The Principles and Limits of Algorithm-in-the-Loop Decision Making

BEN GREEN, Harvard University, USA
YILING CHEN, Harvard University, USA

The rise of machine learning has fundamentally altered decision making: rather than being made solely by people, many important decisions are now made through an "algorithm-in-the-loop" process where machine learning models inform people. Yet insufficient research has considered how the interactions between people and models actually influence human decisions. Society lacks both clear normative principles regarding how people should collaborate with algorithms as well as robust empirical evidence about how people do collaborate with algorithms. Given research suggesting that people struggle to interpret machine learning models and to incorporate them into their decisions—sometimes leading these models to produce unexpected outcomes—it is essential to consider how different ways of presenting models and structuring human-algorithm interactions affect the quality and type of decisions made.

This paper contributes to such research in two ways. First, we posited three principles as essential to ethical and responsible algorithm-in-the-loop decision making. Second, through a controlled experimental study on Amazon Mechanical Turk, we evaluated whether people satisfy these principles when making predictions with the aid of a risk assessment. We studied human predictions in two contexts (pretrial release and financial lending) and under several conditions for risk assessment presentation and structure. Although these conditions did influence participant behaviors and in some cases improved performance, only one desideratum was consistently satisfied. Under all conditions, our study participants 1) were unable to effectively evaluate the accuracy of their own or the risk assessment's predictions, 2) did not calibrate their reliance on the risk assessment based on the risk assessment's performance, and 3) exhibited bias in their interactions with the risk assessment. These results highlight the urgent need to expand our analyses of algorithmic decision making aids beyond evaluating the models themselves to investigating the full sociotechnical contexts in which people and algorithms interact.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Social and professional topics** → *Computing / technology policy*; • **Applied computing** → *Law, social and behavioral sciences.*

Additional Key Words and Phrases: ethics; fairness; risk assessment; behavioral experiment; Mechanical Turk

ACM Reference Format:
Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 50 (November 2019), 24 pages. https://doi.org/10.1145/3359152

1 INTRODUCTION

People and institutions increasingly make important decisions with the aid of machine learning systems: judges use risk assessments to determine criminal sentences, municipal health departments use algorithms to prioritize inspections, and banks use models to manage credit risk [3, 36, 64].

Authors' addresses: Ben Green, bgreen@g.harvard.edu, Harvard University, USA; Yiling Chen, yiling@seas.harvard.edu, Harvard University, USA.

2573-0142/2019/11-ART50 $15.00
https://doi.org/10.1145/3359152

These "algorithm-in-the-loop" settings involve machine learning models that inform people, with a person rather than an algorithm making the final decision [37].

This trend represents a fundamental shift in decision making: where in the past decision making was a social enterprise, decision making today has become a sociotechnical affair. These novel algorithm-in-the-loop decision making processes raise two questions—one normative, one empirical—that require answers before machine learning should be integrated into some of society's most consequential decisions:

(1) What criteria characterize an ethical and responsible decision when a person is informed by an algorithm?
(2) Do the ways that people make decisions when informed by an algorithm satisfy these criteria?

Both of these questions lack clear answers. While there exist many standards, policies, and studies related to the decisions made by people and institutions, our normative and empirical understanding of algorithm-in-the-loop decision making is far thinner.

Despite widespread attention to incorporating ethical principles (most notably, fairness, accountability, and transparency) into algorithms, the principles required of the people using algorithms largely remain to be articulated and evaluated. For although calls to adopt machine learning models often focus on the accuracy of these tools [14, 46, 59, 66], accuracy is not only attribute of ethical and responsible decision making. The principle of procedural justice, for instance, requires that decisions be (among other things) accurate, fair, consistent, correctable, and ethical [55]. Even as algorithms bear the potential to improve predictive accuracy, their inability to reason reflexively and adapt to novel or marginal circumstances makes them poorly suited to achieving many of these principles [2]. As a result, institutions implementing algorithmic advice may find themselves hailing the algorithm's potential to provide valuable information while simultaneously cautioning that the algorithm should not actually determine the decision that is made [74].

In practice, algorithm-in-the-loop decision making requires synthesizing the often divergent capabilities of people and machine learning models. Despite this imperative, however, research and debates regarding algorithmic decision making aids have primarily emphasized the models' statistical properties (e.g., accuracy and fairness) rather than their influence on human decisions [3, 21]. Thus, even as institutions increasingly adopt machine learning models in an attempt to be "evidence-based" [15, 50, 66, 73], relatively little is actually known about how machine learning models affect decision making in practice. This lack of evidence is particularly troubling in light of research which suggests that people struggle to interpret machine learning models and to incorporate algorithmic predictions into their decisions, often leading machine learning systems to generate unexpected and unfair outcomes [37, 67].

In this paper, we explore both the normative and empirical dimensions of algorithm-in-the-loop decision making. We focused on risk assessments—machine learning models that predict the probability of an adverse outcome—which are commonly used in algorithm-in-the-loop decisions in settings such as the criminal justice system.

We began by articulating a framework with which to evaluate human-algorithm interactions, positing three desiderata that are essential to effective and responsible decision making in algorithm-in-the-loop settings. These principles relate to the accuracy, reliability, and fairness of decisions. Although certainly not comprehensive, these desiderata provide a starting point on which to develop further standards for algorithm-in-the-loop decision making.

We then ran experiments using Amazon Mechanical Turk to study whether people satisfy these principles when making predictions about risk. We explored these decisions in two settings where risk assessments are increasingly being deployed in practice—pretrial release hearings and financial loan applications [15, 50, 64]—and under several conditions for presenting the risk assessment or structuring the human-algorithm interaction. This experimental setup allowed us to evaluate algorithm-in-the-loop decision making as a function of risk assessment presentation and to compare outcomes across distinct prediction tasks. Although these experiments involved laypeople rather than practitioners (such as judges or loan officers), meaning that we cannot take the observed behaviors to be a direct indication of how risk assessments are used in real-world settings, our results highlight potential challenges that must be factored into considerations of risk assessments.

People's behavior in the experiments reliably satisfied only one of our three principles for algorithm-in-the-loop decision making. While almost every treatment improved the accuracy of predictions, no treatment satisfied our criteria for reliability and fairness. In particular, we found that under all conditions in both settings our study participants 1) were unable to effectively evaluate the accuracy of their own or the risk assessment's predictions, 2) did not calibrate their reliance on the risk assessment based on the risk assessment's performance, and 3) exhibited racial bias in their interactions with the risk assessment. Further research is necessary to determine whether the practitioners who use risk assessments exhibit similar behaviors.

These results highlight the urgent need to more rigorously study the impacts of risk assessments, focusing on the full set of mechanisms through which potential outcomes come to pass. Risk assessments have the potential to improve decision making, but can also lead to unintended outcomes as they are integrated into human decision making processes and broader political contexts; evaluations must therefore be grounded in rigorous sociotechnical analyses of the downstream impacts [35]. As this study indicates, one essential component that shapes these outcomes is the quality and reliability of human-algorithm interactions. Continued research into how people should and do collaborate with machine learning models is necessary to inform the design, implementation, and governance of algorithmic decision making aids being deployed across society.

2 RELATED WORK

A core component of integrating a technical system into social contexts is ensuring that people recognize when to rely on the tool and when to discount it. As technology is embedded into critical human decisions, the stakes of human trust and reliance on technology rise, such that "poor partnerships between people and automation will become increasingly costly and catastrophic" [51]. Recent breakdowns in the human-automation partnership in self-driving cars and airplane autopilot have led to disaster [5, 39]. In many contexts, designing effective human-machine collaborations hinges as much (if not more) on presenting guidance that is tailored to human trust and understanding as it does on providing the technically optimal advice [26, 51].

Significant research in human-computer interaction has considered how to develop systems that effectively integrate human and computer intelligence [40, 45]. In the context of algorithm-assisted human decision making, prior research has explored topics such as what interactions can facilitate the development of machine learning models [9, 29, 47], how to improve human performance with an algorithm's assistance [12, 37, 48], and the ways in which laypeople perceive algorithmic decisions [4, 28, 52]. Research related to human-algorithm interactions when making predictions can be summarized into two broad categories of findings.

2.1 People struggle to interpret and effectively use algorithms when making decisions

Several experimental studies have uncovered important issues that arise when people use algorithms to inform their decision making. People often discount algorithmic recommendations, preferring to rely on their own or other people's judgment and exhibiting less tolerance for errors made by algorithms than errors made by other people [22, 56, 76]. This may be due in part to the fact that people struggle to evaluate their own and the algorithm's performance [37, 48]. Although people appear in some contexts to follow correct predictions more than incorrect ones [48], other studies suggest that people are unable to distinguish between reliable and unreliable predictions [33] or to detect algorithmic errors [62]. Moreover, people have been shown to be influenced by irrelevant information, to rely on algorithms that are described as having low accuracy, and to trust algorithms that are described as accurate but actually present random information [27, 48, 65]. And despite widespread calls for explanations and interpretable models, recent studies have found that simple models do not lead to better human performance than black box models [62] and that varying algorithmic explanations does not affect human accuracy [61].

In turn, although introducing algorithms into decision making can improve human performance, even people who are shown an algorithm's advice underperform the algorithm itself [37, 48]. It remains an open question whether this outcome is fundamental to human-algorithm collaboration or is due to poor interfaces, training, and other factors; notably, despite the assumption that humans and algorithms can productively collaborate, prior research has suggested that the differences between human and algorithmic decision making cannot be leveraged to produce better predictions than either could acting alone [70].

