Proceedings of the Tenth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2022)

Taking Advice from (Dis)Similar Machines: The Impact of Human-Machine Similarity on Machine-Assisted Decision-Making

Nina Grgić-Hlača,[1,2] **Claude Castelluccia,** [3] **Krishna P. Gummadi** [1]

[1] Max Planck Institute for Software Systems
[2] Max Planck Institute for Research on Collective Goods
[3] Inria
nghlaca@mpi-sws.org, claude.castelluccia@inria.fr, gummadi@mpi-sws.org

Abstract

Machine learning algorithms are increasingly used to assist human decision-making. When the goal of machine assistance is to improve the accuracy of human decisions, it might seem appealing to design ML algorithms that complement human knowledge. While neither the algorithm nor the human are perfectly accurate, one could expect that their complementary expertise might lead to improved outcomes. In this study, we demonstrate that in practice decision aids that are not complementary, but make errors similar to human ones may have their own benefits.

In a series of human-subject experiments with a total of 901 participants, we study how the similarity of human and machine errors influences human perceptions of and interactions with algorithmic decision aids. We find that (i) people perceive more similar decision aids as more *useful*, *accurate*, and *predictable*, and that (ii) people are *more likely to take opposing advice* from more similar decision aids, while (iii) decision aids that are less similar to humans have more opportunities to provide opposing advice, resulting in a higher influence on people's decisions overall.

1 Introduction

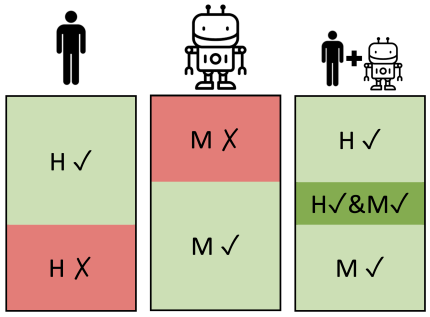
Machine decision aids assist human decision-makers in a variety of scenarios, ranging from medical diagnostics (Esteva et al. 2017) to bail decision-making (Angwin et al. 2016). The potential societal impact of using machine decision aids in real-world settings sparked concerns about their accuracy and fairness (Angwin et al. 2016; Barocas and Selbst 2016). Decades of research on machine learning can be leveraged to optimize machine decision aids for accuracy, while more recent research in the FAccT community proposed methods for alleviating some concerns about their fairness (Chouldechova 2017; Friedler, Scheidegger, and Venkatasubramanian 2016; Grgić-Hlača et al. 2018; Hardt et al. 2016; Kleinberg, Mullainathan, and Raghavan 2017; Zafar et al. 2017a,b), accountability and transparency (Caruana et al. 2015; Datta, Sen, and Zick 2016; Lakkaraju, Bach, and Leskovec 2016; Lakkaraju and Rudin 2017; Lipton 2016; Ribeiro, Singh, and Guestrin 2016; Wachter, Mittelstadt, and Russell 2017).

However, machine decision aids, as the name suggests, do not make the final decisions — they assist human decision makers. Hence, when designing decision aids, it is crucial to consider not only the decision aids' accuracy and fairness, but also how *human decision-makers* take their advice.

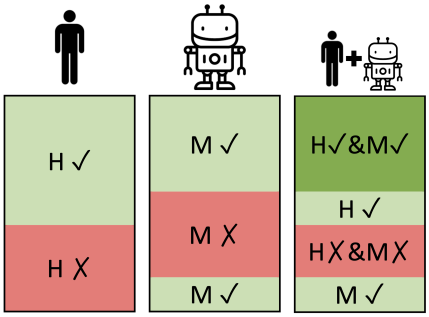
Recent work proposed machine learning algorithms which account for the presence of human agents in their learning procedure (De et al. 2020; Madras, Pitassi, and Zemel 2018; Meresht et al. 2020; Wilder, Horvitz, and Kamar 2020). A common thread underlying much of this research is the idea that designing algorithms with skills complementary to human ones may lead to better decision-making outcomes (Bansal et al. 2019; Horvitz and Paek 2007; Kamar, Hacker, and Horvitz 2012; Tan et al. 2018; Wilder, Horvitz, and Kamar 2020; Zhang, Liao, and Bellamy 2020), in line with the intuition presented in Figure 1.

How would we expect people to react to advice from a decision aid that complements human skills? Daniel Kahneman's notion of a modern Turing test (Kahneman 2021) posits that it is acceptable for a system to make mistakes a human might make, but it needs to avoid making mistakes that people would find to be absurd. Additionally, prior research in social psychology found that people are more receptive to advice from advisors more similar to themselves (Chan et al. 2017; Faraji-Rad, Samuelsen, and Warlop 2015; Suls, Martin, and Wheeler 2000; Yaniv, Choshen-Hillel, and Milyavsky 2011). In this paper, we study the comparative (dis)advantage of using similar or complementary decision aids in machine-assisted decision-making, where human agents make decisions upon receiving machine advice.

Experiment. In this paper, we experimentally test how the *similarity* of the *decision aid's errors* to typical *human errors* influences human advice-taking behavior. We compare decision aids which are equally accurate overall, but differ with respect to the type and distribution of their errors. Namely, they differ with respect to how similar their errors are to human errors: *human-like*, *anti human-like*, or *random*. Decision aids with *human-like* errors make mistakes for inputs where human respondents are also found to be the least accurate on average. Conversely, decision aids with *anti human-like* errors make mistakes complementary to human ones. They are accurate where humans tend to make mistakes, while making mistakes for inputs where most human respondents make accurate predictions. Finally, decision aids with *random* errors are between these two extremes, and have randomly distributed errors.



(a) Decision aid with errors *complementary* to human errors.



(b) Decision aid with errors *similar* to human errors.

Figure 1: Illustration: Examples of decision aids with errors *complementary* (Fig. 1a) or *similar* (Fig. 1b) to human errors. Green shading denotes areas where predictions are accurate, while red denotes inaccurate ones. The three panels in both subfigures show the distributions of human, machine, and joint human and machine errors respectively. *Complementary* decision aids provide a better upper bound on the accuracy of joint human and machine decisions (H✓∪ M✓) in the best case scenario, where humans take all correct advice, but no incorrect advice. On the other hand, *similar* decision aids provide a better lower bound (H✓∩ M✓) in the worst case scenario, where people take all incorrect machine advice, and no correct advice.

I.e., human-like decision aids make mistakes for questions that most people would find difficult to answer correctly, while anti human-like decision aids make mistakes only for questions that the majority of people would find easy to answer correctly. In terms of Kahneman's notion of a modern Turing test (Kahneman 2021), errors made by anti human-like decision aids may hence be perceived as egregious or absurd, since most humans would not make such errors.

To quantify the relationship between machine errors and human advice-taking, we conduct a series of human-subject experiments, in which we present respondents with machine advice and measure how the advice influences their decisions. Specifically, we utilize the Judge-Advisor paradigm (JAS), commonly used to study human advice-taking behavior (Bonaccio and Dalal 2006), where the decision aids serve as advisors, while human respondents retain decision rights.

To test the robustness of our findings across different domains, we consider three distinct decision-making scenarios: age estimation, criminal recidivism prediction, and dating preference prediction. As an additional robustness check, we consider two degrees of decision aid accuracy: decision aids that achieve the same degree of accuracy as typical human decisions, compared to decision aids that vastly outperform typical humans in terms of decision accuracy.

Contributions. We conducted a large-scale online study with 901 participants, exploring the impact of machine errors on human advice taking. We found that people's perceptions and advice taking behavior depend on the similarity of the decision aid's errors to typical human errors. In particular, we find that:

- People perceive decision aids that make errors similar to human ones as more useful, accurate, and predictable.
- People are more likely to take opposing advice from decision aids which make errors more similar to human ones.
- Nevertheless, people are significantly more likely to receive *opposing* advice from complementary decision aids. Hence, despite the lower influence of their opposing advice, complementary decision aids have a higher influence on people's decisions overall.

We follow up this confirmatory analysis with an exploratory analysis of the effects of human-machine similarity on the *accuracy* of people's decisions. We find that the use of complementary decision aids leads to a slightly greater increase in our respondents' accuracy. However, we also find that complementary decision aids are significantly farther from reaching their full potential for improving accuracy, since people are more likely to take opposing advice from decision aids that are more similar to them.

1.1 Related Work

Algorithmic decision aids nowadays advise human decision makers in a plethora of domains, ranging from bail decisions (Angwin et al. 2016) to medical diagnostics (Esteva et al. 2017). Hence, it is not surprising that much recent research has studied people's advice taking behavior in machine-assisted decision-making settings. Our research contributes to this interdisciplinary line of work, building up on prior research in social psychology and computer science.

