable information (e.g., patient demographics or blood test results) and predictive information (e.g., the probability of organ transplant success). Using kidney transplant allocation as a case study, we conduct a randomized experiment to elicit human ethical preferences on scarce resource allocation to understand how human ethical preferences are impacted by the verifiable and predictive information. We find that the presence of predictive information significantly changes how humans take into account other verifiable information in their ethical preferences. We also find that the source of the predictive information (e.g., whether the predictions are made by AI or human doctors) plays a key role in how humans incorporate the predictive information into their own ethical judgements.

CCS CONCEPTS

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

KEYWORDS

ethical preference, AI ethics; preference elicitation

ACM Reference Format:
Saumik Narayanan, Guanghui Yu, Wei Tang, Chien-Ju Ho, and Ming Yin. 2022. How Does Predictive Information Affect Human Ethical Preferences?. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3514094.3534165

AIES'22, August 1–3, 2022, Oxford, United Kingdom
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9247-1/22/08.
https://doi.org/10.1145/3514094.3534165

1 INTRODUCTION

As the capability of artificial intelligence increases, AI systems are increasingly involved in decision making in high stakes domains, such as medical decision making [10, 29, 30, 41, 43], loan applications [8, 18, 22], or legal systems [1, 6]. Meanwhile, the growing prevalence of AI in decision making has raised ethical concerns, as the decisions made by these systems might be biased or might not align with human values [1, 4, 21, 32]. To address these concerns, we would ideally want to have a set of rules specifying what it means for a decision to be *ethical* such that AI researchers and practitioners can incorporate these rules when designing and deploying AI in practice. However, in ethically-sensitive domains, there are often no clear-cut right and wrong decisions. Instead, we are often forced to choose the "lesser of two evils", prioritizing and trading off different ethical values and principles. Moreover, different stakeholders may have different preferences on the priority of ethical principles. Finding a trade-off between ethical principles that everyone agrees on for a given task may be challenging or even impossible.

To explore the above challenges and align the design of AI with human values, one natural approach is to elicit human preferences on ethical principles from relevant populations and incorporate the elicited information in the design of AI systems [2, 16, 25, 31]. In this line of work, during preference elicitation, human participants are presented information on hypothetical scenarios involving moral dilemmas and asked to express their preferences in the scenario. For example, Awad et al. [2] considers the moral dilemmas faced by autonomous vehicles; participants were given hypothetical scenarios in which a vehicle is bound to crash, and were asked to express their preference on sparing the lives of one group of people over another. By varying the demographics and attributes of the two groups, researchers can infer which ethical values (e.g., sparing lives, sparing youth, etc) the population prioritizes. To focus on the trade-offs in the moral dilemmas, the information presented to participants in most prior work has been *verifiable*, meaning that the information only describes the past or present, and there is no uncertainty associated with the presented information. In the meantime, as *predictive* information, which concerns predictions made about the future, is increasingly integrated in ethical decision making (e.g., judges might utilize predictive risk scores in making bail decisions), it is important to understand the influence predictive information has on human ethical preferences.

In this work, we aim to understand how the elicitation of human ethical preferences are impacted by the information shown to humans. We provide a contrast between verifiable information (e.g., patient demographics or blood test results) and predictive information (e.g., the probability of organ transplant success). As predictive information, from either AI or human experts, is increasingly integrated in ethical decision making, we investigate how the *presence* and the *source* of the predictive information affect human ethical preferences. Specifically, we ask the following two research questions:

- **RQ1**: How does the presence of predictive information affect human ethical preferences?
- **RQ2**: How does the source of the predictive information (e.g., predictions by human experts or predictions by AI) affect human ethical preferences?

To answer the above research questions, we conducted randomized online experiments on Amazon Mechanical Turk (MTurk). Using the domain of kidney transplants as a case study, we presented scenarios where two candidates needed a kidney transplant but only one was available, and asked MTurk workers to express their preference on which candidate should receive the kidney first. We designed two sets of experiments, one to answer each research questions. In the first experiment, we investigated how ethical preferences varied between workers who saw only verifiable information and workers who saw both verifiable and predictive information. We find that even when predictions are equal between candidates, the presence of predictions change human ethical preferences. We also find that both the direction and magnitude of differences in predictive information is relevant and important for understanding how human ethical preferences change. In the second experiment, we analyzed how human ethical preferences change based on the source of the predictive information. We find that humans rely more on predictions from AI than predictions from a human doctor, possibly indicating that humans trust AI predictions more than human predictions. Moreover, humans seem to discount the importance of other verifiable information more when an AI prediction is presented, implying that humans are more likely to treat AI predictions as a summary of other verifiable information.

Our findings show that the elicitation of human ethical preferences are impacted by both the presence and source of the predictive information. As predictive information is increasingly integrated in ethical decision making, it is important to conduct more studies to understand how humans take predictive information into account when forming ethical preferences. Moreover, our results suggest that elicited human ethical preferences might not be robust or consistent, as the elicited preferences vary with different elicitation methods. Therefore, it is important to conduct more studies in understanding to what extent the elicited ethical preference is robust to manipulation. We should take this into account when utilizing the elicited information to inform the design of AI systems.

