Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing

NINA GRGIĆ-HLAČA, Max Planck Institute for Software Systems, Germany
CHRISTOPH ENGEL, Max Planck Institute for Research on Collective Goods, Germany
KRISHNA P. GUMMADI, Max Planck Institute for Software Systems, Germany

Much of political debate focuses on the concern that machines might take over. Yet in many domains it is much more plausible that the ultimate choice and responsibility remain with a human decision-maker, but that she is provided with machine advice. A quintessential illustration is the decision of a judge to bail or jail a defendant. In multiple jurisdictions in the US, judges have access to a machine prediction about a defendant's recidivism risk. In our study, we explore how receiving machine advice influences people's bail decisions.

We run a vignette experiment with laypersons whom we test on a subsample of cases from the database of this prediction tool. In study 1, we ask them to predict whether defendants will recidivate before tried, and manipulate whether they have access to machine advice. We find that receiving machine advice has a small effect, which is biased in the direction of predicting no recidivism.

In the field, human decision makers sometimes have a chance, after the fact, to learn whether the machine has given good advice. In study 2, after each trial we inform participants of ground truth. This does not make it more likely that they follow the advice, despite the fact that the machine is (on average) slightly more accurate than real judges. This also holds if initially the advice is mostly correct, or if it initially is mostly to predict (no) recidivism.

Real judges know that their decisions affect defendants' lives. They may also be concerned about reelection or promotion. Hence a lot is at stake. In study 3 we emulate high stakes by giving participants a financial incentive. An incentive to find the ground truth, or to avoid false positive or false negatives, does not make participants more sensitive to machine advice. But an incentive to follow the advice is effective.

Additional Key Words and Phrases: Machine-Assisted Decision Making; Human-Centered Machine Learning; Algorithmic Decision Making; Algorithmic Fairness, Accountability, and Transparency

ACM Reference Format:
Nina Grgić-Hlača, Christoph Engel, and Krishna P. Gummadi. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 178 (November 2019), 25 pages. https://doi.org//10.1145/3359280

1 INTRODUCTION

Research question. In the artificial intelligence literature, man and machine are sometimes portrayed as competitors. Machines (data-driven algorithms) compete with humans, say on playing chess. The machine tries to beat its human counterpart. Yet in multiple practical applications, it is not plausible that human decision-makers are completely replaced by machines. A human-machine symbiosis [48] is much more likely, with humans keeping responsibility for the decisions, but

Authors' addresses: Nina Grgić-Hlača, Max Planck Institute for Software Systems, Saarbrücken, Germany, nghlaca@mpi-sws.org; Christoph Engel, Max Planck Institute for Research on Collective Goods, Bonn, Germany, engel@coll.mpg.de; Krishna P. Gummadi, Max Planck Institute for Software Systems, Saarbrücken, Germany, gummadi@mpi-sws.org.

2573-0142/2019/11-ART178
https://doi.org//10.1145/3359280

relying on machine decision aids. In a typical situation, the machine suggests a decision, but the ultimate human decision-maker keeps the right and the ability to override this suggestion.

On first reading, this collaboration promises an improvement in decision quality. Before making the final decision, the human decision-maker has access to advice. The advice from the machine can exploit the comparative advantages of machines over humans. The advice may for instance exploit a much larger training dataset than a human decision-maker might ever see in her lifetime. The advice might be less prone to certain types of bias, since the machine is not liable to motivated reasoning [10, 47]. Yet the advantage presupposes that human decision-makers use the advice appropriately. We run a series of experiments to investigate whether machine advice actually improves the quality of human decisions. We focus on accuracy.

Legal application. One context in which machine advice is already used regularly is judicial decision-making. Judges in various jurisdictions in the United States are using the COMPAS tool [4] to help them decide when to grant someone bail. They are presented with a machine-generated evaluation of the defendant's risk of committing another crime before being tried. The tool uses machine learning methods to assess this risk. It exploits a large training dataset to distill indicators, and in particular combinations of multiple indicators, to increase the accuracy of the prediction. Yet earlier work has shown that, on average, COMPAS makes correct predictions in only slightly more than 65% of all cases [4, 24].

For this task, full delegation of decision making to the machine would be illegal. 18 U.S. Code § 3145 gives everyone a right to have the decision of detention reviewed by the court having jurisdiction over the offense. In Germany, the right to be heard by a competent judge is even constitutionally protected (Art. 101 GG, BVerfG May 12, 2005, 2 BvR 332/05). Based on such considerations, the Supreme Court of the State of Wisconsin ruled that judges looking at COMPAS scores during sentencing must get warnings about the tool's performance [67]. Are the procedural rights of defendants circumvented by giving the judges machine advice? The answer should depend on (a) the quality of the advice, but also (b) on the way how the judges use the advice. In this project, we focus on the latter.

Empirical strategy. Ecologically fully valid evidence would require decisions taken by real judges about real defendants, with and without the benefit of machine advice. Yet even if we could get data from one jurisdiction where the COMPAS tool is in use, and from another where it is not, this would not suffice to isolate the effect of machine advice. Legal rules, courts, judges, defendants and charges would inevitably differ as well. Were we to find a difference, say in the accuracy of decisions, we would not know whether the difference is indeed driven by the availability of machine advice. There could also be reverse causality: a jurisdiction may have been more concerned about accuracy in the first place, and therefore making the tool available. As usual, observational data would not allow for the identification of causal effects.

Causal evidence requires that otherwise identical decision-makers, faced with identical cases, are randomly assigned to either receive machine advice or not. In principle, this could be achieved in one of two ways. One could either administer the same (hypothetical) cases to real judges, or to lay persons. The former is seemingly appealing as decisions would be taken by individuals competent to take comparable decisions in the field. Yet for that very reason, the approach would be problematic. Real judges either have access to computer advice in their daily practice, or they do not. They have more or less experience with a certain type of case. They would come from jurisdictions, courts or communities with different normative attitudes towards bailing versus jailing. Hence we would again face a concern about omitted variables. By contrast when testing laypersons, we can much more confidently expect idiosyncratic differences to be quasi randomly distributed across treatments.

For these reasons we revert to a method that is common in the social sciences. We run a vignette study. We have a number of crowdworkers each decide on a series of cases. These cases are taken from Broward County in Florida, where (a) the COMPAS tool is in use and (b) ProPublica, relying on freedom of information legislation, has found out whether defendants actually have recidivated. This choice of stimulus material has a number of advantages. The tool is not hypothetical, but actually used for consequential human decisions. The tool is managed by a commercial provider that has every interest in improving accuracy. This allows us to put algorithmic decision aids to a fair test. We increase ecological validity by selecting cases for which we know whether COMPAS predicts recidivism, and whether the defendant has recidivated in the two years after release. We do of course not know whether defendants who have been incarcerated would have committed new crimes when on bail. But for those who have been bailed we also know whether they have been apprehended again before the original procedure terminates. At least for false negatives, we can therefore also investigate whether human decision-makers override the machine recommendation when this would lead to a better choice. We also know whether those who have been incarcerated have committed a new crime in the two years after their release, which gives us an imperfect proxy for correct and false positives.

Real judges at least occasionally learn whether they were right when bailing (and, to a lesser extent, when jailing) the defendant. This in particular holds if cases are not randomly assigned to judges, and if the same defendant again appears before the same judge. This gives judges an (imperfect) opportunity for learning. To see whether learning is critical for benefiting from machine advice, in a second study, after each of the 50 trials, we tell participants whether the defendant actually recidivated. In this study, we also have treatments where participants first see mostly reliable vs. mostly unreliable advice, or are initially mostly advised to predict recidivism or not.

Real judges are aware of the fact that the decision to bail the defendant may be consequential for future victims, and the decision to jail the defendant will be consequential for the defendant. They may also feel external pressure, for instance from the electorate, if they are up for re-election. Hence for real judges the stakes are considerably higher. We emulate high stakes by a third study in which we give participants a (pecuniary) incentive for deciding in line with ground truth. In further treatments, we capture different normative attitudes towards jailing innocent defendants, or bailing recidivists, by differential incentives. Finally we incentivize participants for following the machine's advice.

For the reasons just explained, giving feedback, and adding incentives, increases ecological validity. We do, however, readily admit that the way in which we add these elements to the design of our experiment only emulates feedback and incentives as they occur in the courtroom. Real judges do not receive feedback for each and every case, and they never receive feedback immediately after making a decision. Furthermore, real judges normally do not experience immediate financial consequences of their decision to bail or jail a defendant. This gap between our manipulations and the decisions in the field that we want to understand is the inevitable price we have to pay for identifying causal effects.

Contribution. Prior work has chiefly been interested in the fairness of decisions completely delegated to machines [1, 4, 6, 31] and has extensively studied the distributive [26, 29, 39, 43, 49, 50, 56, 65, 73–75] and procedural [36, 38] fairness of machine decision-making [9, 14, 18, 19, 32, 46], as well as the feasibility of making fair decisions [14, 19, 32, 46], as reviewed in [9, 18]. Several studies also explored human perceptions of the fairness of machine decisions, in the context of self-driving cars [5, 12], targeted-advertising [59], loan approval [62], and bail decisions [36–38, 66].