Research in social psychology has studied how people perceive, react to and utilize machine advice compared to human advice (Madhavan and Wiegmann 2007). The results are mixed, and the findings vary across decision-making tasks (Vodrahalli, Gerstenberg, and Zou 2021). The majority of studies have reported evidence of algorithm aversion, finding that people tend to favor human advice over machine advice (Burton, Stein, and Jensen 2020; Dietvorst, Simmons, and Massey 2015, 2018; Dzindolet et al. 2002; Mahmud et al. 2022; Prahl and Van Swol 2017). People are also found to perceive machine decisions as less fair and trustworthy than human ones in tasks perceived to require human skills (Lee 2018). On the other hand, some studies reported algorithm appreciation (Logg 2017; Logg, Minson, and Moore 2019), finding that people were more receptive to machine advice than to human advice.

The insights related to algorithm aversion from Dietvorst, Simmons, and Massey (2015) are particularly relevant for our research: after observing humans and algorithms make the same mistake, people are found to lose confidence in algorithms more quickly than in humans, and opt for human advice over algorithmic advice. In our work, we hypothesize that this preference for human errors over algorithmic errors may go beyond the *identity* of the advisor, and that it may relate to expectations about the advisor's *behavior*. Namely, we study how people react to observing algorithms make errors that are similar to typical human mistakes compared to mistakes that very few people would make.

Recent work in computer science studied how people take advice from machine learning based decision aids. Green and Chen (2019a,b) studied how machine advice impacts the accuracy and fairness of human decisions. Several studies explored which factors impact the magnitude and quality of a decision aid's influence. For instance, prior work explored the effects of the accuracy of machine advice (Salem et al. 2015; Yin, Wortman Vaughan, and Wallach 2019; Yu et al. 2016, 2017), the interpretability and explainability of machine advice (Poursabzi-Sangdeh et al. 2018; Wang and Yin 2021; Zhang, Liao, and Bellamy 2020), providing warnings about machine limitations (Engel and Grgić-Hlača 2021), and varying the stakes associated with the decision-making task (Grgić-Hlača, Engel, and Gummadi 2019).

The work closest to ours is that of Yin, Wortman Vaughan, and Wallach (2019), which studied the effects of a decision aid's stated and observed accuracy on human advice taking behavior. In their experiments, they found that people were more likely to take advice from decision aids with a higher stated and observed accuracy than from less accurate ones. In this paper, we hypothesize that people's advice taking behavior depends not only on the *amount* of errors the decision aid makes (i.e., its accuracy), but also on the *type* of errors it makes. Specifically, we explore how human advice-taking behavior is influenced by observing algorithms make errors with varying degrees of similarity to typical human errors.

Much recent research in CS highlights the benefits of human-machine complementarity in joint human-machine decision-making (Bansal et al. 2019; Horvitz and Paek 2007; Kamar, Hacker, and Horvitz 2012; Tan et al. 2018; Wilder, Horvitz, and Kamar 2020; Zhang, Liao, and Bellamy 2020), especially in settings where algorithms allocate decision rights, such as the learning to defer framework (De et al. 2020; Madras, Pitassi, and Zemel 2018; Meresht et al. 2020; Wilder, Horvitz, and Kamar 2020). The intuition is straightforward. During the training phase, the algorithm prioritizes ensuring high predictive accuracy for inputs where humans are expected to make mistakes. Algorithms can make decisions when they are confident in their predictions, and otherwise defer decisions to their human collaborators. However, in many real-world settings algorithms are used as decision aids, whereas humans — such as judges or doctors — retain decision rights. Hence, a decision aid's impact depends on people's reactions to the provided advice. In this work, we study if the benefits of complementarity hold in machine-assisted decision-making, and if decision aids more similar to humans also exhibit some desirable properties.

2 Methodology

Hypotheses. Inspired by the notion of a modern Turing test (Kahneman 2021) and prior work in social psychology which found that people are more receptive to advice from similar advisors (Chan et al. 2017; Faraji-Rad, Samuelsen, and Warlop 2015; Suls, Martin, and Wheeler 2000; Yaniv, Choshen-Hillel, and Milyavsky 2011), we build upon prior work in machine-assisted decision-making reviewed in Section 1.1 to form three main hypotheses:

Comparing decision aids of equal accuracy, which differ with respect to the degree of similarity between the decision aid's and typical human errors,

Hypothesis 1: People perceive similar decision aids as more a) *useful*, b) *accurate*, and c) *predictable*.

Hypothesis 2: People are *more likely to take opposing advice* from similar decision aids.

Still, since complementary decision aids have less overlap with human decisions, they have more opportunities to give opposing advice. Hence, even though the likelihood of taking any individual piece of opposing advice from complementary decision aids may be lower (H2), we hypothesize:

Hypothesis 3: Complementary decision aids have a higher *overall influence* on human decisions.

We additionally engage in an exploratory analysis, to investigate the relationship between human-machine similarity and the accuracy of people's decisions.

2.1 Stimulus Material

Vignettes. In our experiments, we consider three different decision-making scenarios: *dating preference* prediction, *criminal recidivism* prediction, and *age* estimation. This set of scenarios covers a broad range of possible applications of machine decision aids. Firstly, these tasks differ with respect to their potential societal impact, with dating preference prediction on one end, and criminal recidivism prediction on the other. Secondly, they differ with respect to the type of thinking required by the decision-maker (Kahneman 2011), with age estimation being close to System 1, or fast thinking tasks, while dating and recidivism prediction being close to System 2, or slow thinking tasks. Finally, these scenarios provide decision makers with different amounts and types of information (images or natural language text).

The *dating preference* prediction task leverages data from a speed dating experiment gathered by Fisman et al. (2006). After being shown a summary of a speed date, our respondents were asked to predict if the speed dating participant wanted to see their date again. The speed date summaries contained information about the participants' demographics, romantic expectations and impressions about their date, as shown in Figure 8a in the Supplementary Material (SM).[1] Yin, Wortman Vaughan, and Wallach (2019) previously used this dataset in their study of machine-assisted decision-making, and the format of our vignettes exactly replicates theirs.

The *criminal recidivism* prediction task uses the ProPublica COMPAS dataset (Angwin et al. 2016). Our respon-

[1]Full paper including the Supplementary Material (SM) available at https://arxiv.org/abs/2209.03821.

dents were shown defendants' profiles, and asked to predict if they will commit a new crime within two years. The defendants' profiles contained information about the defendants' gender, age, and criminal history, as shown in Figure 8c in the SM. This dataset was previously used for studying machine-assisted decision-making (Grgić-Hlača, Engel, and Gummadi 2019) and algorithmic fairness (Dressel and Farid 2018; Zafar et al. 2017a; Grgić-Hlača et al. 2018; Corbett-Davies and Goel 2018; Chouldechova 2017).

Finally, the *age* estimation task relies on the IMDB-WIKI dataset gathered by Rothe, Timofte, and Gool (2015, 2018). We showed respondents images of people's faces, and asked them to estimate if they are above or below the age of 21, as depicted in Figure 8d in the SM.

For each of the three datasets, we selected a random subset of 50 data points to use as vignettes in our experiments. All of the datasets were cleaned, and incomplete or otherwise defective data points were removed prior to the random subset selection process. For the age dataset, which was skewed towards images of people above the age of 21, we additionally enforced class balancing constraints.

Decision Aids. To construct decision aids that have varying degrees of similarity to human errors, we first gathered data about how accurate humans are in our three decision-making scenarios. For each of the scenarios, we recruited approximately 100 respondents (a total of 305 respondents) and asked them to make predictions for all 50 data points from the respective dataset. As a measure of human accuracy for a specific data point, we used the fraction of respondents whose prediction matched the ground truth. For our sample size, the margin of error for human accuracy estimates is 10%, for a 95% confidence level.[2]

To construct the decision aids used in the main experiment, we relied on (i) estimates of the fraction of people who make accurate predictions for a given vignette, and (ii) ground truth labels. For each decision aid, we started with the ground truth labels and flipped a certain fraction of labels to achieve the desired degree of human-machine similarity and accuracy.[3] Human-like decision aids made errors only where people were most likely to make errors as well. More formally, they maximized the similarity between human and machine decisions, subject to an accuracy constraint. On the other hand, anti human-like decision aids made errors only where people were most likely to make correct decisions. I.e., they made mistakes that people are unlikely to make. More formally, they minimized the similarity between human and machine decisions, subject to an accuracy constraint. Finally, random decision aids had errors distributed uniformly at random, given an accuracy constraint.

As a robustness check, we developed decision aids at two levels of accuracy: human majority vote accuracy and superhuman accuracy. The former set of decision aids achieves the same degree of accuracy as the majority vote of human predictions did for the same decision-making scenario (Age: 76%, COMPAS: 60%, Speed Date: 64%). The latter has a superhuman accuracy of 84%, fixed across all three scenarios. We chose an accuracy of 84% to ensure that these decision aids outperform the human majority vote in all decision making tasks, while still making some errors (which is a prerequisite for studying our research questions).