2 RELATED WORK

Our work joins the flourishing line of recent research in aligning the design of AI systems with human values. One natural way to approach this alignment is to elicit real human ethical preferences in scenarios where multiple ethical principles conflict, to determine the relative weights of the principles and to understand in which scenarios, one principle might be favored over another. Correspondingly, there has been a line of work researching the elicitation of human ethical preferences [2, 9, 16, 37]. Among these works, Awad et al. [2] studied human preferences on autonomous driving when faced with an adaptation of the trolley problem, and learned how these ethical preferences vary across worldwide cultures. Smith et al. [38] studied human preferences in moderation of Wikipedia quality prediction. Freedman et al. [16] studied human preferences in the allocation of kidneys for transplants. Our work differs from this line of work in that we focus on discussing the impact of predictive information to human ethical preferences while existing work have mostly utilized verifiable information only. Another related work by Chan et al. [9] also analyzed the elicitation of ethical preferences in the kidney domain. However, they analyzed how assessments of human ethical preferences impacted their ethical decision making, and did not focus on the impact of predictive information to human ethical preferences. As a closely related line of research, if we consider different fairness measures as different ethical principles, our work is also related to the research in understanding human perceptions of different fairness measures [19, 39, 42, 44], especially because it's usually impossible to satisfy all fairness measures simultaneously [7, 11, 12, 23].

Another related line of research is on utilizing participatory design to govern the design and implementation of AI systems [25, 31, 38, 46]. These works looked at the next steps after we have elicited these ethical preferences, namely how to integrate these preferences into the deployment of the AI systems. For example, Yu et al. [46] looked at methods of presenting these preferences to stakeholders, so that they better understand the trade offs that they must make. Noothigattu et al. [31] worked to construct a system where multiple models of ethical preferences vote on which principles should be used for a given scenario, based on pre-elicited human preferences, and Lee et al. [25] explored how such a participatory framework could leverage multiple stakeholders during the decision-making process.

2.1 Background: Ethical Principles for Allocation of Scarce Medical Interventions

In this work, we use the domain of kidney transplants as a case study. There has been extensive literature on the ethical principles in allocating scarce medical interventions [14, 15, 17, 33]. In particular, our task design is based on the work by Persad et al. [33], who list the following four categories of ethical principles for allocating scarce medical resources.

- Promoting and rewarding social usefulness: This principle could be implemented through prioritizing *instrumental value*, e.g., giving medical workers higher priority in receiving vaccines during a pandemic, or *reciprocity*, e.g., giving prior organ donors higher priority to receive a transplant of their own.
- Treating people equally: In this principle, everyone should have equal chance of receiving medical interventions. It can often be implemented using *lottery* or *first-come-first-serve* approaches.

- Favoring the worst-off: This principle could be implemented through deploying the strategy of *sickest first*, prioritizing those who have a more severe disease condition or *youngest first*, prioritizing those who have not lived as many years yet.
- Maximizing total benefits: This principle aims to maximize some definition of utility, e.g., maximizing the number of saved lives or maximizing the increase life-years after intervention.

These categories of ethical principles are widely used, both in academic contexts [14, 24, 33, 45], and in action for real-world medical organizations [34, 36].

3 EXPERIMENT 1 - PREDICTION PRESENCE

In this experiment[1], we investigate our first research question: How does the presence of predictive information affect human ethical preferences? To answer this question, we present recruited workers with scenarios involving ethical dilemmas. We then observe their expressed ethical preferences among candidate choices both when predictive information is presented and when predictive information is not presented. In particular, we have the following two hypotheses:

- **H1**: We hypothesize that human ethical preferences stay the same when predictive information is equal across candidates, compared to when no predictive information is presented.
- **H2**: We hypothesize that human ethical preferences are strengthened when the prediction is aligned with human preferences, compared to when predictions are equal. Correspondingly, we hypothesize that ethical preferences are weakened when the prediction is aligned against human preferences.

To examine the above hypotheses, we conducted a case study on the domain of kidney transplants and designed a randomized experiment. We chose the domain of kidney transplants for our study for two reasons. First, there has been extensive literature on the ethical principles in allocating scarce medical interventions [14, 15, 17, 33]. This allows us to tailor our task design to align with well-established ethical preference frameworks. Second, incorporating machine learning predictions in medical decision making is attracting a great amount of research effort and has significant potential in improving medical outcomes [3, 5, 35]. Understanding the effect of predictive information on human ethical preferences could help us better align the use of predictions with human values.

3.1 Experiment Task

In our experiments, workers were recruited to judge a set of kidney transplant scenarios. In each scenario, workers were presented two patient candidates who both need a kidney transplant, but only one kidney is available. Given information about each of these candidates, workers were asked to express their preference on which candidate should receive the kidney first.

Based on the ethical principles which govern the allocation of scarce medical resources [33], as discussed in Section 2.1, we chose four factors to display to workers. The first three factors concern the present condition and attributes of the candidates, which we denote as *verifiable information*. The fourth factor concerns a future prediction made about the candidates, which we denote as the *predictive information*. Specifically, these factors (along with the corresponding ethical principle) are:

- **Kidney Donor Status** (Promoting social usefulness):
  Whether 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 transplant. 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)}.
- **Post-Transplant Survival Chance** (Maximizing total benefits):
  The predictive probability that the candidate will remain alive after 5 years post-transplant. This feature has possible values between 72% and 98%.

Based on the established ethical principle framework [33], there is a preference ordering on each factor when all other factors are equal. For example, if two candidates share the same values for kidney donor status, kidney disease stage, and post-transplant survival chance, the patient with longer wait time is preferred according to the ethical principle. In our experiments, we presented different scenarios to online workers to understand how humans make tradeoffs on these four factors, mapping to the four corresponding ethical principles.

