MACHINE ADVICE WITH A WARNING ABOUT MACHINE LIMITATIONS: EXPERIMENTALLY TESTING THE SOLUTION MANDATED BY THE WISCONSIN SUPREME COURT

*Christoph Engel** and Nina Grgić-Hlača †

ABSTRACT

The Wisconsin Supreme Court allows machine advice in the courtroom only if accompanied by a series of warnings. We test 878 US lay participants with jury experience on fifty past cases where we know ground truth. The warnings affect their estimates of the likelihood of recidivism and their confidence, but not their decision whether to grant bail. Participants do not get better at identifying defendants who recidivated during the next two years. Results are essentially the same if participants are warned in easily accessible language, and if they are additionally informed about the low accuracy of machine predictions. The decision to grant bail is also unaffected by the warnings mandated by the Supreme Court if participants do not first decide without knowing the machine prediction. Oversampling cases where defendants committed violent crime does not change results either, whether coupled with machine predictions for general or for violent crime. Giving participants feedback and incentivizing them for finding ground truth has a small, weakly significant effect. The effect becomes significant at conventional levels when additionally using strong graphical warnings. Then participants are less likely to follow the advice. But the effect is counterproductive: they follow the advice less if it actually is closer to ground truth.

1. INTRODUCTION

Arguably humans and machines have complementary skills (Tan et al. 2018; Raghu et al. 2019; Wilder et al. 2020). Machines (computers) cannot only

* Max Planck Institute for Research on Collective Goods, Bonn, Germany. Email: engel@coll.mpg.de

† Max Planck Institute for Software Systems, Saarbrücken, Germany, and Max Planck Institute for Research on Collective Goods, Bonn, Germany. Email: nghlaca@mpi-sws.org

We are grateful for very helpful advice by Judge Nicholas J. McNamara (Wisconsin Circuit Court) and by Professor Cecilia M. Klingele (University of Wisconsin Law School), the editors, and two anonymous referees.

doi:10.1093/jla/laab001

process much more information. The effect of each input in machine judgment can also be perfectly observed and reconstructed. This makes it easier to shield machine judgment from unintended bias. But despite intense work on algorithmic explainability, it remains hard to make human addressees understand why machines judge a case the way they do (Doshi-Velez & Kim 2017). Machines are only as good as the data that have been used to train them. Even if explicit discrimination is ruled out, discrimination may sneak in through correlations with seemingly non-discriminatory variables (Dwork et al. 2012; Zafar et al. 2017). Human addressees may be more willing to trust human than machine judgment, an effect known as *algorithmic aversion* (Dietvorst et al. 2015).

All of this now also matters at the core of the legal system. Multiple jurisdictions, in particular in the USA, give judges access to machine predictions. They not only do so for the purpose of choosing between bail and jail. Machine predictions are even used for sentencing. Yet for neither purpose, the legal order completely delegates the decision to an algorithm. Human judges are also not obliged to take the machine assessment for a fact. The machine prediction is merely an input to a decision that ultimately is taken by the competent judge. The machine only gives advice.

Yet this advice has increasingly come under scrutiny. In a recent case, the Supreme Court of the State of Wisconsin had to decide about the constitutionality of this practice.[1] It held that defendant was not denied due process. But it only cleared the use of such advice if judges have been properly warned about its limitations. In this article, we use experimental methods to test whether this institutional intervention delivers on its promises. Does the warning specified by the court change how judges decide? More importantly even: does the warning help them follow the advice where the machine got it right, and discard the advice where the machine got it wrong?

For obvious reasons we cannot test how Wisconsin judges sentence real defendants in the light of the advice commanded by the Supreme Court. This would require randomly assigning some defendants to a judge who receives the warning, and others to judges who are not warned. Moreover, not all criminal cases are perfectly comparable. All Wisconsin judges of course know about the Supreme Court ruling. As they repeatedly decide criminal cases, all of them have seen the warnings in other cases. We revert to a method that is prevalent in research on judge and jury decision-making. We expose randomly selected laypeople to a series of case sketches and ask them how they would decide.

To come as close to the real situation as possible, we select participants from the USA who report to have served on a jury. All of them get the same advice.

1 Supreme Court of Wisconsin, 2016 WI 68. For detail *see* Section 2.

This is advice from the exact software used in Wisconsin. In all cases we happen to know ground truth, i.e., whether the defendant recidivated in the two years after release (from not being put into jail in the case of bail, and from jail when incarcerated.) We randomly select half of all participants to receive the list of warnings mandated by the Wisconsin Supreme Court, using the exact same wording as the court administration in Wisconsin. We have four measures: the choice between bail and jail, the estimate of recidivism risk, confidence, and accuracy (do choices coincide with ground truth?). Participants first decide on their own. Then they receive the advice which the machine has given in this specific case, and are given the opportunity to revise their responses.

The Wisconsin Supreme Court ruled on the use of machine advice in sentencing. In this one respect, we depart from the ruling. We do not test our participants on sentencing, but on the choice between bail and jail. We do so for a reason of data availability. We have the good fortune to exploit that ProPublica has followed up 7,214 convicts in Broward County, Florida (Angwin et al. 2016). For all of them, the competent judge had access to an assessment and a recommendation from the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software, that Wisconsin also uses for sentencing. We know this assessment and recommendation. Most importantly, for all cases, we know whether defendants were rearrested during the two years after release. We use this as a reasonably reliable proxy for recidivism.

There has been an extensive debate about the ruling of the Wisconsin Supreme Court. Some commentators have been skeptical whether the warnings would deliver on their promises (*see* Section 2). But to the best of our knowledge, no research has related the solution proposed by the Wisconsin court to the extensive psychological literature on the effect of warnings, let alone tested the effect of the warning empirically. This is our contribution.

Results are sobering. We do find a small effect of warnings on the estimate of the likelihood that the defendant will recidivate during the next two years. If warned, participants change their estimates a little less. They also get a small bit less confident. Yet the warnings have no effect on the normatively important outcomes: participants are no less influenced by machine advice, and their choices do not become more accurate, in the sense of being in line with ground truth.

We put this result into a series of robustness checks. In the main experiment, participants first decide before receiving machine advice and have the opportunity to revise their responses in the light of the advice. This procedure makes it possible to see the effect of advice at the individual level. But theoretically, it might introduce a commitment effect: participants might not want to contradict their earlier responses. This is why we repeat the

experiment, but drop the first phase. This does not change outcomes. Replacing the legalistic jargon of the warnings with warnings in plain English, or additionally informing participants that COMPAS is only correct in 68 percent of all cases, has practically no effect either.

It is not very clear whether Wisconsin law qualifies the risk of recidivism before trial and requires a risk of violent crime. We test this in two steps. We had randomly drawn the original fifty cases from the ProPublica dataset. Regarding the general risk of recidivism, these cases are well balanced. But only six defendants committed new violent crimes in the two years after release. We would have had a heavily biased set of cases. We therefore compose a new set of fifty cases where we oversample violent recidivism. In two new treatments we check whether results differ if we use these cases, but still give participants access to the general recidivism score generated by COMPAS. In another two treatments, we replace the machine advice and give participants COMPAS' violent recidivism score. Neither change significantly affects outcomes.

In the final step, we try a whole battery of changes that might potentially make warnings more powerful: we give participants a monetary incentive for finding ground truth; after each case, we give them feedback about ground truth, and hence provide them with an optimal learning environment; we make it explicit that the decision hinges on recidivism risk; we remove two sentences from the instructions that were meant to be fully transparent, but might have biased participants, and we replace the warnings mandated by the Wisconsin Supreme Court by very powerful graphical warnings. This battery of changes yields a small effect: participants are a bit less likely to follow the warning. Yet the effect is counterproductive. They become less likely to take the advice when it actually is good, in that the COMPAS score is closer to ground truth than their own likelihood rating.

The remainder of the article is organized as follows: in Section 2, we introduce the ruling of the Wisconsin Supreme Court, and relate it to the legal debate about algorithmic input to criminal law decision making. In Section 3, we derive hypotheses from the psychological literature on advice, and in particular on the effect of warnings. Section 4 introduces the design of the main experiment. Section 5 reports results from this study. Section 6 reports results if participants immediately see the machine prediction. Section 7 reports on the treatments testing alternative warnings, without legalistic jargon or with additional information about the machine's accuracy. Section 8 covers treatments that replace general with violent recidivism. In Section 9, we show in which ways results change if we try to get an effect of warnings, with the listed battery of changes, or additionally with graphic warnings. Section 10 concludes.

2. THE LEGAL DEBATE

In plea bargaining, Eric Loomis pleads guilty for two minor offenses. The trial court rules out probation and sentences the defendant to six years in prison. Among the reasons, the court notes that the defendant scores high on all three scales provided by the software Correctional Offender Management Profiling for Alternative Sanctions (COMPAS): recidivism, violence, and failure to appear in court.[2] The subsequent rulings by the Court of Appeals,[3] and the Supreme Court of Wisconsin[4] are concerned with this justification.

The Wisconsin Supreme Court upholds the decision of the trial court, but formulates a series of conditions that must be fulfilled if scores from this software weigh in on sentencing. The scores may be taken into consideration, but may not be "aggravating" (¶87) or "determinative" (¶88, ¶94). The trial court is obliged to explain in which ways the scores have been relevant for its decision (¶99). In the concrete case, the Supreme Court accepts that the trial court would have come to the same conclusion even when not taking the scores into consideration, so that no harm was done (¶106).

This article brackets these supplementary measures and focuses on the key safeguard. The Supreme Court requires that the scores resulting from applying the program come with the following list of warnings (¶100, *also see* ¶66):

- The proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are determined.
- Because COMPAS risk assessment scores are based on group data, they are able to identify groups of high-risk offenders—not a particular high-risk individual.
- Some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism.
- A COMPAS risk assessment compares defendants to a national sample, but no cross-validation study for a Wisconsin population has yet been completed. Risk assessment tools must be constantly monitored and renormed for accuracy due to changing populations and subpopulations.

2 Cir. Ct. No. 2013CF98; for more detail about the case see the appeals case 2015 WL 5446731, the ruling of the Wisconsin Supreme Court 2016 WI 68, and the documents filed to the Supreme Court of the United States regarding the writ for certiorari, https://www.scotusblog.com/case-files/cases/loomis-v-wisconsin/, in particular the petition, No. 16-6387.

3 2015 WL 5446731.

4 2016 WI 68.

- COMPAS was not developed for use at sentencing, but was intended for use by the Department of Corrections in making determinations regarding treatment, supervision, and parole.

Certiorari was denied by the Supreme Court of the United States,[5] probably following the suggestions by the State of Wisconsin and the US Department of Justice that the issue needs more time "to percolate."[6]

The ruling has been rather critically reflected in the legal literature (Freeman 2016; Harvard 2016; Carlson 2017; Israni 2017; Beriain 2018; Collins 2018; Washington 2019). Critics have pointed to the low accuracy of COMPAS predictions (Beriain 2018). They do not find it acceptable that the data and the algorithm are proprietary, and have called for open source software (Freeman 2016; Carlson 2017; Israni 2017; Beriain 2018; Nishi 2019). They also see the risk that the provider's profit interest clashes with due process (Freeman 2016). They criticize the composition of the norm groups in comparison to which the scores are calculated (Freeman 2016, 81 f.). They make it clear why group recidivism rates do not predict individual recidivism risk (Starr 2014). They argue that trial courts could be overly influenced by the seeming precision of the machine predictions (Freeman 2016; Harvard 2016, p. 1536), and lack the expertise to properly assess the probative value of machine predictions (Harvard 2016, p. 1535; also *see* Rizer & Watney 2018, 224; Stevenson 2018) These critiques tie into the broader legal debate over using machine predictions for decision making in criminal procedure (Starr 2014; Cino 2017; Eaglin 2017; Simmons 2017; Berman 2018; Deskus 2018).

Some critics have also addressed the proposed solution, i.e., the mandatory warning, and have not found it satisfactory (Freeman 2016, pp. 78, 95; Harvard 2016, p. 1531; Liu et al. 2019). The warning does not give the trial court guidance *how* to discount the prediction (Harvard 2016, p. 1534). Critics also relate the issue to *United States v. Rodriguez*,[7] where the court has seen instructions as ineffective that ask the jury to ignore non-permissible information (Freeman 2016).

The warning requirement in the ruling of the Wisconsin Supreme Court is motivated by concerns regarding data quality. Different US States use different software tools to assess recidivism risk (Casey et al. 2014; Desmarais et al. 2016; Berk 2019). All of them are far from perfect (for a discussion of comparative pros and cons *see* Berk & Bleich 2013). A recent metastudy finds performance

5 137 S. Ct. 2290 (2017).

6 *See* the references in footnote 2.

7 585 F.2d 1234, 1244 (5th Cir. 1978).

rates between 0.64 and 0.74, with a maximum of 1; 0.5 would be chance level (Desmarais et al. 2016, Table 4).[8] COMPAS is at 0.67, i.e., at the lower end. Several studies have tried to compare performance with and without the aid of prediction software. In Kentucky, there was a small, transient effect: there was better discrimination in bailing between low- and high-risk cases (Stevenson 2018). A study using data from New York compares actual bailing decisions of human judges with machine predictions, and finds a clear benefit in terms of crime reduction (Kleinberg et al. 2017).

Yet another study compares COMPAS predictions with the predictions made by human laypersons that only have access to very limited information, and finds no difference in performance (Dressel & Farid 2018, but *see* Holsinger et al. 2018). A recent study replicates Dressel & Farid (2018), but finds that human triers do worse when deprived of feedback, and when the most diagnostic information is presented in conjunction with less relevant facts (Lin et al. 2020). Parole decisions also rely on an assessment of recidivism risk. Berk (2017) exploits that machine advice has been deployed to the Pennsylvania authorities deciding on parole in a staggered manner. He finds a small reduction in the frequency of parole after machine advice has been made available. But an equivalent effect obtains for authorities that did not (yet) have access to the advice. Hence the actual advice has not been causal for the change.