We developed and ran experiments with 18 distinct decision aids ($\{$COMPAS, Age, Speed Date$\} \times \{$human-like, random, anti human-like$\} \times \{$majority vote accuracy, superhuman accuracy$\}$). E.g., for the Age dataset, an anti human-like decision aid with human majority vote accuracy has an accuracy of 76%. It makes incorrect predictions for 24% of data points, which were selected according to the accuracy of human predictions. Specifically, since it is the anti human-like decision aid, it makes incorrect predictions for the data points where the observed human accuracy was the highest. An example of mistakes made by the human-like and anti human-like decision aids for the COMPAS dataset is shown in Figure 10 in the SM.

2.2 Experimental Design

Experimental Procedure. Our experiment consisted of two phases: the (first) *test-drive phase* and the (second) *prediction phase*, in line with the design of Yin, Wortman Vaughan, and Wallach (2019) and Dietvorst, Simmons, and Massey (2015). In each phase, respondents answered questions about 25 vignettes — a total of 50 vignettes (Figures 8a, 8c, and 8d in the the SM). For each vignette, respondents made a preliminary pre-advice prediction, before observing machine advice, and then making their final post-advice prediction (Figure 8b in the SM). Before commencing the experiment, participants were shown an introductory text which described the decision-making task and the details of the experimental setup (Figures 7a and 7b in the SM).

In the *test-drive phase*, after each post-advice prediction, respondents received feedback about the accuracy of their own prediction and the decision aid's advice, thereby getting a chance to build a mental model of the decision aid. To incentivize respondents to put effort into building a mental model of the machine's predictions in the test-drive phase of the experiment, we informed them that they could earn monetary rewards in the prediction phase.

In the *prediction phase*, we explored how the mental models formed in the test phase influence people's advice-taking in the second phase. We did not provide feedback about the respondent's and the decision aid's accuracy after every question, in order to minimize the risk of the respondents updating their mental model of the machine's errors (Figure 8b in the SM, without the last paragraph). Following the approach of Dietvorst, Simmons, and Massey (2015), we used monetary incentives only in the prediction phase.

[2] After conducting the main experiment, we evaluated whether these human accuracy estimates coincided with the accuracy of people's pre-advice decisions in the main experiment. For each vignette, we calculated the difference between the fraction of people who made an accurate prediction in the first study and in the second study, prior to receiving machine advice. We found that for all three datasets the difference between people's accuracy in the two studies was not significant, with a mean close to 0.

[3] We opted for this synthetic approach to generating machine advice in order to ensure that we fully control the structure of the errors, to test our hypotheses. In real-world applications, ML based decision aids would be trained, as discussed in the Design Implications subsection of Section 4.

For each correct prediction, we rewarded respondents with a $0.10 bonus, and penalized them the same amount for each incorrect prediction. Similar financial incentives have been shown to encourage respondents to provide accurate responses (Chittilappilly, Chen, and Amer-Yahia 2016; Harris 2011).

At the end of each phase of the experiment, we gathered data about the respondents' perceptions of the decision aid's performance (Figures 9a, 9b and 9c in the SM). Specifically, we asked people to estimate their own and the decision aid's accuracy, and to assess their ability to predict the machine's predictions, as well as the machine's usefulness. We gathered this data twice, once at the end of each phase of the experiment. Immediately afterwards, we informed the participants how well they and the machine decision aid actually performed on the previous 25 questions.

Experimental Conditions. Our experiment has a between-subjects, randomized, full-factorial design with three factors: human-machine similarity (3 levels), decision-making scenario (3 levels), and decision aid accuracy (2 levels). We employ a repeated measures design, where each respondent makes 50 predictions. Each decision-making scenario considers a fixed set of 50 data points. Following the design of Yin, Wortman Vaughan, and Wallach (2019), the data points are split into two subsets, ensuring that each decision aid achieves the same predictive accuracy on both subsets respectively. In each phase, we show participants one of the subsets, and vary the order of subsets across participants. To prevent order-bias (Redmiles et al. 2017), we show the vignettes in random order in each phase.

Dependent Variables. We used data about people's pre and post-advice predictions to measure the *influence* of machine advice on peoples *decisions* and their *accuracy*. We consider two measures of influence on decisions: (i) *overall* influence, defined as the difference between post-advice and pre-advice agreement with machine advice, and (ii) *conditional* influence, defined as the overall influence for data points where people's pre-advice decisions disagreed with machine advice, i.e., the influence of opposing advice. Similarly, we define the *overall* and *conditional* influence on accuracy by comparing the accuracy of the respondents' pre and post-advice predictions. We also used data about the decision aid's perceived *usefulness*, *accuracy*, and *predictability*, gathered at the end of each experimental phase.

2.3 Data Collection

We recruited respondents using the online crowd-sourcing platform Prolific. Prolific is an alternative to MTurk, commonly used for recruiting participants for online human-subject studies in academic research (Palan and Schitter 2018). We recruited respondents who have self-reported to be US nationals, had an approval rate of at least 95% on previous studies, and have completed at least 100 studies so far.

In our experiments, we included two simple instructed response items as attention check questions (e.g., "Please respond to this question by selecting *Somewhat disagree* as the answer."). Similar instructed response items are commonly employed for identifying careless responses in survey data (Meade and Craig 2012). In our analysis, we discarded the responses of all respondents who did not complete the survey, or did not successfully complete both attention check questions. Our final sample consists of 901 respondents who successfully answered both attention check questions.

Demographic Attribute	Sec. 3.1	Sec. 3.2 - 3.4	Census
Female	56%	48%	51%
Asian	10%	14%	6%
Black	7%	7%	13%
Hispanic	8%	7%	18%
White	70%	69%	61%
Other	5%	3%	4%
$>$ Bachelor's degree	57%	56%	30%
$<$ 35 years	63%	59%	46%
$35-54$ years	31%	31%	26%
$55+$ years	6%	10%	28%
Liberal	54%	59%	33%†
Conservative	12%	15%	29%†
Moderate	27%	23%	34%†
Other	7%	4%	4%†

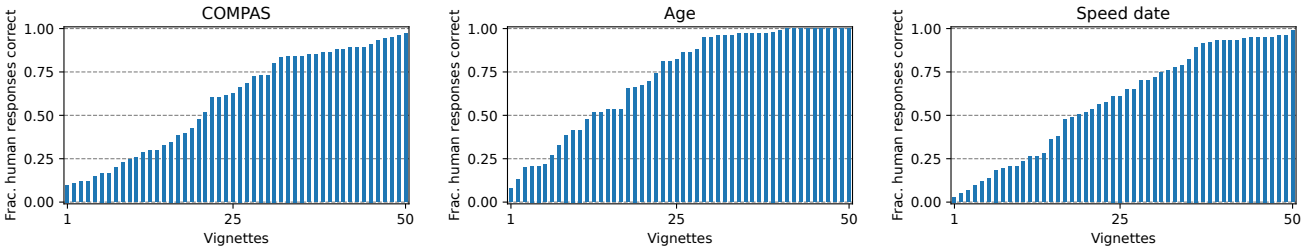
Table 1: Survey Samples. We targeted respondents who have self-reported to be US nationals, and have completed at least 100 studies on Prolific with an approval rate $\geq$ 95%. The respondents' demographics are compared to the 2016 U.S. Census (U.S. Census Bureau 2016), and Pew data on political leaning (Pew Research Center 2016) (marked with †).

We ran two sets of experiments. In the first experiment we gathered data for modeling the accuracy of human decisions (Section 3.1). 305 out of 320 respondents (95%) correctly answered the attention check questions. The average completion time for this set of surveys was 14 minutes, and respondents were paid a base fee of £2 for taking part in this experiment (i.e., slightly above $11 per hour).

In the second experiment we gathered data for testing our hypotheses (Sections 3.2 - 3.4). 92% of respondents answered both attention check questions correctly, resulting in a sample of 596 respondents. On average, participants took 21 minutes to complete the survey. Respondents were paid a base fee of £2.5 for taking part in the experiment (i.e., slightly above $9.3 per hour). Additionally, the respondents could earn bonus payments based on their performance, as described in Figure 7b in the SM.

To assess participant satisfaction and identify any issues with the experiments, participants were asked to answer a series of questions about their experience. In both experiments participants reported that they found the survey interesting (mean ratings of 4.3 and 4.6 on a 5-point Likert scale for the first and second experiment respectively), that they found the questions easy to understand (mean ratings of 4.7 and 4.8 respectively), and that they would like to take part in a similar survey in the future (4.6 and 4.8 respectively).

In Table 1 we report the detailed demographics of the samples we used to model the accuracy of human decisions (1st column) and to test our hypotheses (2nd column). Since the respondents are US nationals, we additionally compared the sample demographics to the 2016 US Census (U.S. Census Bureau 2016), and Pew Research Center's data on political leaning (Pew Research Center 2016) (3rd column). On a



(a) COMPAS (b) Age (c) Speed Date

Figure 2: Distribution of human errors for the COMPAS [left], Age [center], and Speed date [right] dataset. The y-axis shows the percent of human predictions that were correct for a given vignette. The vignettes are sorted increasingly w.r.t. y-axis values.

high-level, our respondents are more educated, younger, and more liberal than the general US population. Also, our sample consists of more white and Asian respondents, and fewer black and Hispanic respondents than the US population.

