Homophily and Incentive Effects in Use of Algorithms

Riccardo Fogliato (rfogliat@andrew.cmu.edu)
Department of Statistics and Data Science, Carnegie Mellon University

Sina Fazelpour (s.fazel-pour@northeastern.edu)
Department of Philosophy and Religion & Khoury College of Computer Sciences, Northeastern University

Shantanu Gupta (shantang@andrew.cmu.edu), Zachary Lipton (zlipton@cmu.edu)
Machine Learning Department, Carnegie Mellon University

David Danks (ddanks@ucsd.edu)
Halıcıoğlu Data Science Institute and Department of Philosophy, University of California, San Diego

Abstract

As algorithmic tools increasingly aid experts in making consequential decisions, the need to understand the precise factors that mediate their influence has grown commensurately. In this paper, we present a crowdsourcing vignette study designed to assess the impacts of two plausible factors on AI-informed decision-making. First, we examine homophily—*do people defer more to models that tend to agree with them?*—by manipulating the agreement during training between participants and the algorithmic tool. Second, we considered incentives—*how do people incorporate a (known) cost structure in the hybrid decision-making setting?*—by varying rewards associated with true positives vs. true negatives. Surprisingly, we found limited influence of either homophily and no evidence of incentive effects, despite participants performing similarly to previous studies. Higher levels of agreement between the participant and the AI tool yielded more confident predictions, but only when outcome feedback was absent. These results highlight the complexity of characterizing human-algorithm interactions, and suggest that findings from social psychology may require re-examination when humans interact with algorithms.

Keywords: human-AI interaction; social learning; homophily; decision support; machine learning

Introduction

In a variety of sensitive domains, predictive models and algorithms have been adopted to aid expert decision makers. This trend has motivated an extensive body of work aimed at characterizing the performance of these models with respect to their accuracy, fairness, and a variety of other desiderata. However, the majority of this work has focused on the models in isolation. More recently, researchers have recognized that the quality of AI-assisted decisions critically depends on human cognition: How do the relevant humans interpret, process, and integrate these predictions into their decision-making processes?

Prior work on human use of algorithmic recommendations has largely focused on the decision context (Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2018; Dietvorst, Simmons, & Massey, 2015; De-Arteaga, Fogliato, & Chouldechova, 2020), or features of decision subjects (Green & Chen, 2019b). In this paper, we instead focus on the "social" interactions between humans and algorithms: Do people regard algorithms similarly to other people?

Social psychological research (e.g., Hoppitt and Laland (2013); Turner (1991)) has revealed two factors that frequently shape social learning and interactions. First, people use a variety of heuristic strategies to estimate the reliability of an information source, and hence the value of the information it provides (Hoppitt & Laland, 2013; Kendal et al., 2018). One common heuristic is based on *homophily*: People use the *perceived similarity* of an information source to themselves as an indicator of the source's reliability. In social interactions, these perceptions of similarity are moderated by factors such as agreements in prior interactions (O'Connor & Weatherall, 2019; Zollman, 2015) as well as deeper identity-based considerations, e.g., shared disciplinary backgrounds (Phillips & Loyd, 2006) and sociocultural identities (Turner, 1991).

Second, people's decision *uncertainty* (Toelch & Dolan, 2015; Kendal et al., 2018) and *costs* (Hoppitt & Laland, 2013; Boyd & Richerson, 1988) are thought to moderate reliance on *social* (as opposed to individual) learning. Information from social sources can be particularly critical when navigating novel, ambiguous, risky, or high-consequence environments. More generally, there is evidence that incentive structures may shape people's predictive and decision-making behavior (Evans, Over, et al., 2004).

While people's use of information from other humans is shaped by homophily and perceived costs, it is unknown whether people conceptualize algorithms in ways that would produce similar influences (though see Lu and Yin (2021)). Both of these factors can be characterized in information-theoretic terms, so they plausibly might extend to algorithms as well. We thus conducted an experiment that explicitly manipulated these factors in the human-AI interaction context, including measures of decisions and perceptions.

