Learning When to Advise Human Decision Makers*

Gali Noti[1,2] , **Yiling Chen**[1]
[1]Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University
[2]The School of Computer Science and Engineering, The Hebrew University of Jerusalem
{galinoti, yiling}@seas.harvard.edu

Abstract

Artificial intelligence (AI) systems are increasingly used for providing advice to facilitate human decision making in a wide range of domains, such as healthcare, criminal justice, and finance. Motivated by limitations of the current practice where algorithmic advice is provided to human users as a constant element in the decision-making pipeline, in this paper we raise the question of *when should algorithms provide advice?* We propose a novel design of AI systems in which the algorithm interacts with the human user in a two-sided manner and aims to provide advice only when it is likely to be beneficial for the user in making their decision. The results of a large-scale experiment show that our advising approach manages to provide advice at times of need and to significantly improve human decision making compared to fixed, non-interactive, advising approaches. This approach has additional advantages in facilitating human learning, preserving complementary strengths of human decision makers, and leading to more positive responsiveness to the advice.

1 Introduction

Artificial intelligence (AI) is increasingly used to support human decision making in high-stake settings in which the human operator, rather than the AI algorithm, needs to make the final decision. For example, in the criminal justice system, algorithmic risk assessments are being used to assist judges in making pretrial-release decisions and at sentencing and parole [NJ-Courts, 2020; PJI, 2019; Northpointe, 2019; Cohen *et al.*, 2018]; in healthcare, AI algorithms are being used to assist physicians to assess patients' risk factors and to target health inspections and treatments [Musen *et al.*, 2021; Garcia-Vidal *et al.*, 2019; Tomašev *et al.*, 2019; Kononenko, 2001]; and in human services, AI algorithms are being used to predict which children are at risk of abuse or neglect, in order to assist decisions made by child-protection staff [Vaithianathan *et al.*, 2017; Chouldechova *et al.*, 2018].

In such systems, decisions are often based on risk assessments, and statistical machine-learning algorithms' abilities to excel at prediction tasks [Meehl, 1954; Dawes *et al.*, 1989; Grove *et al.*, 2000; Obermeyer and Emanuel, 2016; Mullainathan and Spiess, 2017] are leveraged to provide predictions as advice to human decision makers [Kleinberg *et al.*, 2015]. For example, the decision that judges make on whether it is safe to release a defendant until his trial, is based on their assessment of how likely this defendant is, if released, to violate his release terms, i.e., to commit another crime until his trial or to fail to appear in court for his trial. For making such risk predictions, judges in the US are assisted by a "risk score" predicted for the defendant by a machine-learning algorithm [NJ-Courts, 2020; PJI, 2019].

Research on such AI-assisted decision making has mostly addressed two questions. The first is what advice should AI systems provide? The line of research that addresses this question places emphasis on the machine-learning algorithms and focuses on optimizing and evaluating their success in comparison to human predictions, based on statistical metrics such as prediction accuracy and fairness [Kleinberg *et al.*, 2017; Angwin *et al.*, 2016; Haenssle *et al.*, 2018; Chouldechova, 2017]. The implicit expectation is that better algorithmic advice will lead to better human decisions.

The second question is how to present algorithmic advice to human decision makers? This question has been addressed in a recent line of work that emphasizes the role of the human as the one who eventually makes the actual decision. Instead of evaluating the algorithmic performance in isolation, these works concentrate on studying the effect of the algorithmic input on the decisions that humans make [Green and Chen, 2019a; Green and Chen, 2019b; Albright, 2019; Tschandl *et al.*, 2020; Lai and Tan, 2019; Zhang *et al.*, 2020; Bansal *et al.*, 2019; Yin *et al.*, 2019] and hence term the perspective "AI-in-the-loop human decision making"[Green and Chen, 2019a]. These studies typically show—both with human experts such as judges or clinicians and with non-experts in experimental settings—that providing the algorithmic assessment indeed significantly improves human decision makers' prediction performance, and that different ways of providing the algorithmic input to human decision makers, as well as different algorithmic accuracy or error patterns, can have a significant impact on their decisions.

*The full version of this paper appears on ArXiv, at: https://arxiv.org/abs/2209.13578. The data are available on the authors' websites.

Situated in the framework of AI-in-the-loop human decision making, this work aims to answer a different important question: *when should algorithms provide advice?* The current practice in applications and in prior studies, is that algorithms provide advice to the human decision maker in every prediction problem. We explore whether AI systems can be trained to automatically identify the cases where advice is most useful, and those where the human decision maker is better off deciding without any algorithmic input, and whether such an approach that provides the algorithmic advice only when it is needed indeed manages to assist humans in improving their decisions.

1.1 Background and Related Work

Our approach is motivated by several observations from prior work and current practice. First, in prior studies on AI-assisted human decision making, the AI component is completely oblivious of the human decision maker: the human always receives advice from an algorithm, but, importantly, the algorithm is not aware of its human counterpart and whether its advice may actually be helpful to him. This is despite the fact that human decision makers have their own strengths and sometimes reach better decisions on their own, without the algorithmic input, and the computational methods have their own limitations and can have errors and biases (as was studied in recent literature on human-AI complementary performance [Groh *et al.*, 2022; Wilder *et al.*, 2020; Madras *et al.*, 2018; Kamar *et al.*, 2012; Bansal *et al.*, 2021; Steyvers *et al.*, 2022]), and so algorithmic advice may not always be helpful.

E.g., in a recent experiment that studied human prediction in the pretrial-release decision setting [Green and Chen, 2019b], the most accurate human predictions were achieved in an "Update" treatment, in which the human decision makers first made a risk prediction on their own and only then observed the algorithmic prediction and were allowed to update their prediction if they wished. However, in this dataset we found that in 66% of the predictions, the human's initial prediction (before observing the algorithmic input) was already equal to or more accurate than the algorithm's prediction. Moreover, in 36% of the predictions, humans' initial prediction was strictly more accurate than the algorithm's, and after showing them the algorithmic prediction their prediction performance deteriorated 32% of these times.

An additional important point that arises when a human decision maker is assisted by an (inevitably) imperfect AI system, is that the human is de-facto expected to monitor the algorithm, i.e., to identify when the algorithm is wrong so as to override its prediction [Green, 2022]. However, there is a large body of empirical evidence showing that such monitoring is a challenging task for humans: recent studies demonstrate that people do poorly in judging the quality of algorithmic predictions and determining when to override those predictions, and that these judgments are often incorrect and biased [Green and Chen, 2019a; Green and Chen, 2019b; Grgić-Hlača *et al.*, 2019; Tschandl *et al.*, 2020; Yeomans *et al.*, 2019; Bansal *et al.*, 2021; Van Swol and Sniezek, 2005]. This suggests to consider alternative designs of decision pipelines in which the monitoring task is transferred to the AI.