2.2 People often use algorithms in unexpected and biased ways

A particular danger of breakdowns in human-algorithm collaborations is that the application of an algorithm will lead to unintended behaviors and decisions. Ethnographic studies have documented how the uses of algorithms in practice can differ significantly from the planned and proclaimed uses, with algorithms often being ignored or resisted by those charged with using them [7, 11]. In other cases, the application of algorithms has prompted people to alter their behavior, becoming overly fixated on the algorithm's advice or focusing on different goals [6, 42].

Criminal justice risk assessments represent a notable example of algorithms that are highly indeterminate and often do not generate the intended or expected results [34]. Although these algorithms are typically adopted with the explicit goal of reducing detention rates, in many cases they have had only negligible impacts because judges ignore the majority of recommendations for release. Risk assessments used in Kentucky and Virginia have thus far failed to produce significant and lasting increases in pretrial release, as judges often overrode the risk assessment when it recommended release and reduced their reliance on the risk assessment over time [67, 68]. Similar results have been found in Cook County, Illinois [58] and in Santa Cruz and Alameda County, California [41].

There is also evidence that people's interactions with risk assessments are fraught with racial biases. An experimental study found that people using a risk assessment engaged in "disparate interactions," responding to the model's predictions in biased ways that disproportionately led to higher risk predictions about black criminal defendants than white ones [37]. Similarly, analyses have observed that judges in Broward County, Florida penalized black defendants more harshly than white defendants for crossing into higher risk categories [16] and that judicial use of a risk assessment in Kentucky increased racial disparities in pretrial outcomes [1].

3 PRINCIPLES FOR ALGORITHM-IN-THE-LOOP DECISION MAKING

An algorithm-in-the-loop framework provides a new approach to studying algorithmic decision making aids: rather than evaluating models like risk assessments simply as statistical tools of prediction, we must consider them as sociotechnical tools that take shape only as they are integrated into social contexts [37]. In other words, risk assessments are technologies of "social practice" that "are constituted through and inseparable from the specifically situated practices of their use" [69]. This means that a risk assessment's statistical properties (e.g., AUC and fairness) do not fully determine the risk assessment's impacts when introduced in social contexts. Given that the outcomes are ultimately more important than the statistical properties, a greater emphasis on the relationship between risk assessments and their social impacts is necessary.

Although arguments in favor of risk assessments often focus on the predictive accuracy of these tools [14, 46, 59, 66], many important decisions require more than just accuracy. For example, the principle of procedural justice requires that decisions be (among other things) accurate, fair, consistent, correctable, and ethical [55]. While many institutions have a long history of pursuing these goals and creating procedures to ensure that they are satisfied, achieving these goals in algorithm-in-the-loop settings requires new definitions, designs, and evaluations. Notably, although algorithms often make more accurate predictions than people do, their inability to reason reflexively and adapt to novel or marginal circumstances makes them poorly suited to achieving many principles of responsible and ethical decision making [2]. Algorithm-in-the-loop decision making thus requires synthesizing the often divergent capabilities of people and machine learning models.

As a starting point toward this end, we suggest three principles of behavior that are desirable in the context of making predictions (or decisions based on predictions) with the aid of machine learning models. Our three desiderata are as follows:

Desideratum 1 (Accuracy). People using the algorithm should make more accurate predictions than they could without the algorithm.

Desideratum 2 (Reliability). People should accurately evaluate their own and the algorithm's performance and should calibrate their use of the algorithm to account for its accuracy and errors.

Desideratum 3 (Fairness). People should interact with the algorithm in ways that are unbiased with regard to race, gender, and other sensitive attributes.

Desideratum 1 is the most straightforward: the goal of introducing algorithms is typically to improve predictive performance [14, 46, 59, 66].

Desideratum 2 is important for algorithm-in-the-loop decision making to be reliable, accountable, and fair. If people are unable to determine the accuracy of their own or the algorithm's decisions, they will not be able to appropriately synthesize these predictions to make reliable decisions. Such evaluation is essential to correcting algorithmic errors: "overriding" the risk assessment is commonly recognized as an essential feature of responsible decision making with risk assessments [43, 50, 73, 74]. This principle is also important to ensuring the fairness of decisions, since algorithms are prone to making errors on the margins [2] and minority groups are often less well represented in datasets. Moreover, if people are unable to evaluate their own or an algorithm's decisions, they may feel less responsible and be held less accountable for the decisions they make.

Finally, Desideratum 3 connects to fundamental notions of fairness: decisions should be made without prejudice related to attributes such as race and gender. This is particularly important to consider given evidence that people engage in disparate interactions when making decisions with the aid of a risk assessment [37].

These three principles guided our analyses of the experimental results: we evaluated the participant behaviors according to each desideratum, demonstrating how all three can be quantitatively evaluated.

4 STUDY DESIGN

Our study consisted of two stages. The first stage involved creating risk assessments for pretrial detention and financial lending. The second stage consisted of running experiments on Amazon Mechanical Turk to evaluate how people interact with these risk assessments when making predictions. The full study was reviewed and approved by the Harvard University Institutional Review Board and the National Archive of Criminal Justice Data.

4.1 Risk assessments

We began our study by creating risk assessments for pretrial detention and financial lending. Our goal was not to develop optimal risk assessments, but to develop risk assessments that resemble those used in practice and that could be presented to participants during the Mechanical Turk experiments.

4.1.1 Pretrial detention. When someone is arrested, courts can either hold that person in jail until their trial or release them with a mandate to return for their trial (many people are also released under conditions such as paying a cash bond or being subject to electronic monitoring). The higher the perceived risk that a defendant, if released, would fail to return to court for their trial or would commit any crimes, the more likely that a court is to detain that person until their trial.

To create our pretrial risk assessment, we used a dataset collected by the U.S. Department of Justice that contains court processing information pertaining to 151,461 felony defendants who were arrested between 1990 and 2009 in 40 of the 75 most populous counties in the United States [71]. The data includes information about the arrest charges, the defendant's demographic characteristics and criminal history, and the outcomes of the case related to pretrial release (whether the defendant was released before trial and, if so, whether they were rearrested before trial or failed to appear in court for trial). We cleaned the dataset to remove incomplete entries and restricted our analysis to defendants who were at least 18 years old, whose race was recorded as either black or white, and who were released before trial (and thus for whom we had ground truth data about whether that person was rearrested or failed to appear).

This yielded a dataset of 47,141 defendants (Table 1). The defendants were primarily male (76.7%) and black (55.7%), with an average age of 30.8 years. Among these defendants (all of whom were released before trial), 15.0% were rearrested before trial, 20.3% failed to appear for trial, and 29.8% exhibited at least one of these outcomes (which we defined as violating the terms of pretrial release).

We then trained a model using gradient boosted trees [31] to predict which defendants would violate pretrial release (i.e., which defendants would be rearrested before trial or fail to appear in court for trial), based on five features about each defendant: age, offense type, number of prior arrests, number of prior convictions, and previous failure to appear. We excluded race and gender from the model to match common practice among risk assessment developers [50]. For every defendant, we used the `xgboostExplainer` package to determine the log-odds influence of each attribute on the risk assessment's prediction [30].

We evaluated the model using five-fold cross-validation and found an average test AUC of 0.66 (ranging from 0.655 to 0.673 across the five folds). This indicates comparable accuracy to COMPAS [43, 49], the Public Safety Assessment [18], and other risk assessments [19]. According to a recent meta-analysis of risk assessments, our model has "Good" predictive validity [20]. We also evaluated the risk assessment for fairness and found that it is well calibrated. Given these evaluations, our pretrial risk assessment resembles those used within U.S. courts. We selected the highest performing of the five models (along with its corresponding training and test sets) for use in our experiments.

We selected from the test set a sample of 300 defendants whose profiles would be shown to participants during the Mechanical Turk experiments (Table 1). To protect defendant privacy, we could present information about only those defendants whose displayed attributes were shared with at least two other defendants in the full dataset. Although this restriction meant that we could not select a uniform random sample from the full population, we found in practice that sampling from the restricted test set with weights based on each defendant's risk score yielded a sample population that resembled the full set of released defendants across most dimensions.

Table 1. Summary statistics for all of the defendants who were released before trial and for the 300-defendant sample used in the Mechanical Turk experiments, broken down by defendant race. A violation means that the defendant was rearrested before trial, failed to appear for trial, or both.

	All	**Black**	**White**	**Sample**	**Black**	**White**
	N=47,141	N=26,246	N=20,895	N=300	N=178	N=122
Background						
Male	76.7%	77.7%	75.4%	85.7%	87.6%	82.8%
Black	55.7%	100.0%	0.0%	59.3%	100.0%	0.0%
Mean age	30.8	30.1	31.8	27.7	27.4	28.2
Drug crime	36.9%	39.2%	34.0%	44.3%	49.4%	36.9%
Property crime	32.7%	30.7%	35.3%	36.0%	32.0%	41.8%
Violent crime	20.4%	20.9%	19.8%	14.7%	14.0%	15.6%
Public order crime	10.0%	9.3%	10.8%	5.0%	4.5%	5.7%
Prior arrest(s)	63.4%	68.4%	57.0%	55.0%	66.9%	37.7%
# of prior arrests	3.8	4.3	3.1	3.6	4.6	2.2
Prior conviction(s)	46.5%	51.2%	40.7%	39.7%	50.0%	24.6%
# of prior convictions	1.9	2.2	1.6	2.2	2.8	1.3
Prior failure to appear	25.1%	28.8%	20.4%	23.7%	30.3%	13.9%
Outcomes						
Rearrest	15.0%	16.9%	12.6%	19.0%	24.2%	11.5%
Failure to appear	20.3%	22.6%	17.5%	23.3%	28.1%	16.4%
Violation	29.8%	33.1%	25.6%	32.3%	39.9%	21.3%

4.1.2 Financial loans. When someone applies for a financial loan, it is common for the potential lender to assess the risk that the borrower will fail to pay back the money (known as "defaulting" on the loan). The more likely that someone appears to pay off the loan, the more likely the lender is to provide money to that person.