Related Work

Our experiment engages with several bodies of prior research. First, many researchers have aimed to characterize the various sources and types of biases that arise throughout the machine learning pipeline (Fazelpour & Danks, 2021; Mitchell, Potash, Barocas, D'Amour, & Lum, 2021), and in the uptake of algorithmic information by users in particular (Logg, Minson, & Moore, 2019; Dietvorst et al., 2015; Mosier et al., 1998). Previous research has identified factors that can result in biases of overreliance (automation bias; e.g., Skitka, Mosier, and Burdick (2000)) or instead underreliance (algorithm aversion; e.g., Dietvorst et al. (2015)) on algorithmic

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

recommendations. For example, the perceived difficulty of understanding the algorithm (Yeomans, Shah, Mullainathan, & Kleinberg, 2019) or use of sensitive information about decision subjects, such as race (Green & Chen, 2019a) or socioeconomic status (Skeem, Scurich, & Monahan, 2020), can influence use of algorithmic predictions.

Second, we connect with research on the proper design of algorithmic tools. For example, Duan, Ho, and Yin (2020) are inspired by work that highlights the benefits of team diversity to find ways to counteract biases in crowdwork. The advantages of complementarity in teams has also motivated work on the design of decision support tools that best complement human capabilities (Wilder, Horvitz, & Kamar, 2020; Kamar, 2016; Bansal, Nushi, Kamar, Horvitz, & Weld, 2021). Similarly, research on determinants of trust in organizational settings play an increasingly key role in understanding factors that shape human trust in algorithmic tools (Glikson & Woolley, 2020). Prior work by Lu and Yin (2021) is the closest in approach to our study. Lu and Yin (2021) also examine and find evidence for the use of *agreement* as a proxy for algorithmic reliability by human users, though they focus on human-AI disagreement on cases where humans are confident in their predictions. In addition, they provide only binary algorithmic outputs. We instead focus on low-confidence cases and provide the model's likelihood estimates, which help human trust calibration (Zhang, Liao, & Bellamy, 2020).

Third, our study belongs to a line of work that aims to extrapolate general human behavioral patterns in the presence of algorithmic tools via crowdsourcing experiments (Fogliato, Chouldechova, & Lipton, 2021). Much of this work has focused on understanding how human trust and reliance can vary with the underlying properties of these tools, and also with the type of algorithmic recommendations that are communicated (Yin, Wortman Vaughan, & Wallach, 2019; Zhang et al., 2020), such as the presence of explanations (Bansal, Wu, et al., 2021; Dodge, Liao, Zhang, Bellamy, & Dugan, 2019; Poursabzi-Sangdeh, Goldstein, Hofman, Wortman Vaughan, & Wallach, 2021; Lai & Tan, 2019; Wang & Yin, 2021). Other common themes include the impact of the tools on the predictive and fairness properties of human predictions (Green & Chen, 2019b, 2019a). Our two-step elicitation process is inspired by the findings of Buçinca, Malaya, and Gajos (2021), who concluded that eliciting predictions from participants before revealing the model's recommendation decreased their overreliance on the tool. Similarly, Green and Chen (2019b) noted that participants achieved higher predictive performance when they were asked to pre-register their predictions made without the model. We draw on these lessons in the design of our experiment.

Experiment

The experiment was designed to test three key hypotheses:

- **[H1]** *Influence of agreement on trust and reliance.* Higher agreement between algorithmic recommendations and the participant, particularly on cases with high predictive uncertainty, will increase the participant's subsequent trust in, and reliance on, the model.
- **[H2]** *Influence of incentives.* The incentive structure will have influences on both: (a) participants' predictions, which will be skewed towards outcomes with higher monetary incentives; and (b) participant's reliance on the algorithm, which will increase as overall costs of error increase.
- **[H3]** *Influence of feedback on agreement-driven reliance.* Receiving outcome feedback will reduce the impact of agreement as a reliability approximation heuristic, and so will reduce participant's reliance on the model.

We make a further prediction that is not central to the study design (so is investigated through exploratory analyses): Higher level of agreement may lead participants to be more confident about their predictions, especially when outcome feedback is absent. We tested these hypotheses through a vignette study in which participants interacted with recommendations generated by two different algorithmic tools.

Method

Algorithm and vignette construction The vignettes for the experiment all involved descriptions of criminal defendants. Participants and algorithms aimed to predict their future re-arrest outcomes. The vignettes were populated using a dataset of defendants sentenced in Pennsylvania's federal criminal courts between 2004 and 2006 ($N = 117,464$), including whether they were rearrested in the three years following release from prison or imposition of community supervision (Fogliato, Chouldechova, & Lipton, 2021).[1] We focused on a subset of $N = 3,523$ defendants for possible inclusion in a vignette, stratified for race, sex, age, and re-arrest (and assuming reasonable numbers in each group).