Moreover, even if the algorithm were perfect, it is not clear whether the constant advising approach used in prior work is the optimal way to interact with human decision makers and to inform them so as to improve their decisions. Specifically, it may be that providing the advice in every prediction will result in advice discounting or even disregard in the decision maker's judgment. Such behaviors have been demonstrated in other settings of users' interactions with technology [Kalsher and Williams, 2006; Anderson *et al.*, 2014], and are related to the study of habituation [Rankin *et al.*, 2009], but have not been studied in behavioral literature on advice utilization.

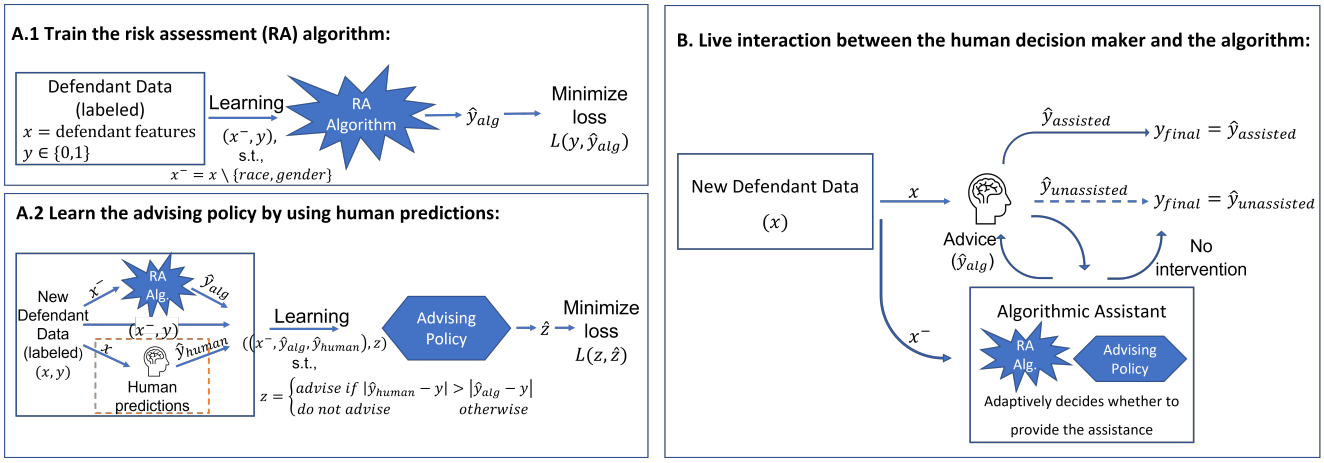
Finally, in the experiment of [Green and Chen, 2019b] mentioned above, it is intriguing to see that while humans made significantly better predictions when they were assisted by the algorithm compared to making predictions without any algorithmic assistance, their performance was still far worse than that of the algorithm alone. This is despite the fact that the human decision makers constantly received the algorithmic prediction, and, in principle, could just have adopted its predictions and reached the algorithmic performance. This observation that a human assisted by an algorithm is still inferior to the algorithm alone is in fact typical in AI-assisted human decision making settings (e.g., [Lai and Tan, 2019; Lai *et al.*, 2020; Jacobs *et al.*, 2021]) and suggests that there is room to improve and extract more value from the interaction between the human and the AI. For a discussion of further related work, see Appendix A in the full version of the paper [Noti and Chen, 2023].

1.2 Our Approach

We propose to replace the constant advising approach with a responsive advising system (an "algorithmic assistant") that interacts with the human decision maker and takes an active part in the decision making process, aiming to improve the human's decisions. Specifically, our algorithmic assistant applies a *learned advising policy* that depends on input from the human decision maker and provides advice only when it is likely to improve his decision. Thus, in this human-AI team, information does not only flow from the algorithm to the human as in prior work, but instead there is a *two-sided interaction*: the algorithmic assistant's advice depends on the human's input, and the human's final decision, in turn, depends on the input he receives from his algorithmic assistant.

We consider a simple form of these two-sided interactions, in which the input from the human to the algorithmic assistant is the human's (initial, unassisted) risk prediction, and the algorithmic assistant's advising policy determines whether or not to advise the human, providing advice only when it identifies that its advice is likely to improve the human's prediction. Thus, our human-AI collaboration is designed such that the human decision maker is operating on his own and makes predictions, while the learned advising policy is there to optimize the added value that the human can extract from his interaction with the AI advising system.

Figure 1 presents a diagram of the advising-policy approach that we take for AI-assisted decision making, which we demonstrate in the pretrial-release decision setting. We first learn an advising policy by using predictions that humans



(a) Training the algorithmic components

(b) Live interaction

Figure 1: A responsive advising approach for AI-assisted human decision making.

made in previous experiments about how likely a criminal defendant is to violate his release terms if released. Then, we conduct a large-scale experiment on Mechanical Turk to evaluate human prediction performance when decision makers interact with our learned advising policy. Our experimental results show that an advising policy is indeed learnable from data and that humans assisted by this learned advising policy make significantly more accurate predictions than human decision makers assisted by the constant advising policy, and achieve comparable performance to the risk-assessment algorithm, thus improving over the state-of-the-art. We further explore how our responsive-advising approach affects human learning, human decision makers' responsiveness to algorithmic advice, and the performance of human decision makers with respect to defendants' racial groups.

2 Responsive Algorithmic Advising

The decision pipeline that we consider is illustrated in Figure 1b. It is composed of a human decision maker and an algorithmic assistant. The algorithmic assistant is in turn composed of a risk-assessment algorithm and an advising policy.[1] When a new criminal defendant arrives, the human observes a description of the defendant and predicts the defendant's likelihood to violate his release terms if released, $\hat{y}_{unassisted}$. Then, given the description of the defendant (excluding race and gender to match common practice among risk-assessment developers and previous experiments [Lowenkamp, 2009; Cohen *et al.*, 2018; Green and Chen, 2019a; Green and Chen, 2019b]) and the prediction that the human made, the algorithmic assistant generates an algorithmic risk assessment, $\hat{y}_{alg}$, and provides the assessment to the human according to the advising policy. The advising policy that we wish to learn aims to provide the algorithmic risk assessment to the human only when it is likely to improve the human's prediction. In cases in which the advice is not provided, the final prediction $\hat{y}_{final}$ is set to the human's unassisted prediction $\hat{y}_{unassisted}$, while in cases that the advice is provided, the human observes the advice and can update his prediction if he wishes (to any prediction value), and the final prediction is then set to the updated value $\hat{y}_{assisted}$.