2.4 Analysis

Throughout the paper, we utilize descriptive statistics to summarize the basic information about our data. We corroborate these findings with statistical hypothesis testing.

To study the causal effect of the experimental manipulations on people's perceptions and behavior, we employ linear mixed models. We account for repeated measures by including crossed random effects terms for respondents and questions. For H1, we rely on a multivariate multiple linear regression with three dependent variables: perceived usefulness (5-point Likert scale from -2 to 2), accuracy ($[0, 1]$) and predictability (5-point Likert scale from -2 to 2). For H2 and H3, we utilize a multiple linear[4] regression with the overall and conditional influence of advice as dependent variables respectively. For each of the three hypotheses we build two models, one including only human-machine similarity as an independent variable, and one which also includes the control variables (experimental phase, accuracy, and dataset).

Human-machine similarity is a categorical variable with 3 levels: Human-like, Random, and ¬Human-like. In all of the models, we use the Random level as the reference category. I.e., the estimated regression coefficients for Human-like and ¬Human-like treatments convey information about the effect of the respective treatment compared to the Random treatment. We additionally perform Wald tests on the estimated regression coefficients of the Human-like and ¬Human-like treatment to directly compare their effects.

3 Results

3.1 Designing the Decision Aids

The first step in studying the effects of human-machine similarity on people's advice taking behavior was designing machine decision aids with varying degrees of similarity to human errors. To do so, we gathered data about human decisions, to understand for which inputs people typically make accurate predictions, and for which they make mistakes.

In Figure 2, we show the distribution of human errors for each of the three datasets separately. For each vignette, we show the percent of respondents who made an accurate prediction (without algorithmic assistance). We find that people's accuracy varies between datasets and between vignettes. Each dataset consists of data points where people are overwhelmingly correct (on the right of the x-axis), and data points where most respondents made incorrect decisions (on the left of the x-axis). In other words, each dataset has vignettes with varying degrees of human accuracy, which will enable us to design decision aids with varying degrees of similarity of human and machine errors.

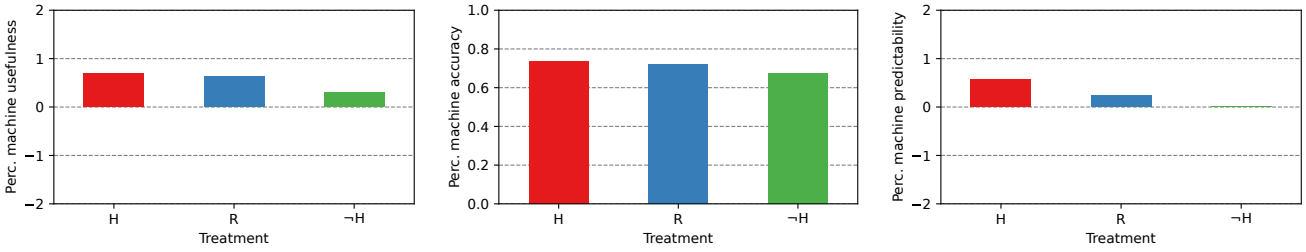
For the COMPAS dataset, the mean accuracy of people's responses was 0.59. For 60% of the vignettes, more than half of respondents made an accurate prediction. People's performance in terms of accuracy on the COMPAS dataset was comparable to that reported in prior work (Dressel and Farid 2018; Grgić-Hlača, Engel, and Gummadi 2019). For the Speed date dataset, people had a mean accuracy of 0.59, and an accurate majority vote for 64% of the vignettes. Finally, for the Age dataset, people were more accurate, with a mean accuracy of 0.72, and an accurate majority vote prediction for 76% of the vignettes. We used this data to construct human-like (H), random (R) and anti human-like (¬H) decision aids, as described in Section 2.

3.2 Human Perceptions of Machine Performance

In this section, we present our results on people's perceptions about the decision aids' performance. Specifically, we explore the relationship between human-machine error similarity, and the decision aids' perceived *usefulness*, *accuracy* and *predictability*.

Descriptively, Figure 3 shows that human-machine similarity is positively correlated with all three perceptions of performance. In Figure 3a, we see that human-like decision aids are on average rated as more useful than random ones, which are in turn rated as more useful than anti human-like decision aids. The same pattern holds for the perceived accuracy in Figure 3b, and predictability in Figure 3c. The mean

[4]We utilize a linear regression for ease of interpretation of coefficients, especially of interaction effects. For a discussion on the applicability of linear models for binary dependent variables, please refer to Hellevik (2009). Our results remain qualitatively the same when a logistic regression is applied on the binary dependent variables in H2 and H3.



(a) Usefulness (b) Accuracy (c) Predictability

Figure 3: Perceptions of machine predictions for decision aids with human-like (H), anti human-like (¬H), and randomly distributed (R) errors. The perceived usefulness and predictability were rated on a 5-point Likert scale. On the plots, -2 corresponds to the lowest rating of usefulness and predictability, and 2 corresponds to the highest. The perceived accuracy was reported as the estimated number of correct predictions, as shown in Figure 9a in the SM, and was converted to a value $\in [0, 1]$ in our analysis. Human-like decision aids are perceived as more useful, accurate, and predictable than anti human-like decision aids.

	Model 1	**Model 2**
Usefulness		
Human-like	0.0674 (0.074)	0.0500 (0.071)
¬ Human-like	-0.328*** (0.074)	-0.330*** (0.071)
Second Phase		-0.0419 (0.057)
High Acc.		0.673*** (0.058)
Age D.		-0.0351 (0.071)
Speed Date D.		0.0199 (0.071)
Intercept	0.630*** (0.052)	0.320*** (0.076)
Accuracy		
Human-like	0.0135 (0.015)	0.00942 (0.013)
¬ Human-like	-0.0471** (0.015)	-0.0488*** (0.013)
Second Phase		-0.00450 (0.006)
High Acc.		0.145*** (0.011)
Age D.		0.0588*** (0.013)
Speed Date D.		0.00929 (0.013)
Intercept	0.723*** (0.011)	0.630*** (0.013)
Predictability		
Human-like	0.338*** (0.072)	0.333*** (0.071)
¬ Human-like	-0.225** (0.072)	-0.230** (0.071)
Second Phase		-0.134* (0.057)
High Acc.		0.217*** (0.058)
Age D.		0.197** (0.071)
Speed Date D.		-0.0309 (0.071)
Intercept	0.232*** (0.051)	0.136 (0.076)

Table 2: Dependent variables: Perceived usefulness, accuracy and predictability. For non-binary variables, the reference categories are the Random Treatment and the COMPAS Dataset. Standard errors in parentheses. * symbols next to coefficients indicate their statistical significance as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. N = 1192. Wald tests show that the coefficients associated with the human-like and anti human-like treatment are significantly different for all three perceptions, with $p < 0.001$.

perceived usefulness and predictability of human-like decision aids are 0.4 and 0.6 points higher (on a 5-point Likert scale) than of anti-human like ones, and human-like decision aids are also perceived as 6% more accurate.

Model 1 in Table 2 and the subsequent Wald test on its coefficients show that the observed difference between the perceptions of human-like decision aids (H) and anti human-like ones (¬H) is statistically significant with p<0.001. However, the difference between the random decision aid (R) and human-like decision aids (H) is significant only for the perceived predictability. This gives us

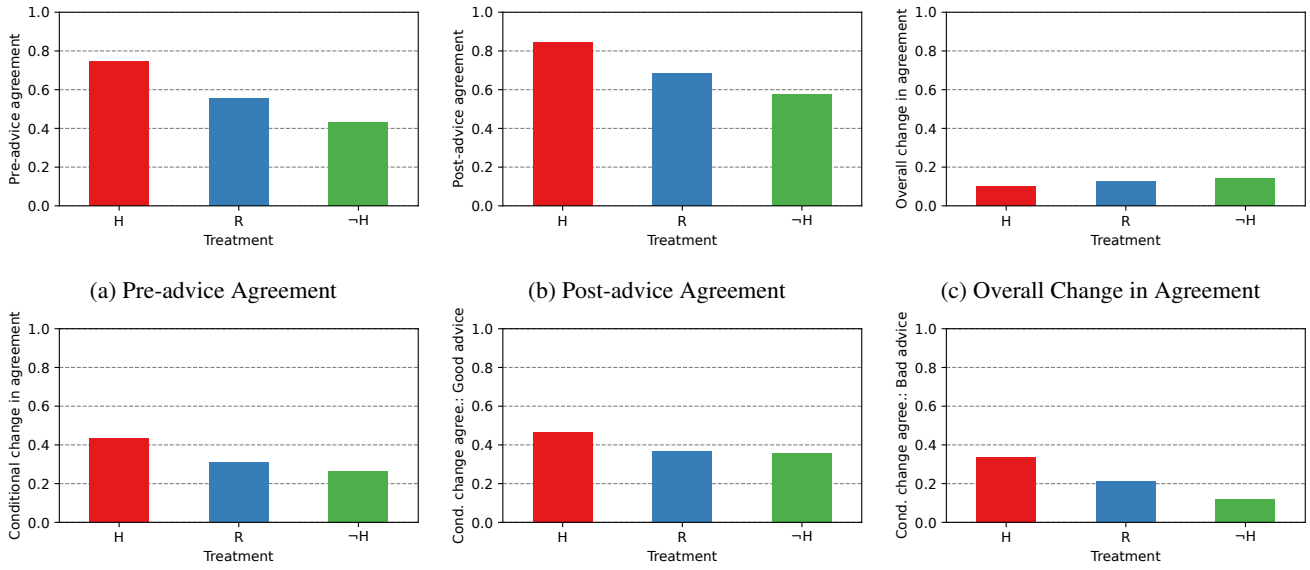
> **Result 1.** People perceive human-like decision aids as more useful, accurate, and predictable than anti human-like decision aids.