COMPAS is used by Wisconsin, California, Michigan, New Mexico, New York, South Carolina, Wyoming (Casey et al. 2014, A-20), and Florida (Blomberg et al. 2010) as well as New Jersey (Fass et al. 2008). The performance of COMPAS has been validated for California (Farabee & Zhang 2007; Skeem & Eno Louden 2007; Zhang et al. 2014), Florida (Blomberg et al. 2010), New Jersey (Fass et al. 2008), New York (Lansing 2012), but not for Wisconsin. Triggered by a large-scale investigation by ProPublica (Angwin et al. 2016; Larson et al. 2016), in recent years there has been an intense debate about racial bias (Dieterich et al. 2016; Chouldechova 2017; Spielkamp 2017; Huq 2019; Hamilton 2019a). Recent studies have also found a strong age bias, to the detriment of the young (Stevenson & Slobogin 2018), and a gender bias to the detriment of female offenders (Hamilton 2019b).

In an earlier experimental study, we have found that participants only rarely change their prediction upon receiving machine advice. Giving them feedback about ground truth does not increase accuracy. Financial incentives for accuracy are also ineffective, while a financial incentive to follow the advice is effective (Grgić-Hlača et al. 2019).

8 The performance measure is the AUC, the area under the curve, where the rate of false positive is on *x*-axis, and the rate of true positives on the *y*-axis.

3. HYPOTHESES

For the solution mandated by the Wisconsin Supreme Court to be effective, human decision-makers must decide differently when just receiving machine advice, compared with receiving machine advice that comes with the prescribed warnings. The effect is normatively desirable if the warning increases accuracy.

To the best of our knowledge, it has not been tested whether human–machine interaction improves with warnings about machine limitations; this is the purpose of our experiment. To formulate expectations, in this section, we review the (mainly psychological) literature on the effect of warnings.

In legal language, the psychological literature on warnings has been interested in an institutional intervention (the warning), not so much in the behavioral channels that explain why, in some contexts, a warning has proven effective or not.[9] We can therefore only use the literature to support the plausibility of analogous effects.

Warnings have been tested in different domains. The most intensely investigated domain is eye witnesses (*see* the metastudy by Blank & Launay 2014). In the prototypical design, participants first receive information about an event, for instance via a movie clip. Then their memory is disturbed, for instance, by hearing a narrative about the event that is not perfectly correct. In the final step, they have to report the event. The accuracy of the report is the dependent variable (Deese 1959; Roediger & McDermott 1995). If participants are warned appropriately, the false recall effect is reduced (*see, e.g.*, Echterhoff et al. 2005), and can even disappear completely (*see, e.g.*, Oeberst & Blank 2012). In important respects, this task differs from a court deciding about a sentence, or whether to release the defendant on bail. The court decides. Its assessment of the recidivism risk is only an input to this decision. The judgment task is a prediction task, not mere recall. The intended effect of the "disturbance" is to improve, not to deteriorate (judgment and) decision quality. Hence the desired outcome is not to immunize the judge against the influence, but to make sure they discriminate between helpful and unhelpful inputs. Given these differences, no direct inferences can be drawn from the findings on warning eye witnesses. This literature is nonetheless instructive as it is so rich. This makes it possible to hypothesize which moderators make warnings particularly powerful.

9 For exceptions, *see in particular* Schul (1993); Wright (1993). Yet their theories about memory processes have no bearing on the question of this study.

A second active domain is food labels (*see* the metastudy by Argo & Main 2004). It can generate large-scale natural experiments (*see, e.g.*, Kaskutas 1993). This paradigm generates a decision, not just a judgment task. The warning is essentially advice, meant to divert behavior from an individually or socially detrimental path. In this task, the valence of the warning is thus exclusively positive. Addressees decide on behalf of themselves, not on behalf of third parties, like a judge. The same holds for warnings against attempts at tilting consumer choice through subliminal advertising (Verwijmeren et al. 2013), or warnings against being inattentive to the risks involved in running a chemical experiment (Wogalter et al. 1989). Warnings have also been investigated for attempts at influencing political opinions (*see, e.g.*, Deaux 1968). This domain is even more remote, since the task is (at least chiefly) judgmental.

Despite these differences, previous results might help predict why a warning in the context of criminal judicial decision-making might be effective. For eyewitnesses, the warning has proven more effective if they are warned before being exposed to the disturbance, rather than only afterward (*see, e.g.*, McCabe & Smith 2002). The Wisconsin Supreme Court mandates that the warning is attached to the information about the COMPAS scores. Being warned before the "disturbance" is more effective as the addressee need not counteract a mental effect that has already been achieved (Schul 1993). The same facilitating effect should be present if the potentially misleading information and the warning arrive simultaneously.

In opinion research, warnings against the risk of manipulation have proven more effective if the cause is of direct personal interest for the addressee (Apsler & Sears 1968). Arguably, holding judicial office makes a judge personally involved; the judge knows that they have power to decide about the defendant's life. In the laboratory task, the warning has been more effective if a confederate of the experimenter complies who is perceived as a peer of the subject (Wogalter et al. 1989). The Wisconsin Supreme Court mandates all trial judges to (repeatedly) receive the same warning in every pertinent criminal case. Arguably this creates peer pressure, which should increase the effectiveness of the warning.

Warnings have been shown to be more effective if they come with a meaningful explanation (Wogalter et al. 1987). If addressees are informed about the underlying psychological effect, the effect of the disturbance can even be completely neutralized (Oeberst & Blank 2012). The warning mandated by the Wisconsin Supreme Court does not go that far. But the cautioning and

debiasing intention can even be derived from the wording of the warning. And trial judges in Wisconsin know that the warning has been prescribed by their Supreme Court, in a published and widely cited ruling. If in doubt, they can therefore relatively easily learn even more about the underlying motives. This suggests that the warning stands an even better chance to be effective.

While direct analogies are not possible, taken together these findings suggest:

> **Hypothesis 1.** *Decision-makers who receive the warnings mandated by the Wisconsin Supreme Court*
> (a) choose differently between bail and jail,
> (b) assess the likelihood of recidivism differently,
> (c) are less confident.

From a normative perspective, the most important issue is not *whether* decision-makers react to the warnings, but *how*. If the hopes that the Wisconsin Supreme Court created with its ruling are well-founded, we should see

> **Hypothesis 2** *Decision-makers who receive the warnings mandated by the Wisconsin Supreme Court discriminate more effectively between cases where the defendant actually recidivated and where they did not.*

4. DESIGN

4.1 Structure of the Experiment

We want to test whether the presence of the warnings mandated by the Wisconsin Supreme Court, separately or jointly, induces participants who receive machine advice

- to make different decisions
- to make better decisions, by better discriminating between cases where the defendant actually recidivated and where they did not.

This invites a straightforward design: participants receive

- the facts of a case and information about the defendant
- the COMPAS prediction about the risk that the defendant recidivates
- depending on treatment
  - no warning
  - the exact warnings mandated by the Wisconsin Supreme Court

4.2 Case Selection

For obvious ethical reasons, we cannot randomly give some Wisconsin trial judges some warning, and others another, or none. We, therefore, use historical cases as stimulus material.

We exploit the fact that we have access to the ProPublica dataset (Angwin et al. 2016), which contains data about 7,214 defendants tried in Broward County, Florida, in 2013–2014. ProPublica relied on Freedom of Information legislation to follow up these defendants in public records. This is why we know the ground truth. Specifically we know whether the defendant was charged with a new crime in the two years after release. As defendants who were jailed were incapacitated, for them the clock counts from the moment on when they have been released from prison. For defendants who were put on bail, the clock counts from that moment on. This imbalance is inherent in the nature of the data.

In our experiments, we randomly select fifty from the 1,000 ProPublica cases which were also covered in the follow-up study by Dressel & Farid (2018). Figure 1 shows that the accuracy of the COMPAS scores is fairly limited. The only unequivocal score is the maximum score of 10. All four (of fifty) defendants with this score have indeed been rearrested during the next two years after release from prison. Yet two of the three defendants with scores 8 and 9 have actually not been rearrested. Two of the nine defendants with the lowest score 1 have been rearrested, as have three of seven with score 2, and six of twelve with score 3. In another paper, one of us has shown that, on average, the accuracy of a decision aid trained using the ProPublica COMPAS dataset is as low as 68 percent (Grgić-Hlača et al. 2018).[10]

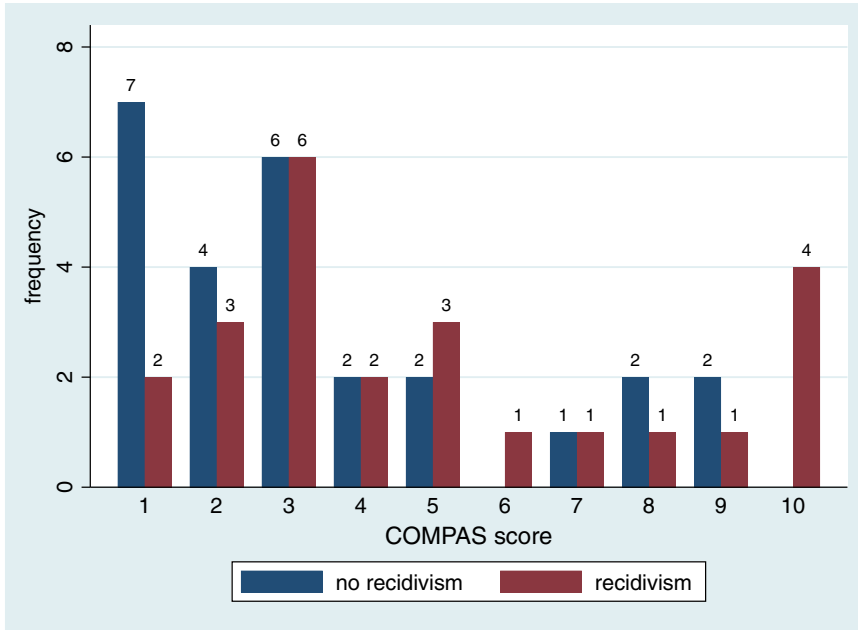
4.3 Survey Instrument

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

> Judges across the United States use a computer program named COMPAS 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 the same computer program.
>
> 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, and whether you would grant bail.

10 COMPAS classifies defendants with scores of 4 or less as "low risk," scores of 5–7 as "medium risk," and scores of 8 or more as "high risk," COMPAS Practitioner's Guide 2019, 11. The study interprets a score of 5 or higher as the prediction that the defendant will recidivate.

Figure 1. COMPAS scores and ground truth



In this survey, you will see the COMPAS tool's predictions in the same way as judges in one of the US states see them.

< Begin: additionally in the warning Treatment > The Supreme Court of this state mandates that the COMPAS prediction be accompanied with a document explaining the technique it uses, and warning against the potential for misuse. Throughout the survey, we will show you an excerpt from the document judges in this state receive together with the COMPAS prediction. < End: additionally in the warning Treatment >

Note, however, that < Begin: additionally in the warning Treatment> the Supreme Court ruling directly addresses the use of COMPAS predictions in sentencing, while our study is concerned with the use of these predictions when deciding whether to grant bail. Finally, < End: additionally in the warning Treatment > our study is confined to one of three scores provided by the COMPAS software, the "General Recidivism Risk". We omit the "Violent Recidivism Risk" and "Pretrial Release Risk" scores.

The stimulus material consists of vignettes, as used in Dressel & Farid (2018). For each of the fifty cases we consider, participants receive a short description of the criminal case, and of the defendant, of the following type:

> 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 (Groves et al. 2011; Redmiles et al. 2017). In the Appendix, we make the complete set of vignettes available in Table A3 in Appendix.

Participants are asked three questions:

> How likely do you think it is that this person will commit another crime within 2 years?
>
> (Likert scale with 5 levels, running from "extremely unlikely" to "extremely likely").
>
> Do you think this person should be granted bail? (Yes/No)
>
> How confident are you in your answer about granting this person bail?
>
> (Likert scale with 5 levels, running from "completely guessing" to "completely confident")

Afterward, in the *Baseline*, participants receive the general recidivism score for the case in question, and are asked the same three questions again, each time reminding them of the answer they had given before knowing the COMPAS score.

In the *Treatment*, participants receive the information in exactly the format used by the court administration in Wisconsin.[11]

11 We are grateful to Judge McNamara for making the precise wording and layout available.

Quoting from the document that judges in this US state receive together with the COMPAS tool's predictions:

Assessment Considerations

COMPAS (Correctional Offender Management & Profiling for Alternative Sanctions) is a validated actuarial assessment tool that predicts the general likelihood that a person will engage in subsequent criminal behavior in comparison to others with a similar history of involvement in the criminal justice system. . . .

It is important to remember that while risk scores may assist in informing sentencing decisions based on the risk principle by categorizing medium and high risk individuals who are appropriate for intervention, they should never be the sole and deciding factor in determining the severity of the sentence or whether an offender should be incarcerated.

Functioning only as a general risk assessment instrument, COMPAS does not attempt to specifically predict the likelihood that an individual offender will commit a certain type of offense within the followup period. Rather, offense-specific instruments may be used to provide additional insight.

The proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are determined. . . .

Risk assessment tools should be constantly monitored and re-normed for accuracy due to changing populations and subpopulations. Despite being validated in other states and jurisdictions, the statewide COMPAS implementation in [the state in question] will include a commitment to continuous research. COMPAS was normed on a [the state in question] population in February of 2016. Likewise, it has been exposed to significant inter-rater reliability testing and measurement under a continuous quality improvement framework. Some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism. The [the state in question] Department of Corrections will conduct independent validation studies of COMPAS that will examine general predictive validity as well as disparity across race and will remain committed to replicating these studies over time.