Our main focus is on learning such advising policies and evaluating their impact on the predictions made by human decision makers. As the risk-assessment component of the algorithmic assistant, we use the model of [Green and Chen, 2019b] that was trained on 47,141 defendant cases from a dataset collected by the U.S. Department of Justice [DOJ., 2014] (top diagram in Figure 1a), all of which were released before trial and we thus know the ground-truth information about their pretrial-release outcome. That is, we know for each case whether eventually the defendant violated his release terms. Among the defendants in this dataset, 29.8% violated their pretrial-release terms. The model gets as input a description of a criminal defendant and outputs a risk assessment $\hat{y}_{alg} \in [0, 1]$ that represents the algorithm's predicted likelihood that this defendant will violate his release terms if released. In [Green and Chen, 2019b] it is shown that this model achieves comparable performance to widely used risk-assessment tools like COMPAS [Northpointe, 2019] and the Public Safety Assessment [Desmarais *et al.*, 2016]. See Appendix B in the full paper [Noti and Chen, 2023] for more details on the dataset and the model.

For learning the advising policy, we train a random-forest model on experimental data of human predictions from [Green and Chen, 2019a]. See the bottom diagram in Figure 1a. Given a defendant case, the algorithmic risk assessment, and the prediction that the human made, the policy determines whether or not to advise the human. In the training

[1] Note that in principle, an algorithmic assistant could be implemented as a single algorithmic component trained end-to-end. However, such an approach would not allow us to isolate the contribution of the advising policy to human prediction performance. Furthermore, an important advantage of decoupling the risk-assessment algorithm from the advising policy is in reliability: such a separation constrains the risk-assessment algorithm to be trained only to optimize the quality of its risk assessments, rather than providing biased assessments that aim to affect the human decision maker.

	Learned N=218	Random N=200	Omniscient N=200	No Advice N=258	Update N=220
Advising policy accuracy	74.1% (±0.9%)	58.4% (±1.5%)	100.0% (±0.0%)	52.5% (±2.0%)	42.0% (±2.1%)
Quadratic score	0.781 (±0.005)	0.755 (±0.007)	0.825 (±0.006)	0.719 (±0.008)	0.770 (±0.007)
Algorithm's quadratic score	0.801 (±0.003)	0.800 (±0.003)	0.803 (±0.003)	0.805 (±0.003)	0.802 (±0.003)
Linear score	0.622 (±0.006)	0.578 (±0.009)	0.653 (±0.007)	0.560 (±0.011)	0.578 (±0.008)
Algorithm's linear score	0.603 (±0.003)	0.602 (±0.004)	0.604 (±0.004)	0.607 (±0.003)	0.603 (±0.003)
Advice influence	0.810 (±0.036)	0.769 (±0.040)	0.787 (±0.041)	–	0.321 (±0.040)
Advice acceptance rates	0.735 (±0.043)	0.683 (±0.048)	0.723 (±0.045)	–	0.305 (±0.043)
Human initial risk prediction that is at least as accurate as the algorithmic risk assessment	62.47% (±1.66%)	56.84% (±2.14%)	60.60% (±1.89%)	52.46% (±1.99%)	58.00% (±2.06%)
KL divergence between the distributions of human initial risk prediction and the algorithmic risk assessment	0.47	0.52	0.41	0.92	0.32
False-positive rates (FPR)	22.88% (±1.82%)	39.39% (±3.45%)	19.63% (±2.30%)	52.64% (±4.18%)	38.04% (±3.39%)
False-negative rates (FNR)	51.39% (±1.99%)	43.55% (±2.89%)	37.51% (±2.72%)	36.28% (±3.33%)	41.83% (±2.62%)
Classification disparity	0.145	0.102	0.138	0.094	0.152

Table 1: Experimental results: Overview of main metrics.

process, the label of each such a prediction example is set to 1 (i.e., do advise) if the algorithm's prediction is more accurate than that made by the human, and to 0 (i.e., do not advise) otherwise. The training data consist of 6,250 predictions made by 250 human participants, for 500 defendant cases. Each participant was asked to predict for a series of 25 defendants, the defendants' risk to violate their release terms if released. In this dataset, in 33.31% of the predictions the algorithm's prediction was more accurate than the human's prediction. To better adapt to our target domain, which is a new experiment in which humans interact with a learned advising policy rather than predict independently from it as in our training data, we train our model on an augmented version of the dataset. For more details on the learning process, see Appendix C in the full paper [Noti and Chen, 2023].

3 Experimental Setup

We conducted an experiment on Mechanical Turk to evaluate the quality of human predictions when assisted by our learned advising policy. In the experiment, each participant was randomly assigned to one of the experimental treatments (see below), and was asked to predict the risk for a series of 50 defendants (from 0% to 100%, in 10% intervals) to violate their release terms if released, according to the decision pipeline described above. Overall, there were 1,096 participants in the experiment, who made a total of 54,800 predictions. The experimental data are available on the authors' website.

Our experimental design compares human prediction performance in five experimental treatments. The first three treatments compare human performance when assisted by advising policies of different learning quality: **"Learned,"** in which humans were assisted by the learned advising policy described above; **"Random,"** in which the subset of defendant cases for which the human received the algorithmic advice was chosen at random, in the same frequency in which the learned advising policy provided advice on the training data; **"Omniscient,"** in which humans were assisted by an advising policy that showed the advice exactly in those cases where the algorithmic risk assessment was more accurate than their initial (unassisted) prediction, based on the ground truth of the defendant case (i.e., whether the defendant eventually violated his release terms). This provides an upper bound for performance improvement that may be achieved by improving the learning quality of our advising policy.

In addition, we ran a **"No Advice"** treatment in which humans made the predictions on their own without observing the algorithmic risk assessment, and the **"Update"** treatment from [Green and Chen, 2019b], in which humans first made the prediction on their own and then always observed the algorithmic prediction and were allowed to update their prediction if they wished. The prediction structure in this Update treatment led to the best human prediction performance in [Green and Chen, 2019b], consistently with findings in other recent studies (e.g., [Buçinca *et al.*, 2021; Groh *et al.*, 2022]) and with prior behavioral research that suggest the importance of forming a pre-advice independent opinion [Van Swol and Sniezek, 2005; Bonaccio and Dalal, 2006; Sniezek and Buckley, 1995].