Control Variables. This pattern holds when we control for the decision-making task, the decision aid's accuracy, and the experimental phase. Figure 11 in the SM shows the results from Figure 3 separated by these three control variables. Model 2 in Table 2 shows that the differences between the perceived usefulness, accuracy and predictability of H and ¬H are statistically significant even when controlling for the aforementioned variables.

Model 2 in Table 2 also provides interesting insights about the effects of the control variables. The accuracy of the decision aid significantly impacted all three perceptions. More accurate decision aids were perceived as more useful, accurate, and predictable. The effects of the experimental phase and the dataset were not consistent across all three perceptions. The experimental phase had a borderline significant effect only on predictability. In the second phase, where people did not observe feedback about the decision aid's performance after every question, people rated decision aids as less predictable. The Speed Date dataset did not lead to significantly different perceptions than the COMPAS dataset. However, for the Age dataset, where the baseline human (and hence machine) accuracy was significantly higher than in the other two datasets, people rated the decision aids as more accurate and predictable.

3.3 Human Receptiveness to Machine Advice

After discussing the effect of human-machine similarity on human *perceptions*, we move on to commenting on human advice taking *behavior*. As a first sanity check, we test if machine advice had an effect in the expected direction. As shown in Table 3, we found that many people updated their



(a) Pre-advice Agreement (b) Post-advice Agreement (c) Overall Change in Agreement

(d) Conditional Change in Agreement (e) Conditional Change: Good Advice (f) Conditional Change: Bad Advice

Figure 4: Influence of machine advice on respondents' agreement with the advice, for decision aids with human-like (H), anti human-like (¬H), and randomly distributed (R) errors. Pre and post-advice agreement, as well as the conditional change in agreement are higher for human-like decision aids than for anti human-like decision aids. On the other hand, for the overall change in agreement, the opposite pattern holds. Figures 4e and 4f compare the conditional change in agreement for good and bad advice separately.

Agreement		**Treatments**		
Pre	Post	H	R	¬H
0	0	0.145	0.306	0.418
0	1	0.111	0.137	0.150
1	0	0.012	0.009	0.006
1	1	0.732	0.549	0.426

Table 3: Distribution of the respondents' decisions with respect to the four possible agreement configurations, for each of the three treatments. The first two columns show the four configurations, where 0 denotes disagreement and 1 denotes agreement with machine advice.

	Model 1		**Model 2**	
Human-like	-0.0296*	(0.012)	-0.0308**	(0.011)
¬ Human-like	0.0163	(0.012)	0.0173	(0.011)
Second Phase			0.0277***	(0.006)
High Acc.			0.0749***	(0.010)
Age D.			-0.0743***	(0.015)
Speed Date D.			-0.0104	(0.016)
Intercept	0.128***	(0.010)	0.105***	(0.013)

Table 4: Dependent variable: Overall agreement change. For non-binary variables, the reference categories are the Random Treatment and the COMPAS Dataset. Standard errors in parentheses. * symbols next to coefficients indicate their statistical significance as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. N = 29800. Wald tests show that the coefficients associated with the human-like and anti human-like treatment are significantly different, with $p < 0.001$.

decisions after receiving opposing advice (the likelihood of doing so varies across treatments, as discussed below), but very few people changed their pre-advice decision after observing that the machine agrees with them. Namely, in the third row we see that very few respondents ($\leq 1\%$) initially agreed with machine advice and then switched their decision and disagreed with it after observing confirming advice. I.e., machine advice affects people's decisions in the expected direction. Next, we explore the variation across treatments.

Figure 4a shows how the agreement of human *pre-advice* decisions with machine advice varies across decision aids. As intended by the design of our decision aids, pre-advice agreement is positively correlated with human-machine similarity. Using a linear mixed model and a subsequent Wald test, we confirmed that this effect is statistically significant, with a p-value < 0.001. This sanity check demonstrates that our decision aids successfully achieve their designated degree of similarity to human decisions.

After this sanity check, we move on to analyzing participants' *post-advice* decisions. Figure 4b show that people's post-advice decisions follow the same pattern as their pre-advice decisions. I.e., even after receiving machine advice, people's decisions are more similar to the human-like decision aids than to anti human-like decision aids. Using a linear mixed model and a subsequent Wald test, we found that this effect is also significant, with a p-value < 0.001.

In Figure 4c, we focus on the *overall change in agree-*

	Model 1		Model 2	
Human-like	0.124^{***}	(0.032)	0.102^{***}	(0.030)
¬ Human-like	-0.0453	(0.025)	-0.0436	(0.022)
Second Phase			0.0736^{***}	(0.012)
High Acc.			0.184^{***}	(0.020)
Age D.			-0.111^{***}	(0.028)
Speed Date D.			-0.0305	(0.030)
Intercept	0.309^{***}	(0.020)	0.231^{***}	(0.024)

Table 5: Dependent variable: Conditional agreement change, i.e., agreement change for instances where pre-advice decisions disagreed with machine advice. For non-binary variables, the reference categories are the Random Treatment and the COMPAS Dataset. Standard errors in parentheses. * symbols next to coefficients indicate their statistical significance as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. N = 12578. Wald tests show that the coefficients associated with the human-like and anti human-like treatment are significantly different, with $p < 0.001$.

ment, calculated as the difference between post and pre-advice agreement. While both pre and post-advice decisions showed a positive correlation between human-machine similarity and agreement, the change in agreement exhibits the opposite pattern. Human-like decision aids lead to a slightly lower (by 5 percentage points) overall increase in agreement than anti human-like ones. This finding is corroborated by the regression and Wald test in Table 4, Model 1, with a p-value < 0.001. This gives us

Result 2. Anti human-like decision aids have a higher overall influence than human-like decision aids.

However, it is important to keep in mind that the pre-advice agreement rates for anti human-like decision aids were significantly lower than for human-like ones. Hence, people had more opportunities to receive, and consequently take, opposing advice from anti human-like decision aids than from human-like ones. Next, we account for this.

In Figure 4d, we calculate the *conditional change in agreement* as the difference between post and pre-advice agreement, for instances where people initially disagreed with machine advice. I.e., this measure captures the amount of received opposing advice that was taken. With this, we control for the differences in pre-advice agreement rates across treatments. Even though the overall change in agreement was higher for anti human-like decision aids, the conditional change is higher for human-like decision aids. Specifically, the likelihood of taking opposing advice from human-like decision aids was 17 percentage points higher than for anti human-like ones (0.43% vs 0.26%). This result is corroborated by the regression and Wald test shown in Table 5, Model 1, with a p-value < 0.001. This leads to

Result 3. Human-like decision aids have a higher conditional influence than anti human-like decision aids.

Control Variables. All of the findings hold when we control for the decision-making task (i.e., for the dataset), the decision aid's accuracy, and the experimental phase. This can be seen in Figure 12 in the SM, where the results from Figure 4 are separated by the three aforementioned control variables. We show that our results also remain statistically significant when we introduce these control variables in Tables 4 and 5, Model 2.

	Model	
Human-like Treatment	0.127^{**}	(0.042)
¬ Human-like Treatment	-0.0883^{**}	(0.028)
Advice Correct	0.156^{***}	(0.028)
Human-like # Advice Correct	-0.0265	(0.041)
¬ Human-like # Advice Correct	0.0788^{**}	(0.030)
Intercept	0.210^{***}	(0.025)

Table 6: Dependent variable: Conditional agreement change. This Model builds up on Model 1 from Table 5 by accounting for the correctness of machine advice. Interactions are denoted by #. The reference category is the Random treatment. Standard errors in parentheses. * symbols next to coefficients indicate their statistical significance as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. N = 12578.

As was the case for perceptions, the control variables again have a significant effect on advice taking behavior. As seen in Tables 4 and 5, Model 2, the control variables have a consistent effect on both overall and conditional agreement. The coefficients indicate that people were more likely to take machine advice in the second experimental phase.[5] Also, people were more likely to take advice from more accurate decision aids. There was no significant difference between advice taking for the COMPAS and Speed Date dataset, for which people have a similar baseline accuracy. However, for the Age dataset, where their baseline accuracy was higher, people were less likely to take machine advice.