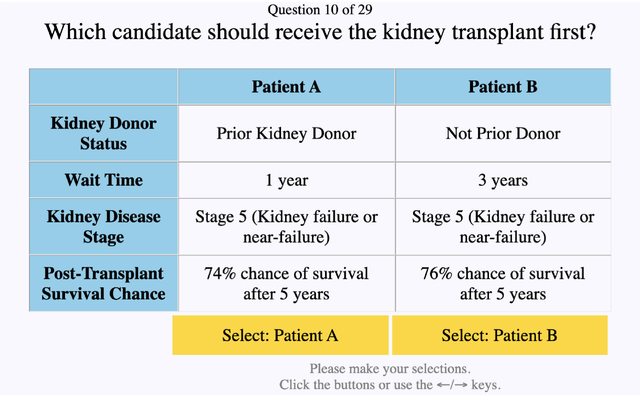
3.2 Experiment Design

To understand the effect of predictive information on human ethical preferences, we conducted a randomized behavioral experiment with two treatments.

- **Treatment 1 (Verifiable Only)**: This treatment group was shown the three factors of verifiable information. This represents the human priors on human ethical preferences, and gives us a baseline to measure the effects of the predictive factors against.
- **Treatment 2 (Verifiable and Predictive)**: The treatment group was shown both the three verifiable factors, and one factor based on predictive information. We did not present the source, explanation, or any other information about this predictive factor.

Each recruited worker were asked to express their ethical preference in 29 scenarios (the choice of the scenarios is described later). In each scenario, workers were presented with two candidate profiles and were asked to provide their preference on which candidate should receive the kidney transplant first. We show an example of what a worker in the second treatment (verifiable and predictive) saw in Figure 1. Workers in the first treatment (verifiable only) saw the same design, except they were not shown the predictive information of post-transplant survival chance in the last row.

[1] All experiments in this study are approved by our institution's IRB.



Question 10 of 29

Which candidate should receive the kidney transplant first?

	Patient A	Patient B
Kidney Donor Status	Prior Kidney Donor	Not Prior Donor
Wait Time	1 year	3 years
Kidney Disease Stage	Stage 5 (Kidney failure or near-failure)	Stage 5 (Kidney failure or near-failure)
Post-Transplant Survival Chance	74% chance of survival after 5 years	76% chance of survival after 5 years

Please make your selections.
Click the buttons or use the ←/→ keys.

Figure 1: The task interface for treatment 2 (verifiable and predictive). The interface for treatment 1 (verifiable only) is similar but does not contain the information of post-transplant survival chance.

3.2.1 Scenario Selection. In the first treatment (verifiable only), workers were only presented verifiable information about the candidates. Each of the three verifiable factors are ordinal, and we have two candidates presented in each scenario, which we label as A and B. This gives us three possible orderings for each factor: candidate A is preferred over candidate B, candidate B is preferred over candidate A, and both candidates are equally preferred. Because we have three factors and three orderings, we get 27 total scenarios of factor orderings to assign. We discard the one scenario where both candidates share the same values for all factors and are left with 26 scenarios. Each worker in the first treatment group will view each of these 26 combinations once. Each combination is realized with randomly generated values. If we want donor status to be equal, we may display both patients as "Prior Kidney Donor", or both "Not Prior Donor". If we want the wait time of A to be higher than B, we may show 2 years and 1 year, or 5 years and 3 years, or any other pair of values as long as the difference is no more than two years. After the worker views the first 26 scenarios, we randomly choose three of the scenarios shown and show these scenarios to the worker again, with the exact same realization of the factor values. We do this as a consistency check, so we can determine the quality of a particular worker's data by how consistent their preferences are over these three repetitions of scenarios. To minimize the potential presentation bias caused by the ordering of the scenarios, we randomize the first 26 scenarios. To minimize the potential bias caused by the ordering of the candidates, we randomize the order of the candidates independently for each scenario.

In the second treatment (verifiable and predictive), workers were presented both verifiable and predictive information about the candidates. Note that the additional predictive factor is also ordinal, with three directions. In the second treatment, each worker was also presented 29 scenarios. To generate the combinations for the second treatment group, we take the same 26 combinations as in the first treatment, but when we present this to workers, we randomly select a direction for the predictive information (whether the predicted survival chance of one candidate is larger than, equal to, or smaller than the other), and show this to workers. As with the wait time feature, we randomly select a pair of values for each scenario, where values can be between 72% and 98%, and constrain the difference to be no more than 6%. We then again add three repeated scenarios randomly drawn from the first 26 scenarios for consistency check. We then apply the randomization procedure for the ordering of the first 26 scenarios and the presentation order of the two candidates in each scenario.

To examine our first hypothesis, we compare workers' preferences in the first treatment with workers' preferences in the second treatment on the scenarios where the two candidates have the same predicted survival chance. Given the number of scenarios in the second treatment for the above comparison is only one-third of the number of scenarios in the first treatment (as we randomly draw the ordering of predictive information from the three possible orderings), during random treatment assignment, we assigned three times more workers in the second treatment compared with the number of workers assigned to the first treatment. To examine our second hypothesis, we split the workers' preference data collected from the second treatment into three groups, based on the direction of predictive information, and analyze how this direction affects their ethical preferences.

3.2.2 Experiment Procedure. For this experiment, we recruited participants by posting a HIT on Amazon Mechanical Turk (MTurk). The HIT was only open to U.S. workers, and workers were paid $0.80 after completing the job. The median hourly pay was $10.19. In the preview page of the HIT, workers first viewed an instruction page, a sample scenario, and the consent form. Workers need to agree to the consent form to accept the HIT and participate in the experiments. After accepting the HIT, workers were randomly assigned to one of the treatments, with 25% chance of being assigned to the first treatment and 75% chance of being assigned to the second treatment. Workers were then shown a background page explaining the factors used for determining which candidate would receive a kidney. Workers were only presented the explanations on the factors used in their corresponding treatments. Afterwards, the workers began to evaluate kidney transplant scenarios. While evaluating scenarios, workers were still able to reference the background information on transplants. Finally, workers were asked to complete a short demographic survey.