For comparability with the experimental results of [Green and Chen, 2019b], we used in the experiment the same set of 300 defendant cases that they used, which were sampled

from the heldout dataset of the risk-assessment algorithm's training process, and followed their experimental setup and procedure. 200 people or more participated in each of the experimental treatments. For more details, see Appendix D in the full paper [Noti and Chen, 2023].

4 Results

Next, we describe the main experimental results. Table 1 provides an overview of the results according to the main metrics. All p-values and confidence intervals are generated on distribution of performance at the participant level, unless otherwise stated.

4.1 Learning Performance

Our analysis starts by evaluating the extent to which our learned advising policy managed to generalize from the fixed training data to the new domain of our experiment, which includes new participants, new defendant cases, and importantly, live interaction between the advising policy and the human decision maker. The experimental results show that our learned advising policy managed to provide the advice in the correct times, i.e., when the algorithmic risk assessment was more accurate than the human's initial risk prediction, significantly more frequently than all other treatments (except, of course, from the Omniscient treatment, which by definition has perfect accuracy), obtaining accuracy of 74.1±0.9%. This is compared with 58.4±1.5% accuracy in the Random treatment in which the advice is given at random times; with 42.0±2.1% accuracy in the Update treatment which can be thought of as an "always advising policy;" and with 52.5±2.0% accuracy in the No Advice treatment which can be thought of as a "never advising policy." In the Learned treatment, in 37.5% of the predictions the algorithmic risk assessment was more accurate than the human's initial risk prediction, and our learned advising policy provided the advice in 37.0% of the predictions, thus achieving calibrated advice frequency. For further details, see Appendix E in the full paper [Noti and Chen, 2023].

4.2 Impact on Human Prediction Performance

We turn to look at the actual impact of our learned advising policy on the quality of the final predictions of the human decision makers. For each risk prediction $\hat{y} \in \{0, 0.1, ..., 1\}$ with ground truth $y \in \{0, 1\}$ (0 for not violating the release terms or 1 otherwise), the prediction error is defined as $error = |y - \hat{y}|$. We evaluate the prediction performance primarily according to two measures that capture different error patterns: the linear score (i.e., $1 - error$) and the quadratic score (i.e., $1 - error^2$), which is a proper scoring rule [Gneiting and Raftery, 2007]. Evaluation according to additional measures gives qualitatively similar results (see Appendix E).

Figure 2 shows the prediction performance of the human participants in the experiment and the algorithmic prediction performance, according to the linear score (left panel) and quadratic score (right panel). See the full paper for the full performance distributions. According to both score measures, human predictions in the Learned treatment have a clear and statistically significant advantage over the No Advice, Random, and Update treatments, and specifically the ranking of performance, from best to worst, is: Omniscient, Learned, Update, Random, and No Advice. The advantage of the Learned treatment over the constant-advising Update treatment demonstrates the usefulness of our learned advising policy approach that considers input from the human decision maker, and provides advice that is focused only on those places where it is likely to be useful. The performance of the Random treatment shows that providing advice only in part of the predictions does not lead in itself to an improvement in the quality of human predictions, and that the learned advising approach is important to achieve this improvement. The large gap of the performance of Omniscient above all other treatments shows the potential for further improvement of human predictions by improving the learning quality of the advising policy (e.g., by utilizing more advanced computational methods or larger datasets).

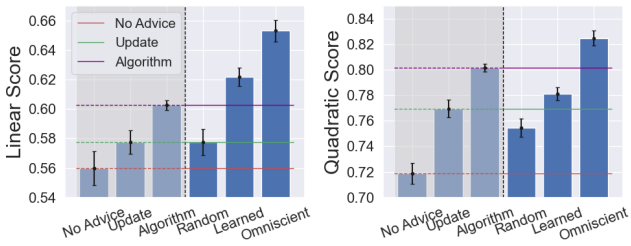


Figure 2: Participant performance in the experimental treatments, and the algorithm's performance, according to both linear and quadratic scores. Error bars show 95% confidence intervals. The No Advice, Update, and algorithmic benchmarks are presented on the left of each figure, and for ease of comparison, their means continue as horizontal lines. The algorithm's performance is computed over its predictions for the cases given to participants in Learned.

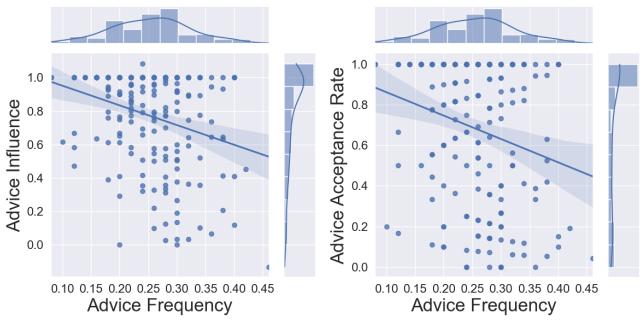
A comparison with the algorithmic performance shows that according to the linear score human decision makers in the Learned treatment outperformed the algorithm, while according to the quadratic score the algorithm had better performance.[2] Thus, we conclude that the prediction performance of human decision makers when assisted by our learned advising policy was on par with the performance of the algorithm. Humans in the Omniscient treatment outperformed the algorithm by a large gap according to both measures, which again shows the potential for further gains from improving the learning quality. Reaching the algorithmic performance is a notable improvement compared to prior advising methods in the human-AI collaboration in decision making setting that we consider that requires human agency, and in particular compared with the constant-advising Update treatment, in which human prediction performance is typically significantly inferior to that of the algorithm.

4.3 Human-Algorithm Interaction

Human Responsiveness to the Algorithmic Advice

We now look at the responses of the human decision makers to the algorithmic advice. We measure human responsiveness

[2]Note that the algorithm we use was trained to optimize quadratic score, and thus it could be expected that it will have an advantage according to this measure compared to other measures.