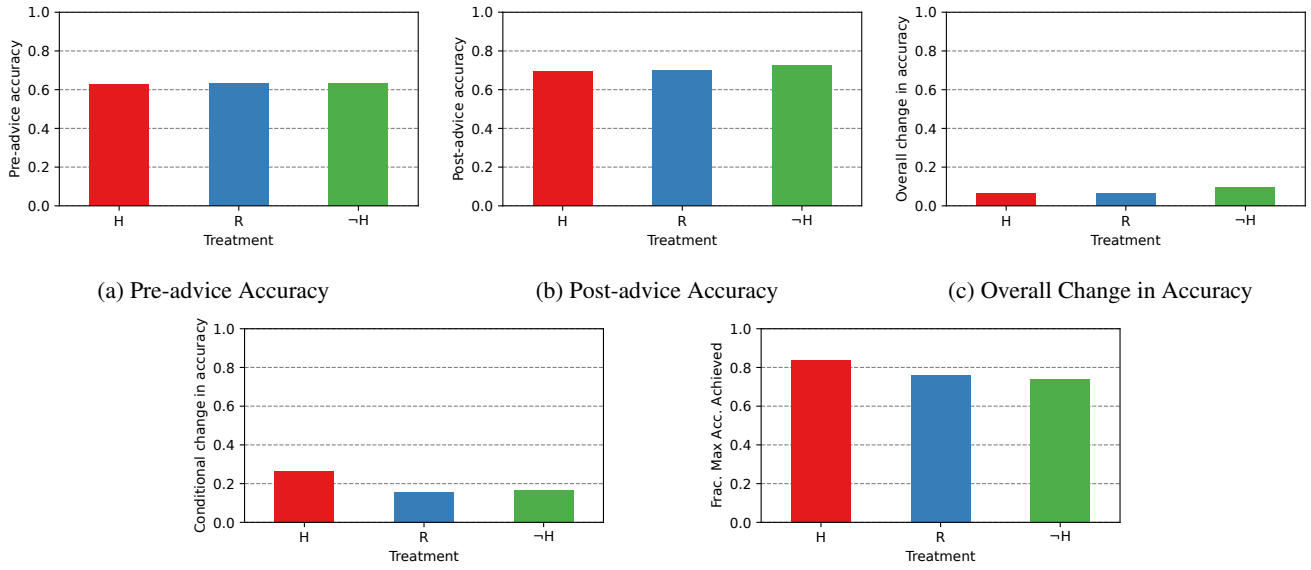
3.4 Impact of Advice on Human Accuracy

In this Section, we conduct an exploratory study, and describe the impact of human-machine similarity on the accuracy of people's decision in our experiments.

Comparing the Influence of Good and Bad Advice. The impact of people's advice taking behavior on the accuracy of their decisions depends not only on the amount of advice that is taken, but also on the quality of the taken advice. In Figures 4e and 4f, we show the conditional change in agreement for correct and incorrect advice separately. Specifically, Figure 4e depicts the influence of good advice, i.e., the fraction of times people change their incorrect pre-advice decision after observing correct machine advice. On the other hand, Figure 4f shows how people react to bad advice, i.e., the fraction of times people update their correct pre-advice decision after receiving incorrect advice.

Firstly, and reassuringly, we observe that good advice was more influential than bad advice, for all of the decision aids. However, we see differences in the magnitude of influence across treatments. Human-like decision aids have a higher conditional influence (i.e., people exhibit a higher propen-

[5]This difference could be caused by the introduction of monetary incentives and removal of feedback about performance in the second phase, or other factors, such as learning or fatigue effects.



(a) Pre-advice Accuracy

(b) Post-advice Accuracy

(c) Overall Change in Accuracy

(d) Conditional Change in Accuracy

(e) Fraction of Potential Achieved

Figure 5: Influence of machine advice on respondents' accuracy, for decision aids with human-like (H), anti human-like (¬H), and randomly distributed (R) errors. Pre and post-advice accuracy, and the conditional change in accuracy are higher for human-like decision aids than for anti human-like decision aids. Human-like decision aids were also closer to reaching their full potential in terms of improving the respondents' accuracy. However, for the overall change in accuracy, the opposite holds.

sity to take opposing advice), both for correct advice and for incorrect advice. I.e., people are more likely to take good advice from human-like decision aids, but they are more likely to reject bad advice from anti human-like decision aids. These observations are corroborated by the regression in Table 6. Hence, our results do not show evidence that varying the degree of human-machine similarity might help people distinguish between good and bad machine advice.

Impact on Accuracy. Figure 5a shows that, descriptively, people's *pre-advice* accuracy is on average equal across treatments. I.e., as expected, before observing machine advice, people were equally accurate in all of our treatments. Using a linear mixed model, we found that the difference across treatments is indeed not significant, with a Wald test p-value of 0.7982. However, the *post-advice* accuracy slightly differs across treatments, as visible in Figure 5b. Using a linear mixed model, we found that post-advice accuracy is higher for the anti human-like treatment, with a borderline significant Wald test p-value < 0.05.

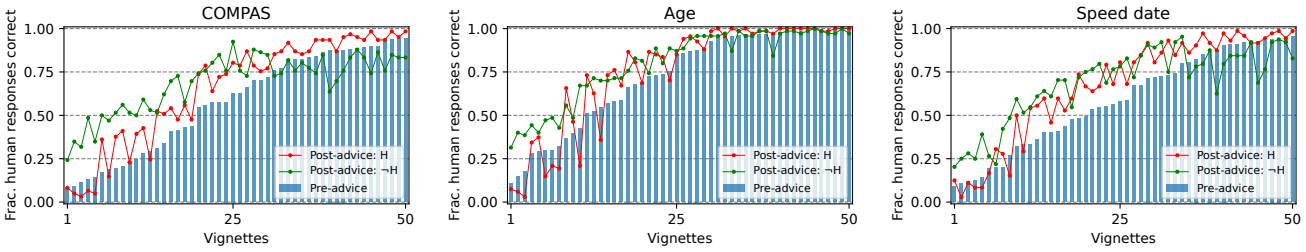
In Figure 5c, we focus on the *overall change in accuracy*, calculated as the difference between post and pre-advice accuracy. Consistently with our finding about the post-advice accuracy, we observe that the anti human-like treatment leads to a slightly higher overall change in accuracy. Using the same statistical analysis as above, we found that this effect is also borderline significant, with a p-value < 0.05.

Next, we discuss the *conditional change in accuracy*, calculated as the difference between post and pre-advice accuracy, for instances where a respondent's pre-advice decision disagreed with machine advice. In Figure 5d, we observe that human-like decision aids lead to a higher conditional change in accuracy. Again, this effect is found to be borderline statistically significant, with a p-value < 0.05.

All of the findings related to people's accuracy are in line with the findings on people's advice taking behavior, from Section 3.3. Since the decision aids are on average more accurate than humans (i.e., they provide more good advice than bad advice), and good advice was more influential than bad advice, machine advice positively impacted the accuracy of people's responses in line with their advice taking propensity. The overall influence of anti human-like advice is higher, and so is the overall change in accuracy. On the other hand, the conditional influence of human-like advice is higher, and so is the conditional change in accuracy.

While Figure 5 shows the decision aids' impact on accuracy averaged across all vignettes, in Figure 6 we show how the decision aids impacted people's accuracy for specific vignettes (averaged across the two experimental phases and degrees of accuracy). We observe that for vignettes where most people made errors (left side of the plots), the human-like decision aid (red line) also made errors and hence the anti human-like (green) decision aid had a more positive impact on the respondents' accuracy. However, for instances where most people made correct predictions (right side of plots), the human-like (green) decision aids helped many of the remaining respondents correct their decisions, unlike the anti human-like (green) decision aid which steered some of the respondents away from their initially correct responses.

Potential for Improving Accuracy. Next, we go back to the illustration from Figure 1. Using the same reasoning as the one applied in the illustration, we can conclude that anti human-like decision aids have the potential to improve the



(a) COMPAS (b) Age (c) Speed Date

Figure 6: Influence of machine advice on respondents' accuracy. The bars show the respondents' pre-advice accuracy, while the lines show their post-advice accuracy, for decision aids with human-like (red line) and anti human-like (green line) errors. On the x-axis, the 50 vignettes are sorted increasingly w.r.t. the fraction of respondents' whose pre-advice decision was correct. We observe that on the left side of the plots (where people are more prone to errors) the anti human-like decision aid impacts the respondents' accuracy more positively on average. On the right side of the plots (where people are less likely to make errors) the human-like decision aid improves accuracy, while the anti human-like one reduces people's accuracy.

accuracy of human decisions more than human-like decision aids. I.e., for cases where humans make incorrect predictions, anti human-like decision aids are more likely to be able to give people correct advice. In the following analysis, we try to account for this. We explore how close the post-advice accuracy of human decisions is to the upper bound given by the union of the responses where humans were correct and those where the decision aids were correct.