Another series of studies has compared human and machine decision-making. Kleinberg et al. [45] found that algorithms trained on NYC arrest data outperform expert humans judges in predicting

criminal risk. Dressel and Farid [24] compared the accuracy of groups of human laypersons with the accuracy of the COMPAS decision aid, and found no significant difference. Tan et al. [68] studied qualitative differences between human and machine decision-making, again for making bail decisions. Green and Chen [35] compared the performance of human decision makers with and without advice, against the machine's performance, in terms of the accuracy and fairness of the resulting decisions. An experiment comparing the influence of (alleged) human with "statistical" advice found that the latter has a substantially smaller effect on people's choices [52].

There is also a small amount of literature studying the effect of machine advice on human decisions. Feng and Boyd-Graber [30] investigate how different ways of explaining machine advice, in a quiz, improve human decisions. Crandall et al. [20] and Dimitrakakis et al. [23] study algorithms that are able to communicate their strategy to humans, and demonstrate their benefits empirically [20] and theoretically [23]. Christin [15] studies the use of algorithms in the criminal justice and web journalism domains, finding that the intended and actual use of algorithms often differ, and that practitioners often develop strategies for minimizing the impact of algorithmic input in their daily work. Garrett and Monahan [33] focus on the judicial domain and find pronounced variance in the way judges use risk assessment tools. Kirchkamp and Strobel [44] investigate whether sharing decision-making authority with a machine induces human decision makers to become less responsible, in the sense of being more selfish.

We contribute to this literature by systematically investigating the conditions under which machine advice improves the accuracy of human decisions. We provide causal evidence, by a series of experiments. We report three studies. Study 1 uses a within-subjects design. It investigates whether access to machine advice improves human predictions. We find a small effect, which is biased in the direction of predicting no recidivism. From a policy perspective, it may be worrisome that machine advice has little effect, given that deploying the tool is costly. This finding motivates Studies 2 and 3. In Study 2 we test whether giving human decision makers feedback about the performance of the machine moderates the effect; it does not. In Study 3, we emulate higher stakes, by giving participants a financial incentive. It only proves effective if participants gain money by following the advice.

2 STUDY 1: DOES MACHINE ADVICE IMPROVE ACCURACY?

2.1 Hypotheses

Judges may receive machine advice simply because the legislator (or the court administration) wants to ease the case load, and get the same job done by fewer judges. Yet from a normative perspective, one would hope that advice also improves decision quality, and accuracy in particular. Accuracy is a relevant concern, as judges must make decisions under substantial uncertainty. If they wrongly bail a defendant who commits a new crime before being adjudicated, the well-being of the victims is at stake. If judges wrongly jail defendants who would not have reoffended anyway, they unnecessarily curtail personal freedom, and impose a higher cost on the state.

To the best of our knowledge, there is no literature that systematically investigates the effect of machine advice on the accuracy of decisions made by humans. Thus we derive our hypotheses from the social psychology and economics literature on the effect of advice received from a human advisor on a decision taken by a human decision maker. This literature identifies four main dimensions which affect the impact of advice: the properties of the *advice*, *advisor*, *advisee* and the *task*. Properties of the advice which affect the propensity to take the advice include the framing [69], the representation [34, 40] and the timing [27, 64] of the advice. Advice taking is also influenced by traits of the advisor and advisee [11], such as their expertise and confidence, their prior relationship, the frequency of prior interactions and mutual trust. Finally, the properties of

the task affect advice taking through the risk [42, 71] and responsibility [7, 8, 21, 44, 71] associated with the decision making process and outcomes.

If the findings on human advice carry over to machine advice, we expect that machine advice has an effect on human predictions:

Hypothesis 1 After receiving machine advice, humans change their predictions.

Hypothesis 2 Humans shift their predictions in the direction of the machine's advice.

The advice taking literature [11] finds that advice is more likely to be followed if the advisor is confident and accurate. Hence, we hypothesize that:

Hypothesis 3 The less confident the machine was in the advice it gave, the less likely participants are to follow the advice.

Receiving advice in agreement with one's pre-advice opinion is found to increase one's post-advice confidence [72]. This leads to:

Hypothesis 4 When they are in agreement with the advice given by the machine, participants are more confident.

Criminal procedure presumes innocence. This translates into the standard of proof. Conviction is only permitted if the jury is convinced "beyond reasonable doubt" that the defendant is guilty [22]. Strictly speaking, this standard of proof does not apply to bailing decisions. They strike a balance between protecting potential new victims and the burden inflicted upon the defendant by incarceration. Still the culturally entrenched presumption might create an asymmetric effect of machine advice, which leads to

Hypothesis 5 Advice to predict that a defendant will not recidivate has a stronger effect than advice that he will recidivate.

2.2 Design

Stimulus material. We exploit the fact that we have access to the ProPublica dataset [4], which contains data about 7214 defendants subjected to the COMPAS tool and tried in Broward County, Florida, in 2013 and 2014. Using this data, we train a logistic regression classifier to predict whether a defendant will recidivate within two years. The classifier is implemented using the Python Scikit-learn package [55]. Comparing predictions with the ground truth retrieved by ProPublica, we learn that the accuracy of the decision aid we trained is 68% , close to the accuracy of the COMPAS tool [4, 24].

In our experiments, we randomly select 50 of the 1000 ProPublica cases which were studied by Dressel and Farid [24].[1] For these defendants, we know whether they have recidivated in the two years after release from prison (if they have been put into jail) or after being put on bail. For the 50 defendants we consider in our experiments, the reconstructed classifier predicts that 18 defendants to reoffend when put on bail, while police records show that 24 defendants actually reoffended. 75% of the misclassified cases are false negatives: the machine predicted that the defendant will not commit a new crime within two years, but they did.

Survey Instrument. After completing a consent form, respondents are shown the following text:

> *Judges in Broward County, Florida use a computer program to help them decide if a defendant can be released on bail before trial. To help you make bail decisions in this survey, we will use a similar computer program.*

[1] Focusing on the cases studied by Dressel and Farid [24] makes it possible to compare the performance of our respondents with the performance of their respondents, on the same set of cases.

> *In each question, we will describe a defendant and tell you if the computer program predicts that this defendant will commit a crime in the next 2 years or not. We will then ask **you** to tell us what you believe.*
> *Your answer does not need to match the computer's prediction, because these predictions are not always correct. This computer program correctly predicts if a person will commit a crime in the next 2 years for **68% defendants**.*

Vignettes. Cases are presented as vignettes [cf. 24], with the following token:

> *The defendant is a <sex> aged <age>. They have been charged with: <crime>. This crime is classified as a <misdemeanor / felony>. They have been convicted of <non-juvenile prior count> prior crimes. They have <juvenile felony prior count> juvenile felony charges and <juvenile misdemeanor prior count> juvenile misdemeanor charges on their record.*

The vignettes are shown in random order, to mitigate the effects of order bias [61]. After reading a vignette, participants are asked whether they believe the defendant will commit another crime in the next 2 years or not. They are also asked how confident they are about their answer.[2] After answering these questions, participants receive the following information:

> *A computer program similar to the one used by judges in Broward County, Florida, estimates that this defendant **<will / will not>** commit a crime in the next 2 years. This computer program correctly predicts if a person will commit a crime in the next 2 years for **68% of defendants**.*

Participants are able to observe their previous answer, and are asked to make a new prediction, and to report their confidence in their new prediction.

Design choices. Two design choices warrant explanation. Real judges are not making predictions; they decide whether to put the defendant on bail. However, the law wants this decision to be based on the risk of recidivism, and the machine advice exclusively covers this risk. We test laypersons, who are likely to vary in their attitudes towards crime, and incapacitation in particular. We do not know these attitudes. To overcome this potential confound and directly ask respondents to predict recidivism risk.

COMPAS produces a "risk score", on a scale from 0 to 10. We use the risk scores to assess COMPAS / machine confidence (i.e., inverse of the difficulty) in giving advice for defendants on a scale of 0 to 4. Specifically, we map defendants with extreme risk scores of 0 and 10 to the highest confidence (least difficulty) score of 0, and map the defendants with a middle risk score of 5 to the lowest confidence (highest difficulty) score of 4.

However, COMPAS does not reveal how it calculates the risk score. It is therefore not clear what exactly the predicted risk score represents. We replace it with the reconstructed classifier, which we know to predict precisely what we ask our participants to predict, namely the risk that defendant recidivates within two years. Given this classifier is trained on the ground truth data that ProPublica has retrieved, we are also able to exactly measure its accuracy rate. This is how we measured the accuracy rate of 68%, which we reported to the participants in the stimulus material. Hence with this design choice, we test machine advice as it truly performs in the field.

Conduct. We run the experiment on Prolific – an online crowdworking platform, similar to Amazon's Mechanical Turk (MTurk). Unlike MTurk, Prolific was explicitly designed for online subject recruitment for the scientific community [53]. Prolific's advanced pre-screening capabilities allow us to select participants from the United States that have self-reported to have served on a jury. For study 1, we randomly recruit 103 participants from this pool of participants in their

[2]For the exact wording and layout of these questions, please see the screenshots in the appendix.