To create our loans risk assessment, we used a dataset about loans from the financial company Lending Club, which posts anonymized loan data on its website [53]. The data contains records about all 421,095 loans issued during 2015, including information such as the loan applicant's job, annual income, and credit score; the loan amount and interest rate; and whether the loan was paid off. The data did not include any demographic information about loan applicants such as their age, race, or gender. We classified credit scores into one of five categories (Poor, Fair, Good, Very Good, and Exceptional) defined by FICO [60] and limited the data to loans that have been either fully paid or defaulted on.

This yielded a dataset of 206,913 issued loans (Table 2). The average loan was for $15,133.51; the average applicant had an income of $78,093.47 and a "Good" credit score. Approximately three-quarters of these loans were fully paid.

We trained a model using gradient boosted trees to predict which loan applicants would default on their loans. Our model considered seven factors about each loan: the applicant's annual income, credit score, and home ownership status; the value and interest rate of the loan; and the number of months to pay off the loan and the value of each monthly installment. Finally, we used the `xgboostExplainer` package to determine the log-odds influence of each attribute on the risk assessment's prediction about each loan.

Table 2. Summary statistics for all approved loans in 2015 and for the 300-loan sample used in the Mechanical Turk experiments. Numbers in parentheses represent standard deviations.

	All N = 206,913	**Sample** N = 300
Applicant		
Annual income	\$78,093.47 (\$73,474.56)	\$83,190.08 (\$83,681.52)
Credit score	695.3 (30.5)	693.9 (30.3)
"Good" credit score	71.2%	70.7%
Home owner	10.2%	10.0%
Renter	40.1%	40.3%
Has mortgage	49.7%	49.7%
Loan		
Loan amount	\$15,133.51 (\$8,575.05)	\$15,377.75 (\$8,520.84)
36 months to pay off loan	70.5%	73.3%
60 months to pay off loan	29.5%	26.7%
Monthly payment	\$448.49 (\$251.44)	\$462.19 (\$253.86)
Interest rate	12.9% (4.5%)	13.05% (4.5%)
Outcomes		
Fully paid	74.1%	74.0%
Charged off	25.9%	26.0%

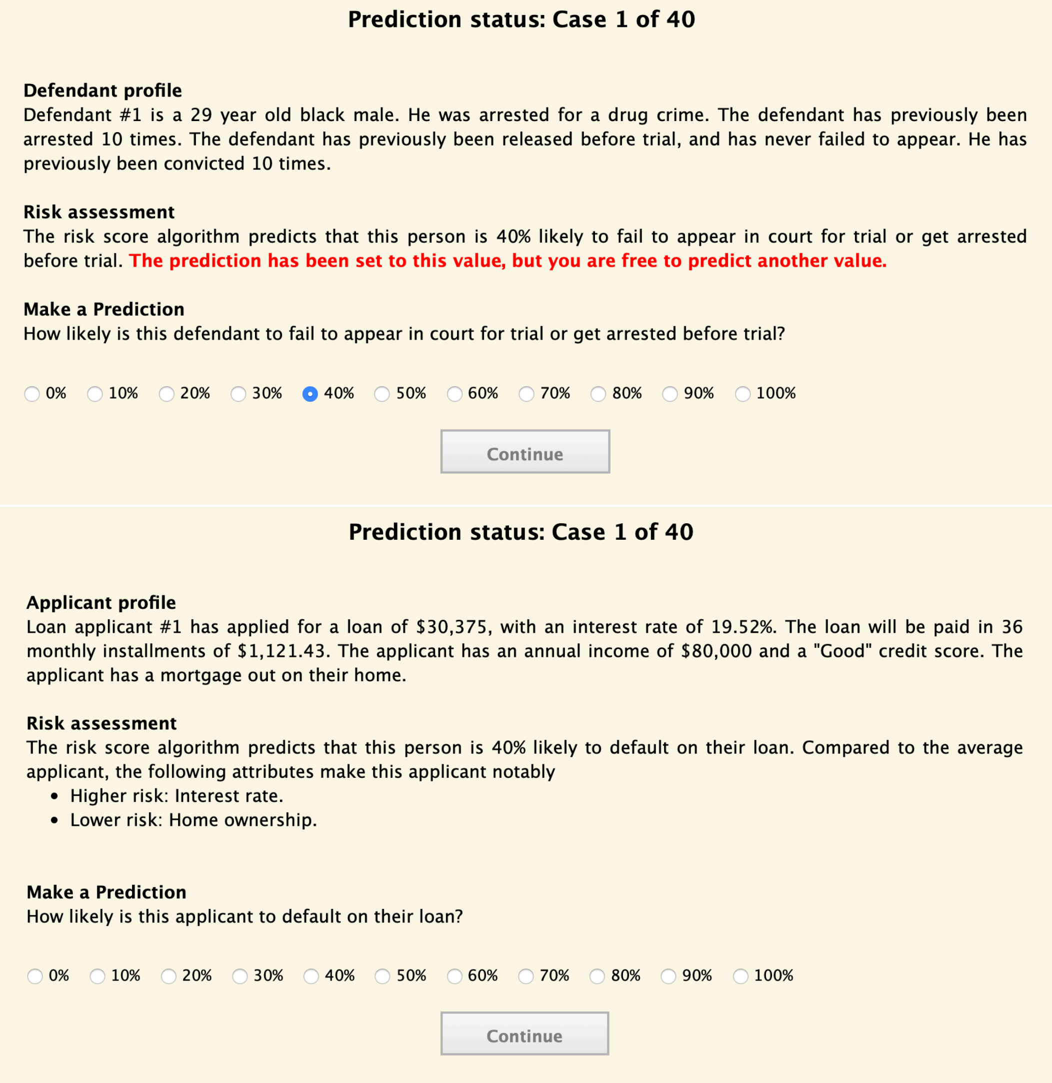
We evaluated the model using five-fold cross-validation and found an average test AUC of 0.71 (ranging from 0.706 to 0.715 across the five folds). This is similar to the performance of other loan default risk assessments that have been developed [72] and suggests "Excellent" performance [20]. We selected the highest performing of the five models (along with its corresponding training and test sets) for use in our experiments.

We selected a uniform random sample of 300 loans from the test set that would be presented to our experiment participants (Table 2).

4.2 Experimental setup

The second part of the study involved conducting behavioral experiments on Amazon Mechanical Turk to determine how people interact with these two risk assessments when making predictions. Each trial consisted of a consent page, a tutorial, an intro survey (to obtain demographic information and other participant attributes), the primary experimental task comprising a series of predictions, and an exit survey (to obtain participant reflections on the task, in the form of both multiple choice and free response questions). Both the intro and exit surveys included a simple question designed to ensure that participants were paying attention; we ignored data from participants who failed to answer both of these questions correctly. We also included a comprehension test with several multiple choice questions at the end of the tutorial; we ignored data from participants who required more than three attempts to answer every question correctly. We restricted the task to Mechanical Turk workers inside the United States who had an historical acceptance rate of at least 75%.

When participants entered the task, they were randomly sorted into one of two settings: pretrial or loans. Participants in the pretrial setting were required to estimate the likelihood that criminal



Prediction status: Case 1 of 40

Defendant profile
Defendant #1 is a 29 year old black male. He was arrested for a drug crime. The defendant has previously been arrested 10 times. The defendant has previously been released before trial, and has never failed to appear. He has previously been convicted 10 times.

Risk assessment
The risk score algorithm predicts that this person is 40% likely to fail to appear in court for trial or get arrested before trial. **The prediction has been set to this value, but you are free to predict another value.**

Make a Prediction
How likely is this defendant to fail to appear in court for trial or get arrested before trial?

○ 0% ○ 10% ○ 20% ○ 30% ● 40% ○ 50% ○ 60% ○ 70% ○ 80% ○ 90% ○ 100%

Prediction status: Case 1 of 40

Applicant profile
Loan applicant #1 has applied for a loan of $30,375, with an interest rate of 19.52%. The loan will be paid in 36 monthly installments of $1,121.43. The applicant has an annual income of $80,000 and a "Good" credit score. The applicant has a mortgage out on their home.

Risk assessment
The risk score algorithm predicts that this person is 40% likely to default on their loan. Compared to the average applicant, the following attributes make this applicant notably

- Higher risk: Interest rate.
- Lower risk: Home ownership.

Make a Prediction
How likely is this applicant to default on their loan?

○ 0% ○ 10% ○ 20% ○ 30% ○ 40% ○ 50% ○ 60% ○ 70% ○ 80% ○ 90% ○ 100%

Fig. 1. Examples of the prompts presented to participants in two of the six treatments. The top example is from the Default treatment (note that the "40%" bubble is already filled in, following the risk assessment's prediction) in the pretrial setting, while the bottom example is from the Explanation treatment in the loans setting.

defendants will be arrested before trial or fail to appear in court for trial. Participants in the loans setting were required to estimate the likelihood that a loan applicant will default on their loan (Figure 1). In both settings, participants were presented with narrative profiles about a uniform random sample of 40 people drawn from the 300-person sample populations and were asked to predict their outcomes on a scale from 0% to 100% in intervals of 10%. Profiles in the crime setting included the five features that the risk assessment incorporated as well as the race and gender of each defendant (we included these latter two features in the profiles despite not including them in the risk assessment because judges are exposed to these attributes in practice). Profiles in the loans setting included the same seven features that were included in the model. So that participants could look up background information and the definitions of key terms, the tutorial was visible at the bottom of the screen throughout the entire prediction task. Each worker was allowed to participate in each setting only once.