(a) (b)

Figure 3: The scarcity effect. The figures show for participants in the Random treatment, scatter plots of the advice influence (Figure 3a) and the advice acceptance rate (Figure 3b) vs. the advice frequency, alongside the marginal distributions and the regression lines.

to the algorithmic advice, in those cases in which the advice was given, by two measures:

1. *Advice influence* [Green and Chen, 2019b]: for each prediction $\hat{y}^k_{unassisted}$ for which an advice $\hat{y}^k_{alg}$ was given, the influence is defined by $I^k = (\hat{y}^k_{assisted} - \hat{y}^k_{unassisted})/(\hat{y}^k_{alg} - \hat{y}^k_{unassisted})$. This measure quantifies the extent to which the human prediction after observing the advice changed from its initial value in the direction of the value of the advice. It is similar to the "weight of advice" measure [Yaniv, 2004]: when the final (assisted) prediction falls within the initial (unassisted) prediction and the advice, the influence reflects the weight that a participant assigns to the advice. Influence of 0 means that the participant ignored the advice, while an influence of 1 means that the participant adopted the advice exactly. The influence values ranged in $[-6, 5]$, with 86.2% of the predictions in $[0, 1]$.
2. *Advice acceptance rate*: considering predictions where the initial prediction is different than the algorithmic risk assessment, the advice acceptance rate is the frequency in which the advice is exactly followed. I.e., $Pr(\hat{y}^k_{assisted} = \hat{y}^k_{alg} | \hat{y}^k_{unassisted} \neq \hat{y}^k_{alg}$ and $\hat{z} = 1)$.

We observe a clear pattern (Figure 3), which we term a "scarcity effect": as the advice is given less frequently, it tends to be followed by a stronger response on the human decision maker's part. Specifically, we look at the Random treatment, in which advice is given at random times and thus there is a natural variance in the frequencies in which participants received the advice. We find that the advice frequency is negatively correlated with the responsiveness of participants to the advice, as measured by the advice acceptance rate ($\rho = -0.23$, $p < 0.001$) and by the advice influence measure ($\rho = -0.29$, $p < 0.0001$). Additionally, we observed that human responsiveness to the advice in the partial-advising treatments (Random, Learned, and Omniscient) was stronger, by a large gap, than the responsiveness in the Update treatment in which algorithmic risk assessment was provided for all predictions (Figure 10 in Appendix E in the full paper). While the scarcity effect we observed in the Random treatment is sufficiently strong to explain such a gap (by extrapolating the correlation pattern to an advice frequency of 100%), this gap could also result from other factors, and our experimental design does not isolate the sources for the gap in human responses between these treatments. See the full paper for more details [Noti and Chen, 2023].

Indication of Human Learning

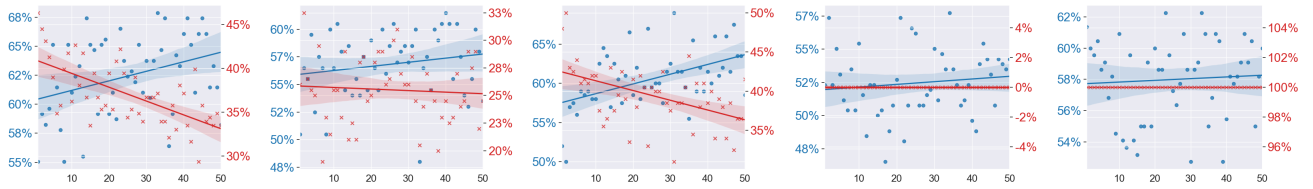
In order to see whether our human decision makers managed to learn and improve over the course of the experiment, we analyze the quality of participants' initial (i.e., unassisted) prediction in comparison with the algorithm's prediction (which is the only type of feedback that the participants received in the experiment). Note that in all treatments participants had the same information when making their initial predictions, and so differences between treatments in the quality of these predictions are a result of some learning process from the interaction with the different advising policies.

The results show that in all experimental treatments participants managed to learn and improve their initial predictions relative to the No Advice benchmark, and suggest that the informed advising policies, namely Learned and Omniscient, better facilitate human learning. Specifically, our first indication of human learning is that the overall frequency in which the human initial prediction was at least as accurate as that of the algorithm, was significantly higher than No Advice in all experimental treatments, and was the highest in the Learned and Omniscient treatments (Figure 11a in Appendix E in the full paper). Second, looking over time, we find that this frequency significantly increased with prediction period only in the Learned and Omniscient treatments (Figure 11b and analysis in Appendix E). While these observations show a clear learning effect with respect to the algorithmic feedback that participants received, we find that this effect was only weakly translated to an improvement in the quality of the initial predictions with respect to the ground truth. See the full paper for further details [Noti and Chen, 2023].

Figures 4a and 4c show the learning effect in the Learned and Omniscient treatments alongside the response of the advising policies to this effect, and demonstrate the advantage of our two-sided interaction approach: as the human initial prediction improves compared to the algorithmic prediction, the learned advising policy identifies more cases in which the algorithmic advice is not needed, and as a consequence provides significantly less advice. We note that the better learning observed in the Learned and Omniscient treatments may result from a combination of several effects, which their impact on human learning is not isolated in our experimental design; e.g., the higher informativeness of the given advice and the higher responsiveness to the advice in these treatments. Further studying the factors that facilitate human learning is a broad and interesting direction for future work.

Tension between Imitating the Algorithm and Preserving Complementary Human Strengths

The results so far show that participants managed to learn from the algorithmic feedback and improve their initial predictions, and that this improvement was more substantial in the Learned and Omniscient treatments than in the Random and Update treatments. Now we turn to look directly at how



(a) Learned (b) Random (c) Omniscient (d) No Advice (e) Update

Figure 4: Human learning. In blue: the average frequency in which the human initial prediction was at least as accurate as the algorithm's prediction, as a function of the prediction period. In red: the average frequency in which the advice was provided as a function of the prediction period. The blue and red lines are the regression curves for the two measures. Only in the Learned and Omniscient treatments these frequencies are significantly correlated with prediction period, and this learning effect resembles a "training wheels" pattern: as the participants' initial predictions improve, the algorithmic advice is useful less often and the frequency in which it is provided decreases.

the distributions of initial predictions differ between the different treatments. We demonstrate that this learning phenomenon raises a tension between the extent to which humans learn from the algorithmic advice on the one hand, and their ability to preserve their own relative prediction strengths on the other hand.

A comparison of the distribution of human initial predictions and the algorithmic predictions shows that, as expected, in the No Advice treatment, in which human predictions are completely independent of the algorithmic predictions, the distance between these two distributions is the largest (as measured by KL divergence [Kullback, 1997], see Table 1). The distribution of initial predictions in the Update treatment was the closest to the algorithmic predictions, and the treatments in which the advice was provided only in part of the predictions had intermediate KL divergence values. This suggests that in the Update treatment, in which participants constantly observed the algorithmic risk assessment, the participants learned to predict similar values to the feedback that they observed, while in the partial advice settings this imitation effect was moderated.