3.2.3 Performance Measure. To measure workers' ethical preferences from collected data, we use conjoint analysis to compute the average marginal component effect (AMCE) of each factor (kidney donor status, wait time, kidney disease stage, and post-transplant survival chance). More concretely, for each factor, we select all scenarios where the factor value is unequal, and aggregate the average number of times that workers select the higher value over the lower value (recall that for each factor, there is an ethically preferred direction). We calculate the percentage of workers who select the higher value and the percentage of workers who select the lower value, and denote the difference between these values as ΔP. For example, to calculate ethical preferences for the kidney donor status, we select all scenarios where one patient is a prior kidney donor and the other patient is not, and measure the difference between the preference of the former and the preference of the latter. This difference is the reported ΔP.

3.3 Experiment Results

We recruited a total of 600 workers, with 184 workers being assigned to the first treatment, and 416 workers being assigned to the second treatment. We discarded workers who were not completely consistent on the three consistency check questions and report the results for the 202 workers who were fully consistent. We have also performed the same analysis on the entire worker pool, and the results are qualitatively the same.

H1: Effect of equal prediction on human ethical preferences. We first examine our Hypothesis 1, which claims that the addition of equal predictions between candidates have no effect on human ethical preference compared with no predictive information. To evaluate this hypothesis, we compare the ethical preference from the first treatment (verifiable only) and the ethical preferences from the subset of samples with equal values in the predictive factor in the second treatment (verifiable and predictive).

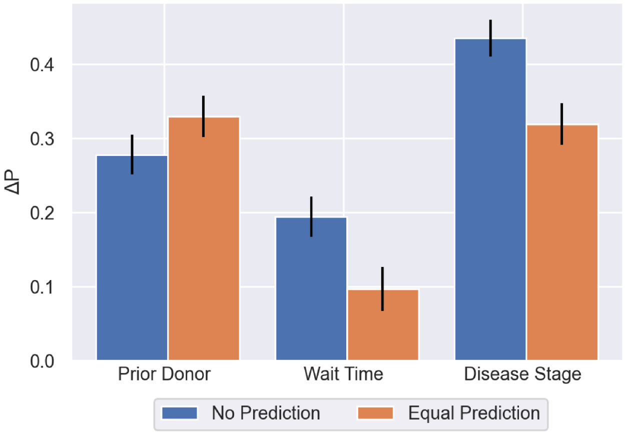


Figure 2: Effect of Equal Prediction. We present ΔP for each verifiable factor and treatment. We also present p-values with Bonferroni correction for multiple comparisons. There is no significant difference between treatments in the Prior Donor factor ($p = .54$). There is a significant difference between treatments in the Wait Time factor ($p = .045$). There is a significant difference between treatments in the Disease Stage factor ($p = .0057$).

The results are shown in Figure 2. We compare ΔP (the difference between preferring the higher value in a factor and preferring the lower value in a factor) for the three factors in verifiable information between the first treatment and the second treatment where the predictive factor was equal between candidates. We also apply Bonferroni correction to our significance tests to account for multiple comparisons. The first treatment represents the baseline of human ethical preferences when no predictive information is available, and the second treatment represents situations where predictive information is shown to humans, but does not favor either candidate. We find that the presence of equal predictive information significantly decreases the ethical preference of Wait Time from 0.194 to 0.097 ($p = .045$), significantly decreases the ethical preference of Disease Stage from 0.435 to 0.319 ($p = .0057$), and increases the ethical preference of Prior Donor from 0.278 to 0.330, though this increase is not significant ($p = .54$). These results reject our first hypothesis, as we have shown that human ethical preferences do change when predictive information is presented and is equal across candidates.

Interestingly, these changes are not consistent for all ethical preferences. We speculate that the reason for this is because humans may create their own predictions about the scenario based on the verifiable information we present, but when we present an externally sourced prediction about the scenario, they no longer create their own predictions and instead use the prediction provided. For example, one possible conjecture for the explanation of the result is that workers might think wait time and disease stage is more predictive of survival outcomes than prior donor status. Therefore, workers in the first treatment without predictive information may have used these in forming their own predictions which influence their ethical preferences. But when we present the prediction, this supersedes their own prediction, and their final preference is weighted less heavily towards wait time and disease stage when predictive information is available.

H2: Effect of aligned prediction on human ethical preferences. We next examine our Hypothesis 2, which claims that the addition of predictions that differ between candidates strengthens human ethical preferences if the predictions are aligned with the preferences. We also test the opposite side of this hypothesis, which claims that the addition of predictions between candidates weakens or even reverses human ethical preferences if the predictions are aligned against with the preferences.

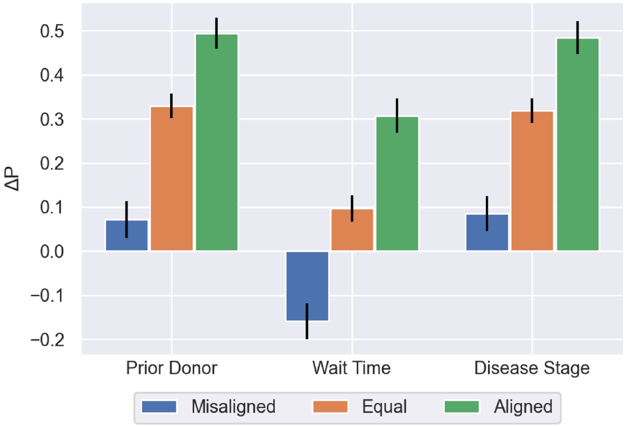


Figure 3: Effect of Prediction Alignment. We present ΔP for each verifiable factor and treatment. We also present p-values after scaling with Bonferroni correction. There is a significant difference between a misaligned prediction and equal prediction for all factors ($p < .001$). There is a significant difference between a equal prediction and aligned prediction for all factors ($p < .003$).