After being sorted into one of the two settings, participants were then randomly sorted into one of six conditions:

Baseline. Participants were presented with the narrative profile, without any information regarding the risk assessment. This condition represents the status quo prior to risk assessments, in which people made decisions without the aid of algorithms, and was one of our two control conditions.

RA Prediction. Participants were presented with the narrative profile as well as the risk assessment's prediction in simple numeric form. This condition represents the simplest presentation of a risk assessment and the typical risk assessment status quo, in which the advice of a model is presented in numerical or categorical form as a factor for the human decision maker to consider. This treatment served as the second control condition against which we evaluated the following four treatments, which represent a core (though not exhaustive) set of potential reforms to algorithmic decision aids.

Default. Participants were presented with the RA Prediction condition, except that the prediction form was automatically set to the risk assessment's prediction (Figure 1). Participants could select any desired value, however. A recent study found that many people followed this strategy when making predictions with the aid of a risk assessment, looking at the algorithm's prediction first and then considering whether to deviate from that value [37]. Moreover, this condition accords with the implementations of risk assessments that treat the model's prediction as the presumptive default and require judges to justify any overrides [13, 73].

Update. Participants were first presented with the Baseline condition; after making a prediction, participants were presented with the RA Prediction condition (for the same case) and asked to make the prediction again. A recent study found that many people first made a prediction by themselves and then took the algorithm into model when making decisions with the aid of a risk assessment [37]. This treatment adds structure to the prediction process (by prompting people to focus on the narrative profile before considering the risk assessment's prediction), which prior research has found improves decision making [44, 54].

Explanation. Participants were presented with the RA Prediction condition along with an explanation that indicated which features made the risk assessment predict notably higher or lower levels of risk (Figure 1).[1] This treatment follows from the many calls to present explanations of machine learning predictions [23, 24, 63]. In addition, by indicating which attributes strongly influenced the risk assessment's prediction, this treatment may prevent people from double counting features that the model had already considered, a problem found in prior research [37].

Feedback. Participants were presented with the RA Prediction condition; after submitting each prediction, participants were presented with an alert indicating the outcome of that case (e.g., whether the loan applicant actually defaulted on their loan). Although in practice immediate feedback on the outcomes of pretrial release or financial loans would not be available, this treatment provides one form of training for the users of machine learning systems, which is

[1] The explanations were derived from the log-odds influence of each factor (calculated in Section 4.1), with a threshold of 0.1 and -0.1 to be included in the lists of positive and negative factors, respectively.

often regarded as an essential ingredient for the effective implementation of risk assessments [8, 43, 73].

We used the same set of 300 cases for all six treatments in both settings, allowing us to directly measure the impact of each treatment on behavior. Because our experiment participants predicted risk in increments of 10%, we rounded the risk assessment predictions to the nearest 10% when presenting them to participants and when comparing the performance of participants and the risk assessments.

Participants were paid a base sum of $2 for completing the study, plus an additional reward of up to $2 based on their performance. We allocated rewards following a Brier score function: $score = 1 - (prediction - outcome)^2$, where $prediction \in \{0, 0.1, \ldots, 1\}$ and $outcome \in \{0, 1\}$. The Brier score is bounded between 0 (worst possible performance) and 1 (best possible performance), and measures the accuracy and calibration of predictions about a binary outcome.[2] We mapped the Brier score for each prediction to a payment using the formula $payment = score * \$0.05$, such that perfect accuracy on all 40 predictions would yield a bonus of $2. Because the Brier score is a proper score function [32], participants were incentivized to report their true estimates of risk. We articulated this to participants during the tutorial and included a question about the reward structure in the comprehension test to ensure that they understood.

4.3 Analysis

We analyzed the behavior of participants using metrics related to three topics: the quality of participant predictions, the influence of the risk assessment on participant predictions, and the extent to which participants exhibited bias when making predictions.

4.3.1 *Prediction performance measures.*

The first set of metrics evaluated the quality of participant predictions across treatments.

We evaluated the quality of each prediction using the Brier score. When presented with a loan applicant who does not default on their loan, for example, a prediction of 0% risk would yield a score of 1, a prediction of 100% would yield a reward of 0, and a prediction of 50% would yield a score of 0.75.

We defined the "participant prediction score" as the average Brier score attained among the 40 predictions that each participant made. Similarly, the "risk assessment prediction score" is the average Brier score attained by the risk assessment. These two metrics were used to evaluate the performance of each participant and the risk assessment.

We defined the performance gain produced by each treatment t as the improvement in the participant prediction score achieved by participants in treatment t over participants in the Baseline condition, relative to the performance of the risk assessment:

$$Gain_t = \frac{S_t - S_B}{S_R - S_B} \tag{1}$$

where S_t, S_B, and S_R represent the average prediction scores of participants in the treatment t, of participants in Baseline, and of the risk assessment, respectively. By definition, the gain of the Baseline condition is 0 and the gain of the risk assessment is 1.

4.3.2 *Risk assessment influence measures.*

The second set of metrics evaluated how much the risk assessment influenced participant predictions.

We measured the influence of the risk assessment by comparing the predictions made by participants who were shown the risk assessment with the predictions about the same case made by

[2] Because the sample populations were restricted to defendants who were released before trial and loans that were granted, we have ground truth data about the binary outcome of each case.

participants who were not shown the risk assessment. That is, the influence of the risk assessment on the prediction p_i^k by participant k about case $i \in \{1, \ldots, 300\}$ is

$$I_i^k = \frac{p_i^k - b_i}{r_i - b_i} \tag{2}$$

where b_i is the average prediction about that case made by participants in the Baseline treatment and r_i is the prediction about that case made by the risk assessment. For participants in Update, b_i is b_i^k: participant k's initial prediction about case i before being shown the risk assessment's prediction. This is akin to the "weight of advice" metric that has been used in other contexts to measure how much people alter their decisions when presented with advice [57, 75]. To obtain reliable measurements, when evaluating risk assessment influence we excluded all predictions for which $|r_i - b_i| < 0.05$.

Given an influence I_i^k, we can express each prediction as a weighted sum of the risk assessment and baseline predictions, where $p_i^k = (1 - I_i^k)b_i + I_i^k r_i$. $I = 0$ means that the participant ignored the risk assessment, $I = 0.5$ means that the participant equally weighed their initial prediction and the risk assessment, and $I = 1$ means that the participant relied solely on the risk assessment.

4.3.3 Disparate interaction measures. The third set of metrics evaluated whether participants responded to the risk assessment in a racially biased manner. Following prior work, we evaluated "disparate interactions" by comparing the behaviors of participants when making predictions about black and white criminal defendants [37].[3] We measured disparate interactions in two ways.

Our first measure of disparate interactions compared the influence of the risk assessment on predictions made about black and white defendants. We divided the data based on whether the risk assessment prediction r_i was greater or less than the baseline prediction b_i (and thus whether the risk assessment was likely to pull participants toward higher or lower predictions of risk). For each of these two scenarios, we measured the risk assessment's influence on predictions about black defendants and white defendants; for example, we defined the influence on predictions about black defendants when $r_i > b_i$ as $I_{black,>} = \text{mean}\{I_i^k | \forall k, Race_i = \text{black}, r_i > b_i\}$. We then defined the *RA influence disparity* as follows:

$$\begin{aligned} \textit{RA influence disparity}_{>} &= I_{black,>} - I_{white,>} \\ \textit{RA influence disparity}_{<} &= I_{black,<} - I_{white,<} \end{aligned} \tag{3}$$

RA influence disparity$_{>} > 0$ means that when $r_i > b_i$, participants were more strongly influenced to increase their predictions of risk when evaluating black defendants than when evaluating white defendants.

Our second measure of disparate interactions compared the extent to which participants deviated from the risk assessment's suggestion when making predictions. For each prediction p_i^k by participant k about defendant i, we measured the participant's deviation from the risk assessment as $d_i^k = p_i^k - r_i$ (i.e., $d_i^k > 0$ means that participant k predicted a higher level of risk than the risk assessment about defendant i). We used this metric to measure the average deviation for each race; for example, the average deviation for all predictions about black defendants is $D_{black} = \text{mean}\{d_i^k | \forall k, Race_i = \text{black}\}$. We then defined the *Deviation disparity* as follows:

$$\textit{Deviation disparity} = D_{black} - D_{white} \tag{4}$$

Deviation disparity > 0 means that participants were more likely to deviate positively when evaluating black defendants than when evaluating white defendants.

[3] Because we did not possess demographic characteristics about the loan applicants, we applied this analysis only to the pretrial setting.

5 RESULTS

We conducted trials on Mechanical Turk over the course of several weeks in March 2019. Filtering out workers who failed at least one of the attention check questions, who required more than three attempts to pass the comprehension test, and who participated in the experiment more than once[4] yielded a population of 1156 participants in the pretrial setting and 732 participants in the loans setting (Table 3). Across both settings, a majority of participants were male, white, and have completed at least a college degree. We asked participants to self-report their familiarity with the U.S. criminal justice system, financial lending, and machine learning on a Likert scale from "Not at all" (1) to "Extremely" (5). The average reported familiarity with the three topics in each setting was between "Slightly" (2) and "Moderately" (3), with little variation across treatments.