A closer look suggests that the partial advice has an advantage in preserving human prediction behavior that is complementary to the predictions of the algorithm. A notable example is that in the always-advising Update treatment participants learned to almost never predict a certain low-risk value of zero – a value that was never predicted by the algorithm,[3] but was predicted by human participants in the No Advice treatment in 10.5% of all predictions. This is despite the fact that in the subset of instances in which humans predicted a zero risk, their predictions were significantly more accurate than their average prediction performance. By contrast, in the Learned treatment participants preserved this relative strength and predicted a risk of zero for 11.0% of the predictions, and similarly to the No Advice treatment, with a higher accuracy in those predictions relative to their average performance. See more details in the full paper [Noti and Chen, 2023].

4.4 Fairness

We further examine how human decision makers assisted by our advising policies perform with respect to defendants' racial groups. We start by comparing the false-positive rates (FPR) and false-negative rates (FNR) in our experiment for black and white defendants (see Figure 5 and a summary in Table 1, as well as Figure 13 in Appendix E in the full paper). For each racial group, the group FPR is the rate in which defendants from that group did not violate their release terms but were wrongly classified as high-risk defendants, and the group FNR is the rate in which defendants from the group violated their release terms but were wrongly classified as low-risk defendants. The decision threshold is the value that optimizes F-score [Zou *et al.*, 2016], which is 0.3 for each treatment as well as for the algorithm, so that predictions above 0.3 are classified as high-risk decisions and otherwise are classified as low-risk decisions.[4]

First, the results show that the FPR in the Learned treatment, for both black and white defendants, is substantially lower than the FPR in the Update, Random, and No Advice treatments, but at the cost of higher FNR (see Figure 5 as well as more details in Appendix E in the full paper). Second, in terms of FPR and FNR disparities between black and white defendants, we find that the learned advising policy is comparable to the Update treatment, and has a significant advantage compared with the risk-assessment algorithm. For the details, see the full version of the paper [Noti and Chen, 2023].

Figure 5 shows that according to both FPR and FNR, all treatments have an error that is biased to the same direction which gives harsher predictions for black defendants. We quantify this discrimination by defining "classification disparity" for a treatment as: $Pr(Y = 0)(FPR_{Black} - FPR_{White}) + Pr(Y = 1)(FNR_{White} - FNR_{Black})$. The classification disparity weighs the discrimination of black compared to white non-risky defendants (the first term), and the bias in favor of white compared to black risky defendants (the second term). Equivalently, the classification disparity can be interpreted in terms of utility from the point of view of the defendant: the first term is the utility gap for non-risky defendants and the second term is the utility gap for risky defendants, both in favor of white defendants and are weighted by the overall frequencies of risky and non-risky defendants in the population.

[3] Recall that the algorithm was optimized for minimizing quadratic error, and so avoiding predictions of extreme values is a typical outcome of such an optimization process.

[4] The same threshold is also obtained by taking the fraction of high-risk defendants, which is 0.326 in our dataset (and since risk predictions are in multiples of 0.1).

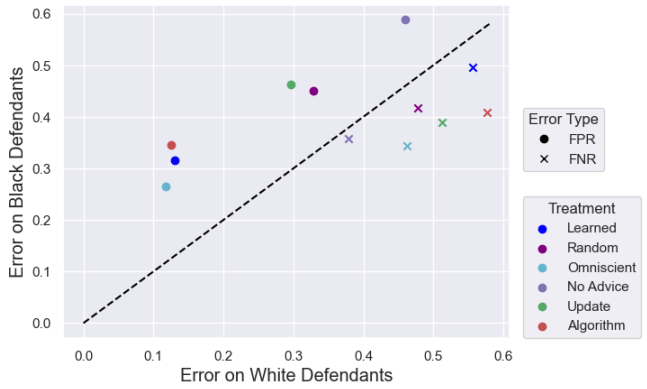


Figure 5: False-positive rates (FPR) and false-negative rates (FNR) for black and white defendants in each experimental treatment and for the algorithm's predictions.

We find that, interestingly, the algorithm has the highest classification disparity (this is despite the fact that the algorithm did not directly observe the defendant's race whereas the human participants did), the No Advice and Random treatments have the least classification disparity, and Omniscient, Learned, and Update have intermediate disparity levels with an advantage to Omniscient and Learned (see Table 1). When considering the accuracy-fairness tradeoff, the results show that Learned and Omniscient Pareto dominate the Update treatment (Figure 14 in Appendix E in the full paper). This tradeoff suggests that the use of informed advising policies in Learned and Omniscient allowed the human decision makers on the one hand to extract gains from the high performance of the algorithm, while on the other hand to moderate its racial disparity.

Finally, in the full paper we analyze interaction disparity according to the two measures studied in [Green and Chen, 2019b], to evaluate whether participants responded to the risk assessment in a racially biased manner. In [Green and Chen, 2019b], every experimental treatment exhibited disparate interactions, including the Update treatment (which is identical to Update in our experiment) that yielded the smallest disparity. Our experiment replicates the results for the first "influence disparity" measure for the Update treatment, but this influence disparity was eliminated in the Learned and Omniscient treatments. The second "deviation disparity" observed in [Green and Chen, 2019b] was not replicated in our experiment, including in our Update treatment.

5 Discussion

What is the best way to use algorithms to advise human decision makers? Existing methods constantly provide advice and focus on optimizing the algorithmic advice itself or its presentation or explanation to the human user. Motivated by limitations of the constant-advising approach that frequently advises the users in redundant or even harmful times, and by the complementary abilities of humans and algorithms that have been demonstrated in many settings, this paper proposes a responsive advising approach, in which the algorithm interacts with the human user and provides advice only when it is most beneficial for the human in making their decision.

We analyzed over fifty thousand human predictions in five experimental treatments that compared our new responsive-advising approach to the constant-advising approach and to other benchmarks. Our analysis shows that people assisted by responsive-advising policies succeeded in making predictions that are more accurate and better preserve human relative strengths, compared with people who constantly received the algorithmic advice. Also, human predictions when assisted by our advising policy achieved comparable performance to the algorithm's predictions in terms of accuracy, which, as discussed, is a significant improvement over prior methods, and had a significant advantage over the algorithm in terms of fairness measures. Importantly, we showed that such an advising policy that identifies when (and when not) to provide advice to the human user, based on input from the user, can be automatically learned from existing data.