The results are shown in Figure 3, in which we compare the difference in ΔP in each factor based on the three possible directions of prediction alignment from the samples in the second treatment. We also apply Bonferroni correction to our significance tests to account for multiple comparisons. For each factor, we first select all scenarios where the factor value is unequal in the second treatment. We then split the samples into three groups (Aligned, Equal, or Misaligned), depending on how the preference of the prediction aligns with the preference of the verifiable factor. We then calculate the values of ΔP, the difference between the ratio of workers choosing the higher value and the ratio of workers choosing the lower value, for each factor and each group. We find that for all factors, there is a significant ($p < .001$) difference between misaligned prediction and equal prediction, and that there is a significant ($p < .003$) difference between equal prediction and aligned prediction.

These results support our second hypothesis that human ethical preferences are strengthened when predictions are aligned with the human preferences, and weakened when predictions are oppositely aligned with the preferences.

Exploratory analysis. We performed additional exploratory analysis on the collected data to gain more insights on how human ethical preferences are affected by predictive information. In particular, we expand our analysis for Hypothesis 2 and look at the impact of not just the direction of the preference in predictions, but the magnitude of prediction differences. Moreover, instead of looking at individual factors, we look at how the predictive information impacts human preferences as a whole. More concretely, using the data collected in the first treatment (verifiable only), we can determine the *prior preferred* candidate, the candidate who is more preferred for each scenario (i.e., a pair of candidates with different combination of factor differences) on the population-level in the first treatment. We then split the scenarios in the second treatment (verifiable and predictive) into 7 groups, where the difference between the survival chance of the prior preferred candidate and the unpreferred candidate is $\{-6, -4, -2, 0, 2, 4, 6\}$. We then measure ΔP of the overall candidate preference (as opposed splitting up by dimension) for each group to understand the impact of the prediction magnitude on the prior preference.

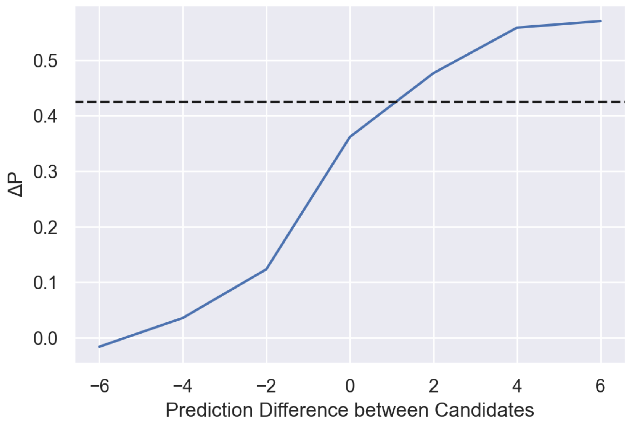


Figure 4: Effect of Prediction Magnitude. We present ΔP for each magnitude of prediction difference, in blue. We also present ΔP for the verifiable only treatment group, in black.

From the results in Figure 4, we can see how ΔP changes for various magnitudes of prediction value difference. This trend is monotonic, which makes sense intuitively, as we would expect that a larger difference in prediction values has a bigger effect on ethical preferences than a smaller difference in prediction values. However, when the predictions are equal between candidates, workers' ethical preferences decreases compared with the verifiable only group. This result again supports our first two hypotheses that adding predictive information could impact human ethical preferences, even when the predictive information does not seem to provide differentiating information between candidates.

4 EXPERIMENT 2 - PREDICTION SOURCE

In this experiment, we investigate our second research question: How the effect of predictive factors on human ethical preferences changes based on the source of the prediction. Specifically, we aim to find if there are differences if we tell workers that the prediction was generated by a human doctor or an AI system. In particular, we have the following hypothesis:

- **H3**: We hypothesize that the predictive factor has a stronger effect on ethical preferences when the prediction is made by a human doctor instead of an AI system.

This hypothesis aligns with our intuition, as well as prior research which suggests that workers trust AI more than humans for objective decisions, but trust humans more than AI for subjective decisions [27]. Though we aren't directly measuring trust of the prediction, one potential interpretation of a stronger reaction to predictions made by one source over another is that workers trust that source more, and we would expect workers to judge ethical kidney allocation as a subjective decision.

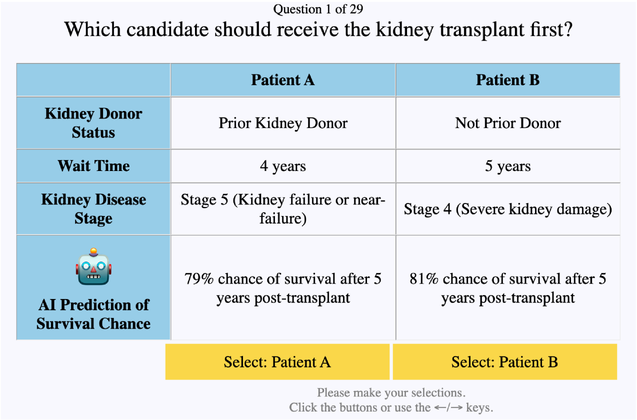
4.1 Experiment Design

The overall experiment procedure is similar to Experiment 1. We again run our experiment by posting a HIT for U.S. workers on Amazon Mechanical Turk. Workers were paid $0.80 after completing the task, and the median hourly pay was $9.12.