Participants reported in the exit survey that the experiment paid well, was clear, and was enjoyable. Considering both the base payment and the bonus payment, participants in the pretrial setting earned an average wage of $15.20 per hour and participants in the loans setting earned an average wage of $17.18 per hour. Participants were also asked in the exit survey to rate how clear and enjoyable the experiment was on a Likert scale from "Not at all" (1) to "Extremely" (5). More than 90% of participants in both settings reported that the experiment was "Very" or "Extremely" clear, and more than half of participants in both settings stated that the experiment was "Very" or "Extremely" enjoyable.

In response to exit survey questions asking how they made predictions, participants reported a variety of strategies for using the risk assessment:

- Follow the risk assessment in most or all cases (e.g., "i mostly trusted the algorithm to be more objective than i was.").
- Use the risk assessment as a starting point and then adjust based on the narrative profile (e.g., "It served as a jumping off point for my prediction.").
- Rely on the risk assessment only when unsure about a particular prediction (e.g., "I put my trust into the algorithm's predictions for when I felt like I wasn't too sure.").
- Make a prediction without the risk assessment and then adjust based on the risk assessment (e.g., "I tried not to look at it until I came to my own conclusion and then I rated my score against the computers.").
- Ignore the risk assessment (e.g., "I don't think the algorithm can be relied on").

Participants in the pretrial setting also reported diverging approaches with regard to race: while 4.4% of participants reported that they considered race when making predictions, 2.2% of participants reported explicitly ignoring race. These opposing strategies reflect differences in the perceived relationship between race and prediction: participants in the first category saw race as a factor that could improve their predictive accuracy, while participants in the second category saw race as a factor that should not be incorporated into predictions of risk (e.g., "I tried to ignore race").

5.1 Desideratum 1 (Accuracy)

Desideratum 1 states that people using the algorithm should make more accurate predictions than they could if working alone. We found that every treatment except Feedback reliably improved performance over the Baseline treatment and that the Update treatment yielded the best performance across both settings.

Across all predictions in the pretrial setting, the average participant prediction score was 0.768 and the average risk assessment prediction score was 0.803. Aside from Feedback (whose performance was not statistically distinct from that of Baseline), every treatment yielded a performance that was

[4] A server load issue prevented us from recognizing all repeat users when they entered the experiment.

Table 3. Attributes of the participants in our experiments.

	Pretrial N=1156	**Loans** N=732
Demographics		
Male	55.3%	53.0%
Black	7.1%	7.2%
White	77.2%	77.6%
18-24 years old	8.4%	7.9%
25-34 years old	42.4%	44.5%
35-59 years old	45.0%	43.2%
60+ years old	4.2%	4.4%
College degree or higher	70.9%	71.7%
Criminal justice familiarity	2.8	2.9
Financial lending familiarity	2.7	2.9
Machine learning familiarity	2.4	2.5
Treatment		
Baseline	16.5% (N=191)	15.3% (N=112)
Risk Assessment	17.3% (N=200)	16.9% (N=124)
Default	16.9% (N=195)	17.6% (N=129)
Update	16.1% (N=186)	17.9% (N=131)
Explanation	15.1% (N=174)	16.8% (N=123)
Feedback	18.2% (N=210)	15.4% (N=113)
Outcomes		
Participant hourly wage	\$15.20	\$17.18
Experiment clarity	4.4	4.4
Experiment enjoyment	3.5	3.7

statistically significantly greater than Baseline and lower than the risk assessment. Compared to RA Prediction, which had an average prediction score of 0.774, two treatments (aside from Baseline) had statistically significant differences: Feedback had a lower average prediction score of 0.751 ($p < 10^{-6}$, Cohen's $d = 0.08$), while Update had a higher average score of 0.782 ($p = 0.041$, $d = 0.03$). The gain produced by each non-Baseline treatment (Equation 1) ranged from 0.011 for Feedback to 0.603 for Update, while RA Prediction achieved a gain of 0.464 (Figure 2). Update produced a prediction score that was 1.0% greater and a gain that was 30.0% larger than RA Prediction.

A similar pattern emerged in the loans setting. Across all predictions in the pretrial setting, the average participant prediction score was 0.793 and the average risk assessment prediction score was 0.823. Compared to RA Prediction, which had an average prediction score of 0.802, two treatments (aside from Baseline) had statistically significant differences: Feedback had a lower average prediction score of 0.779 ($p < 10^{-4}$, $d = 0.09$), while Update had a higher average score of 0.813 ($p = 0.019$, $d = 0.05$). The gain produced by each non-Baseline treatment ranged from 0.327 for Feedback to 0.821 for Update, while RA Prediction achieved a gain of 0.682 (Figure 2). In other words, Update produced a prediction score that was 1.4% greater and a gain that was 20.4% larger than RA Prediction.

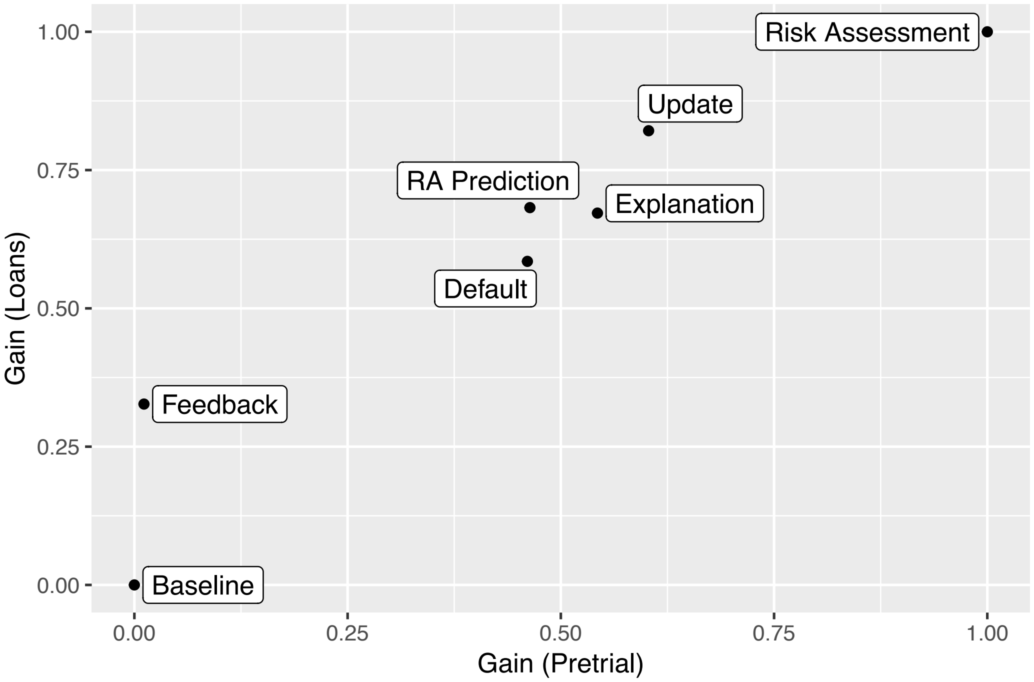


Fig. 2. The relative performance gain (Equation 1) achieved by each experimental condition across the pretrial and loans settings. In both settings, the Update treatment performed statistically significantly better than RA Prediction and the Feedback treatment performed statistically significantly worse. Across the two settings, the gain of the conditions was highly correlated.

The relative performance of each treatment was similar across the two settings (Figure 2): the gain of the five non-Baseline treatments had a Pearson correlation of 0.96 ($p = 0.010$) and a Spearman correlation of 0.9 ($p = 0.083$). In both settings, Feedback yielded significantly worse performance than RA Prediction, while Update produced significantly better performance.

To evaluate the relationship between model performance and model presentation, we measured how much more or less accurate the risk assessment would have needed to be for RA Prediction to yield the same performance as the other treatments. Taking all of the predictions made by participants in RA Prediction, we regressed the participant prediction score on the risk assessment's prediction score to determine how participant performance depends on model performance. In both cases the slope was close to 1 (1.14 in pretrial, 0.98 in loans) and was significant with $p < 10^{-15}$. In the pretrial setting, Update was equivalent to RA Prediction with a risk assessment that performs 0.91% better than the actual risk assessment while Feedback was equivalent to RA Prediction with a risk assessment that performs 2.52% worse (a range of 3.43%). In the loans setting, Update was equivalent to RA Prediction with a risk assessment that performs 1.35% better than the actual risk assessment while Feedback was equivalent to RA Prediction with a risk assessment that performs 2.91% worse (a range of 4.26%).

We observed several patterns that can partially account for the different performance levels observed. The average participant prediction score in each treatment was closely related to the rate at which participants matched their prediction to the risk assessment's prediction: the more often participants in a treatment followed the risk assessment's advice, the better the average participant prediction score in that treatment ($p = 0.012$ in pretrial, $p = 0.055$ in loans).

Although we were unable to ascertain clear explanations for why participants matched the risk assessment at different rates in every treatment, a striking pattern emerged in the Feedback treatment, which had by far the lowest match rate in both settings: the match rate declined drastically after the first prediction. In the pretrial setting, for example, the match rate of the first prediction in Feedback was 42.9%, whereas the match rate for the following 39 predictions ranged between 22.9% and 31.4% (average=26.4%). This was due to a shift in participant predictions toward the extremes (0% and 100%). For instance, the rate at which participants predicted 0% risk increased by a factor of 1.8 and 2.8 after the first prediction in the pretrial and loans settings, respectively. This indicates that many participants responded to the feedback presented after the first prediction (this feedback was necessarily binary, since the outcome either did or did not occur) by treating their own predictions as binary. This change in behavior led to a decrease in the performance of participants in the Feedback treatment.