Figure 5e shows the results of this analysis. This plot shows the fraction of correct post-advice decisions, amongst all of the decisions where either the decision aid's advice was correct or the human respondent's pre-advice decision was correct. This analysis aims to capture the fraction of the upper bound on accuracy that was achieved. Here we found that human-like decision aids reached more of their full potential than anti human-like decision aids. This difference was confirmed to be statistically significant, with a p-value < 0.001. I.e., even though the overall increase in accuracy was slightly higher for anti human-like decision aids (Fig. 5c), they were significantly farther from reaching their full potential for improving human decisions.

4 Discussion

Limitations and Future Work. In our experiments, we considered synthetic decision aids with three degrees of similarity to human decisions, three binary prediction tasks, and two degrees of predictive accuracy. Future work could study additional variables or more fine-grained levels of these variables. E.g., the human-like decision aids we studied made mistakes only where humans made them as well, while the anti human-like decision aids made mistakes only where most humans were accurate. These differences may be less severe for real-world decision aids, and it is hence worth conducting further research varying the degree of human-machine similarity beyond the three degrees we studied.

We compared decision aids that make mistakes for different inputs. In a multi-class (e.g., emotion recognition) or regression (e.g. real-estate price or continuous age estimation), setting, one could also compare decision aids that make different types of mistakes for the same input. E.g., people and human-like decision aids might underestimate the true value of a certain vignette in a regression, while anti human-like decision aids might overestimate it.

Future studies could also consider other respondent samples, e.g., representative of the US population or other non-US populations. For decision aid used by professionals in the real world (e.g., the COMPAS tool for predicting criminal recidivism used by judges in the US legal system), it may also be interesting to test whether our findings hold for the relevant population of real world decision makers.

In our work we focused on experimentally testing the effect of human-machine similarity on human perceptions and utilization of machine decision aids. Past research has identified other factors that also influence how people take machine advice, such as the interpretability and explainability of machine advice (Poursabzi-Sangdeh et al. 2018; Wang and Yin 2021; Zhang, Liao, and Bellamy 2020), and future research could explore if these factors moderate the effects of human-machine similarity. Finally, future work in social psychology could study the psychological mechanisms which underlie the observed effect of human-machine similarity on advice taking. As a first step in this direction, we provide a brief discussion on this below.

Underlying Psychological Mechanisms. As reported in Section 3.2, people perceive human-like decision aids as more predictable, accurate and useful. The predictability of others' future actions has long been recognized as a crucial component of interpersonal trust (Rempel, Holmes, and Zanna 1985), and was also shown to be critical for human trust in automation (Hoff and Bashir 2015; Madhavan and Wiegmann 2007). People's receptiveness to advice was also shown to be correlated with the inferred quality and accuracy of an advisor's advice (Bonaccio and Dalal 2006). Finally, the perceived usefulness of information technologies is highly predictive of the adoption of such technologies (Davis 1989; Karahanna and Straub 1999). Hence, people's perceptions about the comparative advantages of human-like decision aids in terms of their predictability, accuracy and

usefulness may lead to the observed higher receptiveness to opposing advice given by such systems.

One crucial mediator of the effect of human-machine similarity on advice taking could be trust, which is found to positively affect receptiveness to advice (Bonaccio and Dalal 2006). As mentioned in Section 1.1, literature on algorithmic aversion has found that people tend to lose trust in algorithms more quickly than in human advisors after observing them make the same mistakes (Dietvorst, Simmons, and Massey 2015). We comment on two concepts which are found to affect trust resilience, which go beyond the human/algorithmic identity of the advisor and may explain our observations: error severity and anthropomorphism.

The magnitude of a system's errors was shown to be correlated with the magnitude of users' loss of trust (Rossi et al. 2017; Weun, Beatty, and Jones 2004). It is possible that human-like mistakes are perceived as less severe, while anti human-like errors are perceived as egregious, hence leading to lower trust resilience for complementary decision aids. Anthropomorphism refers to the tendency to ascribe human-like characteristics to non-human agents (Epley, Waytz, and Cacioppo 2007). Prior research has found that a robot's behavior affects the degree of anthropomorphism (Duffy 2003; Złotowski et al. 2015). More anthropomorphic machine advisors were observed to exhibit higher degrees of trust resilience (De Visser et al. 2016). Hence, decision aids which make mistakes more similar to human ones may also be anthropomorphized more, and in turn be trusted more.

Perceptions vs Behavior. Next we briefly comment on the relationship between people's perceptions about the decision aids (covered in Section 3.2) and their advice taking behavior (discussed in Sections 3.3 and 3.4).

People perceived human-like decision aids as more *useful* than anti human-like decision aids. To explore if people's perceptions of usefulness correspond to the observed usefulness of the decision aids in practice, it is necessary to define what constitutes usefulness in this setting. One possible definition could be that machine advice is useful if it leads to an increase in the accuracy of people's decisions.[6] In our experiments, human-like decision aids — which were perceived as more useful — were more successful in reaching their potential for improving accuracy. However, the overall increase in accuracy was slightly larger for anti human-like decision aids, and hence perceptions about the decision aids' usefulness arguably did not coincide with the observed impact of the advice on the accuracy of people's decisions.

Human-like decision aids were also perceived as more *accurate* than anti human-like decision aids. Since all of the compared decision aids were equally accurate, people's perceptions did not coincide with the factual reality.

Finally, in terms of the perceived *predictability* of machine advice, human-like decision aids again received higher ratings. Given the higher overlap between people's decisions and those of human-like decision aids, it seems plausible that people were able to predict human-like machine advice better. In future research, it would be interesting to test whether people are actually better in predicting such advice.

Design Implications. With the increasing popularity of ML algorithms that aim to complement human skills, it is important to understand the effects of human-machine complementarity on machine-assisted decision-making. Our findings about the effects of human-machine similarity on people's perceptions and utilization of machine advice have important implications for the design of decision aids, particularly in settings where human agents are the final decision makers, while algorithmic decision aids serve as advisors.

Depending on the normative goals of utilizing machine assistance, it may be beneficial to use decision aids with different degrees of human-machine similarity. To ensure that a decision aid has a high influence for specific pieces of advice (e.g., a set of especially important or sensitive decisions), one may opt for human-like decision aids. The same holds if the normative goal is to ensure that people perceive the decision aid more favorably in terms of its usefulness, accuracy and predictability. On the other hand, if the goal is to maximize the overall influence of machine advice, a decision aid complementary to humans might be a better choice.

To tailor the degree of human-machine similarity to the normative goals of interest it is necessary to have access to models of human decision-making. Such models can be trained using data about people's past decisions in the relevant decision context. With access to models of human decision-making, developers can control the similarity between people's and algorithmic mistakes in various ways. One simple approach would entail leveraging predictive multiplicity (Marx, Calmon, and Ustun 2020). Namely, when multiple competing decision aids exhibit similar degrees of accuracy, one could select the decision aid based on the degree of human-machine similarity. Alternatively, when only one decision aid is available, one could manipulate the perceived similarity of human and algorithmic errors by selectively choosing when the algorithm provides advice (e.g., by avoiding to give advice for inputs where most people are predicted to (dis)agree with the machine advice). Finally, future research may enable the development of algorithms which in addition to optimizing for predictive accuracy can also optimize for the degree of similarity of human and machine decisions. This line of research fits well within existing efforts on developing machine learning algorithms which account for the presence of human agents in their learning procedure (De et al. 2020; Madras, Pitassi, and Zemel 2018; Meresht et al. 2020).

Conclusion. Our work contributes to the growing body of research on machine-assisted decision-making by studying a specific factor that may influence people's advice taking behavior: the degree of similarity between the decision aid's errors and typical human errors. In a series of large-scale online experiments, we experimentally show that human perceptions and utilization of algorithmic advice are in fact influenced by the similarity of human and machine errors. We invite future interdisciplinary research in social psychology and computer science, that will both provide deeper insights about the psychological mechanisms underlying our findings, and promote the development of algorithmic decision aids which may account for human-machine similarity.

[6] Other possible definitions of usefulness might aim to capture the ease and speed of decision-making, or other similar factors.

Ethics Statement

The design of our human-subject experiments has been approved by our institution's Ethical Review Board. The data was collected and stored in line with relevant ethical guidelines, which promote respondents' dignity, autonomy, and privacy (including but not limited to: paying respondents a fair wage, gathering respondents' informed consent, and avoiding storing personally identifiable data).

The purpose of our research is to identify factors which influence how people perceive and utilize machine decision aids, in order to support the design of decision aids that will effectively assist human decision-making. However, it is important to note that, as is the case for much research on human perceptions and behavior, these findings could also be utilized with malicious intent to steer people towards perceiving and utilizing decision aids in undesirable ways (e.g., over-relying on low quality decision aids, or under-relying on high quality decision aids).

Acknowledgements

This research was supported in part by a European Research Council (ERC) Advanced Grant for the project "Foundations for Fair Social Computing", funded under the European Union's Horizon 2020 Framework Programme (grant agreement no. 789373).

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed: 2022-01-15.