Bearing these considerations in mind, research suggests criminal justice officials will be positioned to make more informed decisions at all

decision points, including the sentencing event, as a result of understanding risk/need information.

Then participants are reminded of the answers they have given before receiving this piece of advice, and are asked to answer all three questions again.

4.4 Sample

We ran 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 (Palan & Schitter 2018). One of the reasons we choose Prolific over MTurk is their advanced pre-screening capabilities, which allowed us to focus our study on participants with relevant demographics and prior experiences. For each treatment, we recruited sixty-seven participants, randomly selected from a pool of Prolific participants from the United States, which have self-reported to have served on a jury in their Prolific user profiles. This design gives us power to detect an effect of medium size (0.5) at the conventional values of is $\alpha = 0.05$; $\beta = 0.2$.[12] This choice of sample size was motivated by external validity. An even smaller effect would not be meaningful from the (policy) perspective of the Wisconsin Supreme Court.

After taking part in the experiments, participants answered a set of questions about their demographics. Their answers are shown in Table 1.

Participants received £2.5 for their participation, but were not incentivized for the choices they made. The experiment took approximately thirty-five minutes. In order to make sure that participants take the task seriously, they had to answer two attention check questions, shown in Figure A1 in Appendix. We discard the answers of five respondents who have not successfully answered both attention check questions.

We have preregistered the experimental design and our hypotheses with the Open Science Framework (Engel & Grgić-Hlača 2020a). The design of the experiment has been approved by our institution's Ethical Review Board.[13]

4.5 Design Choices

The following design choices warrant justification. The data would have higher external validity, had we tested Wisconsin trial judges, rather than laypersons. Yet Wisconsin is a relatively small state, so that it would have been almost impossible to find enough judges to fill the cells. More importantly, we could not

12 We have used software G*Power for these calculations.

13 Ethical Review Board of University of Saarland, December 13, 2019.

Table 1. Demographics

Demographic attribute	Study 1	Census
Total respondents	139	–
Passed attention checks	134	–
Female	53%	51%
African American	8%	13%
Asian	5%	6%
Caucasian	78%	61%
Hispanic	7%	18%
Other	1%	4%
Bachelor's Degree or above	95%	30%
Liberal	55%	33%[a]
Conservative	10%	29%[a]
Moderate	29%	34%[a]
Other	5%	4%[a]
Jury duty experience	72%	27%[b]

Notes: Comparison between survey sample for Study 1 (Section 5) and 2016 US Census (US Census Bureau 2016). Attributes marked with a [a] were compared to Pew data (Pew Research Center 2016) for political leaning. Attributes marked with a [b] were compared to DRI 2012 National Poll data (DRI 2012). 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. Demographics reported for respondents who passed the attention checks.

possibly have hoped to get all judges to take our survey simultaneously. If not, we would have had to worry that judges talk to each other, which would have caused dependence between observations. Most importantly, all Wisconsin trial judges have regularly seen the mandated warning in every pertinent case since 2016. We could therefore not plausibly argue that we have randomly assigned some of them to receiving no warning.

The ruling of the Wisconsin Supreme Court is valid for COMPAS scores being used in sentencing. We test our participants on the choice between bail and jail. We have a pragmatic reason. We would not only have to give participants sentencing ranges that differ from case to case. With sentencing, the fact would also have weighed in that our participants do not have sentencing experience. Participants would neither have had a sense of the acceptable level of punishment, nor would they have had known how to map the charge, and other features of the case, to the severity of the sanction. This lack of experience might have induced choices to be erratic. At the least, we would have had to worry about learning, and therefore order effects. Compared with sentencing, the binary choice between bail and jail is much easier. Most importantly, for sentences, we would not know the ground truth, while we do for bailing, as the only normative reason for granting or refusing bail about which participants receive information is the risk of recidivism, which we know.

Recidivism risk is not the only criterion when choosing between bail and jail. Art. I (8) Wisconsin Constitution is also concerned that the defendant might not appear in court, or that they might intimidate witnesses. We bracket these concerns and do not give participants any information on them, neither in the abstract nor per case. This is why we can be confident that they will focus on the one concern to which the COMPAS score speaks, i.e., recidivism. Due to this design choice, we are able to define "accuracy" narrowly, in the sense of preventing new crime.

It finally is worth noting that the COMPAS score assesses the risk that the defendant commits a new crime during the next two years. Even when used for decisions about bail, the match with the normatively relevant prediction is imperfect. For granting bail, only the risk matters that the defendant commits a new crime before tried for the crime for which they have presently been apprehended. Usually, this period of time will not precisely be two years.[14]

5. RESULTS OF THE MAIN EXPERIMENT

5.1 Conflict between Human and Machine Assessment

Machine advice is important in our sample.[15] In Figure 2, cases are ordered by the average propensity of participants to incarcerate the defendant before they have had access to the COMPAS scores. As the left panel shows, their assessment and the recommendation given by COMPAS differ widely.[16] This also holds for the estimated likelihood that the defendant commits a new crime during the next two years (right panel).

5.2 Effect of Machine Advice on Choices and Confidence

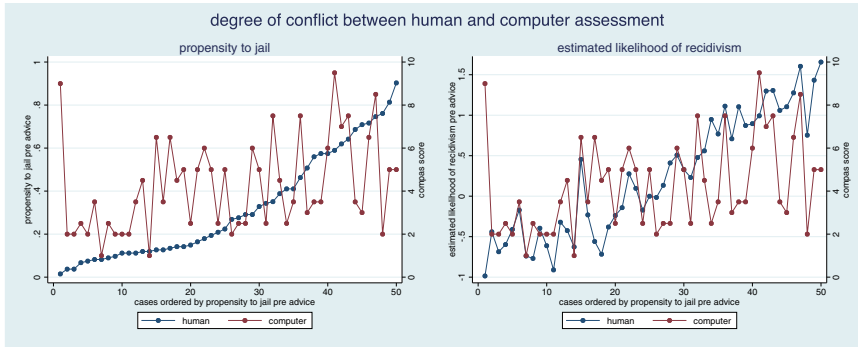
Machine advice matters. The effect on the propensity to incarcerate the defendant is normatively most important, and quite pronounced. As Figure 3 shows, in a fair number of cases, participants indicate a different decision after having seen the COMPAS assessment (the green and red dots are far away from the blue dots). In line with our earlier finding (Grgić-Hlača et al. 2019), participants are more likely to change their decision in the direction of releasing the

14 On the question whether only the risk of violent crime matters, or whether any crime is relevant, *see* Section 8.

15 As results are then easier to read and understand, we only report tests of our hypotheses, i.e., effects of warnings, in the end of this section. Technically the remaining results are exploratory.

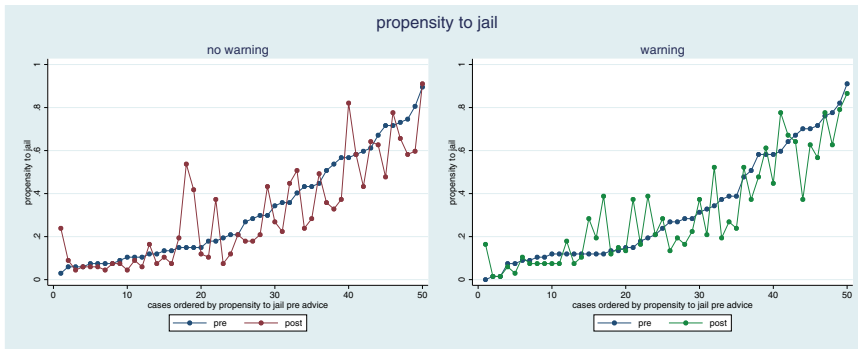
16 As explained in the Design section, COMPAS makes no explicit recommendation. For the purposes of this figure, we follow the literature and classify scores of 5 or larger (which COMPAS flags as "medium risk" or "high risk") as the recommendation to jail the defendant.

Figure 2. Human vs. computer assessment



x-axes: cases ordered by the mean propensity to jail; right *y*-axis: COMPAS score; left *y*-axis left panel: mean propensity to jail before receiving machine advice; left *y*-axis right panel: mean likelihood rating, on a 5-point Likert scale running from −2 (very unlikely) through 2 (very likely).

Figure 3. Effect of machine advice on propensity to jail



defendant on bail. Machine advice has a tendency to radicalize the assessment (in Figure 4 blue dots on the one hand, and green and red dots on the other hand, frequently point into the same direction, but upon receiving advice participants either deem it even more likely that the defendant will, or that they will not, recidivate). Overall receiving machine advice makes participants more confident. This effect is particularly pronounced if, before receiving the advice, they had predominantly intended to release the defendant on bail, or to put them into jail. This suggests that machine advice affects confidence most in easy cases (Figure 5).

Figure 4. Effect of machine advice on estimated likelihood of recidivism

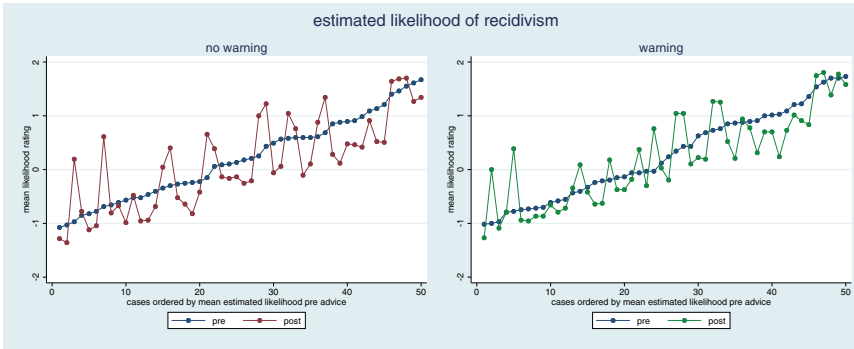
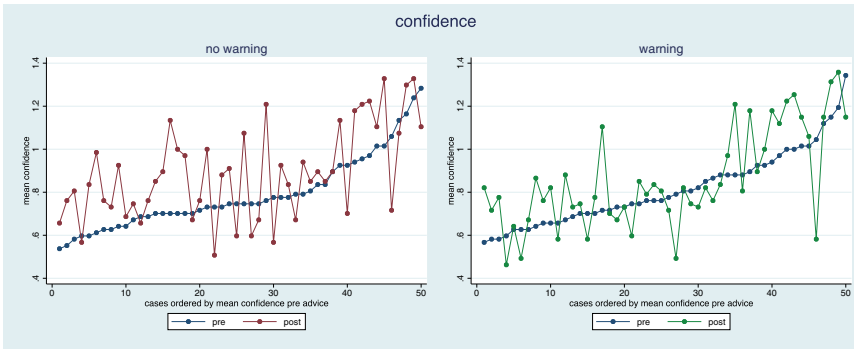


Figure 5. Effect of machine advice on confidence



5.3 Statistical Model

Table 2 tests Hypothesis 1. As Figure 3 shows, the reaction to machine advice is not uniform. In some cases, more participants shift from bail to jail (the green and red dots are above the respective blue dot), whereas in other cases more participants shift from jail to bail (the green and red dots are below the blue ones). Were we to test the direction of the shift, shifts toward bail and shifts toward jail would cancel out. To see to which degree machine advice moves participants' choices, we, therefore, work with the absolute of the shift. Actually, in 11.73 percent of all cases, a participant changes their choice after receiving machine advice.

For the same reason, we do not use the directed change in the estimated likelihood that the defendant will recidivate during the next two years, but work

Table 2. Effect of machine advice on verdict, likelihood rating, and confidence

Propensity to Jail	
Warning	−0.011
	(0.021)
Cons	0.117***
	(0.018)
Likelihood of recidivism	
Warning	−0.088**
	(0.027)
Cons	0.531***
	(0.021)
Confidence	
Warning	−0.053†
	(0.027)
Cons	0.097***
	(0.022)
N	6,700

Notes: The first two dvs are the absolutes of first differences, i.e., $|dv_{post} - dv_{pre}|$. Confidence is a first difference, i.e., $conf_{post} - conf_{pre}$. Warning: COMPAS predictions came with the warnings mandated by the Wisconsin Supreme Court. Linear multivariate (structural) model. Standard errors with random effects for fifty defendants and for 134 respondents in parenthesis. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{\dagger}p < 0.1$.

with the absolute of the change. This makes sure shifts toward a lower and toward a higher likelihood do not cancel out. Given the variable has five expressions (−2 .. 2), the maximum shift is four units (from extremely unlikely to extremely likely, or vice versa). The mean shift in the data is 0.531 units. Averaged over all choices, upon receiving machine advice, participants adjust the likelihood rating by about half a point.

For our third dependent variable, confidence, there is no need for an analogous correction. We are interested to learn whether machine advice makes participants more or less confident. Confidence is rated on a scale from −2 ("completely guessing") to 2 ("completely confident"). Changes in confidence in reaction to machine advice do therefore range from −4 .. 4. The average change is positive; participants become more confident. Yet the degree of the change is on average only 0.097, and hence very small.

Each of our 134 participants reacts to each of the fifty cases. We capture this potential source of dependence by a participant random effect. This random effect removes participants' idiosyncrasies from the estimation of population effects. Moreover, each participant, for each case, generates all three dependent variables. We capture this additional source of dependence by a multivariate statistical model. We thus simultaneously estimate effects on all three variables. Finally, each case is presented as a vignette that sketches the situation and the

action of the defendant in question. We are ultimately not interested in the reaction of participants to individual vignettes. We want to see the overall effect of machine advice. We, therefore, add a second random effect for defendants (cases), and thus estimate a multivariate (linear) statistical model with crossed random effects.[17]

5.4 Effect of Warnings on Verdicts, Likelihood Ratings, and Confidence

As all our dependent variables are first differences, the constants reported in Model 1 of Table 2 measure whether machine advice matters. This turns out to be the case. As the statistical model controls for the presence of the warnings mandated by the Supreme Court, the constants are the predicted effect of machine advice when not being warned. The regression predicts that machine advice will change the choice between jail and bail in about 12 percent of all cases; that the expected likelihood rating (on a 5-point scale) will change by about half a point; and that machine advice will make participants slightly more confident. Hence all descriptive effects are highly significant and therefore credible.