One basic explanation for the advantage of our approach in terms of accuracy, is that our learned advising policies managed to utilize, for every given prediction problem, both the input from the algorithmic risk assessment and the input from the human user. This resulted in providing the advice in more informative times, and specifically, providing the advice when the algorithmic assessment was more accurate than the human initial prediction and refraining from providing misleading advice. Notably, in our implementation, the input from the user was composed of solely the human's initial prediction. The results show that this single additional bit of information that the AI system received already enabled this significant advantage. Future work will determine whether more complex inputs from the human users can further improve the quality of the advising policies and their usefulness for the users (e.g., by using active queries to the users, individualized analyses of their historical behavior, or signals that indicate their levels of confidence or engagement).

Aside from the direct impact on performance, our analysis raises two concerns about the longer-term impact on human decisions from constantly receiving input from an algorithm. First, in the treatment where humans received algorithmic advice all the time, this advice was followed by a weak response, and importantly humans often failed to identify those cases where this advice was especially useful for them. By contrast, in the treatments where advice was given only in selected times, this advice was followed by higher responsiveness on the side of the human decision makers. Indeed, in a within-treatment analysis in the Random treatment, we find a clear connection between the frequency in which the advice is provided and human responsiveness to the advice, which we term the "scarcity effect": When advice is given less frequently, it tends to be followed by stronger responses. We conjecture that observing advice more frequently leads to habituation in human responses, while scarce advice are perceived as more valuable, however further research is needed in order to explain the behavioral source of this effect. More broadly, our study suggests the importance of studying the effect of partial or conditional advising on advice utilization, which in contrast to various other factors (see, e.g., review in [Bonaccio and Dalal, 2006]) has not yet received attention in behavioral literature.

Second, our analysis of the initial predictions given by the participants in the experiment (i.e., before observing the algorithmic risk assessment), shows that people who constantly observe the algorithmic advice learn to imitate the algorithm's past predictions. Arguably, this is instead of focusing on forming their own judgments for the problems at hand. By contrast, in the interactive advising treatments, in which advice was provided in only about one third of the times, this imitation effect was moderated, and the distributions of human-predicted risk assessments preserved features that were unique to human judgments (and not to the algorithm), which almost completely disappeared in the constant-advising treatment. This empirical observation of the imitation effect raises a concern, which may in fact be inherent to any algorithmic advising setting: on the one hand algorithmic advice assists humans to improve their decisions, but on the other hand, through repeated exposure to the algorithm, human decision makers may also internalize its biases and weaknesses into their own judgments. Our results suggest that the advising-policy approach that we propose manages to balance this tradeoff to a good extent.

A potential limitation of the present study is that the findings are based on predictions made by Mechanical Turk workers in controlled experimental settings, rather than on decisions made in practice by real human experts like judges or clinicians. While controlled experiments with lay decision makers are useful in isolating and suggesting human behavioral tendencies that are then identified in practice [Guthrie *et al.*, 2000; Barberis, 2013], the effects in the "real world" may differ from those in experimental settings due to the experimental abstracted context and the decision makers' domain knowledge and levels of expertise [Tschandl *et al.*, 2020; Zhang *et al.*, 2020]. Thus, continued research is important in order to study the extent to which the findings generalize to human-algorithm interactions in practice, and specifically to test the usefulness of our responsive algorithmic-advising approach in real AI-assisted decision making scenarios.

Algorithmic advising systems are becoming increasingly prevalent in situations in which human judgment is important and cannot be replaced by an algorithm. Such systems provide advice to human decision makers in high-stake domains ranging from criminal justice to finance and healthcare, as well as in day-to-day applications such as personal assistants and recommendation systems. The ways in which we choose to design such algorithmic advising tools shape our lives and may have broad implications to society. The present study points to the importance of asking *when* algorithms should provide advice. The findings show that a responsive approach that considers input from the human user and provides advice that is *focused* on those places where it is most needed can better assist humans in making their decisions. Future work will study how to best apply this approach in current AI-assisted decision making systems, aiming to create better human-AI collaboration that will efficiently harness AI strengths to assist humans in making better decisions.

Ethics Statement

This study deals with the use of artificial intelligence (AI) to aid human decision making. It was reviewed and approved by the Harvard University Institutional Review Board (IRB) (under the reference number IRB21-0851) and the National Archive of Criminal Justice Data (NACJD)(data usage agreement DUA21-0771) to ensure that ethical considerations were addressed during the research process.

AI-assisted decision making has the potential to bring many benefits, but it also poses ethical risks. One concern that may arise is about the potential misuse of the technology to bias people to make decisions in directions that are not beneficial for themselves. Our approach attempts to mitigate this risk at the algorithmic level by decoupling the advising policy from the risk-assessment algorithm, instead of training a single end-to-end algorithmic component. This separation constrains the risk-assessment algorithm to be trained to optimize the quality of its risk assessments independently of the human decision maker, rather than learning to provide biased assessments that aim to affect their decisions. Our advising policy aligns well with the decision maker's objective, providing advice only when it deems the algorithmic assessment is more accurate than the human prediction. Another potential risk is that the AI assistance may unintentionally lead to unfair outcomes and contribute to discrimination against certain groups of people. We analyzed our results for such potential biases with respect to defendants' racial groups. The analysis suggests that the informed advising policies manage to balance the impact of algorithmic advice on human predictions by gaining from the algorithm's high performance while reducing its racial disparity.

The use of AI technology in decision making more broadly, may raise concerns about removing human agency and responsibility from the decision-making process. When algorithms alone are making the decisions, it may be difficult to hold anyone accountable for the outcomes. Moreover, lack of transparency in how these algorithms work can make it difficult to understand the rationale behind specific decisions. In this study, we focus on the setting in which the algorithm only provides advice, but the human makes the final decision. Furthermore, we use a decision structure that encourages humans to form an independent opinion: the human first makes a prediction on his own and only then observes the algorithmic advice. However, our present study does not address the issue of transparency, except for a high-level explanation of the machine learning algorithm used. Explanatory information from the algorithm regarding its decisions in combination with our advising-policy approach is an interesting extension for future studies.

Acknowledgments

This project is partially supported by U.S. National Science Foundation under grant No. IIS 2007887 and grant No. IIS 2147187. The project has also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 740282).

References

[Albright, 2019] Alex Albright. If you give a judge a risk score: evidence from kentucky bail decisions. *Harvard John M. Olin Fellow's Discussion Paper*, 85:16, 2019.

[Anderson *et al.*, 2014] Bonnie Anderson, Anthony Vance, Brock Kirwan, David Eargle, and Seth Howard. Users aren't (necessarily) lazy: Using neurois to explain habituation to security warnings. 2014.

[Angwin *et al.*, 2016] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016.

[Bansal *et al.*, 2019] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *HCOMP 2019*, 2019.

[Bansal *et al.*, 2021] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *CHI 2021*, 2021.