In experimentally manipulating the levels of human-AI agreement, we needed to ensure that disagreements are not perceived as an indicator of inaccuracy (e.g., if an algorithm disagrees with a user over trivial cases). We were thus particularly interested in using cases of human-AI agreement (disagreement) that typically produced high (low) confidence judgments from humans. We identified the cases in the dataset of Fogliato, Chouldechova, and Lipton (2021) where more than 80% (less than 60%) of the participants made the same prediction (*high (low) confidence*).

Moreover, to ensure that we used realistic algorithmic predictions, we developed two different predictive models that had comparable overall performance while occasionally disagreeing on a subset of cases (that could be used to manipulate agreement). We used stratified 70/30 train/test sets of the

[1] Criminal justice data in the U.S. are highly biased, and heavily affected by measurement issues (Bao et al., 2021; Fogliato, Xiang, Lipton, Nagin, & Chouldechova, 2021; Pierson et al., 2020; Goel, Rao, & Shroff, 2016). In particular, re-arrest and re-offense are only imperfectly correlated, and those biases are (unavoidably) reflected in the AI tools trained and tested on such data (Fogliato, G'Sell, & Chouldechova, 2020). In an effort to minimize these issues, we had both participants and algorithms predict only the directly observable (though highly biased) outcome of re-arrest.

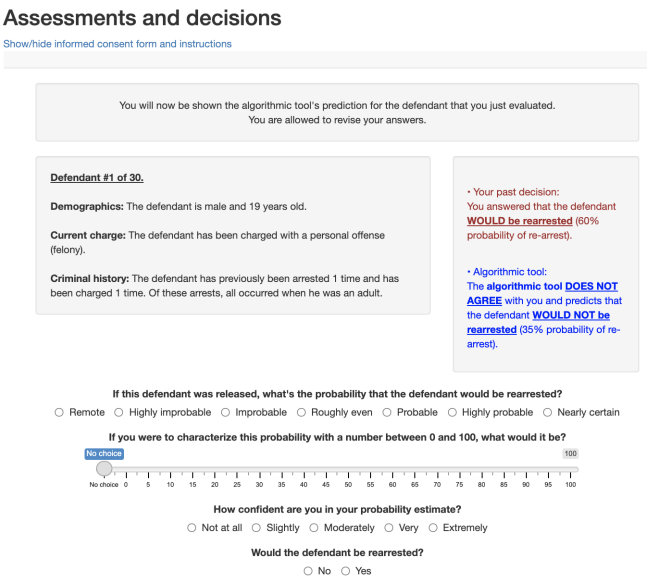

Figure 1: Sample vignette in the study. The panel on the right, which contained the AI's recommendation, became visible only once participants had made an initial prediction.

full dataset (minus our presentation subsample), and trained a model using XGBoost (Chen & Guestrin, 2016), and another using logistic Lasso (Tibshirani, 1996). Both models take defendant's demographics, the current charge type, and criminal history information as input. The probability values outputted by the models were converted into binary predictions using a threshold of 0.5. On the test set, the two models were well-calibrated and had virtually identical predictive performance (AUC of 0.71 and accuracy of 66%). This performance is comparable, if not superior, to the performance of many models deployed in real-world criminal justice settings (Desmarais, Zottola, Duhart Clarke, & Lowder, 2020).

Given these two algorithms, we selected the instances of model-agreement as cases where the models both predicted the likelihood of re-arrest to be either above 70% or below 30% (48 cases), and the instances of model-disagreement as those where their binary predictions differed (34 cases). This sample of "easy" (high confidence, model agreement) and "hard" (low confidence, model disagreement) cases was then used to construct the vignettes.

Each vignette (see Figure 1) contained a description of the defendant with exactly the same set of variables that were used for model training. Vignettes also possibly (see Experimental design below) included the output from one of the two algorithms—the estimated probability of re-arrest (e.g., 78%) and binary prediction (re-arrest vs. no re-arrest).

Given a vignette, participants were asked to estimate the likelihood that the defendant would be rearrested in the three years following release. Participants responded using a 7-point Likert scale ("Remote" – "Nearly certain"), and also a $[0, 100]$ slider of probabilities (in increments of 5%). They had to provide their confidence in the likelihood judgment (5-point Likert scale "Not at all" – "Extremely") and, finally, they were also asked to predict whether the defendant would be rearrested or not ("no" vs. "yes").

Experimental design and procedures Tests of our three hypotheses require independent manipulation of the level of agreement between the participant and algorithm (to test for homophily effects); incentive structure (to test for cost effects); and availability of feedback (to control for learning effects). In order to avoid cross-condition learning, we used a fully between-participants design with $18 = 3\ \{\text{Homophily}\} \times 3\ \{\text{Incentive}\} \times 2\ \{\text{Feedback}\}$ conditions.