Demographic Attribute	Study 1	Study 2	Study 3	Census
Total respondents	108	112	108	-
Passed attention checks	103	100	100	-
Male	49%	55%	56%	49%
African American	5%	11%	14%	13%
Asian	6%	4%	6%	6%
Caucasian	82%	68%	68%	61%
Hispanic	6%	12%	6%	18%
Other	1%	5%	6%	4%
Bachelor's Degree or above	60%	58%	58%	30%
Liberal	54%	50%	47%	33%†
Conservative	26%	17%	16%	29%†
Moderate	17%	29%	29%	34%†
Other	3%	4%	8%	4%†
Jury Duty Experience	74%	79%	72%	27%‡

Table 1. Demographics of our survey samples, compared to the 2016 U.S. Census [70]. Attributes marked with a † were compared to Pew data [58] for political leaning. Attributes marked with a ‡ were compared to DRI 2012 National Poll data [25]. We also report the total number of respondents who participated in our studies, as well as the number of respondents who successfully answered both attention check questions. The reported demographics concern the respondents who passed the attention checks.

Prolific user profiles. For each of the five treatments reported in study 2, and for each of the five treatments reported in study 3, we have 20 participants, selected in the same way.

After taking part in the experiment, participants answer a set of questions about their demographics. Their answers are shown in Table 1. Our 3 samples consists of more Caucasians than the U.S. population (82%, 68% and 68%, compared to 61%). Our participants are also more highly educated, with close to 60% holding at least a bachelor's degree, compared to 30% in the 2016 U.S. Census. Also, our participants are more liberal leaning (54%, 50% and 47% compared to 33%). The skews in education and ethnicity are consistent with the findings from prior studies on crowdworking platform demographics [41, 54, 57]. Even though we targeted only participants that have self-reported that they have served on a jury in their Prolific profiles, only 74% (study 1), 79% (study 2) and 72% (study 3) of them reported that this is the case in our demographic survey.

In Study 1, participants receive £2 for their participation, but are not incentivized for the predictions they make. The experiment takes approximately 25 minutes. In order to make sure that participants take the task seriously, they have to answer 2 attention check questions. We discard the answers of all of the respondents who did not successfully answer both attention check questions. The design of the experiment has been approved by our institution's Ethical Review Board.

2.3 Results

Figure 1 summarizes predictions made in Study 1. While for certain defendants only very few participants predict that the defendant will commit a new crime, in other cases all of the participants predict recidivism, and decisions in the intermediate cases are well balanced. Descriptively, machine advice does not have a strong effect. For most cases, the fraction of participants who predict recidivism is very similar with and without advice. There are exceptions though, in particular for cases that are closer to the endpoints of the distribution.

Our participants achieved an accuracy of 60.2%, which is close to the accuracy of 62.1%, reported by [24] in the same setting. Our respondents are less accurate when responding without observing

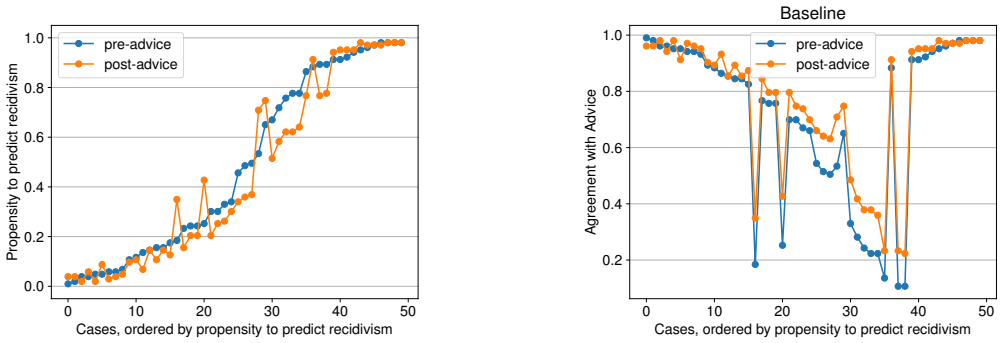

Fig. 1. Predictions Made With and Without Machine Advice. Cases ordered by fraction of participants who predict that the defendant will recidivate when not having machine advice (blue line). Fraction of participants who predict recidivism upon receiving advice superimposed (red line). [Left] Y-axis: fraction of participants who predict recidivism. [Right] Y-axis: fraction of participants who make a prediction which corresponds to the machine's advice.

	Model 1		**Model 2**		**Model 3**	
recidivism prediction	.593***	(.012)	.475***	(.021)	.684***	(.028)
ground truth			-.019	(.014)	-.059***	(.014)
recidivism prediction * ground truth			.167***	(.027)	.179***	(.027)
difficulty					.072***	(.005)
recidivism prediction * difficulty					-.121***	(.009)
cons	.268***	(.012)	.275***	(.013)	.176***	(.014)

Table 2. Effect of advice, ground truth and COMPAS confidence. dv: dummy that is 1 if participant predicts that defendant will recidivate; recidivism prediction: dummy that is 1 if machine predicts recidivism; ground truth: dummy that is 1 if defendant recidivates; difficulty (to give advice) on a scale from 0 .. 4. Linear probability model with participant random effect. Number of observations, cases, participants are 5150, 50, and 103 respectively. Standard errors in parenthesis, *** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

advice (60.2%), than they are with advice (61.6%), and in both cases they are less accurate than the decision aid (68%), which is also consistent with the findings of [35].

Upon receiving advice, in 390 of 5150 cases the participants change their prediction. For statistical analysis, we take into account the fact that each participant makes 50 predictions. This is why we estimate a model with a random effect for participants. We estimate linear probability models, as coefficients can then be directly interpreted, and as we keep the ability to add interaction terms [2].

For testing Hypothesis 1 we estimate a constant only model, and use a dummy as the dependent variable that is 1 if the participant has changed her prediction upon receiving machine advice. This regression estimates that, in the population, 7.6% will change their decision in the light of machine advice (p < .001).[3] This gives us

Result 1: In a minority of cases, participants react to machine advice by changing whether they predict recidivism.

[3]The result stays the same if we additionally add a random effect for cases to the statistical model, to capture potential dependence across participants that originates in the fact that cases trigger systematically different responses.

Machine \ Respondent	pre-advice		post-advice	
	recidivism	no recidivism	recidivism	no recidivism
recidivism	1339	309	1419	229
no recidivism	1149	2353	939	2563

Table 3. Contingency tables, describing the relationship between machine advice and the respondents' pre-advice and post advice-decisions. Number of observations, cases, participants are 5150, 50, and 103 respectively.

The fact that machine advice has little influence might, at least partly, be due to the fact that we make the relatively low accuracy of the advice explicit. Participants might read this as a hint not to take the advice seriously. However, the accuracy rate of only 68% closely matches the accuracy of the COMPAS tool [4, 24], and in a recent decision the Supreme Court of Wisconsin has mandated that such information be given to judges if they receive advice from the COMPAS tool (Wisconsin Supreme Court 2016 WI 68). We are thus testing the effect that the advice is likely to have in the field.

As noted, our respondents change their prediction for 390 out of the 5150 cases. However, our respondents had a high rate of agreement with the machine advice to start with. The agreement rates differ amongst respondents, and are detailed in Figure 1 [Right]. Namely, before observing machine advice, 3692 out of the 5150 responses (71.7%) concur with the advice, while 1458 differ. After observing machine advice, 3982 out of the 5150 responses (77.3%) concur with the advice. In other words, observing advice resolves 290 out of the 1458 pre-advice disagreements (19.9%). The detailed contingency tables can be found in Table 3.

In 340 of the 390 cases in which participants change their prediction after receiving machine advice, they follow the advice, while they chose the opposite of the advice in 50 cases. We test individual average switches of these 103 participants against the equal split, and find that they are more likely to follow the advice (signrank test, $p < .001$). We thus have support for Hypothesis 2, which gives us

Result 2: Participants are more likely to change their prediction in the direction of the received machine advice.

As we predicted in Hypothesis 3, participants are indeed able to infer the quality of the machine advice from the information about the offence and the defendant. We focus on Model 3 of Table 2 as this turns out to be most informative.[4] If the machine predicts that the defendant will not recidivate, the machine is perfectly confident, and the defendant has actually not recidivated according to the ground truth data, 17.6% of the participants nonetheless predict that the defendant will recidivate (constant). If the machine is one out of five steps less confident in its advice, 24.8% predict recidivism (main effect of difficulty). Interestingly, if the machine has predicted no recidivism, and has been perfectly confident, but the defendant has actually recidivated according to the ground truth data, participants are only 11.7% likely to predict recidivism (main effect of ground truth). Yet if the available information suggests that the machine might be less confident by one step, the counterintuitive effect is already neutralized, and participants are 18.9% likely to predict recidivism (main effect of ground truth and difficulty).

Upon receiving advice, in 130 cases participants shift towards predicting recidivism, and in 260 cases they shift towards predicting no recidivism. As predicted by Hypothesis 5, the advice that the defendant will not recidivate has a stronger effect than the advice that he will recidivate. The

[4]The effects of ground truth, and its interaction with advice to jail, are insignificant if we add a case random effect to the specification. All other effects remain significant, with $p < .015$.