In order to examine whether there are differences if we tell the user that the prediction was generated by a human doctor or AI, we created two treatment groups with varying prediction sources.

- **Treatment 1 (Doctor)**: The first treatment group was shown the three demographic factors, the predictive factor, and an explanation saying that the prediction was generated by a human doctor.
- **Treatment 2 (AI)**: The second treatment group was shown the three demographic factors, the predictive factor, and an explanation saying that the prediction was generated by an AI system.

Each recruited worker was asked to express their ethical preference in 29 scenarios. The choice of the 29 scenarios is the same as the second treatment in Experiment 1, with the addition of the prediction source, which is given along with the predictive value. The first 26 scenarios reflect all combinations of factors in verifiable information and a random draw of predictive information. The last three scenarios are randomly drawn from the first 26 for checking



Question 1 of 29

Which candidate should receive the kidney transplant first?

	Patient A	Patient B
Kidney Donor Status	Prior Kidney Donor	Not Prior Donor
Wait Time	4 years	5 years
Kidney Disease Stage	Stage 5 (Kidney failure or near-failure)	Stage 4 (Severe kidney damage)
🤖 AI Prediction of Survival Chance	79% chance of survival after 5 years post-transplant	81% chance of survival after 5 years post-transplant
	Select: Patient A	Select: Patient B

Please make your selections.
Click the buttons or use the ←/→ keys.

Figure 5: The task interface for treatment 2 (AI). The interface for treatment 1 (Doctor) is similar except that "AI Prediction" in the final row is replaced with "Doctor Prediction".

worker consistency. We show an example of what a worker in the second treatment (AI) saw in Figure 5. Workers in the first treatment (doctor) saw the same design, except they were told that the prediction is made by a doctor, and were presented an image of a doctor instead of a robot.

4.2 Experiment Results

We recruited a total of 300 workers, with 156 workers being assigned to the first treatment, and 144 workers being assigned to the second treatment. We discarded workers who were not completely consistent on the three consistency check questions and report the results for the 127 workers who were fully consistent. We have also conducted the same analysis on the entire worker pool, and the results are qualitatively the same.

H3: Effect of prediction source on human ethical preferences. We examine our Hypothesis 3, which claims that the effect of predictive information on ethical preferences varies based on the source of the prediction (AI vs Human Doctor), and that a prediction sourced from a Human Doctor has a stronger effect than a prediction sourced from an AI.

In Figure 6, we see how ΔP changes based on the source of the prediction for each human ethical preference factor. We also apply Bonferroni scaling to our significance tests. We see that changing the prediction source from AI to Doctor significantly decreases the ethical preference of the prediction ($p = .0316$). This result rejects our third hypothesis, as we actually see evidence suggesting that human ethical preferences from a prediction are weakened when the prediction source is a human doctor, and strengthened when the prediction source is an AI. We see that changing the prediction source from AI to Doctor increases the preference of Prior Donor, Wait Time, and Disease Stage, though not significantly. Combining both observations, one plausible conjecture is that workers might believe that AI predictions are generated by incorporating all verifiable information. Therefore, their preferences are influenced more by AI predictions instead of doctor predictions. Moreover, when AI predictions are available, workers put a smaller weight on other factors as they might be incorporated in AI predictions already. Our results suggest that how humans process predictions might vary when the predictions are from different sources.

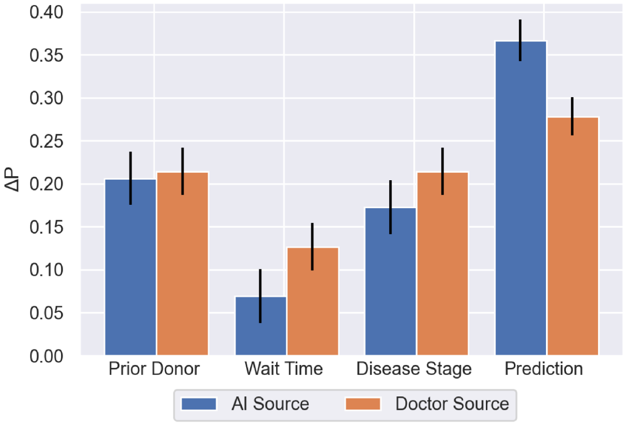


Figure 6: Effect of Prediction Source. We present ΔP for each factor and treatment. We also present p-values after Bonferroni correction for multiple comparisons. There is no significant difference between treatments in the Prior Donor factor, Wait Time factor, or Disease Stage. There is a significant difference between treatments in the Predictive factor ($p = .0316$).

Exploratory analysis. In our post-scenario survey, we asked workers to report the perceived trustworthiness of the predictive information on a five-point scale, as well as demographic information on age, gender, race, education level, and political leanings. We find that workers in the doctor treatment rated perceived trustworthiness of the prediction as 1.85/5, and workers in the AI treatment rated the perceived trustworthiness of the prediction as 1.96/5. This aligns with our results which showed that human preferences were more influenced by AI predictions than human predictions, and prior literature which suggests that humans trust AI more than human experts [28]. It is interesting to note that the relative values of perceived trustworthiness were so low for both, especially considering that the workers involved were laypeople, and prior research shows that experts trust algorithms less than lay people do [26–28].

We find that perceived trust is negatively correlated with levels of education. Workers with a bachelor's degree report 0.37 lower perceived trust in predictions than workers without a bachelor's degree. Workers with a graduate degree report 0.36 lower perceived trust in predictions than workers with a bachelor's degree or lower. This trends hold when we split workers by treatment (Doctor vs AI). We speculate that cause for this trend is that as humans believe themselves to be more capable, they tend to rely less on advice from others [13].