Bansal, G.; Nushi, B.; Kamar, E.; Lasecki, W. S.; Weld, D. S.; and Horvitz, E. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, 2–11.

Barocas, S.; and Selbst, A. D. 2016. Big Data's Disparate Impact. *California Law Review*.

Bonaccio, S.; and Dalal, R. S. 2006. Advice Taking and Decision-Making: An Integrative Literature review, and Implications for the Organizational Sciences. *Organizational Behavior and Human Decision Processes*.

Burton, J. W.; Stein, M.-K.; and Jensen, T. B. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2): 220–239.

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; and Elhadad, N. 2015. Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *KDD*.

Chan, I. C. C.; Lam, L. W.; Chow, C. W.; Fong, L. H. N.; and Law, R. 2017. The effect of online reviews on hotel booking intention: The role of reader-reviewer similarity. *International Journal of Hospitality Management*, 66: 54–65.

Chittilappilly, A. I.; Chen, L.; and Amer-Yahia, S. 2016. A survey of general-purpose crowdsourcing techniques. *IEEE Transactions on Knowledge and Data Engineering*, 28(9): 2246–2266.

Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*.

Corbett-Davies, S.; and Goel, S. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv preprint arXiv:1808.00023*.

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, 598–617. IEEE.

Davis, F. D. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319–340.

De, A.; Koley, P.; Ganguly, N.; and Gomez-Rodriguez, M. 2020. Regression under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2611–2620.

De Visser, E. J.; Monfort, S. S.; McKendrick, R.; Smith, M. A.; McKnight, P. E.; Krueger, F.; and Parasuraman, R. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3): 331.

Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1): 114.

Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3): 1155–1170.

Dressel, J.; and Farid, H. 2018. The Accuracy, Fairness, and Limits of Predicting Recidivism. *Science Advances*.

Duffy, B. R. 2003. Anthropomorphism and the social robot. *Robotics and autonomous systems*, 42(3-4): 177–190.

Dzindolet, M. T.; Pierce, L. G.; Beck, H. P.; and Dawe, L. A. 2002. The perceived utility of human and automated aids in a visual detection task. *Human Factors*, 44(1): 79–94.

Engel, C.; and Grgić-Hlača, N. 2021. Machine Advice with a Warning about Machine Limitations: Experimentally Testing the Solution Mandated by the Wisconsin Supreme Court. *Journal of Legal Analysis*, 13(1): 284–340.

Epley, N.; Waytz, A.; and Cacioppo, J. T. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4): 864.

Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; and Thrun, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639): 115–118.

Faraji-Rad, A.; Samuelsen, B. M.; and Warlop, L. 2015. On the persuasiveness of similar others: The role of mentalizing and the feeling of certainty. *Journal of Consumer Research*, 42(3): 458–471.

Fisman, R.; Iyengar, S. S.; Kamenica, E.; and Simonson, I. 2006. Gender differences in mate selection: Evidence from a speed dating experiment. *The Quarterly Journal of Economics*, 121(2): 673–697.

Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. 2016. On the (im)possibility of Fairness. *arXiv:1609.07236*.

Green, B.; and Chen, Y. 2019a. Disparate interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *FAT**.

Green, B.; and Chen, Y. 2019b. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–24.

Grgić-Hlača, N.; Engel, C.; and Gummadi, K. P. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–25.

Grgić-Hlača, N.; Zafar, M. B.; Gummadi, K. P.; and Weller, A. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *AAAI*.

Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of Opportunity in Supervised Learning. In *NeurIPS*.

Harris, C. 2011. You're hired! an examination of crowdsourcing incentive models in human resource tasks. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, 15–18. Hong Kong, China.

Hellevik, O. 2009. Linear versus logistic regression when the dependent variable is a dichotomy. *Quality & Quantity*, 43(1): 59–74.

Hoff, K. A.; and Bashir, M. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3): 407–434.

Horvitz, E.; and Paek, T. 2007. Complementary computing: policies for transferring callers from dialog systems to human receptionists. *User Modeling and User-Adapted Interaction*, 17(1-2): 159–182.

Kahneman, D. 2011. *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D. 2021. A Conversation on Human and Machine Intelligence. https://nips.cc/virtual/2021/invited-talk/22284. Accessed: 2022-04-23.

Kamar, E.; Hacker, S.; and Horvitz, E. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *AAMAS*, volume 12, 467–474.

Karahanna, E.; and Straub, D. W. 1999. The psychological origins of perceived usefulness and ease-of-use. *Information & management*, 35(4): 237–250.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *ITCS*.

Lakkaraju, H.; Bach, S. H.; and Leskovec, J. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1675–1684. ACM.

Lakkaraju, H.; and Rudin, C. 2017. Learning cost-effective and interpretable treatment regimes. In *Artificial Intelligence and Statistics*, 166–175.

Lee, M. K. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1): 2053951718756684.

Lipton, Z. C. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Logg, J. M. 2017. Theory of machine: When do people rely on algorithms? *Harvard Business School working paper series# 17-086*.

Logg, J. M.; Minson, J. A.; and Moore, D. A. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151: 90–103.

Madhavan, P.; and Wiegmann, D. A. 2007. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8(4): 277–301.

Madras, D.; Pitassi, T.; and Zemel, R. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 6150–6160.

Mahmud, H.; Islam, A. N.; Ahmed, S. I.; and Smolander, K. 2022. What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175: 121390.

Marx, C.; Calmon, F.; and Ustun, B. 2020. Predictive multiplicity in classification. In *International Conference on Machine Learning*, 6765–6774. PMLR.

Meade, A. W.; and Craig, S. B. 2012. Identifying careless responses in survey data. *Psychological methods*, 17(3): 437.

Meresht, V. B.; De, A.; Singla, A.; and Gomez-Rodriguez, M. 2020. Learning to switch between machines and humans. *arXiv preprint arXiv:2002.04258*.

Palan, S.; and Schitter, C. 2018. Prolific.ac – A Subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance*.

Pew Research Center. 2016. 2016 Party Identification Detailed Tables.

Poursabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; Vaughan, J. W.; and Wallach, H. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*.

Prahl, A.; and Van Swol, L. 2017. Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6): 691–702.

Redmiles, E. M.; Acar, Y.; Fahl, S.; and Mazurek, M. L. 2017. A Summary of Survey Methodology Best Practices for Security and Privacy Researchers. Technical report.

Rempel, J. K.; Holmes, J. G.; and Zanna, M. P. 1985. Trust in close relationships. *Journal of personality and social psychology*, 49(1): 95.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144. ACM.

Rossi, A.; Dautenhahn, K.; Koay, K. L.; and Walters, M. L. 2017. How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario. In *International Conference on Social Robotics*, 42–52. Springer.

Rothe, R.; Timofte, R.; and Gool, L. V. 2015. DEX: Deep EXpectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*.

Rothe, R.; Timofte, R.; and Gool, L. V. 2018. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4): 144–157.

Salem, M.; Lakatos, G.; Amirabdollahian, F.; and Dautenhahn, K. 2015. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 1–8. IEEE.

Suls, J.; Martin, R.; and Wheeler, L. 2000. Three kinds of opinion comparison: The triadic model. *Personality and Social Psychology Review*, 4(3): 219–237.

Tan, S.; Adebayo, J.; Inkpen, K.; and Kamar, E. 2018. Investigating Human+ Machine Complementarity for Recidivism Predictions. *arXiv preprint arXiv:1808.09123*.

U.S. Census Bureau. 2016. American Community Survey 5-Year Estimates.

Vodrahalli, K.; Gerstenberg, T.; and Zou, J. 2021. Do Humans Trust Advice More if it Comes from AI? An Analysis of Human-AI Interactions. *arXiv preprint arXiv:2107.07015*.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2): 2018.

Wang, X.; and Yin, M. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, 318–328.

Weun, S.; Beatty, S. E.; and Jones, M. A. 2004. The impact of service failure severity on service recovery evaluations andpost-recovery relationships. *Journal of services marketing*.

Wilder, B.; Horvitz, E.; and Kamar, E. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582*.

Yaniv, I.; Choshen-Hillel, S.; and Milyavsky, M. 2011. Receiving advice on matters of taste: Similarity, majority influence, and taste discrimination. *Organizational Behavior and Human Decision Processes*, 115(1): 111–120.

Yin, M.; Wortman Vaughan, J.; and Wallach, H. 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.

Yu, K.; Berkovsky, S.; Conway, D.; Taib, R.; Zhou, J.; and Chen, F. 2016. Trust and reliance based on system accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, 223–227.

Yu, K.; Berkovsky, S.; Taib, R.; Conway, D.; Zhou, J.; and Chen, F. 2017. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 307–317.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2017a. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *WWW*.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2017b. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. *arXiv preprint arXiv:2001.02114*.

Złotowski, J.; Proudfoot, D.; Yogeeswaran, K.; and Bartneck, C. 2015. Anthropomorphism: opportunities and challenges in human–robot interaction. *International journal of social robotics*, 7(3): 347–360.