The statistical model shows that the warnings mandated by the Supreme Court do have a significant effect on likelihood ratings and a weakly significant effect on confidence, but do not have an effect on verdicts.

We thus only have partial support for Hypothesis 1 and conclude

Result 1. When they receive the warnings mandated by the Wisconsin Supreme Court,

(a) participants do change their likelihood ratings less than without the warnings,
(b) participants are less confident,
(c) participants do not change their choice between jail and bail.

5.5 Effect of Warnings on Accuracy

From a normative perspective, the most important issue is accuracy. Do the warnings mandated by the Wisconsin Supreme Court make it more likely that

17 We cannot run a fixed effects model as the explanatory variable differs between, not within participants. For richer models with additional controls that vary within participants, the Hausman test never turns out significant. These richer models are available from the authors upon request. In several respects, this empirical strategy improves over the analysis plan in our preregistration. In that document, we had proposed a procedure that would neglect the multivariate nature of the data, and defendants as an additional source of dependence. We also had proposed specifications with levels, rather than the absolute of first differences, which would neglect that the direction of the change is not meaningful.

participants release the defendant on bail if they actually have not recidivated, and that they incarcerate them if they have recidivated?

Now COMPAS cautiously leaves the interpretation of the score to the user. We do therefore not have a direct machine recommendation for the choice between bail and jail. This is where it helps that we have not only elicited choices, but also likelihood ratings. The COMPAS score is a likelihood rating as well. We can compare it with the likelihood rating the participant has indicated before learning the COMPAS score.[18]

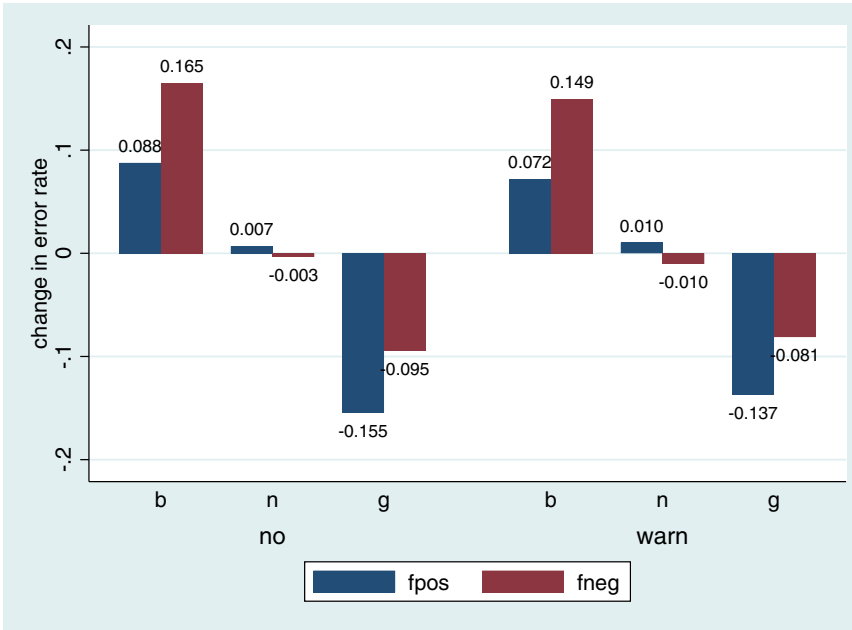
This makes it possible for us to say whether, in the respective case, machine advice has been helpful. We classify the advice as "good" if before receiving the advice the participant's verdict has been a false positive (the participant wants to jail a defendant who has not recidivated two years after release) or a false negative (the participant wants to release a defendant on bail who has recidivated two years after release) and the COMPAS score is closer to ground truth than the participant's rating before seeing this score. Likewise, we classify the advice as "bad" if before receiving the advice the participant's verdict has been a true positive (the participant wants to jail a defendant who has recidivated two years after release) or a true negative (the participant wants to release a defendant on bail who has not recidivated two years after release) and the COMPAS score is further away from ground truth than the participant's rating before seeing this score. Otherwise, we classify machine advice as neutral.

The normative assessment may differ between false positives (the defendant is sent to jail, although they did not recidivate in the two years after release) and false negatives (the defendant is released although they have been arrested for another crime during the next two years). We therefore separately assess whether the availability of the warnings mandated by the Supreme Court reduces the incidence of either mistake.

Figure 6 reports descriptives. If the advice is in line with the participant's own assessment before receiving the advice, unsurprisingly it has little effect. "Good" advice, in the sense defined above, matters, in the desirable direction. Descriptively, its effect is even more pronounced without the warnings mandated by the Supreme Court. The reduction in false positives is more pronounced than the reduction in false negatives. "Bad" advice has a comparable effect. The effect of "bad" advice on false negative outcomes is more pronounced than on false positives. Warnings have practically no effect on the relevance of "bad" warnings.

18 In the experiment, we use a less fine-grained scale, with only five, rather than ten, levels. We correct for this in the comparison. We note a limitation inherent in the nature of the data. COMPAS defines its decile scores relative to the complete training data which it uses to generate its predictions. We ask our participants for an absolute likelihood score. Yet given the breadth of the training data that COMPAS reports to use, this limitation should only be minor.

Figure 6. Effect of warnings after having decided without machine advice on accuracy



In terms of false positives (fpos) and false negatives (fneg). The effects of bad (b), neutral (n), and good (g) machine advice are shown separately for the treatments without warnings (no), and with warnings (warn).

Yet statistically we only find that good advice does indeed significantly and substantially reduce the incidence of both false positives and false negatives (Models 2 and 3 of Table 3). Bad advice has the opposite effect.[19] However, all interactions between the warnings mandated by the Supreme Court and the fact that advice is good or bad (and not neutral) are insignificant. All we find is a small main effect of warnings on false negatives, Model 3. Since this model controls for advice being good or bad, this is an effect of advice that is neutral. The COMPAS prediction does not inform participants that their assessment of the likelihood of recidivism was too optimistic, or too pessimistic for that matter.

We conclude

Result 2. Receiving the warnings mandated by the Wisconsin Supreme Court

19 Models with crossed random effects do not converge, which is why we revert to two-way clustering of standard errors.

Table 3. Effect of machine advice on accuracy

	False positives			False negatives		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
Warning	0.003	0.002	0.004	−0.004	−0.005	−0.007*
	(0.005)	(0.005)	(0.004)	(0.006)	(0.006)	(0.003)
Good advice		−0.154***	−0.161***		−0.081*	−0.091*
		(0.037)	(0.042)		(0.035)	(0.039)
Warning × good advice			0.014			0.020
			(0.031)			(0.015)
Bad advice		0.071*	0.081*		0.164***	0.168**
		(0.032)	(0.036)		(0.038)	(0.043)
Warning * bad advice			−0.019			−0.009
			(0.014)			(0.032)
Cons	−0.007	0.007†	0.007†	0.013	−0.004	−0.003†
	(0.012)	(0.004)	(0.004)	(0.014)	(0.004)	(0.002)

Notes: dvs are first differences: did the participant, in the case in question, and upon receiving machine advice, shift toward making a false positive/false negative verdict (coded as 1), shift away from a false positive/false negative verdict (coded as −1), or did their verdict not change (coded as 0)? Warning: COMPAS predictions came with the warnings mandated by the Wisconsin Supreme Court. Good/bad advice if the COMPAS score is closer to/further away from ground truth than the participant's likelihood rating before receiving machine advice and the participant's verdict before receiving advice was at variance with/in line with ground truth. Linear model. Standard errors clustered for fifty defendants and for 134 respondents in parenthesis. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$.

(a) does not make it more likely that a participant changes their verdict in the direction of ground truth if the COMPAS score has been closer to ground truth than their own likelihood rating before learning the COMPAS score,

(b) does not make it less likely that a participant changes their verdict in the direction away from ground truth if the COMPAS score has been further away from ground truth than their own likelihood rating before learning the COMPAS score.

6. IMMEDIATE CHOICE

In the conditions reported thus far, for each defendant (case) participants have first decided without knowing the computer's assessment. They have then learned the COMPAS prediction, and have been asked to revisit their earlier choices. This procedure has the advantage that we can, separately per participant and case, compare choices without and with the benefit of advice. Yet participants might want to choose consistently. The fact that they have already

decided once might therefore bias the effect of warnings downwards. The observation that warnings did have no significant effect on the choice between bail and jail, and that they did not make choices more accurate, might be an artifact of our design.

To test whether this concern is relevant, we rerun the experiment, but now ask for each decision only once. This of course requires a new baseline. In the new treatment, participants receive exactly the warnings mandated by the Supreme Court. We have tested another 134 participants. Demographics are reported in Table A2 in Appendix A.1. We have also preregistered the research question, design, hypotheses, and the analysis plan of these treatments with the Open Science Framework OSF (Engel & Grgić-Hlača 2020b). The design of the experiment has been approved by our institution's Ethical Review Board.[20] Participants again received £2.5 for their participation, but were not incentivized for the choices they made. The experiment took approximately twenty-five minutes.

Descriptives are summarized by Figure A3 in Appendix A.4. The regression in Table 4 shows that a desire to decide consistently did not bias results. The same way as in Table 2, we do not find a significant effect of warnings on verdict. Yet the regression finds a small, but significant effect on likelihood ratings (if they have been warned, participants are slightly more likely to predict that the defendant will be rearrested during the next two years). And unlike with the original design, when participants are warned, confidence (strongly) increases.

We thus conclude[21]

Result 3. When they receive the warnings mandated by the Wisconsin Supreme Court without having provisionally decided absent machine advice,

(a) participants deem it more likely that the defendant will be rearrested during the next two years,
(b) participants are more confident.

In these treatments, we cannot use the procedure introduced in Section 5 for defining whether machine advice is good or bad, as participants only decide once. In the treatment where participants receive the warnings mandated by the Supreme Court, the incidence of false positive verdicts (the participant decides for jail although the defendant did not recidivate) and false negative

20 Ethical Review Board of University of Saarland, December 13, 2019.

21 In the previously discussed treatments, the dependent variables were first differences between pre- and post-advice decisions. Here, we only gather post-advice decisions, and hence use this data as dependent variables. Due to this difference between the dependent variables, Result 3 is not directly comparable with Result 1.

Table 4. Effect of warning when machine advice is immediate on verdict, likelihood rating, and confidence

Propensity to jail	
Warning	0.033
	(0.032)
Cons	0.283***
	(0.039)
Likelihood of recidivism	
Warning	0.094*
	(0.040)
Cons	0.156***
	(0.043)
Confidence	
Warning	0.193***
	(0.038)
Cons	0.740***
	(0.042)
N	6,700

Notes: All dvs are absolutes. Warning: COMPAS predictions came with the warnings mandated by the Wisconsin Supreme Court. Standard errors with random effects for fifty defendants and for 134 respondents in parenthesis. ***$p < 0.001$, *$p < 0.05$.

verdicts (the participant decides for bail although the defendant did recidivate) is not significantly different from the treatment without a warning.[22]

7. ALTERNATIVE WARNINGS

We originally had intended to proceed sequentially. In the first step, we wanted to check whether the full set of warnings affects how our participants decide. Provided this yields significant and meaningful results, we had planned to split the warnings up and test their effect one by one. As the full set of warnings turns out immaterial for choices and accuracy, we have refrained from implementing this second step. We rather have added the supplementary treatments reported below, meant to test whether the warning can be made effective by a different design.

Warnings might be ineffective in the main experiment because they are not formulated in the most powerful way. If this were true, the policy implication would be straightforward: the presentation of the warnings should be improved. With two additional treatments, we test whether there is room for improvement in this respect.

22 These regressions are available from the authors upon request.

7.1 Simplified Warnings

The warnings mandated by the Wisconsin Supreme Court are couched in legalistic terms. The wordy and somewhat involved presentation might weaken their effect. This is why, in the first additional treatment, we replace them with the following *simple* (plain English) version:

> Judges should make their decision based on the evidence presented in court, and not only on the COMPAS risk score.
>
> COMPAS cannot be used to predict whether someone will commit a specific crime.
>
> COMPAS is proprietary software. The company does not reveal how it arrives at its prediction.
> The prediction quality hasn't been revaluated since February 2016.
> Predictions may be biased against African American or Hispanic people.

Except for these changes, the design is the same as in the main experiment. We have again pre-registered the research question, design, hypotheses, and the analysis plan with the Open Science Framework OSF (Engel & Grgić-Hlača 2020b). The design of the experiment has been approved by our institution's Ethical Review Board.[23] We had sixty-eight participants in this additional treatment who have passed both attention checks. Sample composition is reported in Table A2 in Appendix A.1. Participants again received £2.5 for their participation, but were not incentivized for the choices they made. The experiment took approximately thirty-three minutes.

Descriptives are summarized and compared with the results from the original experiment, in Figure A4 in Appendix A.5. As the first panel of the regression in Table 5 shows, simplified warnings do not influence verdicts. Compared with the baseline, in which participants do not receive any warnings, the effect of warnings is not significantly different from zero.[24] We also find no difference between the baseline with no warnings and receiving the simplified warnings on likelihood ratings, as shown in the middle panel of Table 5. However, as the lower panel of Table 5 shows, simplified warnings slightly reduce the effect of machine advice on confidence. This effect is weakly significant ($p = 0.051$).