After consenting, all participants were provided with instructions that described the task, showed an example case, and explained that the algorithm's recommendations had been generated by a model that was well-calibrated, including an explanation of what "calibration" meant.

The first phase consisted of 15 cases, randomly drawn from the previously-described subset, with 10 hard cases and 5 easy cases. The re-arrest rate for these cases was matched to the re-arrest rate in the dataset (around 40%). For each case, participants were first asked to answer the four rating and prediction questions. After answering, the algorithm's recommendation was shown and participants were allowed to revise their answers. Two attention checks were randomly inserted in this sequence. At the end of the first phase, participants were asked to provide their confidence in their predictive ability (5-point Likert scale from "Very low" – "Very high"), and whether they agreed that the model would help them make predictions (7-point Likert scale from "Strongly disagree" – "'Strongly agree").

The Homophily manipulation modified the share of first-phase cases in which the algorithm's binary predictions matched the participant's binary predictions. This manipulation focused on the hard cases, as the participant should be most uncertain and the models also generated different predictions. For those 10 cases, we ensured that the algorithm matched the participant's binary prediction on 9 (High homophily), 5 (Medium), or only 1 (Low) cases. Since these cases were ones where the models disagreed, we simply switched between models depending on which made a different binary prediction from the participant (see also Lu and Yin (2021, Experiment 3)).[2] This manipulation ensured that the model's accuracy, as inferred by the participant, was orthogonal to the level of agreement.

We also manipulated whether outcome feedback was available during the first phase. In particular, after receiving the algorithm's recommendation and potentially revising their predictions, participants in the Feedback condition were informed of whether the defendant was actually rearrested, while those in the No feedback condition were not told anything about the eventual outcome. No participant received any feedback in the second phase of the survey.

For the second phase of the experiment, we drew another

[2]This design enabled us to truthfully tell participants that all algorithm predictions came from a well-calibrated model.

random sample without replacement of 15 cases (with 40% re-arrest rate), this time with 8 hard and 7 easy cases. Only the predictions of the Lasso model were shown in this phase. The Incentive manipulation determined participant compensation based on their performance in this phase. We used three different structures, all of which provided no reward for an incorrect (binary) prediction:

- High true positive (High TP): \$0.36 for a true positive prediction (i.e., re-arrest) and \$0.18 for true negatives.
- Neutral: \$0.27 for each correct prediction.
- High true negative (high TN): \$0.36 for true negative predictions (i.e., no re-arrest) and \$0.18 for true positives.

Note that this incentive manipulation should only impact binary predictions, as those determine the payoffs. Participants were presented with a detailed description of the incentives, including comprehension questions before and after the second phase. The incentive structure was also listed in each vignette as a reminder.

At the end of the second phase, participants again rated confidence in their predictive ability, and whether they agreed that the model helped them make their predictions. A subset of participants[3] were also asked whether they thought that their predictions had been influenced by the incentives.

For data analysis purposes, we used participants' judgments and ratings, and also three additional measures adopted from prior work (Yin et al., 2019; Green & Chen, 2019b):

- *Agreement fraction*: the fraction of cases in which the participant's binary prediction matched the algorithm's.
- *Switch fraction*: out of the cases for which participant and algorithm initially disagreed, the fraction of cases where the participant changed binary prediction.
- *Influence*: the median (per participant) difference between the revised and initial numerical likelihood estimates made by the participant, divided by the difference between the model and the participant's initial likelihood estimate.

Our statistical analysis employs the average of each metric computed at the participant level.

Participants A sample of 862 participants was recruited on MTurk, all with HIT approval rating >90%, >500 completed HITs, and physically present in the US. Participants were paid variable amounts depending on performance; mean compensation was \$4.40 (sd=\$0.60), translating to average payment of slightly more than \$10 per hour. 369 participants failed the attention checks,[4] resulting in a final sample of $N = 493$.

[3] A bug in the survey meant that only a (random) subset of participants were asked this question.

[4] Because of the complexity of the task, we wanted to ensure that participants were actually paying attention. We thus used more severe attention checks (i.e., not simply "click here to continue"). As a result, we had higher-than-normal failure rates.