In our main analysis, we analyzed the difference in ΔP values between the AI prediction and Doctor prediction. For context, the total pool of workers have an average ΔP difference of 0.089. This value can be considered as a proxy of the difference between humans' reliance on AI prediction and the reliance on doctor prediction. To understand whether there exist individual differences, we break this down by demographic. We find that workers above the age of 40 have a ΔP difference of 0.027, while workers below the age of 40 have a ΔP difference of 0.122, suggesting that the majority of difference in overall workers is based on age, where younger workers' preferences are more influenced by AI predictions than doctors' predictions, compared to older workers. We find that male workers have a ΔP difference of 0.072, while female workers have a ΔP difference of 0.080, which does not suggest a strong contrast according to gender. We find that liberal workers have a ΔP difference of 0.058, while conservative workers have a ΔP difference of 0.101. Interestingly, conservative workers have higher values of ΔP than liberal workers regardless of source, with ΔP values of 0.407 and 0.276, respectively. While the presented results are not causal, the results as a whole suggest that there are individual differences in how humans incorporate AI/doctor predictions, and it would be an interesting future direction to further explore these individual differences.

5 DISCUSSION

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

Limitations and generalizability. Our study has a few limitations. First, our work has used the domain of kidney transplants as a case study to investigate how predictive information affects human ethical preferences. We believe this domain is representative of the family of problem domains involving allocating scarce medical interventions, e.g., organ transplants, vaccine distributions, or ventilator allocation. Relaxing the application beyond medical domains, our problem domain is in the family of domains involving allocation of scarce societal resources, such as allocating homelessness resources to people in need. We conjecture that the results of our study are very likely to generalize to the domains of medical resource allocation and are also likely to generalize to scarce societal resource allocation. However, it is also possible that our results will not directly generalize to these domains due to the uniqueness of the domain of kidney transplantation. Therefore, more future studies should be conducted to examine the generalizability of our results in other domains thoroughly.

We have conducted our experiments on Amazon Mechanical Turk. Due to the distributed nature of crowd work, we can not guarantee that workers have sufficiently engaged with the tasks and expressed their true preferences. While we have checked their answer consistency to remove potential noisy responses, the *hypothetical* nature of the presentation of the moral dilemma (as also adopted in prior works) might not provide a true reflection of what human ethical preferences would be when facing the scenarios in real life. Moreover, we have surveyed the ethical preferences from a general population of laypeople, who might also have different interpretations of the moral dilemma (e.g., whether they think another kidney will be available soon). It might be interesting/helpful to survey the preferences from relevant domain stakeholders. For example, in the domain of kidney transplants, we might want to also elicit preferences from medical doctors or policy makers. In the domain of autonomous vehicles, we might want to elicit preferences from car manufacturers, drivers, or pedestrians.

Implications of our results. Despite the limitations, our findings suggest a few important implications. First, our results suggest that the inclusion of predictive information impacts human ethical preferences in a nontrivial manner. Humans might consider what other factors might have already been incorporated in generating the predictive information and adjust their ethical preferences accordingly. We do not have a definite answer on how humans process predictive information. However, as predictive information is becoming increasingly involved in ethical decision making, it is important to understand how humans incorporate predictive information in forming their ethical preferences. Moreover, as shown in our exploratory analysis in Section 4.2, there exist individual differences in how people process predictive information. It is therefore important to take this into account when utilizing the elicited information to inform the design of AI systems.

Another important implication is on the robustness of elicited ethical preferences. Our results demonstrate that human ethical preferences could change significantly depending on how information is presented to them (e.g., highlighting the source of predictive information). This suggests that the elicited human ethical preferences might not be entirely robust and might be subject to information manipulation. While the growing literature on participatory design [25, 31, 46] have attempted to involve stakeholders in shaping the design of AI systems, our results suggest that, using the techniques from the literature on information design [20, 40], the advantageous party (e.g., the party that performs the elicitation) might strategically choose the information presentation to lead populations to express preferences that align with their objective. It is therefore important to understand under what conditions and to what extent we might rely on these elicited human preferences to guide the design with the goal of aligning AI with human values.

Future work. Our work has presented interesting findings on how predictive information affects human ethical preferences. However, there are still a lot of open questions that deserve future study. For example, how do human ethical preferences change when the presented predictive information becomes more accurate? If we explain how the predictive information is generated, does it impact how humans incorporate the information into their ethical preferences? Again, as predictive information becomes more ubiquitous, it is important to have a better understanding on how the presence and presentation of the predictive information impact humans. Moreover, as brought up by the above discussion on the limitations and implications, more studies on different problem domains and the populations surveyed would help us understand the generalizability of the results. It is also important to study how to leverage this elicited information to inform the design of AI systems and whether the elicited information is robust against potential manipulations.