We further analyzed the Update treatment by evaluating the quality of participants' initial predictions, which they made before being shown the risk assessment for that case. Surprisingly, despite making predictions under the same condition as participants in Baseline, participants' initial predictions in Update outperformed the predictions made in Baseline (pretrial: 0.772 vs. 0.750, $p < 10^{-5}$; loans: 0.799 vs. 0.757, $p < 10^{-14}$). This appeared to be due to the risk assessment serving a training role for participants: the initial predictions in Update improved over the course of the 40 predictions in the pretrial setting[5] ($p = 0.015$) and exhibited a sharp improvement after the first prediction in the loans setting, suggesting that being shown an algorithm's prediction about some cases can help people make more accurate predictions about future cases. The final predictions in Update, made with the benefit of the risk assessment's advice, provided further improvement over the initial predictions (pretrial: 0.782 vs. 0.772, $p = 0.014$; loans: 0.813 vs. 0.799, $p = 0.002$). These results suggest that the improvement produced by the Update treatment was twofold: first, it trained participants to make more accurate predictions in general, and second, it provided the risk assessment's prediction for the particular case at hand.

5.2 Desideratum 2 (Reliability)

Desideratum 2 states that people should accurately evaluate their own and the algorithm's performance and should calibrate their use of the algorithm to account for its accuracy and errors. This principle involves two components: first, the ability to evaluate performance, and second, the ability to calibrate a decision based on the algorithm's performance. We found that participants could not reliably exhibit either of these behaviors in any treatment.

5.2.1 Evaluation. We assessed whether participants could evaluate their own and the risk assessment's performance by comparing participant exit survey responses to the actual behaviors that they exhibited and observed (Table 4). Participants were asked to respond to each question on a Likert scale from "Not at all" (1) to "Extremely" (5).

To measure perceptions of their own performance, all participants were asked "How confident were you in your decisions?" We evaluated whether participants' self-reported confidence in their performance was related to their actual performance. The average participant confidence was 3.1 in pretrial and 3.2 in loans. Within each treatment in both settings, we regressed confidence on performance, controlling for each participant's demographic information and exit survey responses, along with the risk assessment's performance (Table 4). Across both settings, the only statistically

[5] In only one other treatment across the two settings did participant performance improve statistically significantly over time.

Table 4. Summary of participant abilities to evaluate performance (first two columns) and to calibrate their predictions (third column). The columns measure the relationships between between participant confidence and actual performance (Confidence), participant estimates of the algorithm's performance and its actual performance (RA Accuracy), and participant reliance on the risk assessment and the risk assessment's performance (Calibration). + signifies a positive and statistically significant relationship, - signifies a negative and statistically significant relationship, and 0 signifies no statistically significant relationship. In all cases, + means that the desired behavior was observed.

	Confidence		**RA Accuracy**		**Calibration**	
	Pretrial	Loans	Pretrial	Loans	Pretrial	Loans
RA Prediction	0	0	0	0	-	0
Default	0	-	0	-	0	0
Update	0	-	-	-	0	0
Explanation	0	0	0	0	-	+
Feedback	0	0	0	0	-	0

signifiant relationships between a participant's confidence and performance emerged as negative negative associations in Default and Update in loans ($p = 0.03$ and $p = 0.047$, respectively). In none of the treatments could participants reliably evaluate their performance, in some cases actually performing less well as they became more confident.

To measure participant evaluations of the risk assessment's performance, we asked every participant who was shown the risk assessment "How accurate do you think the risk score algorithm is?" and analyzed whether participant responses reflected the risk assessment's accuracy.[6] The average report of algorithm accuracy was 3.1 in pretrial and 3.3 in loans. Within each treatment in both settings, we regressed the participant evaluations of the risk assessment's accuracy against the risk assessment's actual performance, controlling for each participant's performance, demographic information, and exit survey responses (Table 4). In the Update treatment in both settings ($p = 0.04$ in pretrial and $p < 10^{-3}$ in loans) and in the Default treatment in loans ($p = 0.01$), participant evaluations of the risk assessment were negatively associated with the risk assessment's actual performance. In no treatment or setting were participants able to accurately evaluate the risk assessment's performance.

5.2.2 Calibration. To evaluate whether participants calibrated their use of the risk assessment to the risk assessment's performance, we compared the influence of the risk assessment on each prediction (Equation 2) with the quality of the risk assessment's predictions. Within each treatment, we regressed the risk assessment's influence on each participant prediction on the risk assessment's score for that prediction (Table 4). Across all settings and treatments, only the Explanation treatment in the loans setting had a positive and statistically significant relationship in which people relied more strongly on the risk assessment as its performance improved ($p = 0.006$); in pretrial, however, Explanation, RA Prediction, and Feedback had a negative relationship in which people relied less strongly on the risk assessment as its performance improved ($p \leq 0.04$). In the six other treatments across the two settings, participants did not differentiate their reliance on the risk assessment based on how it actually performed.

[6] Although all participants were presented with predictions from the same model, each participant was presented with a different set of 40 predictions. As a result of this variation, each participant observed a different level of risk assessment quality.

5.3 Desideratum 3 (Fairness)

Desideratum 3 states that people should interact with the algorithm in ways that are unbiased with regard to race, gender, and other sensitive attributes.

To assess whether this desideratum was satisfied, we analyzed if any "disparate interactions" [37] emerged in the various treatments. Because Desideratum 3 concerns bias with respect to sensitive attributes and the loans data did not contain any such attributes about applicants, we applied this analysis only in the pretrial setting. Following prior work [37], we analyzed disparate interactions along two framings: first, comparing the risk assessment's influence on participants when making predictions about black and white defendants, and second, comparing the participant deviations from the risk assessment when making predictions about black and white defendants. In both cases, we found that every treatment exhibited disparate interactions and that the Update treatment yielded the smallest disparate interactions.

5.3.1 Influence of the risk assessment. For each treatment, we compared the influence of the risk assessment on predictions about black and white defendants (Equation 3). We broke down the analysis based on whether the risk assessment's prediction was greater or less than the average Baseline participant prediction for that defendant ($r_i > b_i$ and $r_i < b_i$, respectively).

In cases where $r_i > b_i$, the risk assessment exerted a larger influence to increase risk on predictions about black than white defendants in every treatment (Figure 3). These differences were statistically significant in three of the five treatments: RA Prediction ($p = 0.001$), Update ($p < 10^{-4}$), and Feedback ($p = 0.02$). The largest disparities of 0.38 occurred in Feedback and RA Prediction; in the latter, for example, the influence for black defendants was 0.50 (meaning that participants equally weighed their own and the risk assessment's judgments) and the influence for white defendants was 0.12 (meaning that participants only slightly considered the risk assessment's judgments). The smallest disparity of 0.07 occurred in Update. Thus, although the *RA influence disparity*$_>$ was positive in Update, the disparity was reduced by 81.5% compared to RA Prediction.

The inverse pattern emerged in cases where $r_i < b_i$: in every treatment, the risk assessment exerted a greater influence to reduce risk when participants were evaluating white defendants. The discrepancies between black and white defendants were reduced, however, and were significant only in the Update treatment, which had a disparity of 0.05 ($p = 0.02$).

5.3.2 Deviation from the risk assessment. For each treatment, we compared the extent to which participants deviated from the risk assessment when making predictions about black versus white defendants (Equation 4). In every treatment, participants on average deviated positively (toward higher risk) for black defendants and negatively (toward lower risk) for white defendants. Aside from Update ($p = 0.053$), these deviation disparities were statistically significant in every treatment ($p < 10^{-6}$). The largest gap in average deviations (of 4.1%) came in Feedback, where the average deviation was +1.3% for black defendants and -2.8% for white defendants. The smallest disparity (of 0.6%) came in Update, where the average deviation was +0.4% for black defendants and -0.2% for white defendants. Compared to RA Prediction, which had a disparity of 2.3%, Update reduced the *Deviation disparity* by 73.9%.

6 DISCUSSION

This study explored the normative and empirical dimensions of algorithm-in-the-loop decision making, with a focus on risk assessments in the criminal justice system and financial lending. We first posited three desiderata as essential to facilitating accurate, reliable, and fair algorithm-in-the-loop decision making. We then ran experiments to evaluate whether people met the conditions of these principles when making decisions with the aid of a machine learning model. We studied how

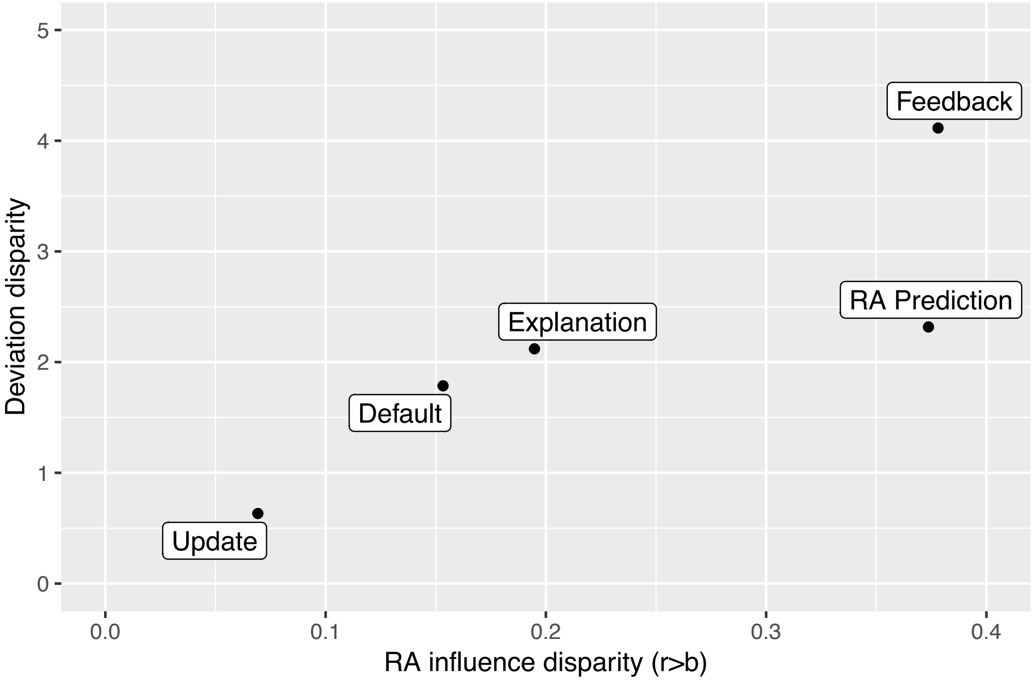


Fig. 3. The disparate interactions present in each treatment in the pretrial setting, measured by the disparities in risk assessment influence (Equation 3) and in participant deviations (Equation 4) for black versus white defendants. In both cases, values closer to 0 indicate lower levels of bias. The Update treatment yielded the smallest disparate interactions along both metrics, reducing the disparities (compared to RA Prediction) by 81.5% and 73.9%, respectively.