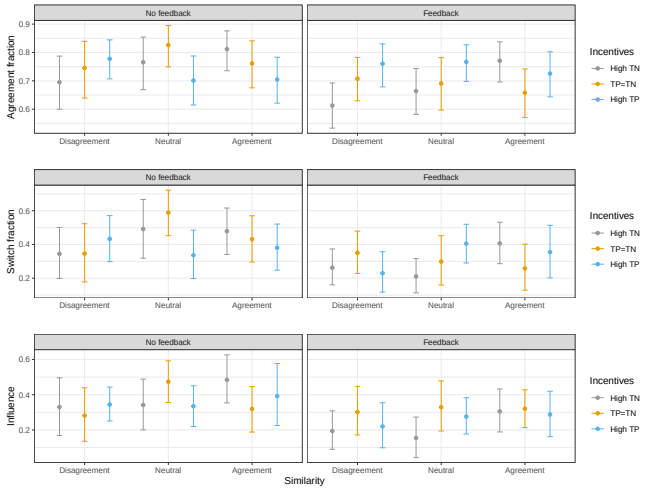

Figure 2: Second-phase agreement fraction (top), switch fraction (middle), and mean influence (bottom) for agreement (horizontal axis), incentives (color), and feedback (panels) manipulations (means and 90% confidence intervals computed via nonparametric bootstrap and percentile method).

Results

The two measures of likelihood judgments—Likert scale and probability slider—were highly correlated ($\rho = 0.77$, $p < 0.001$), so we analyze only the probability judgments. We first consider the targeting effectiveness of our homophily manipulation. In particular, we assess the impact of this manipulation on participants' judgements in the first phase of the survey. We should expect to observe higher agreement between the participants and the model's predictions on easy cases. Consistently, the average agreement fractions for participants' initial and revised binary predictions on these easy cases were both above 80%. 319 participants (64%) agreed with the algorithm on all 5 easy cases, and 53 (11%) agreed for 4-of-5 cases. Participants' predictions were more accurate for easy cases than difficult ones (classification accuracies were 73% vs. 45% respectively). Participants also reported being more certain about their predictions on the easy cases (3.9 (out of 5) vs. 3.5; p-value of paired t-test<0.01). Lastly, participants spent longer on difficult cases (mean = 45s) than on easy cases (mean = 40s) for initial predictions (paired t-test, $p < 0.01$).

The cleanest tests of our hypotheses are based on responses in the second phase of the experiment, as the initial manipulations should have had an impact, and there should not be further learning since participants receive no feedback. Figure 2 shows the agreement fraction, switch fraction, and influence for participants' predictions across treatments in the second phase. We fitted two separate linear regressions—one to predict agreement fraction and one to predict switch fraction—with `homophily` and `incentives` as independent factors (the latter dichotomized into symmetric vs. asymmetric incentives to reflect H2b), interacted with the two `feedback` interven-

tions. We use robust standard errors of the coefficients estimates for hypothesis testing.

For the agreement fraction regression, the coefficients relative to homophily and incentives were all virtually zero and not statistically significant. However, the regression results indicate that, controlling for incentives effects, the presence of outcome feedback drastically reduced the agreement fraction across all conditions, e.g., by 0.1 (0.76 vs. 0.66) in the low-homophily manipulation (Wald test of difference: $p < 0.04$). This pattern is also visible in Figure 2. In the switch fraction regression, we find that the medium-homophily manipulation led to higher reliance on the model compared to the low-homophily one, but only when outcome feedback was absent (increase of 0.09 with one-sided $p < 0.07$). While the presence of feedback substantially decreased reliance also according to this metric, the homophily interventions did not appear to impact reliance, e.g., the effect of medium homophily (vs. low) was even negative. We also find that the coefficients for the incentives effects were close to zero and not statistically significant. A parallel analysis for influence yields similar results: Neither higher levels of homophily nor asymmetric incentives increased reliance. However, these regression results again indicate that the presence of outcome feedback decreased reliance.

Small effects combined with the limited sample size of our study may explain the null results discussed in the previous paragraph. As a confirmatory analysis, we investigated whether any sizable effects could be detected on difficult cases, i.e., those where predictive uncertainty was high and participants were not confident about their predictions. Any effects should be most salient on these subsets of cases, so we performed the same regression analyses for the three metrics. The conclusions from these analyses are analogous to those we have discussed above: Reliance increased from low to medium homophily only when outcome feedback was absent (here effects were statistically significant across all three metrics), and providing feedback decreased reliance. No other effects were statistically significant.