	Model 1		Model 2	
shift	1.068***	(.066)	.326***	(.060)
advice to predict recidivism	.564***	(.033)		
shift * advice to predict recidivism	-2.440***	(.134)		
agreement			1.005***	(.034)
shift * agreement			-.285+	(.163)
cons	5.233***	(.062)	4.639***	(.067)

Table 4. Effect of Machine Advice on Confidence. dv: confidence, on a scale from 1 to 7, in prediction, after seeing machine advice; shift: participant changes prediction upon receiving machine advice (-1 = shift towards predicting no recidivism, 1 = shift towards predicting recidivism, 0 = no change) ; advice to predict recidivism: dummy that is 1 if machine predicts recidivism; agreement: participant's pre-advice prediction concurs with machine advice. Linear probability models with participant random effect. Number of observations, cases, participants are 5150, 50, and 103 respectively. Standard errors in parenthesis, *** p < .001, ** p < .01, * p < .05, + p < .1

probability that a participant shifts towards predicting no recidivism upon receiving the advice to do so is 8.45%. By contrast the probability that a participant shifts towards predicting recidivism upon receiving the advice to do is only 5.7%. This difference is significantly different from zero.[5] We conclude

Result 3: Participants are more likely to change their decision upon receiving machine advice if the machine predicts that defendant will not recidivate.

We further test the effects of ground truth and machine confidence on participants' decisions. Even though the participants do not observe this information, it seems that they are able to infer it. If the machine advises to predict recidivism and is perfectly confident, but this is at odds with the ground truth, participants are nonetheless 86% likely to predict that the defendant will recidivate (main effect of advice to predict recidivism). Yet if the machine is 1 out of 5 steps less confident, this probability is reduced to 81.1% (main effects of advice and of difficulty and their interaction). If participants are advised to predict recidivism, and the machine is perfectly confident, and this advice is in line with the ground truth, participants are 98% likely to follow this advice (main effects of advice and ground truth and their interaction). The lower the machine's confidence, the more this probability is reduced (by 4.9% for each degree of uncertainty, main effect of difficulty and interaction). We conclude

Result 4: Participants' decisions are affected by ground truth and machine confidence. The more difficult a case is for COMPAS to decide, the more participants are likely to deviate from the advice.

As the regressions of Table 4 show, participants are pretty confident overall.[6] If the machine advises to predict that a defendant will not recidivate and this is also the participant's prediction, participants indicate a confidence level of 5.233, on a scale from 1 to 7. Their confidence reduces by 1.068 points if they initially predicted recidivism. If they follow the machine advice and shift towards predicting recidivism, their confidence is reduced by .808 (main effect of advice to predict recidivism and interaction). Model 2 further analyses the effect of receiving advice which concurs or does not concur with the respondents' pre-advice prediction on their reported post-advice confidence. If the advice is in line with their original prediction, this makes them more confident. These regressions support Hypothesis 4, which gives us

[5]In a statistical model with a participant random effect, p < .001; if we further add a random effect for cases, p = .089.
[6]For a graphical representation, see the Appendix.

Result 5: If respondents receive advice which concurs with their pre-advice prediction, they become more confident, while receiving opposing advice leads to a decrease in confidence.

3 STUDY 2: IS MACHINE ADVICE MORE EFFECTIVE WITH FEEDBACK?

3.1 Motivation and Hypotheses

For a trial judge, the decision whether to set a defendant free on bail is a routine decision. In the jurisdictions that give the judge access to machine advice, judges thus repeatedly receive advice from the same source. They do not get feedback in each and every case. But it is not unlikely that a judge over time learns whether a defendant has reoffended, for instance as the judge has to decide about now incarcerating him. In some jurisdictions, case assignment is stable over time, and for instance tied to the defendant's name, or to his residence. That gives the judge additional opportunities for learning whether her prediction had been correct. Through such feedback, the judge may over time gain a better sense of the reliability of the COMPAS tool, both in general, and for specific types of cases or defendants. These considerations straightforwardly translate into

> **Hypothesis 6** If recipients receive feedback about the accuracy of machine advice, they are more likely to rely on the advice.

Now as we have repeatedly stressed, our decision aid is only mildly reliable, with average accuracy of 68%, close to the accuracy of the COMPAS tool [4, 24]. The perceived quality of the advice will therefore depend on the sequence of cases that the individual judge had to decide. Arguably if its predictions have mostly been correct in the beginning of the sequence, the judge will be more likely to trust the advice than if she has initially received feedback that mainly ran counter to the machine advice. This gives us

> **Hypothesis 7** If recipients initially mostly receive feedback in line with the machine advice, they are more likely to rely on the advice. If they initially mostly receive feedback in opposition to the machine advice, they are less likely to rely on the advice.

In Study 1, we found support for Hypothesis 5: respondents are more likely to follow the advice if the machine predicts that the defendant will not recidivate. Now this finding might be due to the specific parameters of Study 1. We had only 18/50 cases in which our tool predicts recidivism. To test whether the finding is robust, we did not want to change the set of cases. That would have made the results vulnerable to confounds, as not only the frequency of the recommendation to predict recidivism would change, but also the case characteristics. We avoid the confound by merely changing the order in which participants see the cases. This is analogous to a court that is predominantly presented with defendants with a long criminal history, or with particularly egregious charges. Both may happen if case selection is not random. This yields

> **Hypothesis 8** If recipients initially mostly learn that defendants have recidivated, they are more likely to predict recidivism. If they initially mostly learn that defendants have not recidivated, they are less likely to predict recidivism.

3.2 Design

Study 2 builds on the design of Study 1. It complements the within subjects design of Study 1 with a between subjects design. For testing Hypothesis 6, Study 1 is the *baseline*. We add the *feedback* treatment in which, after having recorded their final prediction, participants learn the ground truth. This is of course more and better feedback than a real judge would ever see. Real judges at best learn for a few select cases whether the defendant recidivated. They are more likely to get feedback if they have released the defendant, in particular if the prosecution presents the defendant again after having apprehended him for a new crime. Feedback will never be immediate. For all these reasons, with this comparison we estimate only an upper bound on learning.

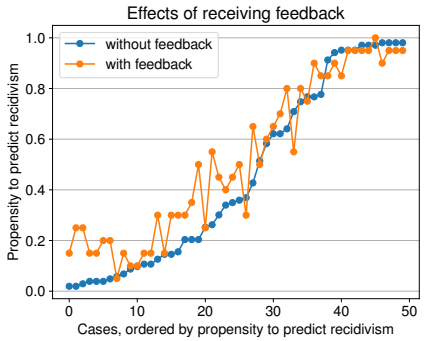

Fig. 2. Effect of receiving feedback. Comparing patterns of post advice predictions between the *baseline* (without feedback) and the *feedback* treatment. X-axis shows defendants, sorted by participants' propensity to predict that they will recidivate (in the baseline). Y-axis: fraction predicting recidivism.

In the *feedback* treatment, participants see the 50 cases in the same sequence as in the *baseline*. In the *good advice* treatment, participants initially see mostly cases in which the machine advice is in line with ground truth. In the *bad advice* treatment, participants initially see mostly cases in which the machine advice is in opposition to ground truth. The precise sequences are reported in the appendix. We intersperse a few cases with incorrect advice in the former, and with correct advice in the latter treatment. Otherwise the sequence would have seemed utterly unrealistic. Also, the participants would have noticed the pattern too easily. In both treatments, we inform participants about the ground truth after every case. This is why we compare these treatments to the *feedback* treatment (with random order).

By the same token, in the *bail advice* treatment, participants initially see mostly cases in which our tool advises to not predict recidivism. In the *jail advice* treatment, participants initially see mostly cases in which our tool advises to predict recidivism. Participants are again informed about ground truth after every case.

In Study 2, participants receive £2.5 for their participation. Due to receiving feedback, this experiment took slightly longer than the one in study 1 – approximately 30 minutes. We ran Study 2 on the same online platform as Study 1, and participants were randomly selected from the same subject pool. Participants only participated in one treatment. For each treatment we had 20 participants. The demographics of the participants are shown in Table 1.

3.3 Results

Figure 2 shows that giving participants feedback has no discernible effect. A fresh group of participants, upon receiving machine advice, essentially makes the same predictions as their predecessors in the *baseline*. We also do not find a significant effect of the *feedback* treatment on participants' propensity to change their prediction upon receiving advice.[7] We thus have no support for Hypothesis 6. This is remarkable. As explained, the learning opportunity is much more powerful than it could ever be in the field. It is also reassuring that fresh participants judge individual cases in very similar ways.

Figure 3 shows that the propensity to predict that defendant will recidivate is also not influenced by the sequence in which participants see the cases. Neither initially mostly seeing correct, or mostly incorrect evidence matter (left panel), nor does initially seeing mostly cases where our tool

[7]The regression is available from the authors upon request.