people made predictions in two distinct settings under six conditions—including four that follow proposed approaches for presenting risk assessments—and found that only the desideratum related to accuracy was satisfied by any treatment. No matter how the risk assessment was presented, participants could not determine their own or the model's accuracy, failed to calibrate their use of the model to the quality of its predictions, and exhibited disparate interactions when making predictions.

These results call into question foundational assumptions about the efficacy and reliability of algorithm-in-the-loop decision making. It is often assumed that, because risk assessments are merely decision making aids, the people who make the final decisions will provide an important check on a model's predictions [43, 50, 74]. For example, in *State v. Loomis*, the Wisconsin Supreme Court mandated that COMPAS should be accompanied by a notice about the model's limitations and emphasized that staff and courts should "exercise discretion when assessing a COMPAS risk score with respect to each individual defendant" [74]. But such behavior requires people to evaluate the quality of predictions and to calibrate their decisions based on these evaluations—abilities that our findings indicate people do not reliably possess. That assumptions about human oversight are so central to risk assessment advocacy and governance is particularly troubling given the inability of algorithms to reason about novel or marginal cases [2]: people may make more accurate predictions on average when informed by an algorithm, but they are unlikely to recognize and discount any errors that arise. Even when people are making the final decisions, using a risk assessment may reduce the capacity for reflexivity and adaptation within the decision making process. These concerns are particularly salient given the persistence of disparate interactions across all of our experimental treatments.

The first step toward remedying these issues is to further develop criteria that should govern algorithm-in-the-loop decision making. If society is to trust the widespread integration of machine learning models into high-stakes decisions, it must be confident that the decision making processes that emerge will be ethical and responsible. Rather than emphasizing only those values which technology is capable of promoting (such as accuracy), society must evaluate technology according to a full slate of normative and political considerations, paying particular attention to the technology's downstream implications [35, 36]. Despite providing initial steps in this direction, the three desiderata proposed here are not comprehensive and may not even be of primary concern in certain contexts. Our three desiderata do not capture broader considerations such as whether the context of a decision is just and whether it is appropriate to incorporate algorithmic advice into that context at all. Existing theories of justice must be more thoroughly adapted to algorithm-in-the-loop decision making and to the contexts in which these decisions arise.

Another important step will be to develop a deeper science of human-algorithm interactions for decision making. Although debates about risk assessments have centered on the statistical properties of the models themselves [3, 21], we found that varying risk assessment presentation and structure affected the accuracy of human decisions to an extent equivalent to altering the underlying risk assessment accuracy by more than 4%. The relative performance of each treatment was similar across two distinct domains, suggesting that our results may reflect general patterns of human-algorithm interactions. But while we were able to explain some of the differences in treatment performance, we lack a comprehensive understanding of how risk assessment presentation affected people's behaviors. Notably, we found several counterintuitive results that challenge assumptions about how to improve human-algorithm interactions. Although it is commonly assumed that providing explanations will improve people's ability to understand and take advantage of an algorithm's advice [23, 24, 63], we found that explanations did not improve human performance, a result that accords with prior work [61, 62]. We also found, counterintuitively, that providing feedback to participants significantly decreased participant accuracy (in one setting leading to predictions that were no better than those made without the advice of a risk assessment at all) and exacerbated disparate interactions.

More broadly, evaluations of algorithm-in-the-loop decision making should consider not just the quality of decisions (the focus of this study) but also how working with an algorithm can change one's perceptions of the task itself. The presentation of models can shape people's responses to the predictions made, prompting people to focus on the predictive dimensions of a complex decision and suggesting particular assumptions. For example, predictive policing systems have prompted police to alter their focus while on patrol [6, 42] and are sometimes displayed in a manner that could exacerbate a militaristic police mindset [36].

The presentation and structure of an algorithm could also diminish someone's sense of moral agency when making predictions. Prior work has found that using automated systems can generate a "moral buffer" that prompts people to feel less responsible and accountable for their actions [17]. For behavior within algorithm-in-the-loop settings to be reliable and accountable, it is essential that human decision makers feel responsibility for their actions rather than deferring agency to the computer. As a corollary, in the face of "moral crumple zones" that place undue responsibility on the human operators of computer systems rather than on the creators of those systems [25], the people developing algorithmic decision aids must feel responsibility and be accountable for how their design choices affect the final decision makers' actions.

With these considerations in mind, an important direction of future work will be to develop design principles for algorithms—as well as for the social and political contexts in which they are embedded—to promote reliable, fair, and accountable decision making. Given that only the accuracy desideratum was satisfied even when various interventions were tested, a great deal of work is clearly required to promote the full slate of desired behaviors. Such work requires a fundamental shift in algorithmic practice that begins with expanding the goals of development and evaluation to include considerations beyond model accuracy. Producing algorithms for use in social contexts means not just designing technology, but designing sociotechnical systems in which human-algorithm interactions, governance, and political discourse are all as central to the outcomes as the model predictions themselves. A thorough understanding of how each of these factors affects the impacts of algorithms is essential to building sociotechnical systems that can reliably produce ethical outcomes.

A critical step along these lines will be to further study human-algorithm interactions in real-world rather than experimental settings. A significant limitation of this paper is that our findings are based on the behaviors of Mechanical Turk workers rather than judges or loan agents, meaning that we cannot assume that the observed behaviors arise in practice. There are several indications that our results accord with real-world outcomes, however: judges suffer from many of same cognitive illusions as other people [38], are skeptical about the benefits of algorithms [10, 11], and exhibit disparate interactions when using risk assessments [1, 16]. Continued research regarding the use of risk assessments in practice (and the relationship between behaviors observed in experimental versus natural settings) will provide vital evidence to inform ongoing debates about what role algorithms can or should play in consequential decisions.

This study was further hindered by the limits of its methodology and scope. Our experiments abstracted human decision making into a series of prediction tasks, thus potentially overstating the importance of accuracy and removing many other important factors from consideration. In the U.S. criminal justice system, for instance, decisions must satisfy due process and equal protection, meaning that defendants must have the right to hear and challenge claims against them, that rules based on accurate statistical generalizations are often rejected in favor of treating people like individuals, and that decisions must be made without discriminatory intent. Because these considerations were not captured by our experimental task or evaluation metrics, experiments such as ours—by nature of how they are designed—fail to provide a holistic evaluation of risk assessments' merits and flaws. Thus, even as future work further develops principles and methods for ethical algorithm-in-the-loop decision making, it is necessary to retain a focus on the broader questions of justice that surround human-algorithm interactions and algorithmic policy interventions.

ACKNOWLEDGMENTS

We thank the two area chairs and two reviewers for their thorough and thoughtful comments. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1745303. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Alex Albright. 2019. If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions. *The John M. Olin Center for Law, Economics, and Business Fellows' Discussion Paper Series* 85 (2019).

[2] Ali Alkhatib and Michael Bernstein. 2019. Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* ACM, New York, NY, USA, Article 530, 13 pages. https://doi.org/10.1145/3290605.3300760

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[4] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).* ACM, New York, NY, USA, Article 377, 14 pages. https://doi.org/10.1145/3173574.3173951

[5] Laura Bliss. 2018. Former Uber Backup Driver: 'We Saw This Coming'. *CityLab* (2018). https://www.citylab.com/transportation/2018/03/former-uber-backup-driver-we-saw-this-coming/556427/

[6] Darwin Bond-Graham and Ali Winston. 2013. All Tomorrow's Crimes: The Future of Policing Looks a Lot Like Good Branding. *SF Weekly* (2013). http://archives.sfweekly.com/sanfrancisco/all-tomorrows-crimes-the-future-of-policing-looks-a-lot-like-good-branding/Content?oid=2827968

[7] Sarah Brayne. 2017. Big Data Surveillance: The Case of Policing. *American Sociological Review* 82, 5 (2017), 977–1008. https://doi.org/10.1177/0003122417725865

[8] Pamela M. Casey, Roger K. Warren, and Jennifer K. Elek. 2011. *Using Offender Risk and Needs Assessment Information at Sentencing: Guidance for Courts from a National Working Group.* National Center for State Courts.

[9] Marco Cavallo and Çağatay Demiralp. 2018. A Visual Interaction Framework for Dimensionality Reduction Based Data Exploration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).* ACM, New York, NY, USA, Article 635, 13 pages. https://doi.org/10.1145/3173574.3174209

[10] Steven L. Chanenson and Jordan M. Hyatt. 2016. The Use of Risk Assessment at Sentencing: Implications for Research and Policy. *Bureau of Justice Assistance* (2016).