Participants were also asked about perceived utility and trust of the AI model in two questionnaires, at the end of the first phase and of the experiment respectively. Results from these ratings somewhat mirror the findings on the objective measures of reliance. We could only detect one significant effect for the medium-homophily intervention which increased perceived utility of the model over the low-homophily one in the first questionnaire in absence of outcome feedback (one-sided $p < 0.04$). However, this effect (or any other) could not be detected in the questionnaire at the end of the survey.

Prior studies have found that differential payment for types of performance can influence people's behaviors (Evans et al., 2004). No such effects were found in our experiment. In particular, participants who received larger rewards for true negatives did *not* make more negative predictions. We regressed proportions of of positive predictions (i.e., of re-arrest) made by each participant before seeing the algorithmic recommendations on incentives manipulations. These shares were virtually identical across the interventions (all between 57% and 59%) and the differences based on Wald tests were not statistically significant. Additional analyses focused solely on low-confidence and high-uncertainty cases had similar results. In the final questions, 90% of participants reported the correct rewards for the two types of successes. Moreover, the conclusions of this analysis do not change even if we exclude the 10% of participants who misreported the rewards structure. 60% of the participants that encountered asymmetric rewards reported that their predictions were either slightly or not impacted at all by the incentives, and another 20% reported that impact had been moderate.

Finally, we analyze participants' confidence in the predictions using the regression analysis described earlier. In the first questionnaire, the medium-homophily intervention positively impacted participants' confidence ratings compared to the low-homophily one when feedback was absent (effect is 0.38 on a 1–5 Likert scale rating; Wald test $p < 0.01$). The high-homophily manipulation also increased confidence compared to the medium-homophily one, but the effect was smaller (0.17, one-sided $p < 0.08$). In presence of outcome feedback, no effect was detected. Similar results were found for the same question inserted in the final questionnaire, although the effects relative to the homophily interventions were slightly smaller. A regression analysis of the confidence ratings reported by participants for each of the predictions (here the average rating by participant) in the second phase delivered similar results: Higher levels of homophily increased confidence in absence of feedback (both one-sided $p < 0.07$), but this was the only effect that could be detected.

Remarks

Almost none of our initial hypotheses were confirmed by this experiment. H1 and H3 imply that participants who experience greater homophily (H1) and no feedback (H3) should have the highest levels of reliance. At the same time, participants should rely more on the recommendations when the cost of prediction errors is asymmetric (H2b). Although some of the results suggest the presence of these effects, there is no clear impact of the sort predicted by the hypotheses. The non-monotonicity of the impact of homophily is particularly surprising, as in multiple instances the largest effects were found for medium levels. One possibility is that the high and low levels led participants to largely disregard the AI recommendations, though for different reasons: The former found the AI to be redundant, while the latter found the AI to be error-prone.

General Discussion

Our results have provided little or no support for the key hypotheses driving our research study: Although homophily and incentives impact human reliance on other people, they do not (in this setting) seem to strongly influence human reliance on AI tools. Participants who were shown AI recommendations that matched their predictions more often did not

appear to rely or trust the tool substantially more than their counterparts (H1). Agreement between the AI tool and participants did translate into higher subjective confidence for the participants, but not higher usage of the AI information. Incentives for different types of predictions did not affect participants' judgments or their reliance on the recommendations (H2). However, the presence of outcome feedback did decrease participants' reliance on the tool (H3).

Homophilic effects are prevalent in social life, influencing individual associations (Golub & Jackson, 2012) and trust relations (Tang, Gao, Hu, & Liu, 2013) in social networks. The absence here of effects of homophily—operationalized as similar decisions over prior cases—could be explained in at least two interrelated ways. One possible explanation is that some factors that impact inter-human trust relations simply do not transfer to the human-AI case (Glikson & Woolley, 2020). From this perspective, homophily is relevant to understanding how agents are influenced by other humans, but not relevant to the algorithmic decision support case. Mahmoodi, Bahrami, and Mehring (2018), for example, have highlighted the significance of *reciprocity* for understanding social informational influences, finding agents to be more open to influence from partners who are expected to *reciprocate* this influence later on. Importantly, Mahmoodi et al. (2018) also find that this dynamic process of reciprocity was "abolished when people believed that they interacted with a computer," potentially because people expected that algorithms would not (or cannot) reciprocate.