We conclude:

Result 4. When they are warned in plain English,

23 Ethical Review Board of University of Saarland, December 13, 2019.

24 In this regression, the explanatory variable varies between, not within participants. We are therefore again not in a position to compare a random effects with a fixed effects model, and to run the Hausman test.

Table 5. Treatment effects of simplified warnings

Propensity to jail	
Simplified warnings	0.004
	(0.025)
Cons	0.117***
	(0.020)
Likelihood of recidivism	
Simplified warnings	−0.021
	(0.030)
Cons	0.531***
	(0.023)
Confidence	
Simplified warnings	−0.060†
	(0.031)
Cons	0.097***
	(0.023)
N	6,750

Notes: The first two dvs are the absolutes of first differences, i.e., $|dv_{post} - dv_{pre}|$. Confidence is a first difference, i.e., $conf_{post} - conf_{pre}$. Standard errors with random effects for fifty defendants and for 135 respondents in parenthesis. $^{***}p < 0.001$, $^{\dagger}p < 0.1$.

(a) participants do change their likelihood ratings less than without the warnings,
(b) participants are less confident.

Accuracy is also virtually identical to the *baseline*. If the advice is "good," i.e., if the COMPAS score is closer to ground truth than the participant's likelihood rating before seeing the score, and the participant's verdict before receiving the advice was at variance with ground truth, the frequency of false positive verdicts is on average reduced by 15 percent with no warnings, while it is reduced by 16 percent when participants receive the simplified warnings. The frequency of false negative verdicts is reduced by 9 percent in both treatments. If the advice is "bad," i.e., if the COMPAS score is further away from ground truth than the participant's likelihood rating before seeing the score, and the participant's verdict before receiving the advice was in line with ground truth, the incidence of false positive verdicts increases by 9 percent in the baseline, and by 7 percent with simplified warnings. The incidence of false negatives increases by 16 percent in the baseline, and by 14 percent with simplified warnings. None of these treatment comparisons reach significance.[25]

25 These regressions are available from the authors upon request.

7.2 Quantitative Information about Accuracy

The warnings mandated by the Supreme Court are confined to qualitative statements. Judges are in particular not informed that the performance of COMPAS is actually pretty poor. As mentioned, in earlier work one of us has shown that the accuracy of a decision aid trained using the ProPublica COMPAS dataset is as low as 68 percent (Grgić-Hlača et al. 2018): in about a third of all cases, the software wrongly predicts recidivism or its absence for that matter. In the second additional treatment, we add this information to the (original) instructions, using the following sentence:

> Predictions are only correct in about 68% of all cases.

Except for these changes, the design is identical to the main experiment. As for the simplified warnings, we have pre-registered the study with OSF (Engel & Grgić-Hlača 2020b) and received our institution's Ethical Review Board approval.[26] Sixty-nine participants passed both attention checks in this treatment, and their demographics are reported in Table A2 in Appendix A.1. The experiment took approximately thirty-eight minutes, and participants were remunerated £2.5 for their participation. No additional monetary incentives were utilized.

As the regressions in Table 6 and Figure A5 in Appendix A.5 show, this manipulation does not have a significant effect on verdicts, but it does induce participants to change their likelihood ratings less upon receiving the advice. Confidence is also reduced.

We conclude:

Result 5. When they receive quantitative information about the low accuracy of COMPAS scores,

(a) participants do change their likelihood ratings less than without the warnings,
(b) participants are less confident.

Descriptively, making the low accuracy of COMPAS scores explicit makes participants slightly more sensitive to "good" advice: the incidence of false positives is reduced by 18 percent, rather than 16 percent when they are not warned (the incidence of false negatives remains unaffected). They also become slightly less sensitive to "bad" advice: the incidence of false negatives only increases by 11 percent, while it increases by 16 percent without warnings (the remaining effects essentially stay constant). Yet none of these treatment effects turns out significant.[27]

26 Ethical Review Board of University of Saarland, December 13, 2019.

27 These regressions are available from the authors upon request.

Table 6. Treatment effects of quantitative information about accuracy

Propensity to jail	
Accuracy information	−0.009
	(0.021)
Cons	0.117***
	(0.018)
Likelihood of recidivism	
Accuracy information	−0.093***
	(0.026)
Cons	0.531***
	(0.022)
Confidence	
Accuracy information	−0.072*
	(0.028)
Cons	0.097***
	(0.022)
N	6, 800

Notes: The first two dvs are the absolutes of first differences, i.e., $|dv_{post} - dv_{pre}|$. Confidence is a first difference, i.e., $conf_{post} - conf_{pre}$. Standard errors with random effects for fifty defendants and for 135 respondents in parenthesis. $^{***}p < 0.001$, $^{*}p < 0.05$.

8. VIOLENT RECIDIVISM

8.1 Wisconsin Law

Judges in Wisconsin receive the complete COMPAS output. It covers three risk scores (for "General Recidivism Risk," "Violent Recidivism Risk," and "Pretrial Release Risk"). In the previous parts of the article, we have focused on general recidivism risk. One may wonder whether violent recidivism risk is more relevant for the choice between bail and jail.

Wisconsin law is not very precise on this question. Art. I (8) of the Constitution reads

> All persons, before conviction, shall be eligible for release under reasonable conditions designed to assure their appearance in court, protect members of the community from serious bodily harm, or prevent the intimidation of witnesses.

Chapter 969.01 (1) 1 Wisconsin Statutes repeats this verbatim. The concern for "serious bodily harm" is also mentioned in Chapter 969.01 (4) 2, 965.035 (3) (c) and 969.035 (6) (b). This might suggest that only violent recidivism is relevant for the choice between bail and jail.

Yet Chapter 969.02 (4) and 969.03 (4) stipulate

> As a condition of release in all cases, a person released under this section shall not commit any crime.

This might suggest that general recidivism is relevant as well. Unfortunately, the question does not seem to be settled in the literature.[28]

8.2 Design

To be on the safe side, we have rerun the experiment, and now have given participants the violent recidivism score, rather than the general recidivism score. From the ProPublica dataset, we also know whether the defendant in question has been rearrested for violent crime during the two years after release, which we use in this version of the experiment as the measure for ground truth.

Now in the original cases, very few defendants have committed violent crime within two years of release—only six out of fifty defendants. Had we used the original cases, but the violent recidivism scores and corresponding ground truth, the set of cases would have been very unbalanced. This imbalance would likely have biased results.[29] This is why, for the new experimental wave, we have selected a different set of cases from the set used by Dressel & Farid (2018), and have oversampled serious cases, such that we now have 50 percent violent recidivism. Hence technically, we have randomly selected twenty-five cases where there actually has been no violent recidivism, and twenty-five cases where there actually has been violent recidivism.[30]

Yet from the perspective of experimental control this means we have two changes, compared with the original experiment: we have shifted from general recidivism to violent recidivism, and we have changed the set of stimulus cases. In the interest of maintaining experimental control, we therefore have added two bridge treatments. In the new cases no warning treatment, we only replace

28 The only pertinent paper we could spot is Woelfel (2016). The author stresses how little guidance the judges get for decisions about pretrial detention. In places, she explains that the courts may impose pretrial detention to preempt "new criminal activity" (e.g., 226, 235), and discusses the "safety of the community" in general (e.g., 229). Yet the author proposes new legislation that would constrain the justification of pretrial detention to "protect members of the community from serious bodily harm" (227).
We have tried to elicit an assessment from a colleague teaching criminal law at the University of Wisconsin-Madison Law School, and from a trial judge from Wisconsin, but did not receive a response.

29 The concern would have been particularly pronounced in the later treatment, building on the present, where, after every round, participants receive feedback about ground truth.

30 For details about case composition, *see* Appendix Table A4.

the original with the new set of cases, but otherwise keep everything as in the *baseline*. We thus give participants COMPAS' general recidivism score, and we define ground truth as any new crime for which the defendant has been apprehended two years after release. In the new cases warning treatment, we add the exact warning as mandated by the Supreme Court. Strictly speaking, we can compare the violent no warning and the violent warning treatments only with the new cases no warning and new cases warning treatments.

As for the previous studies, we have received approval from our institution's Ethical Review Board for the experimental design.[31] Two hundred and sixty-three participants who took part in this study successfully passed both attention checks—131 for the treatments considering general recidivism, and 132 for the violent recidivism ones. The sample compositions are reported in Table A2 in Appendix A.1. Participants received £2.5 for their participation and were not incentivized for the choices they made, as in the previous experiments. The experiment took approximately thirty-seven minutes on average.

8.3 Alternative Cases, General Recidivism

Descriptively, with the alternative cases, we see small effects of the warnings mandated by the Supreme Court on changes in verdicts, likelihood ratios, and confidence, Figure A6 in Appendix A.6. Yet only the effect on likelihood ratings turns out significant. When being warned about the limitations of machine advice, participants are 0.076 (of four) points less likely to change their rating.[32]

With the alternative cases, the effects of the warnings on accuracy are also minimal, and not significantly different from zero.[33] Without warnings, "good" advice makes it 13 percent more likely that participants avoid a false positive verdict. With the warnings, this fraction goes down to 9 percent. Hence descriptively in this respect, the warnings are even slightly counterproductive. In all other respects, warnings have virtually no effect: The frequency of false negative verdicts is reduced by 24 percent upon receiving "good" advice if participants receive no warnings, and by 23 percent if they receive the warnings mandated by the Supreme Court. Warnings have no effect on the sensitivity toward "bad" advice. False positives increase by 3 percent without, and

31 Ethical Review Board of University of Saarland, December 13, 2019.

32 $p = 0.004$; the complete structural model that simultaneously estimates effects on changes in verdicts, likelihood, and confidence are available from the authors upon request.

33 The regressions are available from the authors upon request.

by 2 percent with the warnings. False negatives increase by 18 percent without advice and by 17 percent with the warnings.

8.4 Alternative Cases, Violent Recidivism

Descriptively, if participants are asked to decide upon violent recidivism, the warnings mandated by the Supreme Court radicalize the effect of machine advice on verdicts: either they are less, or they are more likely to change their verdict, Figure A7 in Appendix A.6. Yet statistically these effects cancel out.[34] The effect of machine advice on likelihood ratings and confidence visibly is very similar with and without warnings, and is indeed not significantly different. Warnings do also not significantly affect accuracy.[35] Without and with warnings, "good" advice makes it 11 percent less likely that participants issue a false positive verdict, and 10 percent less likely that they issue a false negative verdict. "Bad" advice without warnings makes it 9 percent more likely that participants issue a false positive verdict. This fraction is 10 percent with warnings. "Bad" advice without warnings makes if 14 percent more likely that participants issue a false negative verdict, and 12 percent if they have received the warnings.

9. A BATTERY OF CHANGES

Our study has been triggered by the ruling of the Wisconsin Supreme Court. The court has been sensitive to normative concerns about machine advice in the courtroom. But the court aims at neutralizing these concerns with the set of mandatory warnings. This strategy is viable only if warnings do have the desired effect. In the data reported thus far, we have found very little. In most specifications, we find an effect of warnings on likelihood ratings, often also on confidence. But we find no effect on verdicts and only very occasional effects on accuracy.[36] Before concluding that warnings are just pointless, we make one last effort.

In this last effort, we simultaneously implement a whole battery of changes. These two treatments build on the *violent* treatments. We are thus using the alternative, more severe cases, and ask for violent recidivism. We combine four additional manipulations:

34 The structural model is available from the authors upon request.

35 The regressions are available from the authors upon request.

36 In Table 3, we find an effect of warnings that are neutral on false negatives. No further effects are significant at conventional levels.

(1) after each case, participants receive correct feedback about ground truth. They are thus informed whether the defendant has indeed been rearrested for a violent crime;
(2) participants receive a monetary incentive for making a prediction in line with ground truth, and they lose money when predicting the opposite (for detail about the incentive scheme, please see Appendix A.7);
(3) we make it explicit in the instructions that the only criterion for deciding the case is the risk of violent recidivism;
(4) we remove two sentences from the instructions that had been meant to be completely transparent with participants, but that might misleadingly have reduced their trust in COMPAS advice (for detail, please see Appendix A.7).

We stress two obvious limitations: between the previous treatments and this one, we have changed (many) more than one element. We are therefore not in a position to isolate the effect of one of these manipulations. All four manipulations likely reduce external validity. In the field, judges only rarely learn whether the present defendant has been recidivating. Unless the case is exceptionally salient and the news spread in the media or in the judiciary, the individual judge only learns if she has again jurisdiction. As many states randomly assign cases to judges, this is not common. At any rate, judges never learn about recidivism on the spot. We thus create an artificially powerful learning environment. Judges are never paid for the decisions they make (they would be corrupt), and nonfinancial incentives for judges are at least not immediate. Again unless the case is particularly salient (so that the disposition comes to the attention of authorities deciding about promotion, or the electorate deciding about reelection), how a judge has chosen between jail and bail in an individual case is unlikely to matter more than tangentially.[37] Also in this respect, the treatment is thus an upper bound of forces that might be at play in the field. Finally, as explained in Section 8, in the field judges do not decide exclusively on the basis of recidivism risk, and hence face a cognitively more demanding choice problem. We use all these manipulations simultaneously as seemingly "nothing works" (Farabee 2005).

While the many manipulations forbid direct comparisons with earlier treatments, in one final dimension, we enable a strict test: We replace the somewhat dry warnings mandated by the Supreme Court of Wisconsin by a set of very

37 For a discussion and experimental tests of some channels, *see* Engel & Zhurakhovska (2017).

graphic warnings. They are represented in Figure A2 in Appendix A.3. This attempt at enhancing the effect is motivated by similar efforts regarding graphical warnings on, e.g., cigarette packages (Hammond et al. 2007; Borland et al. 2009) and sugary drinks (Donnelly et al. 2018).