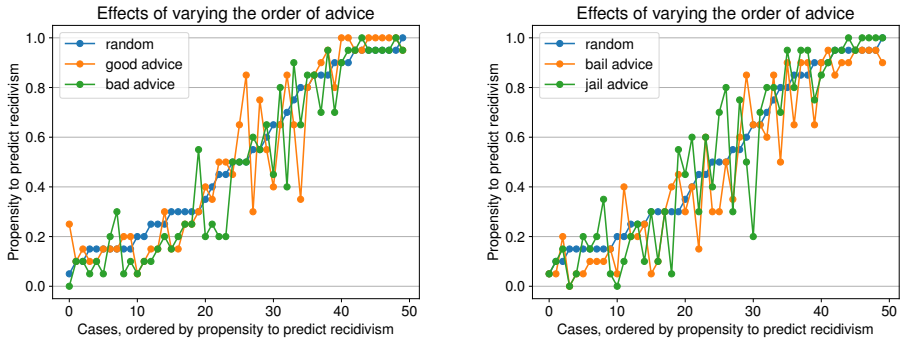

Fig. 3. Effect of varying the order in which the advice is shown. X-axis shows defendants, sorted by participants' post-advice propensity to predict recidivism, in the treatment where the questions are shown in random order. Y-axis: fraction of participants who predict recidivism, after receiving machine advice.

predicts recidivism, or no recidivism (right panel). We also do not find any significant effects of the *good advice*, *bad advice*, *bail advice* or *jail advice* treatments on participants' propensity to change their prediction upon receiving advice. This also holds if we separate effects by the first 20 trials (where the cases are sorted in the direction of the treatment) and the remaining 30 trials (where they are sorted in the opposite direction).[8] Hence we also do not have any support for Hypotheses 7 and 8. Feedback is not critical for the effect of machine advice. Human agents do not have to learn (when) to trust machine advice.

4 STUDY 3: IS MACHINE ADVICE MORE EFFECTIVE WHEN STAKES ARE HIGHER?

4.1 Motivation and Hypotheses

Arguably, policymakers do not only equip judges with machine advice in order to economize on judicial labour. The political intention is not confined to increasing judicial efficiency. Deploying machine advice would be normatively appealing if it (also) increases accuracy. This would help protect victims and defendants alike. From this perspective, two findings from Study 1 may be troublesome: (a) the effect of machine advice is small: most of the time participants stick to the choice they have made before receiving the advice. (b) The effect of machine advice is biased: participants are more likely to follow the advice if the machine recommends to predict that defendant will not recidivate.

Hopefully, judges are predominantly motivated intrinsically. They are aware of the fact that they exercise sovereign powers, in ways that can be very consequential for people's lives. That this expectation is not unfounded can even be shown experimentally [28]. Yet a substantial literature shows that judges are not immune to incentives. Many judges in particular care about advancing their careers [63], and make decisions that they expect to improve their chances for promotion [16] or reelection [13]. Judges have also been shown to be sensitive to more subtle incentives. They care for prestige [17] and reputation [51]. Judges also tend to be averse to their decisions being overturned on appeal [60], and they are wary about the risk that legislators might curtail the judiciary's budget [3].

For obvious reasons, we cannot manipulate any of these incentives in our experiment. But there is a straightforward way of making predictions more consequential for participants: we can give them financial incentives. In the field, incentives are of course more subtle. As with feedback,

[8]These regressions are available from the authors upon request.

	Correct		Incorrect			
			False Positive		False Negative	
	Aligned	Not Aligned	Aligned	Not Aligned	Aligned	Not Aligned
Baseline	0	0	0	0	0	0
Ground Truth	.2	.2	-.2	-.2	-.2	-.2
False Positive	.2	.2	-.5	-.5	-.1	-.1
False Negative	.2	.2	-.1	-.1	-.5	-.5
Weak Alignment	.1	.5	.1	-.2	.1	-.2
Strong Alignment	.1	.2	.1	-.5	.1	-.5

Table 5. Treatments in Study 3. The values indicate the size of the monetary incentives used, in £.

with our manipulation we test the upper bound. If even a financial incentive that is immediate – and directly tied to the prediction made in the concrete case – is not effective, a fortiori the more subtle incentives that policymakers might use as levers will also be ineffective. In Study 3, we thus investigate whether machine advice can be made more effective by increasing stakes.

We ultimately want to learn whether exogenous interventions moderate the effect of machine advice on the predictions participants make. Yet what looks like moderation could effectively be an effect of incentives on predictions, even if they are made without the benefit of machine advice. We therefore begin by testing the following hypothesis:

> **Hypothesis 9** If participants are incentivized to avoid wrongful detention, they are less likely to predict that defendant will recidivate, even before having access to machine advice. Conversely, if they are incentivized to avoid that defendant commits new crime before being tried, they are more likely to predict that he will recidivate, even before having access to machine advice.

Our main interest is, however, in the effect of incentives on the sensitivity towards machine advice, which we expect to work as follows:

> **Hypothesis 10 a** If participants are incentivized to match ground truth, they are more likely to follow machine advice.
> **b** If participants are incentivized to avoid wrongful detention, they are more likely to follow the advice of the machine to predict that the defendant will not recidivate.
> **c** If participants are incentivized to avoid that the defendant commits a new crime before being tried, they are more likely to follow the advice of the machine to predict recidivism.
> **d** If participants are incentivized to follow the machine's advice, they will be more likely to do so, and the stronger the incentive the more they will follow it.

4.2 Design

Study 3 uses the exact same design as Study 1, except that choices are incentivized. Thus, as in Study 2, we complement the within subjects design of Study 1 with a between subjects design. In this between subjects comparison, we treat Study 1 as the *baseline*. Respondents earn a base payment of £2 for completing the survey. At the beginning of the survey, participants are informed that they can earn an additional monetary reward, based on their performance. Table 5 summarizes our manipulations.

In the *ground truth* treatment, for each case participants gain an extra £.2 if their choice is in line with ground truth, i.e. if they predict no recidivism if the defendant did not recidivate, and recidivism if he did. They lose £.2 if they get it wrong. If the bonus earned after answering all 50 questions is negative, no money is deducted from the base earnings of £2.

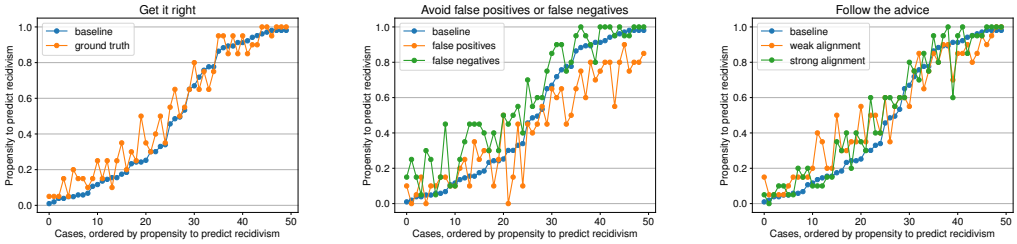

Fig. 4. Effect of Incentives on Predictions pre Advice. Cases ordered by fraction of participants who predict recidivism when not having machine advice and with no incentives (blue line). Fraction of participants who predict recidivism when choice is incentivized superimposed (different colors). Y-axis: fraction of participants who predict recidivism.

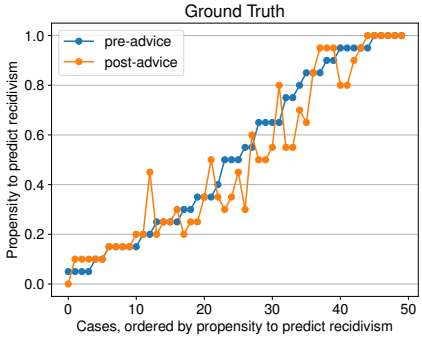

Fig. 5. Predictions Made Without and With Machine Advice with Incentive for Ground Truth. Cases ordered by fraction of participants who predict that defendant recidivates when not having machine advice (blue line). Fraction of participants who predict revidivism upon receiving advice superimposed (red line). Y-axis: fraction of participants who predict recidivism.

In the *false positive* and *false negative* treatments, predictions in line with ground truth also give them an extra £.2. But now different mistakes are incentivized differently. In the *false positive* treatment, £.5 are subtracted from their bonus if they predict recidivism although the defendant would not have recidivated. In contrast, predicting that a defendant will not recidivate who actually has only costs them £.1. In the *false negative* treatment, incentives are swapped, and predicting no recidivism if defendant actually recidivated costs participants £.5, while predicting that a defendant will recidivate who actually has not only costs them £.1.

In the final two treatments, participants always earn an extra £.1 if their prediction is in line with machine advice, irrespective of ground truth. Deviating from machine advice is incentivized two different ways. In the *weak alignment* treatment, participants earn a premium of £.5 when deviating from the advice and they are right (in line with ground truth). If they deviate and are wrong, they lose £.2 (irrespective of whether this is a false positive or a false negative). By contrast in the *strong alignment* treatment, correctly deviating from the advice only gives them an extra £.2, while incorrectly deviating costs them £.5.

We ran Study 3 on the same online platform as Study 1 and Study 2. Participants were randomly selected from the same subject pool. Participants only participated in one treatment. For each treatment we had 20 participants. The demographics of the participants are shown in Table 1.