A second, not exclusive, explanation derives from the many dimensions of "similarity" among individuals, including demographics such as race and gender, underlying values, political affiliations, shared attitudes and beliefs, and more. Most of these factors can potentially drive the emergence of homophilic effects (Monge et al., 2003), but not all of them will be relevant in a specific context (Ahmad, Ahmed, Srivastava, & Poole, 2011). This second, narrower explanation allows for the possibility that people could experience homophilic effects with an algorithm, but only if there were appropriate perceived similarities with the algorithm. For example, this explanation would leave it open that people could experience consistent homophilic effects if they knew that the algorithm had been developed by someone who shared their values. Both of these potential explanations provide avenues for future research. In either case, though, our findings provide reasons to be cautious when transporting findings about interpersonal relations from psychological and organizations sciences to the case of human-AI interaction.

One might also worry about the pool of participants. Researchers have long cautioned against the potentially low quality of data that are collected through crowdsourcing experiments such as ours, particularly those on MTurk (Paolacci, Chandler, & Ipeirotis, 2010; Kennedy et al., 2020). The platform itself incentivizes requesters (i.e., people conducting experiments) to offer low payments and workers (i.e., participants) to exert minimal effort. Thus, workers often adopt a variety of strategies to maximize profit (McInnis, Cosley, Nam, & Leshed, 2016; Chandler, Mueller, & Paolacci, 2014), such as doing multiple HITs simultaneously. This worker strategy would not necessarily be an issue if we were conducting, for example, a quick five-question survey. The present experiment is complicated, however, and requires people to draw relatively fine distinctions. We attempted to minimize these risks by requiring a higher worker approval rating than for many other studies that are similar to ours (Green & Chen, 2019b, 2020). We also employed rigorous attention checks (and had correspondingly higher failures of those checks). We thus expect that we likely filtered out a large share of low-quality responses, and so that possibility is less likely to explain our null results.

Another alternative hypothesis is that our study participants could have performed numerous tasks similar to ours in the past, or have strong prior expectations about the possibility that an AI could be helpful for these kinds of decisions. Their (already mature) beliefs about AI tools may have not been influenced by the short interaction that they had with our tool. However, the positive results around the impact of homophily on *confidence* do not appear to be compatible with these two possibilities. Similarly, the lack of an effect of incentives on predictions could be due to the fact that the offered rewards were too small to nudge participants to change their predictions. Indeed, previous studies have reported that payments do not significantly increase the quality of the data that are collected (Buhrmester, Kwang, & Gosling, 2016). In our experiment, despite information about the base rate and feedback (for some participants), participants may not have realized that their rewards would have been likely higher had they given the same answers in all assessments. Interestingly, however, we could not detect any effect of this manipulation, even on the the cases on which participants were less confident about their predictions.

Future research should carefully take into account and address the key limitations of our study. In particular, our experimental results have shown that small monetary bonuses tied to accuracy do not promote changes in crowdworkers' behavior. We actually found the same result in a pilot study on predictions of loan repayment, and so increased the incentives in this experiment to see whether any effect would be revealed. If our null finding were replicated in other experimental setups, then we would have further evidence that incentives may not represent valid proxies for context-dependent costs in real-world decision-making. Consistent with the findings of Lu and Yin (2021), our experiment has also highlighted that crowdworkers' trust and reliance on AI tools may be insensitive to interventions on their level of agreement with the recommendations generated. Alternative study designs may achieve more promising results, for instance by increasing the duration of the interaction of AI and participant while keeping them fully engaged in the task. Lastly, the connection between confidence and homophily uncovered in our experiment represents another interesting research direction.

References

Ahmad, M. A., Ahmed, I., Srivastava, J., & Poole, M. S. (2011). Trust me, i'm an expert: Trust, homophily and expertise in mmos. In *2011 ieee third international conference on privacy, security, risk and trust and 2011 ieee third international conference on social computing* (pp. 882–887).

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., & Weld, D. S. (2021). Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 11405–11414).

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., ... Weld, D. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 chi conference on human factors in computing systems* (pp. 1–16).

Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., ... Venkatasubramanian, S. (2021). It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498*.

Boyd, R., & Richerson, P. J. (1988). *Culture and the evolutionary process*. University of Chicago press.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). Amazon's mechanical turk: A new source of inexpensive, yet high-quality data?

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods*, *46*(1), 112–130.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Chi conference on human factors in computing systems.*

Desmarais, S. L., Zottola, S. A., Duhart Clarke, S. E., & Lowder, E. M. (2020). Predictive validity of pretrial risk assessments: A systematic review of the literature. *Criminal Justice and Behavior*, 0093854820932959.

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

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K., & Dugan, C. (2019). Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 275–285).

Duan, X., Ho, C.-J., & Yin, M. (2020). Does exposure to diverse perspectives mitigate biases in crowdwork? an explorative study. In *Proceedings of the aaai conference on human computation and crowdsourcing* (Vol. 8, pp. 155–158).