[11] Angèle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society* 4, 2 (2017), 2053951717718855. https://doi.org/10.1177/2053951717718855

[12] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces (IUI '18).* ACM, New York, NY, USA, 329–340. https://doi.org/10.1145/3172944.3172983

[13] Thomas H. Cohen, Bailey Pendergast, and Scott W. VanBenschoten. 2016. Examining overrides of risk classifications for offenders on federal supervision. *Federal Probation* 80, 1 (2016), 12.

[14] Sam Corbett-Davies, Sharad Goel, and Sandra González-Bailón. 2017. Even Imperfect Algorithms Can Improve the Criminal Justice System. *New York Times* (2017). https://www.nytimes.com/2017/12/20/upshot/algorithms-bail-criminal-justice-system.html

[15] New Jersey Courts. 2017. One Year Criminal Justice Reform Report to the Governor and the Legislature. (2017). https://www.njcourts.gov/courts/assets/criminal/2017cjrannual.pdf

[16] Bo Cowgill. 2018. The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities. (2018).

[17] Mary L. Cummings. 2006. Automation and Accountability in Decision Support System Interface Design. *Journal of Technology Studies* (2006).

[18] Matthew DeMichele, Peter Baumgartner, Michael Wenger, Kelle Barrick, Megan Comfort, and Shilpi Misra. 2018. The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky. (2018).

[19] Sarah L. Desmarais, Kiersten L. Johnson, and Jay P. Singh. 2016. Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings. *Psychological Services* 13, 3 (2016), 206–222.

[20] Sarah L. Desmarais and Jay P. Singh. 2013. Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States. (2013).

[21] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. *Northpointe Inc.* (2016).

[22] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114–126. https://doi.org/10.1037/xge0000033

[23] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. 2017. Accountability of AI Under the Law: The Role of Explanation. *arXiv preprint arXiv:1711.01134* (2017).

[24] Lilian Edwards and Michael Veale. 2017. Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review.* 16 (2017), 18–84.

[25] Madeleine Clare Elish. 2019. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society* 5, 0 (2019), 40–60. https://doi.org/10.17351/ests2019.260

[26] Avshalom Elmalech, David Sarne, Avi Rosenfeld, and Eden Shalom Erez. 2015. When Suboptimal Rules. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.* 1313–1319.

[27] Birte Englich, Thomas Mussweiler, and Fritz Strack. 2006. Playing Dice With Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making. *Personality and Social Psychology Bulletin* 32, 2 (2006), 188–200.

[28] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I Always Assumed That I Wasn't Really That Close to [Her]": Reasoning About Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 153–162. https://doi.org/10.1145/2702123.2702556

[29] Jerry Alan Fails and Dan R. Olsen, Jr. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI '03)*. ACM, New York, NY, USA, 39–45. https://doi.org/10.1145/604045.604056

[30] David Foster. 2017. NEW R package that makes XGBoost interpretable. *Medium: Applied Data Science* (2017). https://medium.com/applied-data-science/new-r-package-the-xgboost-explainer-51dd7d1aa211

[31] Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29, 2 (2001), 1189–1232.

[32] Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Statist. Assoc.* 102, 477 (2007), 359–378.

[33] Paul Goodwin and Robert Fildes. 1999. Judgmental Forecasts of Time Series Affected by Special Events: Does Providing a Statistical Forecast Improve Accuracy? *Journal of Behavioral Decision Making* 12, 1 (1999), 37–53.

[34] Ben Green. 2018. "Fair" Risk Assessments: A Precarious Approach for Criminal Justice Reform. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

[35] Ben Green. 2019. Data Science as Political Action: Grounding Data Science in a Politics of Justice. *arXiv preprint arXiv:1811.03435* (2019).

[36] Ben Green. 2019. *The Smart Enough City: Putting Technology in Its Place to Reclaim Our Urban Future.* MIT Press.

[37] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 90–99. https://doi.org/10.1145/3287560.3287563

[38] Chris Guthrie, Jeffrey J. Rachlinski, and Andrew J. Wistrich. 2000. Inside the Judicial Mind. *Cornell Law Review* 86 (2000), 777.

[39] Andrew J. Hawkins. 2019. Deadly Boeing Crashes Raise Questions About Airplane Automation. *The Verge* (2019). https://www.theverge.com/2019/3/15/18267365/boeing-737-max-8-crash-autopilot-automation

[40] Eric Horvitz. 1999. Principles of Mixed-initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*. ACM, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[41] Human Rights Watch. 2017. "Not in it for Justice": How California's Pretrial Detention and Bail System Unfairly Punishes Poor People. (2017). https://www.hrw.org/report/2017/04/11/not-it-justice/how-californias-pretrial-detention-and-bail-system-unfairly

[42] Priscillia Hunt, Jessica Saunders, and John S. Hollywood. 2014. *Evaluation of the Shreveport Predictive Policing Experiment.* RAND Corporation. https://www.rand.org/pubs/research_reports/RR531.html

[43] Northpointe Inc. 2012. COMPAS Risk & Need Assessment System. (2012). http://www.northpointeinc.com/files/downloads/FAQ_Document.pdf

[44] Daniel Kahneman. 2011. *Thinking, Fast and Slow.* Farrar, Straus and Giroux.

[45] Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining Human and Machine Intelligence in Large-scale Crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1 (AAMAS '12)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 467–474. http://dl.acm.org/citation.cfm?id=2343576.2343643

[46] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics* 133, 1 (2017), 237–293. https://doi.org/10.1093/qje/qjx032

[47] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured Labeling for Facilitating Concept Evolution in Machine Learning. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3075–3084. https://doi.org/10.1145/2556288.2557238

[48] Vivian Lai and Chenhao Tan. 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 (FAT* '19)*. ACM, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[49] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica* (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[50] Laura and John Arnold Foundation. 2016. Public Safety Assessment: Risk Factors and Formula. (2016). http://www.arnoldfoundation.org/wp-content/uploads/PSA-Risk-Factors-and-Formula.pdf

[51] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[52] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684. https://doi.org/10.1177/2053951718756684

[53] Lending Club. 2019. Lending Club Statistics. (2019). https://www.lendingclub.com/info/download-data.action

[54] Julia Levashina, Christopher J. Hartwell, Frederick P. Morgeson, and Michael A. Campion. 2014. The Structured Employment Interview: Narrative and Quantitative Review of the Research Literature. *Personnel Psychology* 67, 1 (2014), 241–293. https://doi.org/10.1111/peps.12052

[55] Gerald S. Leventhal. 1980. What Should Be Done with Equity Theory? In *Social Exchange*. Springer, 27–55.

[56] Joa Sang Lim and Marcus O'Connor. 1995. Judgemental Adjustment of Initial Forecasts: Its Effectiveness and Biases. *Journal of Behavioral Decision Making* 8, 3 (1995), 149–168.

[57] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90 – 103. https://doi.org/10.1016/j.obhdp.2018.12.005

[58] Frank Main. 2016. Cook County judges not following bail recommendations: study. *Chicago Sun-Times* (2016). https://chicago.suntimes.com/chicago-news/cook-county-judges-not-following-bail-recommendations-study-find/

[59] Alex P. Miller. 2018. Want Less-Biased Decisions? Use Algorithms. *Harvard Business Review* (2018). https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms

[60] myFICO. 2016. Understanding FICO Scores. (2016). https://www.myfico.com/Downloads/Files/myFICO_UYFS_Booklet.pdf

[61] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *arXiv preprint arXiv:1802.00682* (2018).

[62] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and Measuring Model Interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[63] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[64] Naeem Siddiqi. 2012. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Wiley.

[65] Aaron Springer, Victoria Hollis, and Steve Whittaker. 2017. Dice in the black box: User experiences with an inscrutable algorithm. In *2017 AAAI Spring Symposium Series*.

[66] Sonja B. Starr. 2014. Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Review* 66, 4 (2014), 803–872.

[67] Megan Stevenson. 2018. Assessing Risk Assessment in Action. *Minnesota Law Review* 103 (2018), 303–384.

[68] Megan T. Stevenson and Jennifer L. Doleac. 2018. The Roadblock to Reform. *The American Constitution Society* (2018). https://www.acslaw.org/wp-content/uploads/2018/11/RoadblockToReformReport.pdf

[69] Lucy Suchman, Jeanette Blomberg, Julian E. Orr, and Randall Trigg. 1999. Reconstructing Technologies as Social Practice. *American Behavioral Scientist* 43, 3 (1999), 392–408. https://doi.org/10.1177/00027649921955335

[70] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating Human+ Machine Complementarity for Recidivism Predictions. *arXiv preprint arXiv:1808.09123* (2018).

[71] United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. 2014. State Court Processing Statistics, 1990-2009: Felony Defendants in Large Urban Counties.

[72] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 10–19. https://doi.org/10.1145/3287560.3287566

[73] Gina M. Vincent, Laura S. Guy, and Thomas Grisso. 2012. Risk Assessment in Juvenile Justice: A Guidebook for Implementation. (2012). http://njjn.org/uploads/digital-library/Risk_Assessment_in_Juvenile_Justice_A_Guidebook_for_Implementation.pdf

[74] Wisconsin Supreme Court. 2016. *State v. Loomis*. 881 Wis. N.W.2d 749.

[75] Ilan Yaniv. 2004. Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes* 93, 1 (2004), 1–13. https://doi.org/10.1016/j.obhdp.2003.08.002

[76] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. 2019. Making sense of recommendations. *Journal of Behavioral Decision Making* (2019). https://doi.org/10.1002/bdm.2118

Received April 2019; revised June 2019; accepted August 2019