We again have preregistered the research question, design, hypotheses, and the analysis plan with the Open Science Framework OSF (Engel & Grgić-Hlača 2020c), and received approval from our institution's Ethical Review Board for the experimental design.[38] Two hundred and ten of our participants successfully passed both attention check questions, and the sample composition is shown in Table A2 in Appendix A.1. All participants received a base fee of £2.5 for their participation. However, unlike the previous experiments, they were incentivized for accuracy. Specifically, they were incentivized to provide estimates of recidivism risk aligned with the ground truth. The incentive scheme is detailed in Table A5 in Appendix A.7. The experiment took forty-four minutes on average.

As this is the main point, we focus data analysis on verdicts and accuracy. Again stressing that direct comparisons are not possible, Figure 7 suggests a clear story: In the original data, and with alternative cases but general recidivism risk, verdicts on average almost perfectly coincide whether or not participants receive the warnings mandated by the Supreme Court. Descriptively participants already become less likely to change their verdict upon receiving the advice from COMPAS if violent recidivism is at stake. The difference is a slight bit stronger once we add the battery of additional manipulations. Yet only if, additionally, we replace the warnings mandated by the Supreme Court by the very graphic warnings depicted in Appendix do we see a real shift of the density curve: participants become considerably less willing to change their verdict. This visual impression is supported by statistical analysis, Table 7. With our full battery of manipulations, but the warnings as mandated by the Supreme Court, warnings only have a weakly significant effect ($p = 0.074$) on verdict. Only if, additionally, we replace the warnings with the graphic set represented in the Appendix do we find an effect at conventional levels: decision-makers are predicted to reduce the probability of changing their verdict in the light of COMPAS advice from 14 percent to 8.7 percent, $p = 0.002$.

Is this good news? Figure 8 suggests otherwise. Good and bad advice matter, in the expected direction. If the participant's verdict was in line with ground truth, but the COMPAS score is further away from ground truth than their likelihood rating before seeing the machine advice (i.e., if advice has been bad), whether or not participants are warned, the incidence of false positive and of

38 Ethical Review Board of University of Saarland, December 13, 2019.

Figure 7. Effect of Supreme Court warnings on verdicts

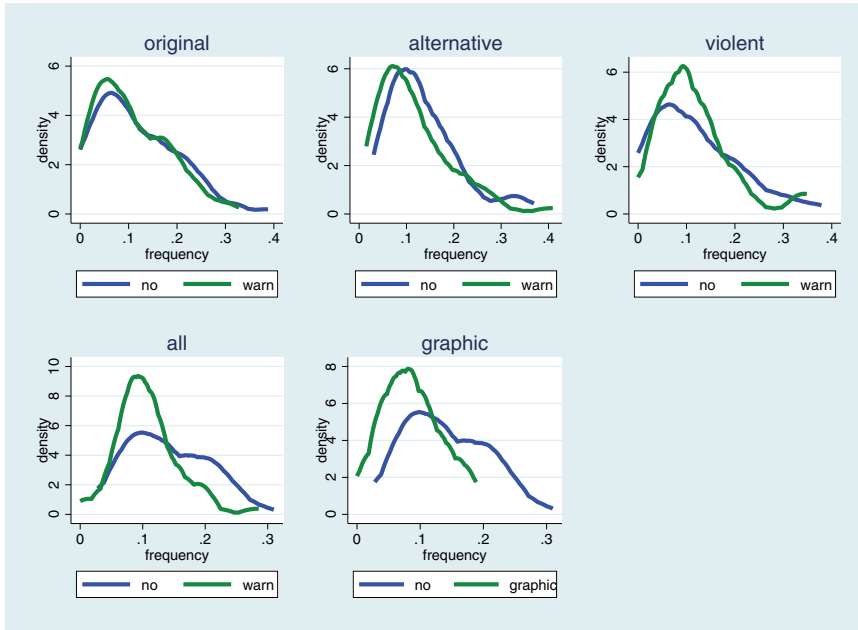


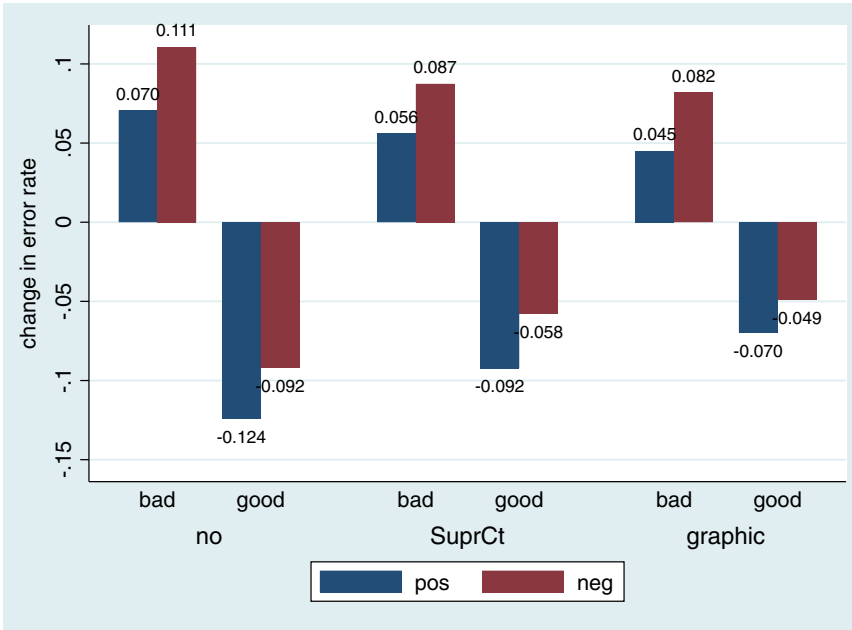
Table 7. Effect of Supreme Court warnings on verdicts

	Original	Alternative	Violent	All	Graphic
Supreme Court	−0.011	−0.018	−0.002	−0.033†	−0.053**
	(0.018)	(0.019)	(0.018)	(0.018)	(0.017)
Cons	0.117***	0.135***	0.119***	0.140***	0.140***
	(0.016)	(0.017)	(0.017)(	(0.014)	(0.013)
N	6,700	6,550	6,600	7,050	7,000

Notes: Dependent variable is absolute of the first difference of verdicts, i.e., $|verdict_{post} - verdict_{pre}|$. Supreme Court: COMPAS predictions came with the warnings mandated by the Wisconsin Supreme Court. Standard errors with random effects for fifty defendants and for $N/50$ respondents in parenthesis. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{\dagger}p < 0.1$.

false negative verdicts increases. Descriptively this detrimental effect is most pronounced without a warning, a little less pronounced with the warnings mandated by the Supreme Court, and least pronounced if participants receive the graphic warnings. Yet in the regressions in Tables 8 and 9 all interaction

Figure 8. Effect of warnings on accuracy with battery of changes



dvs are first differences. pos: change in fraction of false positive, after receiving machine advice; neg: change in fraction of false negatives, after receiving machine advice. bad: COMPAS score has been further away from ground truth than participant's likelihood rating pre-advice, and participant's verdict was in line with ground truth; good: COMPAS score has been closer to ground truth than participant's likelihood rating pre-advice, and participant's verdict was at variance with ground truth. All treatments use full battery of changes. no: participants do not receive a warning; SuprCt: participants receive the warnings mandated by the Supreme Court; graphic: participants receive the warnings in graphic form.

effects are far from significance. Warnings have no documented effect if advice is bad.

This is different for "good" advice, i.e., if the participant's verdict has been at variance with ground truth before seeing the advice, and the COMPAS score is closer to ground truth than the participant's likelihood rating before advice. Overall such advice helps. On average, verdicts move into the normative direction, and both the frequencies of false positives and of false negatives are reduced. Yet the advice reduces the incidence of false positives and of false negatives less if participants have been warned. The normatively desirable effect of "good" advice is reduced. The interactions with good advice in the regressions in Tables 8 and 9 show that this counterproductive effect is significant for false positives and false negatives if the graphic warnings are used, and there is a

Table 8. Effect of warnings on false positives

	Original	Alternative	Violent	All	Graphic
Warning	0.004	0.002	0.005	−0.005	−0.005
	(0.004)	(0.004)	(0.005)	(0.007)	(0.008)
Good advice	−0.161***	−0.134*	−0.112***	−0.144***	−0.144***
	(0.042)	(0.050)	(0.029)	(0.033)	(0.033)
Warning ×	0.014	0.041	−0.013	0.036	0.060*
good advice	(0.031)	(0.026)	(0.019)	(0.025)	(0.025)
Bad advice	0.081*	0.025	0.090*	0.050*	0.050*
	(0.036)	(0.016)	(0.042)	(0.024)	(0.024)
Warning ×	−0.019	−0.013	−0.002	−0.010	−0.020
bad advice	(0.014)	(0.011)	(0.015)	(0.017)	(0.020)
Cons	0.007†	0.006*	0.005†	0.020*	0.020*
	(0.004)	(0.002)	(0.003)	(0.008)	(0.008)

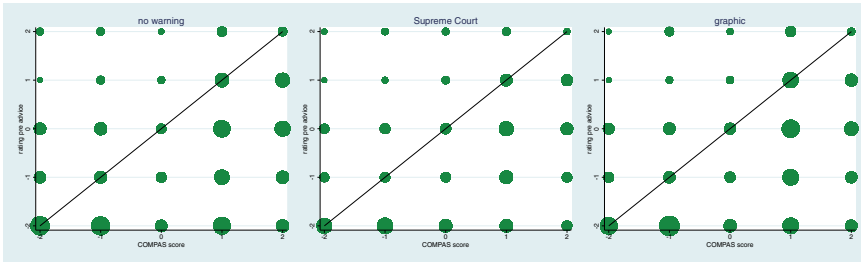
Notes: dvs are first differences: did the participant, in the case in question, and upon receiving machine advice, shift toward making a false positive verdict (coded as 1), shift away from a false positive verdict (coded as −1) or did their verdict not change (coded as 0)? Warning: COMPAS predictions came with the warnings mandated by the Wisconsin Supreme Court/with graphic warnings. Good/bad advice if the COMPAS score is closer to/further away from ground truth than the participant's likelihood rating before receiving machine advice and if participant's verdict was at variance with/in line with ground truth before receiving the advice. Linear model. Standard errors clustered for fifty defendants and for 134, 131, 132, 141 and 140 respondents, respectively, are in parenthesis. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$.

Table 9. Effect of warnings on false negatives

	Original	Alternative	Violent	All	Graphic
Warning	−0.007*	−0.001	0.004	0.008	−0.001
	(0.003)	(0.008)	(0.005)	(0.007)	(0.007)
Good advice	−0.091*	−0.218***	−0.083*	−0.073**	−0.073**
	(0.039)	(0.043)	(0.032)	(0.027)	(0.027)
Warning ×	0.020	0.011	−0.006	0.026†	0.044*
good advice	(0.015)	(0.042)	(0.016)	(0.015)	(0.017)
Bad advice	0.168***	0.201***	0.154***	0.130***	0.130***
	(0.043)	(0.043)	(0.040)	(0.030)	(0.030)
Warning ×	−0.009	−0.007	−0.016	−0.031	−0.028
bad advice	(0.032)	(0.032)	(0.027)	(0.022)	(0.023)
Cons	−0.003†	−0.019**	−0.018**	−0.019*	−0.019*
	(0.002)	(0.007)	(0.006)	(0.008)	(0.008)

Notes: dvs are first differences: did the participant, in the case in question, and upon receiving machine advice, shift toward making a false positive verdict (coded as 1), shift away from a false positive verdict (coded as −1) or did their verdict not change (coded as 0)? Warning: COMPAS predictions came with the warnings mandated by the Wisconsin Supreme Court/with graphic warnings. Good/bad advice if the COMPAS score is closer to/further away from ground truth than the participant's likelihood rating before receiving machine advice and if participant's verdict was at variance with/in line with ground truth before receiving the advice. Linear model. Standard errors clustered for fifty defendants and for 134, 131, 132, 141 and 140 respondents, respectively, are in parenthesis. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$.

Figure 9. Human vs. machine assessment



y-axis: participants' likelihood ratings before receiving machine advice, on a Likert scale from −2 (very unlikely that defendant will recidivate) to 2 (very likely). *x*-axis: COMPAS score, recoded to match scale of participants' ratings. Bubble size indicates frequency.

weakly significant effect on false negatives if the warnings mandated by the Supreme Court are used instead.

This gives us our final

Result 6. When combined with feedback and incentives for being accurate, graphic warnings

(a) make it less likely that participants follow COMPAS advice,
(b) make it less likely that participants' verdicts are in line with ground truth.

How come we find this asymmetry between the effect of (in particular graphic) warnings between "good" and "bad" advice? As we had not expected the asymmetry, we can only offer an ex post rationalization. Figure 9 suggests an explanation. If a rating is on the 45° line, the participant's rating and the COMPAS score coincide. If the rating is above the line, the participant has been more optimistic than COMPAS. If the rating is below the line, the participant has been more pessimistic. Visibly participants have much more often been more pessimistic, not more optimistic. In an earlier experiment, we have shown that participants are more willing to follow COMPAS advice if it suggests releasing the participant on bail, rather than putting them into jail (Grgić-Hlača et al. 2019). Hence warnings weaken the effect of advice when it otherwise would have been most likely to be effective.[39]

39 We have checked that the relation between participants' ratings and COMPAS scores is not an artefact of oversampling severe cases. It is also present in the *baseline*. The mean (transformed) COMPAS score is −0.58, while the mean likelihood rating of participants is 0.202.