[Barberis, 2013] Nicholas C Barberis. Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 2013.

[Bonaccio and Dalal, 2006] Silvia Bonaccio and Reeshad S Dalal. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*, 2006.

[Buçinca *et al.*, 2021] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 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*, 2021.

[Chouldechova *et al.*, 2018] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *FAT* 2018*, 2018.

[Chouldechova, 2017] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 2017.

[Cohen *et al.*, 2018] Thomas H Cohen, Christopher T Lowenkamp, and William E Hicks. Revalidating the federal pretrial risk assessment instrument (ptra): A research summary. *Fed. Probation*, 2018.

[Dawes *et al.*, 1989] Robyn M Dawes, David Faust, and Paul E Meehl. Clinical versus actuarial judgment. *Science*, 1989.

[Desmarais *et al.*, 2016] Sarah L Desmarais, Kiersten L Johnson, and Jay P Singh. Performance of recidivism risk assessment instruments in us correctional settings. *Psychological services*, 13(3):206, 2016.

[DOJ., 2014] US DOJ. State court processing statistics, 1990-2009: Felony defendants in large urban counties. *Inter-university Consortium for Political and Social Research*, 2014.

[Garcia-Vidal *et al.*, 2019] Carolina Garcia-Vidal, Gemma Sanjuan, Pedro Puerta-Alcalde, Estela Moreno-García, and Alex Soriano. Artificial intelligence to support clinical decision-making processes. *EBioMedicine*, 2019.

[Gneiting and Raftery, 2007] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 2007.

[Green and Chen, 2019a] Ben Green and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *FAT* 2019*, 2019.

[Green and Chen, 2019b] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 2019.

[Green, 2022] Ben Green. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45:105681, 2022.

[Grgić-Hlača *et al.*, 2019] Nina Grgić-Hlača, Christoph Engel, and Krishna P Gummadi. Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction*, 2019.

[Groh *et al.*, 2022] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 2022.

[Grove *et al.*, 2000] William M Grove, David H Zald, Boyd S Lebow, Beth E Snitz, and Chad Nelson. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, 2000.

[Guthrie *et al.*, 2000] Chris Guthrie, Jeffrey J Rachlinski, and Andrew J Wistrich. Inside the judicial mind. *Cornell L. Rev.*, 2000.

[Haenssle *et al.*, 2018] Holger A Haenssle, Christine Fink, R Schneiderbauer, Ferdinand Toberer, Timo Buhl, A Blum, A Kalloo, A Ben Hadj Hassen, Luc Thomas, A Enk, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 2018.

[Jacobs *et al.*, 2021] Maia Jacobs, Melanie F Pradier, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry*, 2021.

[Kalsher and Williams, 2006] MJ Kalsher and KJ Williams. Behavioral compliance: Theory, methodology, and results.(chap. 23) in ms wogalter (ed.) handbook of warnings (pp. 313-329), 2006.

[Kamar *et al.*, 2012] Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *AAMAS*, 2012.

[Kleinberg *et al.*, 2015] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 2015.

[Kleinberg *et al.*, 2017] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 2017.

[Kononenko, 2001] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 2001.

[Kullback, 1997] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[Lai and Tan, 2019] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *FAT* 2019*, 2019.

[Lai *et al.*, 2020] Vivian Lai, Han Liu, and Chenhao Tan. " why is' chicago'deceptive?" towards building model-driven tutorials for humans. In *CHI 2020*, 2020.

[Lowenkamp, 2009] Christopher T Lowenkamp. The development of an actuarial risk assessment instrument for us pretrial services. *Fed. Probation*, 2009.

[Madras *et al.*, 2018] David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *NeurIPS 2018*, 2018.

[Meehl, 1954] Paul E Meehl. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. *Minneapolis: University of Minnesota Press*, 1954.

[Mullainathan and Spiess, 2017] Sendhil Mullainathan and Jann Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 2017.

[Musen *et al.*, 2021] Mark A Musen, Blackford Middleton, and Robert A Greenes. Clinical decision-support systems. In *Biomedical informatics*. 2021.

[NJ-Courts, 2020] NJ-Courts. Annual report to the governor and the legislature. *New Jersey Courts*, 2020.

[Northpointe, 2019] Inc. Northpointe. Practitioner's guide to compas core. https://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf, 2019. Accessed: 2023-06-12.

[Noti and Chen, 2023] Gali Noti and Yiling Chen. Learning when to advise human decision makers. *https://arxiv.org/abs/2209.13578*, 2023.

[Obermeyer and Emanuel, 2016] Ziad Obermeyer and Ezekiel J Emanuel. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 2016.

[PJI, 2019] Pretrial Justice Institute PJI. Scan of pretrial practices. *Pretrial Justice Institute*, 2019.

[Rankin *et al.*, 2009] Catharine H Rankin, Thomas Abrams, Robert J Barry, Seema Bhatnagar, David F Clayton, John Colombo, Gianluca Coppola, Mark A Geyer, David L Glanzman, Stephen Marsland, et al. Habituation revisited: an updated and revised description of the behavioral characteristics of habituation. *Neurobiology of learning and memory*, 2009.

[Sniezek and Buckley, 1995] Janet A Sniezek and Timothy Buckley. Cueing and cognitive conflict in judge-advisor decision making. *Organizational behavior and human decision processes*, 1995.

[Steyvers *et al.*, 2022] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human–ai complementarity. *PNAS*, 2022.

[Tomašev *et al.*, 2019] Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 2019.

[Tschandl *et al.*, 2020] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 2020.

[Vaithianathan *et al.*, 2017] Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. Developing predictive models to support child maltreatment hotline screening decisions: Allegheny county methodology and implementation. *Center for Social data Analytics*, 2017.

[Van Swol and Sniezek, 2005] Lyn M Van Swol and Janet A Sniezek. Factors affecting the acceptance of expert advice. *British journal of social psychology*, 2005.

[Wilder *et al.*, 2020] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *IJCAI-20*, 2020.

[Yaniv, 2004] Ilan Yaniv. Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, 2004.

[Yeomans *et al.*, 2019] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. Making sense of recommendations. *Journal of Behavioral Decision Making*, 2019.

[Yin *et al.*, 2019] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *CHI '19*, 2019.

[Zhang *et al.*, 2020] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *FAT* '20*, 2020.

[Zou *et al.*, 2016] Quan Zou, Sifa Xie, Ziyu Lin, Meihong Wu, and Ying Ju. Finding the best classification threshold in imbalanced classification. *Big Data Research*, 5:2–8, 2016.