6 CONCLUSION

In this work, we study the impacts of the presence and the source of predictive information on human ethical preferences. Using kidney transplants as a case study, we conducted randomized online experiments on Amazon Mechanical Turk. We presented scenarios where two candidates needed a kidney transplant but only one was available, and asked MTurk workers to express their preference on which candidate should receive the kidney first. We designed two experiments to examine the impacts of predictive information on human ethical preferences. We find that, when the predictive information is presented, even when the information is equal across two candidates, human preferences on different ethical dimensions change compared to the preferences without predictive information. When the predictive information aligns with existing preferences of the population, the preferences are further strengthened. Moreover, we investigate whether the source of the predictive information (i.e., from AI or from human experts) impacts human ethical preferences. We find that workers overall are influenced by AI predictions more than predictions from a human doctor. Moreover, when predictions from AI are presented, the impact of verifiable information on ethical preferences decreases more compared to when predictions are from a human doctor, possibly suggesting that workers are more likely treat an AI prediction as a summary of other verifiable information. As predictive information is increasingly integrated in ethical decision making, our results suggest that it is important to conduct more studies that involve the presence of predictive information. Moreover, since human ethical preferences are impacted by the presentation of the information (e.g., highlighting the source of the prediction), elicited human ethical preferences might not be robust and consistent across different elicitation methods. It is important to take this into account when utilizing the elicited information to inform the design of AI systems.

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] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23, 2016 (2016), 139–159.

[2] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2019. The Moral Machine Experiment. *Nature* (2019).

[3] Abdullah Awaysheh, Jeffrey Wilcke, François Elvinger, Loren Rees, Weiguo Fan, and Kurt L Zimmerman. 2019. Review of medical decision support and machine-learning methods. *Veterinary pathology* 56, 4 (2019), 512–525.

[4] Alexander Bartik and Scott Nelson. 2016. Credit reports as resumes: The incidence of pre-employment credit screening. (2016).

[5] Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. 2019. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence* 1, 1 (2019), 20–23.

[6] Richard Berk. 2017. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology* 13, 2 (2017), 193–216.

[7] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency.* 514–524.

[8] Philippe Bracke, Anupam Datta, Carsten Jung, and Shayak Sen. 2019. Machine learning explainability in finance: an application to default risk analysis. (2019).

[9] Lok Chan, Kenzie Doyle, Duncan McElfresh, Vincent Conitzer, John P Dickerson, Jana Schaich Borg, and Walter Sinnott-Armstrong. 2020. Artificial Artificial Intelligence: Measuring Influence of AI'Assessments' on Moral Decision-Making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* 214–220.

[10] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems* 29 (2016).

[11] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[12] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[13] Itiel E Dror and David Charlton. 2006. Why experts make errors. *Journal of Forensic Identification* 56, 4 (2006), 600.

[14] 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. *New England Journal of Medicine* 382, 21 (2020), 2049–2055.

[15] Ezekiel J. Emanuel and Alan Wertheimer. 2006. Who Should Get Influenza Vaccine When Not All Can? *Science* 312, 5775 (2006), 854–855.

[16] Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P. Dickerson, and Vincent Conitzer. 2018. Adapting a Kidney Exchange Algorithm to Align with Human Values. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* New York, NY, USA, 115.

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

[18] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2022. Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance* 77, 1 (2022), 5–47.

[19] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference.* 903–912.

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

[21] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security* 7, 6 (2012), 1789–1801.

[22] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. 2018. Discrimination in the Age of Algorithms. *Journal of Legal Analysis* 10 (2018).

[23] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[24] Pius Krütli, Thomas Rosemann, Kjell Y Törnblom, and Timo Smieszek. 2016. How to fairly allocate scarce medical resources: ethical argumentation under scrutiny by health professionals and lay people. *PloS one* 11, 7 (2016).

[25] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proc. ACM Hum.-Comput. Interact.* 3 (2019).

[26] Stephan Lewandowsky, Michael Mundy, and Gerard Tan. 2000. The dynamics of trust: comparing humans to automation. *Journal of Experimental Psychology: Applied* 6, 2 (2000), 104.

[27] Jennifer Marie Logg. 2017. Theory of machine: When do people rely on algorithms? *Harvard Business School working paper series# 17-086* (2017).

[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] Ethan Mark, David Goldsman, Brian Gurbaxani, Pinar Keskinocak, and Joel Sokol. 2019. Using machine learning and an ensemble of methods to predict kidney transplant survival. *PloS one* 14, 1 (2019).

[30] Johan Nilsson, Mattias Ohlsson, Peter Höglund, Björn Ekmehag, Bansi Koul, and Bodil Andersson. 2015. The International Heart Transplant Survival Algorithm (IHTSA): a new model to improve organ sharing and survival. *PloS one* 10, 3 (2015).

[31] Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. 2018. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[32] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[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] Gina M Piscitello, Esha M Kapania, William D Miller, Juan C Rojas, Mark Siegler, and William F Parker. 2020. Variation in ventilator allocation guidelines by US state during the coronavirus disease 2019 pandemic: a systematic review. *JAMA network open* 3, 6 (2020).

[35] Mohammad Pourhomayoun and Mahdi Shakibi. 2021. Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health* 20 (2021), 100178.

[36] Sara J Rosenbaum et al. 2011. Ethical considerations for decision making regarding allocation of mechanical ventilators during a severe influenza pandemic or other public health emergency. (2011).

[37] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public

attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.

[38] C Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[39] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2459–2468.

[40] 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.

[41] Berk Ustun and Cynthia Rudin. 2016. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning* 102, 3 (2016), 349–391.

[42] Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[43] Effy Vayena, Alessandro Blasimme, and I Glenn Cohen. 2018. Machine learning in medicine: addressing ethical challenges. *PLoS medicine* 15, 11 (2018).

[44] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[45] Douglas B White, Mitchell H Katz, John M Luce, and Bernard Lo. 2009. Who should receive life support during a public health emergency? Using ethical principles to improve allocation decisions. *Annals of Internal Medicine* 150, 2 (2009), 132–138.

[46] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. *Keeping Designers in the Loop: Communicating Inherent Algorithmic Trade-Offs Across Multiple Objectives*. Association for Computing Machinery, New York, NY, USA, 1245–1257.