10. CONCLUSION

Humans are fallible. Over thousands of years, the legal system has come to terms with this fallibility. A thick institutional arrangement is meant to reduce the impact of human weakness on judicial decision-making. Judges must be properly trained. They are carefully selected. For many decisions, they decide in benches. The adversarial system gives actors power to intervene that has every interest to spot and counteract this fallibility. Most judicial decisions are open to appeal. Judges are obliged to give explicit reasons. Their reputation, and possibly career, are at stake if they make wrongful decisions.

Machines are deterministic. The output can be precisely traced back to the input. Would it therefore be preferable to have machines decide, rather than humans? While trial by computer is largely still science fiction, computer input into human judicial decision-making is a reality. In many US jurisdictions, this for instance holds for the choice between putting a person apprehended for arguably having committed a crime in jail, or releasing them on bail. The law wants the expectation to weigh in on this decision whether the defendant would commit another crime before tried. The COMPAS software provides a computer-generated prediction. In Wisconsin, this prediction is also used at the sentencing stage.

Yet alas, deterministic machines are fallible as well. Machine learning is powerful. But it only generates predictions. These predictions may turn out to be wrong. Actually, in the specific case of predicting recidivism, they do quite frequently. The COMPAS software is only correct in about 68 percent of all cases.

This pronounced degree of failure is the core of the prominent *Loomis* case. Ultimately, the Supreme Court of Wisconsin has cleared the use of the software regardless. Yet it has mandated that the computer prediction comes with a series of explicit warnings. This is where the present article comes in. It replicates the situation in court with laypersons. They receive sketches of cases that courts have decided, and of which ground truth is known, i.e., whether the defendant has recidivated during the two years after release. Participants are asked to assess the recidivism risk, and to choose between bail and jail.

Results are sobering. If we give participants the exact warnings mandated by the Wisconsin Supreme Court, we find no significant effect on the normatively most important variables: participants do not become less likely to revise their verdict in the light of machine advice. Their verdicts do not become more accurate. They are not more likely to put a defendant on bail that has indeed not recidivated or to put a defendant into jail that has recidivated. This result is not driven by the desire to be consistent with their provisional decision, before

getting access to machine advice. Replacing the legal jargon of the warnings with a version in plain English, and informing participants about the low accuracy rate of the COMPAS software, does not have an effect either. The warnings of the Supreme Court do also not have a significant effect on verdict or accuracy if we test participants on a set of more severe cases, or if we shift from general to violent recidivism.

Yet we do find an effect if we add a whole battery of interventions. We give participants perfect continuous feedback. We give them a monetary incentive to find ground truth. And we make it explicit that the verdict should hinge on the assessment of the recidivism risk. Jointly, these manipulations yield a small, weakly significant effect on verdicts. Participants become a little less likely to change their verdict after seeing the COMPAS score. We see a stronger effect, which is significant at conventional levels, if we additionally give participants very graphic warnings. Yet the effect is counterproductive. Participants become even less accurate. They become less likely to follow the advice when this would bring them closer to the true state of the world.

One should always be cautious when extrapolating from an experiment to the real world. In the interest of cleanly identifying an effect, experiments remove elements of the real-life situation that might matter. We cannot exclude this. But we do not deem it likely. If policymakers are willing to be guided by our results, we have a note of caution. Unless they are made very strong, warnings risk being pointless. We deem it questionable whether policymakers will be comfortable with making warnings as strong as we had to make them to find an effect. And even if policymakers were open to this strategy, they should beware of a counterproductive effect. This is at least what we have found: if warnings work, they make human decisions even less accurate. This suggests that the seeming compromise is problematic. If there are serious concerns about machine advice (as, in our case, not being sufficiently accurate or, in other cases, biased against vulnerable parts of the population), policymakers should rather directly control the algorithm, or even prohibit its use in court.

A. APPENDIX

A.1. PARTICIPANTS' DEMOGRAPHICS

We report the sample composition for Studies 2 through 5, from Sections 6 to 9, in Appendix Table A2.

Table A1. Mean values of participants' responses to questions related to the survey they completed, across all treatments

	Mean Rating
Survey was interesting	1.39
Would participate again	1.68
Questions difficult to understand	−1.51
Warnings difficult to understand	−0.74
Questions difficult to answer	−0.97

The participants responded on a 5-point Likert scale, from "Strongly disagree" (−2) to "Strongly agree" (2)

A.2. PARTICIPANTS' FEEDBACK

At the end of the experiment, we asked our respondents to answer a series of questions related to the survey they just completed. They were asked to tell us how much they agree with the following sentences, on a 5-point Likert scale: (i) This survey was interesting, (ii) I would like to take part in a similar survey in the future, (iii) The questions were difficult to understand, (iv) The quote from the document that judges receive together with COMPAS predictions was difficult to understand, and (v) The questions were difficult to answer. Question (iv) was only shown in treatments where warnings were shown.

In Table A1, we report our respondents' average answers. Descriptively, we find that, in all of our treatments, the respondents found the survey interesting, and reported willingness to participate in a similar experiment in the future. Also, they did not find the survey questions difficult to understand. However, as expected, due to the complex nature of the decision-making task at hand, they found the questions more difficult to answer than to understand. Finally, while participants reported that the warnings were not difficult to understand, across all treatments where warnings were shown, the simplified warnings were rated as easier to understand than the other treatments, with an average rating of −1.25.

A.3. STIMULUS MATERIAL

In Figure A1, we show the two attention check questions that we used to check whether our participants were paying attention to the task. In Figure A2, we show the graphical warnings, which we used in Study 5. In Tables A3 and A4, we list the full set of 100 vignettes used in our experiments.

Table A2. Demographics additional treatments

Demographic attribute	Study 2	Study 3		Study 4		Study 5	
	Immediate choice	Simplified warnings	Quant. info. about acc.	Alt. cases, general rec.	Alt. cases, violent rec.	A battery of changes	Census
Total respondents	138	68	72	131	132	216	–
Passed attention checks	134	68	69	131	132	210	–
Female (%)	37	41	41	51	45	50	51
African American (%)	5	10	9	5	6	9	13
Asian (%)	17	7	17	11	12	12	6
Caucasian (%)	70	74	67	77	72	70	61
Hispanic (%)	7	6	6	5	7	7	18
Other (%)	1	2	1	3	3	2	4
Bachelor's degree or above (%)	94	96	96	93	98	92	30
Liberal (%)	50	51	48	53	54	55	33[a]
Conservative (%)	18	18	23	21	24	20	29[a]
Moderate (%)	32	31	25	23	20	23	34[a]
Other (%)	0	0	4	4	2	1	4[a]
Jury duty experience (%)	67	69	71	65	68	66	27[b]

Notes: Comparison between survey samples for Studies 2–5 and 2016 US Census (US Census Bureau 2016). Attributes marked with a [a] compared to Pew data (Pew Research Center 2016) for political leaning. Attributes marked with a [b] were compared to DRI 2012 National Poll data (DRI 2012). 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. Demographics reported for respondents who passed the attention checks.

Table A3. **Vignettes used as stimulus material in Studies 1–3**

Case ID	Sex	Age	Crime	Misd./ fel.	No. of priors	No. of juv. fel.	No. of juv. misd.	Ground truth	COMPAS score
222	F	43	Driving under the influence	M	0	0	0	0	1
763	M	26	Assault with a deadly weapon	F	7	0	0	1	5
1385	M	28	Battery	M	4	0	0	1	2
1490	M	30	Battery	M	4	0	0	1	4
1870	M	20	Carrying a concealed weapon	F	0	0	0	0	3
2078	M	43	Burglary	F	4	0	0	1	5
2278	M	63	Battery	M	0	0	0	1	1
2438	M	67	Battery	M	0	0	0	0	1
2502	M	45	Grand theft	F	0	0	0	0	1
2533	M	19	Battery	M	0	0	0	1	6
2632	M	56	Possession of cocaine	F	4	0	0	0	1
2898	F	31	Battery	M	0	0	0	0	3
2914	M	29	False imprisonment	F	1	0	0	0	9
3146	M	37	Tampering with a witness	F	0	0	0	1	3
3229	M	26	Driving under the influence	F	2	0	0	0	5
3255	M	51	Driving with a suspended license	M	4	0	0	0	2
3347	F	36	Criminal damage of less than $1,000	F	7	0	0	1	3
3552	M	28	Battery	M	0	0	0	0	3
3679	F	47	Prostitution	M	0	0	0	1	3
3805	M	28	Grand theft	F	0	0	0	1	2
4029	M	28	Battery	M	0	0	0	1	2
4553	M	34	Operating a vehicle without a valid drivers license	M	2	0	0	0	3
4593	M	38	Battery	M	0	0	0	0	1
5057	F	49	Driving with a suspended license	F	1	0	0	0	1
5069	M	24	Prostitution	M	2	1	0	1	10
5128	M	39	Battery	M	1	0	0	0	2
5217	M	37	Battery	F	0	0	0	1	1
5449	M	32	Burglary	F	1	0	0	0	3
5543	M	45	Battery	M	3	0	0	1	3
5921	M	52	Restraining order violation	M	4	0	0	1	3
6000	F	23	Burglary	F	0	0	0	1	5
6094	M	30	Dealing controlled substances	F	7	0	0	1	10
6207	M	38	Extradition of defendants	M	0	0	0	1	8
6658	M	39	Forgery	F	0	0	0	1	4
6908	M	30	Driving with a revoked license	F	21	0	0	1	7
7012	M	28	Possession of Cannabis/ Marijuana	F	0	0	0	0	2
7894	M	20	Burglary	F	0	0	0	0	8
7975	M	23	Possession of Cannabis/ Marijuana	F	11	8	2	1	10
8149	M	27	Domestic violence	M	3	0	0	0	3
8438	F	34	Possession of cocaine	F	12	0	0	1	10
9041	M	36	Driving under the influence	M	0	0	0	0	1
9231	M	35	Possession of cocaine	F	3	0	0	0	4
9303	M	30	Driving with a suspended license	M	0	0	0	0	9
9556	M	30	Assault with a deadly weapon	F	1	0	0	0	2
9953	M	23	Grand theft	F	1	0	0	1	3
10406	M	43	Possession of cocaine	F	2	0	0	0	7
10580	F	22	Domestic violence	M	1	0	0	0	8
10762	F	26	Driving under the influence	M	1	0	0	0	4
10807	M	25	Driving with a suspended license	M	3	0	0	1	9
10946	M	27	Driving under the influence	M	1	0	0	0	5

Table A4. **Vignettes used as stimulus material in Studies 4 and 5**

Case ID	Sex	Age	Crime	Misd./ fel.	No. of priors	No. of juv. fel.	No. of juv. misd.	General ground truth	General COMPAS	Violent ground truth	Violent COMPAS
6690	M	25	Resisting an officer with violence	F	7	0	1	1	10	1	10
8924	M	27	Assault	F	1	0	0	1	4	1	5
5614	M	33	Battery with a deadly weapon	F	2	0	0	1	8	1	7
5217	M	37	Battery	F	0	0	0	1	1	1	1
4643	M	32	Battery	M	12	0	1	1	10	1	6
9392	M	53	Battery	M	1	0	0	0	5	1	5
154	F	39	Assault with a deadly weapon	F	0	0	0	1	1	1	1
3387	M	52	Domestic violence	M	1	0	0	1	1	1	1
7245	M	25	Battery	M	5	0	0	1	8	1	6
6481	F	23	Battery	M	0	0	0	0	4	1	4
8761	M	32	Battery	M	4	0	0	1	2	1	3
10699	M	57	Possession of cocaine	F	28	0	0	1	10	1	6
4703	M	25	Possession of Cannabis/ Marijuana	F	0	0	0	1	2	1	3
3236	M	29	Possession of cocaine	F	12	0	0	1	10	1	7
5646	M	23	Possession of cocaine	F	6	0	0	1	6	1	7
8881	M	40	Possession of cocaine	F	0	0	0	1	10	1	9
2865	M	22	Forgery	F	2	0	0	1	4	1	6
5917	M	32	Driving under the influence	M	1	0	0	1	4	1	2
7807	M	34	Grand theft	F	13	0	0	1	8	1	3
9235	M	56	Driving under the influence	F	1	0	0	0	1	1	1
5593	M	37	Assault	M	4	0	0	1	4	1	3
10660	M	23	Battery	M	0	0	0	1	2	1	4
4082	M	24	Tampering with a witness	F	0	0	0	1	5	1	5
7037	M	26	Possession of a controlled substance	F	6	0	1	1	10	1	9
7123	M	27	Battery	M	0	0	0	1	2	1	3
5754	M	34	Possession of a controlled substance	F	13	0	0	1	4	0	1
7003	M	61	Driving with a suspended license	M	20	0	0	1	2	0	1
4098	M	26	Robbery	F	4	0	0	1	4	0	5
7445	M	31	Assault	F	1	0	0	1	8	0	8
9852	M	21	Grand theft	F	2	0	2	1	10	0	7
7631	M	22	Possession of cocaine	F	0	0	0	1	5	0	5
2396	F	57	Disorderly conduct	M	0	0	0	0	4	0	1
7472	M	80	Battery	M	0	0	0	0	1	0	1
203	F	24	Battery	M	0	0	0	0	2	0	3
3246	M	30	Driving under the influence	M	0	0	0	0	3	0	2
8239	M	28	Grand theft	F	7	0	0	0	10	0	6
6314	M	20	Grand theft	F	0	0	0	1	6	0	9
8140	M	31	Driving with a revoked license	F	16	3	0	1	10	0	8
10919	M	35	Battery	M	12	0	0	1	6	0	3
1841	F	37	Driving under the influence	M	1	0	0	0	3	0	2
8285	M	28	Battery	M	8	0	0	1	6	0	6
6939	M	62	Purchasing cocaine	F	2	0	0	0	1	0	1
8866	F	34	Battery	M	1	0	0	0	3	0	1
7099	M	34	Possession of Cannabis/ Marijuana	F	2	0	0	0	2	0	1
6766	M	29	Possession of a controlled substance	F	8	0	1	1	8	0	4
5852	F	55	Possession of Cannabis/ Marijuana	F	1	0	0	0	1	0	1
9644	M	22	Possession of Cannabis/ Marijuana	M	0	0	0	0	6	0	6
2943	M	42	Assault with a deadly weapon	F	0	0	0	0	1	0	1
7166	F	33	Failure to redeliver hired or leased property	F	1	0	0	0	5	0	2
2329	M	25	Grand theft	F	14	0	4	1	10	0	7