Evans, J. S. B., Over, D. E., et al. (2004). *If: Supposition, pragmatics, and dual processes*. Oxford University Press, USA.

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, e12760.

Fogliato, R., Chouldechova, A., & Lipton, Z. (2021). The impact of algorithmic risk assessments on human predictions and its analysis via crowdsourcing studies. *arXiv preprint arXiv:2109.01443*.

Fogliato, R., G'Sell, M., & Chouldechova, A. (2020). Fairness evaluation in presence of biased noisy labels. *arXiv preprint arXiv:2003.13808*.

Fogliato, R., Xiang, A., Lipton, Z., Nagin, D., & Chouldechova, A. (2021). On the validity of arrest as a proxy for offense: Race and the likelihood of arrest for violent crimes. *arXiv preprint arXiv:2105.04953*.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.

Goel, S., Rao, J. M., & Shroff, R. (2016). Precinct or prejudice? understanding racial disparities in new york city's stop-and-frisk policy. *The Annals of Applied Statistics*, *10*(1), 365–394.

Golub, B., & Jackson, M. O. (2012). How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics*, *127*(3), 1287–1338.

Green, B., & Chen, Y. (2019a). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 90–99).

Green, B., & 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.

Green, B., & Chen, Y. (2020). Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *arXiv preprint arXiv:2012.05370*.

Hoppitt, W., & Laland, K. N. (2013). *Social learning: an introduction to mechanisms, methods, and models*. Princeton University Press.

Kamar, E. (2016). Directions in hybrid intelligence: Complementing ai systems with human intelligence. In *Ijcai* (pp. 4070–4073).

Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social learning strategies: Bridge-building between fields. *Trends in cognitive sciences*, *22*(7), 651–665.

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. (2020). The shape of and solutions

to the mturk quality crisis. *Political Science Research and Methods*, *8*(4), 614–629.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, *133*(1), 237–293.

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 29–38).

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

Lu, Z., & Yin, M. (2021). Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 chi conference on human factors in computing systems* (pp. 1–16).

Mahmoodi, A., Bahrami, B., & Mehring, C. (2018). Reciprocity of social influence. *Nature communications*, *9*(1), 1–9.

McInnis, B., Cosley, D., Nam, C., & Leshed, G. (2016). Taking a hit: Designing around rejection, mistrust, risk, and workers' experiences in amazon mechanical turk. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 2271–2282).

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*, 141–163.

Monge, P. R., Contractor, N. S., Contractor, P. S., Peter, R., Noshir, S., et al. (2003). *Theories of communication networks*. Oxford University Press, USA.

Mosier, K. L., Dunbar, M., McDonnell, L., Skitka, L. J., Burdick, M., & Rosenblatt, B. (1998). Automation bias and errors: Are teams better than individuals? In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 42, pp. 201–205).

O'Connor, C., & Weatherall, J. O. (2019). *The misinformation age*. Yale University Press.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, *5*(5), 411–419.

Phillips, K. W., & Loyd, D. L. (2006). When surface and deep-level diversity collide: The effects on dissenting group members. *Organizational behavior and human decision processes*, *99*(2), 143–160.

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., ... others (2020). A large-scale analysis of racial disparities in police stops across the united states. *Nature human behaviour*, *4*(7), 736–745.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 chi conference on human factors in computing systems* (pp. 1–52).

Skeem, J., Scurich, N., & Monahan, J. (2020). Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Law and human behavior*, *44*(1), 51.

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, *52*(4), 701–717.

Tang, J., Gao, H., Hu, X., & Liu, H. (2013). Exploiting homophily effect for trust prediction. In *Proceedings of the sixth acm international conference on web search and data mining* (pp. 53–62).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Toelch, U., & Dolan, R. J. (2015). Informational and normative influences in conformity from a neurocomputational perspective. *Trends in cognitive sciences*, *19*(10), 579–589.

Turner, J. C. (1991). *Social influence.* Thomson Brooks/Cole Publishing Co.

Wang, X., & 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* (pp. 318–328).

Wilder, B., Horvitz, E., & Kamar, E. (2020). Learning to complement humans. *arXiv preprint arXiv:2005.00582*.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414.

Yin, M., Wortman Vaughan, J., & 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* (p. 279).

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 295–305).

Zollman, K. J. (2015). Modeling the social consequences of testimonial norms. *Philosophical Studies*, *172*(9), 2371–2383.