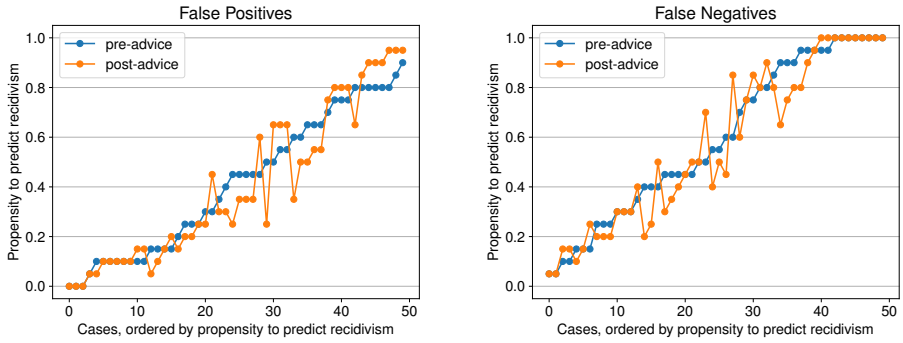

Fig. 6. Predictions Made Without and With Machine Advice with Incentive to Avoid False Positives or False Negatives. Cases ordered by fraction of participants who predict recidivism when not having machine advice (blue line). Fraction of participants who predict recidivism upon receiving advice superimposed (red line).

4.3 Results

The plots of Figure 4 compare the choices the participants have made without the benefit of advice, either with or without incentives. The left panel shows that unbiased incentives have virtually no effect. Participants want to get it right anyways. The incentive has an effect, though, if participants lose money when predicting that a defendant recidivates who actually would not. The middle panel of Figure 4 and a linear regression with participant random effect, explaining prediction before advice with treatment, where the baseline from Study 1 is the reference category, show that the prediction that defendant will recidivate is 6.8% less likely in the *false positive* treatment, p = .030[9]). There is an effect in the opposite direction in the *false negative* treatment, where the prediction that defendant will recidivate is 10.9% more likely (p = .001). Descriptively, whether or not the incentive to align with the advice is *weak* or *strong*, participants are somewhat more likely to predict that defendant will not recidivate (right panel of Figure 4). Yet these effects are far from significance (p = .252 and 239, respectively). We thus have support for Hypothesis 9 and conclude

Result 6: If participants are incentivized to avoid false positive predictions, they are less likely to predict recidivism. If they are incentivized to avoid false negative predictions, they are more likely to predict recidivism.

Comparing Figure 5 with Figure 1 suggests that participants' sensitivity to advice does not change if they are incentivized for finding the ground truth. This is indeed what we find in the regression of Table 6.[10] This runs counter to Hypothesis 10a.

The two panels of Figure 6 show that the incentive to avoid *false positive* or *false negative* choices does not induce participants to rely more on machine advice, counter to our Hypotheses 10 b and 10 c.

On the other hand, incentives to align with machine advice have the effect predicted by Hypothesis 10 d, as shown in Figure 7 and Table 6.[11] However, the effect is not more pronounced if the incentive is stronger (Wald test comparing the two treatment effects, p = .495). We thus have partial support for Hypothesis 10 d and conclude

Result 7: If participants have a (weak or strong) incentive to align with machine advice, they are more likely to change their decision after receiving the advice.

[9]Coefficient and p-value stay the same when adding a case random effect to the specification.
[10]All coefficients and all p-values stay the same if we add a case random effect to the specification.
[11]All significant effects in Table 6 remain significant if we add a case random effect to the specification.

ground truth	-.002	(.019)
false positive	-.016	(.019)
false negative	.010	(.019)
weak alignment	.062***	(.019)
strong alignment	.079***	(.019)
cons	.076***	(.008)

Table 6. Effect of Treatment on Sensitivity Towards Advice. dv: dummy that is 1 if participant changes prediction after learning advice. Linear probability model with participant random effect. Number of observations, cases, participants are 10150, 50, and 203 respectively. Standard errors in parenthesis, *** p < .001, ** p < .01, * p < .05, + p < .1

	no incentive	ground truth	false positive	false negative	weak align.	strong align.
ACC	61.6%	59.6%	61.1%	58.4%	61.3%	64.5%
FPR	.348	.403	.302	.490	.385	.296
FNR	.423	.404	.483	.335	.389	.418

Table 7. Effect of Treatments on Participants' Performance in Terms of Accuracy (ACC), False Positive rate (FPR) and False Negative Rate (FNR).

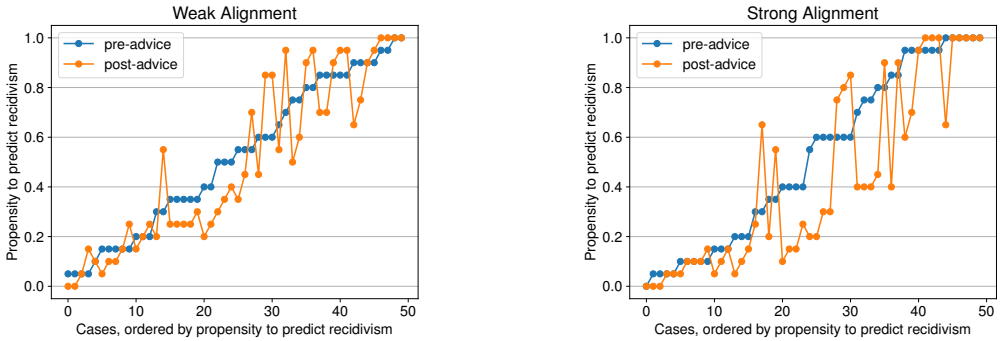

Fig. 7. Choices Made Without and With Machine Advice with Incentive to Align with Advice. Cases ordered by fraction of participants who predict recidivism when not having machine advice (blue line). Fraction of participants who predict recidivism upon receiving advice superimposed (red line).

Table 7 summarizes the effects of different incentive schemes on participants' performance. Incentivizing participants to match the ground truth or to avoid false positive or false negative mistakes does not lead to an increase in accuracy. Yet the intervention leads to a different distribution of mistakes – lower false positive rates for the former (from .348 to .302) and lower false negative rates for the latter (from .423 to .335). Strong alignment incentives lead to an increase in accuracy (from 61.6% to 64.5%), with most of this increase in performance visible in lower false positive rates (from .348 to .296).

5 CONCLUDING DISCUSSION

The more consequential a decision, the less acceptable it is to delegate it to a machine, even if the machine has been trained on a rich dataset, and has been shown to have good accuracy, maybe

even higher accuracy than human decision makers entrusted with the same decision. One such decision is the choice between sending a suspect to jail, versus releasing him on bail. In most legal orders, making this choice is the prerogative of the competent judge. Yet arguably the accuracy of judicial decision making can be improved by giving the judge access to a machine's prediction of the risk that the defendant will recidivate before being tried, if let out on bail. This is not a contrived example; multiple US States give judges access to precisely this prediction, based on the commercial product COMPAS, sold by Northpointe.

We construct vignettes from 50 randomly selected cases from the COMPAS dataset. For all cases we know ground truth. In Study 1, we test laypersons on these vignettes, both without and with the benefit of advice. Advice has a significant, but fairly small effect. Yet this effect is asymmetric: participants are more likely to follow the advice if the machine predicts that the defendant will not recidivate, and hence proposes to put the defendant on bail. In two supplementary experiments we test whether the impact of machine advice can be increased, and whether it can be debiased. Giving participants feedback about ground truth does not prove effective. A financial incentive (meant to emulate higher stakes) only matters if it is tied to following the advice, not if it is tied to the quality of the advice, or to its direction. Using this approach, we are the first to provide causal evidence about the effect of machine advice on the accuracy of human decision making, in the judicial domain.

The main limitation of the study is the subject pool. For obvious ethical reasons, we could not study how real judges decide real cases. It would also have been problematic to test real judges on hypothetical cases, as we could not control for the varied experiences of the different judges. This is why we turn to laypersons who are randomly drawn from the general population. Yet we have made sure that participants are eligible for jury duty, and that most of them have actually served as jury members. They thus have experienced the judicial system from within. We further note that our participants exhibited a prediction accuracy comparable to the one reported in prior work, and that their reported confidence varied across cases. Both suggest that they have taken the task seriously, despite the fact that we could only ask how they would decide, were they to be given authority.

Machine advice is not confined to bailing defendants. Other illustrations include decisions about loans, social security benefits, or hiring applicants. Although it is not normally thought of as machine advice, targeted advertising can also be brought under this rubric. It would be an interesting topic for future work to compare machine advice in such domains with its effect in the judicial domain.

In our study, we have varied the properties of the task, by giving differential feedback, and by changing the monetary rewards and penalties associated with making certain types of decisions. We have made no changes to the decision aid nor to the advice it provided. In future work, we plan to study the effects of varying the properties of the decision aid and of the advice as well. The advice taking literature points us towards properties which seem worth studying. The advice might be differently timed or framed. Additional information, for instance about the performance of the decision aid in specific classes of cases, its confidence in its decisions, or explanations about how it derived the advice, might influence the degree of trust. The degree of trust might also depend on how the advisee has learned this information about the performance of the tool – by having access to statistical evaluations they find differently easy to interpret, or through personal experience. It would also be interesting to bring alternative incentives for accuracy, or for being sensitive to machine advice, to the lab. Another dimension worth studying is the type of information about accuracy which is provided. One might for instance study whether machine advice is processed differently when it comes with accuracy rates per type of crime, or per defendant characteristics (like gender or race). Finally, building up on prior work which found that statistical advice has

a smaller effect than human advice, it would be interesting to look into the root causes of this discrepancy.

Machine advice has to be properly engineered. We see our experiment as a first step towards designing powerful interfaces between advice-giving machines and advice-receiving human decision makers.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Amanda Y Agan and Sonja B Starr. 2016. Ban the Box, Criminal Records, and Statistical Discrimination: A Field Experiment. (2016).
[2] Chunrong Ai and Edward C Norton. 2003. Interaction Terms in Logit and Probit Models. *Economics Letters* (2003).
[3] Gary M Anderson, William F Shughart, and Robert D Tollison. 1989. On the Incentives of Judges to Enforce Legislative Wealth Transfers. *Journal of Law and Economics* (1989).
[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[5] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine Experiment. *Nature* (2018).
[6] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* (2016).
[7] Björn Bartling, Ernst Fehr, and Holger Herz. 2014. The Intrinsic Value of Decision Rights. *Econometrica* (2014).
[8] Björn Bartling and Urs Fischbacher. 2011. Shifting the Blame: On Delegation and Responsibility. *The Review of Economic Studies* (2011).
[9] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* (2018).
[10] David M Bersoff. 2001. Why Good People Sometimes Do Bad Things: Motivated Reasoning and Unethical Behavior. In *The Next Phase of Business Ethics: Integrating Psychology and Ethics*. Emerald Group Publishing Limited.
[11] Silvia Bonaccio and Reeshad S Dalal. 2006. Advice Taking and Decision-Making: An Integrative Literature review, and Implications for the Organizational Sciences. *Organizational Behavior and Human Decision Processes* (2006).
[12] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The Social Dilemma of Autonomous Vehicles. *Science* (2016).
[13] Paul Brace, Melinda Gann Hall, and Laura Langer. 1998. Judicial Choices and the Politics of Abortion: Institutions, Context, and the Autonomy of Courts. *Albany Law Review* (1998).
[14] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* (2017).
[15] Angèle Christin. 2017. Algorithms in Practice: Comparing Web Journalism and Criminal Justice. *Big Data & Society* (2017).
[16] Marc A. Cohen. 1992. The Motives of Judges: Empirical Evidence from Antitrust Sentencing. *International Review of Law and Economics* (1992).
[17] Robert D Cooter. 1983. The Objectives of Private and Public Judges. *Public Choice* (1983).
[18] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv preprint arXiv:1808.00023* (2018).
[19] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *KDD*.
[20] Jacob W Crandall, Mayada Oudah, Fatimah Ishowo-Oloko, Sherief Abdallah, Jean-François Bonnefon, Manuel Cebrian, Azim Shariff, Michael A Goodrich, Iyad Rahwan, et al. 2018. Cooperating with Machines. *Nature Communications* (2018).
[21] John M Darley and Bibb Latané. 1968. Bystander Intervention in Emergencies: Diffusion of Responsibility. *Journal of Personality and Social Psychology* (1968).
[22] Mandeep K Dhami, Samantha Lundrigan, and Katrin Mueller-Johnson. 2015. Instructions on Reasonable Doubt: Defining the Standard of Proof and the Juror's Task. *Psychology, Public Policy, and Law* (2015).

[23] Christos Dimitrakakis, David C Parkes, Goran Radanovic, and Paul Tylkin. 2017. Multi-View Decision Processes: The Helper-AI Problem. In *NeurIPS*.

[24] Julia Dressel and Hany Farid. 2018. The Accuracy, Fairness, and Limits of Predicting Recidivism. *Science Advances* (2018).

[25] DRI. 2012. DRI 2012 National Poll on the Civil Justice System. http://www.dri.org/advocacy/center-for-law-and-public-policy/poll/2012-poll/

[26] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, and Omer Reingold. 2012. Fairness Through Awareness. In *ITCSC*.

[27] Christoph Engel, Andreas Glöckner, and Sinika Timme. 2017. Defendant Should Have the Last Word – Experimentally Manipulating Order and Provisional Assessment of the Facts in Criminal Procedure. *MPI Collective Goods Preprint* (2017).

[28] Christoph Engel and Lilia Zhurakhovska. 2017. You are in Charge: Experimentally Testing the Motivating Power of Holding a Judicial Office. *Journal of Legal Studies* (2017).

[29] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *KDD*.

[30] Shi Feng and Jordan Boyd-Graber. 2019. What can AI do for me: Evaluating Machine Learning Interpretations in Cooperative Play. *IUI* (2019).

[31] Anthony W. Flores, Christopher T. Lowenkamp, and Kristin Bechtel. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks.". (2016).

[32] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of Fairness. *arXiv:1609.07236* (2016).

[33] Brandon L Garrett and John Monahan. 2019. Judging Risk. *California Law Review* (2019).

[34] Gerd Gigerenzer and Adrian Edwards. 2003. Simple Tools for Understanding Risks: From Innumeracy to Insight. *British Medical Journal* (2003).

[35] Ben Green and Yiling Chen. 2019. Disparate interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *FAT**.

[36] Nina Grgić-Hlača, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *WWW*.

[37] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In *NIPS Symposium on Machine Learning and the Law*.

[38] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *AAAI*.

[39] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of Opportunity in Supervised Learning. In *NeurIPS*.

[40] Robin M Hogarth and Emre Soyer. 2015. Providing Information for Decision Making: Contrasting Description and Simulation. *Journal of Applied Research in Memory and Cognition* (2015).

[41] Panagiotis G Ipeirotis. 2010. Demographics of Mechanical Turk. (2010).

[42] Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision Under Risk. In *Econometrica*.

[43] Faisal Kamiran and Toon Calders. 2010. Classification with no Discrimination by Preferential Sampling. In *BENELEARN*.

[44] Oliver Kirchkamp and Christina Strobel. 2019. Sharing Responsibility with a Machine. *Journal of Behavioral and Experimental Economics* (2019).

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

[46] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *ITCS*.

[47] Ziva Kunda. 1990. The Case for Motivated Reasoning. *Psychological Bulletin* (1990).

[48] Joseph CR Licklider. 1960. Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics* (1960).

[49] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. 2017. Calibrated Fairness in Bandits. *arXiv preprint arXiv:1707.01875* (2017).

[50] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. kNN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. In *KDD*.

[51] Thomas J Miceli and Metin M Coşgel. 1994. Reputation and Judicial Decision-Making. *Journal of Economic Behavior & Organization* (1994).

[52] Dilek Önkal, Paul Goodwin, Mary Thomson, Sinan Gönül, and Andrew Pollock. 2009. The Relative Influence of Advice from Human Experts and Statistical Methods on Forecast Adjustments. *Journal of Behavioral Decision Making* (2009).

[53] Stefan Palan and Christian Schitter. 2018. Prolific.ac – A Subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance* (2018).

[54] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making* (2010).
[55] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* (2011).
[56] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware Data Mining. In *KDD*.
[57] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research. *Journal of Experimental Social Psychology* (2017).
[58] Pew Research Center. 2016. 2016 Party Identification Detailed Tables. http://www.people-press.org/2016/09/13/2016-party-identification-detailed-tables/
[59] Angelisa C Plane, Elissa M Redmiles, Michelle L Mazurek, and Michael Carl Tschantz. 2017. Exploring User Perceptions of Discrimination in Online Targeted Advertising. In {*USENIX*} *Security Symposium*.
[60] Kirk A Randazzo. 2008. Strategic Anticipation and the Hierarchy of Justice in US District Courts. *American Politics Research* (2008).
[61] Elissa M Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. 2017. *A Summary of Survey Methodology Best Practices for Security and Privacy Researchers*. Technical Report.
[62] Nripsuta Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David Parkes, and Yang Liu. 2019. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. *AIES* (2019).
[63] Joanna Shepherd. 2011. Measuring Maximizing Judges: Empirical Legal Studies, Public Choice Theory and Judicial Behavior. *University of Illinois Law Review* (2011).
[64] Janet A Sniezek and Timothy Buckley. 1995. Cueing and Cognitive Conflict in Judge-Advisor Decision Making. *Organizational Behavior and Human Decision Processes* (1995).
[65] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices. In *KDD*.
[66] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. *arXiv preprint arXiv:1902.04783* (2019).
[67] State of Wisconsin v. Eric L. Loomis 2016. 2016 WI 68 of July 13.
[68] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating Human+ Machine Complementarity for Recidivism Predictions. *arXiv preprint arXiv:1808.09123* (2018).
[69] Amos Tversky and Daniel Kahneman. 1981. The Framing of Decisions and the Psychology of Choice. *Science* (1981).
[70] U.S. Census Bureau. 2016. American Community Survey 5-Year Estimates.
[71] Michael A Wallach, Nathan Kogan, and Daryl J Bem. 1964. Diffusion of Responsibility and Level of Risk Taking in Groups. *The Journal of Abnormal and Social Psychology* (1964).
[72] Ilan Yaniv, Shoham Choshen-Hillel, and Maxim Milyavsky. 2009. Spurious Consensus and Opinion Revision: Why Might People be More Confident in their Less Accurate Judgments? *Journal of Experimental Psychology: Learning, Memory, and Cognition* (2009).
[73] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *WWW*.
[74] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*.
[75] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. 2013. Learning Fair Representations. In *ICML*.

A APPENDIX

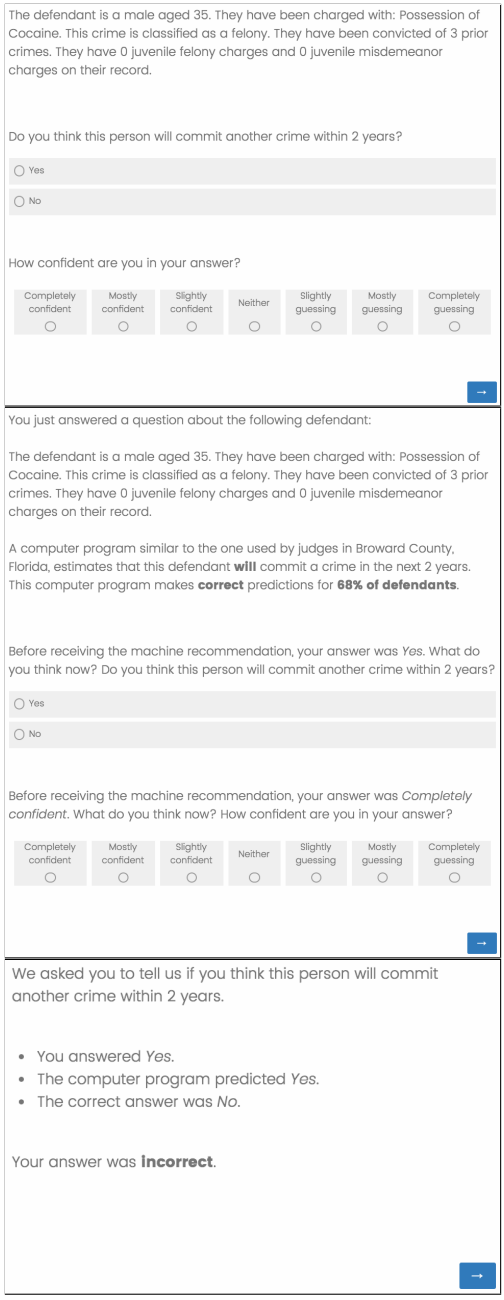

Fig. 8. Screenshot of the survey used in Study 2 to elicit respondents' [Top] pre-advice predictions and confidence, [Center] post-advice predictions and confidence, and [Bottom] provide them with feedback. In Study 1, the feedback shown in [Bottom] was not provided. In Study 3, participants were also informed about their bonus payment.

Case ID	sex	age	crime	misd. / fel.	# priors	# juv. fel.	# juv. misd.	ground truth	machine advice
222	F	43	Driving Under the Influence	M	0	0	0	0	0
763	M	26	Assault with a Deadly Weapon	F	7	0	0	1	1
1385	M	28	Battery	M	4	0	0	1	1
1490	M	30	Battery	M	4	0	0	1	1
1870	M	20	Carrying a Concealed Weapon	F	0	0	0	0	1
2078	M	43	Burglary	F	4	0	0	1	0
2278	M	63	Battery	M	0	0	0	1	0
2438	M	67	Battery	M	0	0	0	0	0
2502	M	45	Grand Theft	F	0	0	0	0	0
2533	M	19	Battery	M	0	0	0	1	0
2632	M	56	Possession of Cocaine	F	4	0	0	0	0
2898	F	31	Battery	M	0	0	0	0	0
2914	M	29	False Imprisonment	F	1	0	0	0	0
3146	M	37	Tampering with a Witness	F	0	0	0	1	0
3229	M	26	Driving Under the Influence	F	2	0	0	0	0
3255	M	51	Driving with a Suspended License	M	4	0	0	0	0
3347	F	36	Criminal Damage of less than $1000	F	7	0	0	1	1
3552	M	28	Battery	M	0	0	0	0	0
3679	F	47	Prostitution	M	0	0	0	1	0
3805	M	28	Grand Theft	F	0	0	0	1	0
4029	M	28	Battery	M	0	0	0	1	0
4553	M	34	Operating a Vehicle without a Valid Drivers License	M	2	0	0	0	0
4593	M	38	Battery	M	0	0	0	0	0
5057	F	49	Driving with a Suspended License	F	1	0	0	0	0
5069	M	24	Prostitution	M	2	1	0	1	1
5128	M	39	Battery	M	1	0	0	0	0
5217	M	37	Battery	F	0	0	0	1	0
5449	M	32	Burglary	F	1	0	0	0	0
5543	M	45	Battery	M	3	0	0	1	0
5921	M	52	Restraining Order Violation	M	4	0	0	1	0
6000	F	23	Burglary	F	0	0	0	1	1
6094	M	30	Dealing Controlled Substances	F	7	0	0	1	1
6207	M	38	Extradition of Defendants	M	0	0	0	1	0
6658	M	39	Forgery	F	0	0	0	1	0
6908	M	30	Driving with a Revoked License	F	21	0	0	1	1
7012	M	28	Possession of Cannabis/Marijuana	F	0	0	0	0	0
7894	M	20	Burglary	F	0	0	0	0	0
7975	M	23	Possession of Cannabis/Marijuana	F	11	8	2	1	1
8149	M	27	Domestic Violence	M	3	0	0	0	1
8438	F	34	Possession of Cocaine	F	12	0	0	1	1
9041	M	36	Driving Under the Influence	M	0	0	0	0	0
9231	M	35	Possession of Cocaine	F	3	0	0	0	1
9303	M	30	Driving with a Suspended License	M	0	0	0	0	0
9556	M	30	Assault with a Deadly Weapon	F	1	0	0	0	0
9953	M	23	Grand Theft	F	1	0	0	1	1
10406	M	43	Possession of Cocaine	F	2	0	0	0	0
10580	F	22	Domestic Violence	M	1	0	0	0	1
10762	F	26	Driving Under the Influence	M	1	0	0	0	0
10807	M	25	Driving with a Suspended License	M	3	0	0	1	1
10946	M	27	Driving Under the Influence	M	1	0	0	0	0

Table 8. Vignettes used as stimulus material in our experiments.

good advice	bad advice	bail advice	jail advice
222	4593	10946	2078
2533	1870	9231	763
8438	5217	10762	1385
10946	5543	9556	1870
10807	9231	5921	1490
10762	2438	9303	222
10406	1385	9041	2278
9953	3679	7894	3347
2278	1490	10580	2533
9556	5921	7012	5069
9303	6207	6207	6000
7975	10580	3255	6094
9041	2502	10406	3679
7894	2078	5543	6908
4029	3146	9953	7975
6908	6094	5449	6658
3805	2898	10807	2632
7012	6658	5217	8149
6000	2914	5128	2898
5449	8149	5057	8438
5128	3229	2438	2914
2632	763	4029	3146
5069	3255	4593	2502
5057	3347	4553	3229
4553	3552	3805	3552
3552	4553	3552	3805
3347	5057	3229	4553
3255	5069	2502	4593
763	2632	3146	4029
3229	5128	2914	2438
8149	5449	8438	5057
2914	6000	2898	5128
6658	7012	8149	5217
2898	3805	2632	10807
6094	6908	6658	5449
3146	4029	7975	9953
2078	7894	6908	5543
2502	9041	3679	10406
10580	7975	6094	3255
6207	9303	6000	6207
5921	9556	5069	7012
1490	2278	2533	10580
3679	9953	3347	7894
1385	10406	2278	9041
2438	10762	222	9303
9231	10807	1490	5921
5543	10946	1870	9556
5217	8438	1385	10762
1870	2533	763	9231
4593	222	2078	10946

Table 9. Sequence in which cases are shown in the treatments *good advice*, *bad advice*, *bail advice*, and *jail advice* in Study 2. The values correspond to the Case ID column in Table 8.

A.1 Stimulus Material

The stimulus material in our experiments consists of vignettes of the following form:

Defendant Description.

> The defendant is a <sex> aged <age>. They have been charged with: <crime>. This crime is classified as a <misdemeanor / felony>. They have been convicted of <non-juvenile prior count> prior crimes. They have <juvenile felony prior count> juvenile felony charges and <juvenile misdemeanor prior count> juvenile misdemeanor charges on their record.

Machine Advice.

> A computer program similar to the one used by judges in Broward County, Florida, estimates that this defendant **<will / will not>** commit a crime in the next 2 years. This computer program correctly predicts if a person will commit a crime in the next 2 years for **68% of defendants**.

In Table 8, we list the full set of 50 vignettes used in our experiments. Screenshots of our survey are shown in Figure 8.

Sequence in which cases are shown.

In Study 1, Study 3 and the *feedback* treatment in Study 2, the participants see the 50 cases in random order. In the remaining four treatments in Study 2, they see the cases in a fixed order. The precise sequences are reported in Table 9.

Received April 2019; revised June 2019; accepted August 2019