Figure A1. Attention check questions

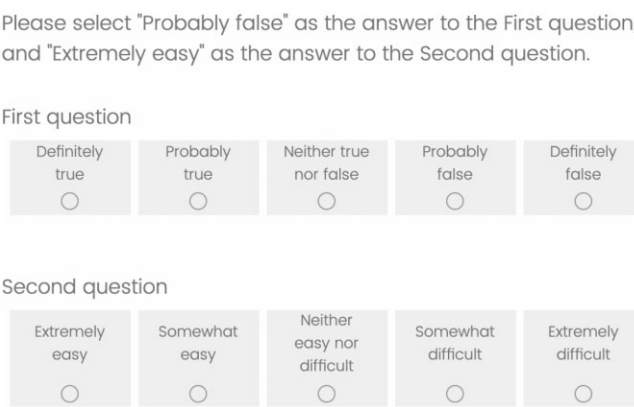


Figure A2. Graphical warnings, as shown for defendant 2329, used in Study 5. The warnings were animated, and shown in a loop

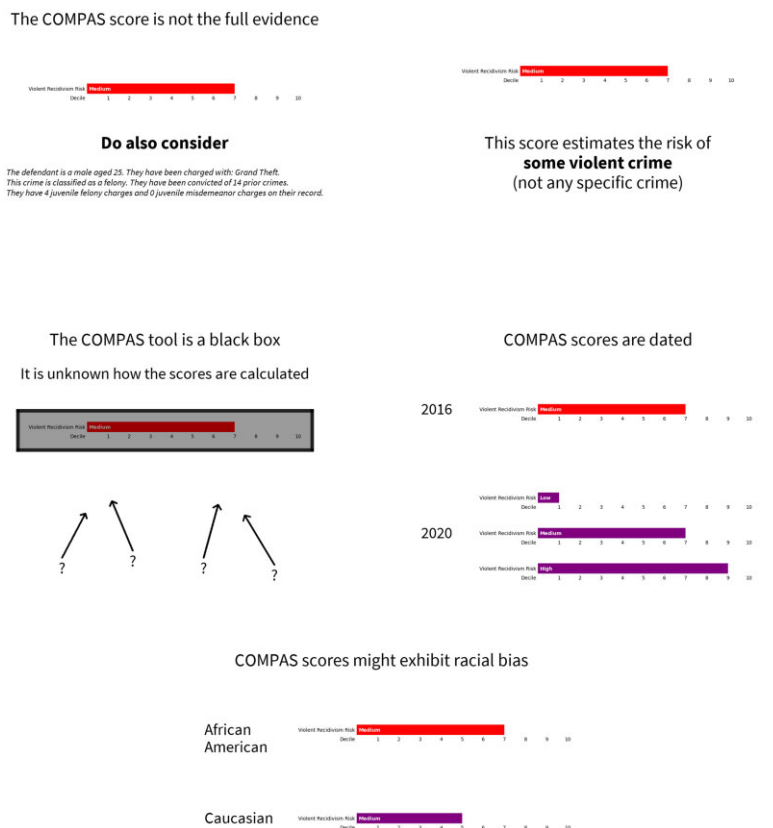
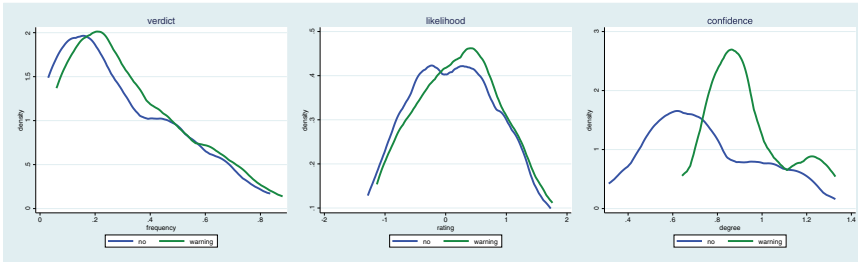
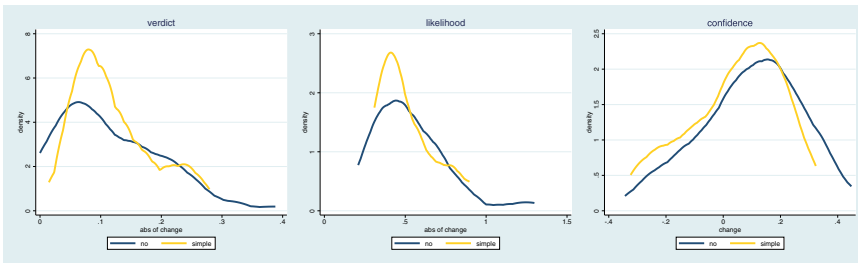


Figure A3. Effect of warning when machine advice is immediate on verdict, likelihood rating, and confidence



x-axis: left panel: mean frequency of jail verdicts per case; middle panel: mean likelihood rating per case (on a 5-point Likert scale running from −2 .. 2); right panel: mean confidence rating per case (on a 5-point Likert scale running from −2 .. 2); *y*-axis: kernel density

Figure A4. Effect of simplified warnings on propensity to jail, likelihood rating, and confidence

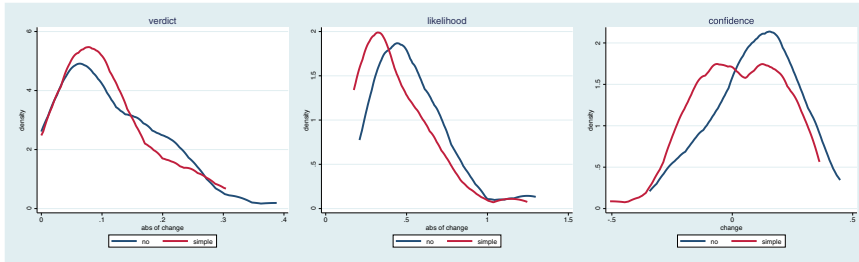


x-axis: left panel: absolute of change in jail verdicts pre- vs. post-advice per case; middle panel: absolute of change in likelihood rating pre- vs. post-advice per case (on a 5-point Likert scale running from −2 .. 2); right panel: change in confidence rating per case (on a 5-point Likert scale running from −2 .. 2); *y*-axis: kernel density

A.4. IMMEDIATE CHOICE—ADDITIONAL DETAILS

Here we provide additional experimental results related to Study 2, discussed in Section 6. Descriptives are summarized in Figure A3. Density plots not only show the respective central tendency, but the complete distribution. As we have each dependent variable only once, we do not report first differences, but mean choices per condition. Visibly, verdicts and likelihood ratings do almost not react to warnings. Not only central tendencies, but even distributions are also almost identical. Confidence however reacts strongly. Interestingly, if they

Figure A5. Effect of quantitative information about accuracy on propensity to jail, likelihood rating, and confidence



x-axis: left panel: absolute of change in jail verdicts pre- vs. post-advice per case; middle panel: absolute of change in likelihood rating pre- vs. post-advice per case (on a 5-point Likert scale running from −2 .. 2); right panel: change in confidence rating per case (on a 5-point Likert scale running from −2 .. 2); *y*-axis: kernel density

are warned against potential limitations of computer advice, participants become even more confident.

A.5. ALTERNATIVE WARNINGS—ADDITIONAL DETAILS

Below we show additional results related to Study 3, presented in Section 7. Results are summarized, and compared with the results from the original experiment, in Figures A4 and A5. We report the effect of receiving machine advice, conditional on treatment. All dependent variables are thus first differences, comparing the dependent variable in question before and after receiving advice. Density plots not only show the respective central tendency, but the complete distribution.

As the left panel of Figure A4 shows, simplified warnings make it descriptively less likely that the advice has an effect (the density plot peaks at about 10 percent of all fifty defendants/cases). However, the regression in Table 5 shows that this effect is not significant. As the middle panel of Figure A4 shows, we also visibly find no difference between the *baseline* with no warnings and receiving the simplified warnings on likelihood ratings. Finally, as the right panel of Figure A4 and the regression in Table 5 show, simplified warnings have a weakly significant effect ($p = 0.051$) on confidence, slightly reducing the effect of machine advice on confidence.

Figure A5 summarizes the effects of quantitative information about accuracy. In line with the regression from Table 3, we visually do not find a difference in

Figure A6. Effect of warnings with alternative set of cases on propensity to jail, likelihood rating, and confidence

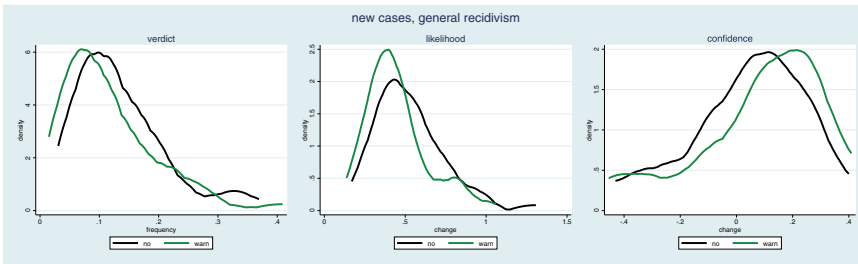
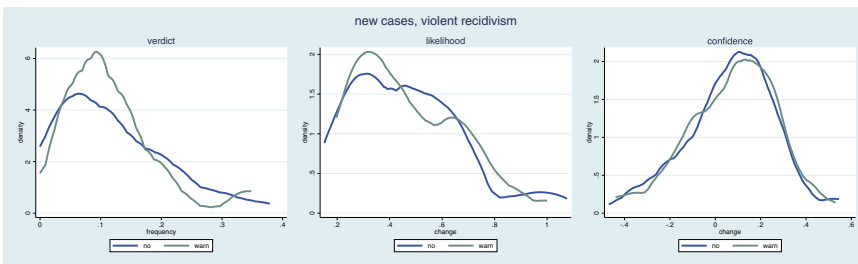


Figure A7. Effect of warnings with violent recidivism scores on propensity to jail, likelihood rating, and confidence



terms of verdicts. However, we see a negative effect both on the change in likelihood ratings and confidence.

A.6. VIOLENT RECIDIVISM—ADDITIONAL DETAILS

In this section, we show additional results related to Study 4, from Section 8. Participants were presented with an alternative set of cases, which differ from those used in Studies 1–3.

Figure A6 summarizes the effect of warnings when participants were asked to decide upon general recidivism. Descriptively, we see small effects of the warnings on all three dependent variables. However, only the effect on likelihood ratings was found to be significant in our regression analysis.

Figure A7 summarizes the effects of warnings when participants were asked to consider violent recidivism, instead of general recidivism. Descriptively, the warnings made the effect of machine advice on verdicts more extreme, making participants either less, or more likely to change their verdict. However, these

Table A5. Monetary incentives for accuracy, used in Study 5

Your answer	Reward if the defendant committed a violent crime within two years	Reward if the defendant did not commit a violent crime within two years
Extremely likely	4	0
Likely	3	1
Neither	2	2
Unlikely	1	3
Extremely unlikely	0	4

effects cancel out, and we observe no statistically significant effect of warnings on verdicts. For likelihood ratings and confidence scores, we do not observe an effect of warnings, neither visually nor statistically.

A.7. A BATTERY OF CHANGES—ADDITIONAL DETAILS

Here we describe the manipulations used in Study 5 from Section 9 in more detail.

We used monetary rewards to incentivize accuracy. Specifically, participants were incentivized to accurately predict criminal recidivism risk. We described the incentive scheme to the participants in the introductory instructions as follow:

> We already know whether these defendants committed a violent crime in the next 2 years or not. Depending on how good you are in predicting whether these defendants will commit a violent crime, you can earn a **bonus payment**.
>
> Throughout the survey, you will be asked to tell us how likely you think it is that the defendants will commit a violent crime, and you will earn rewards based on the table below, in cents:
>
> E.g., if the defendant **committed** a violent crime and you answered that they were **Extremely likely** to commit a crime, you will earn $0.04. On the other hand, if you answered **Extremely unlikely**, you will earn $0.00. If you answer **Neither**, you earn $0.02 with certainty – but miss the opportunity to earn an even higher amount (emphasis in original).

After each case, participants were provided feedback about the ground truth related to the defendant's violent recidivism. They were also reminded of the COMPAS prediction, their own prediction, as well as a summary of how their

response affected their monetary reward, in line with the incentive scheme described above.

Additionally, we made it explicit that the only criterion for deciding between bail and jail in this experiment is the risk of violent recidivism. We updated the introductory instructions by adding the sentence below:

> In this case, the critical question for choosing between bail and jail is your estimate that the defendant will commit violent crime before trial.

Finally, we removed the following two sentences from the instructions, to avoid reducing the participants' trust in the COMPAS tool's predictions:

> Note, however, that the Supreme Court ruling directly addresses the use of COMPAS predictions in sentencing, while our study is concerned with the use of these predictions when deciding whether to grant bail. Finally, our study is confined to one of three scores provided by the COMPAS software, the "General Recidivism Risk". We omit the "Violent Recidivism Risk" and "Pretrial Release Risk" scores.

REFERENCES

Angwin, Julia, Jeff Larson, Surya Mattu & Lauren Kirchner. 2016. Machine Bias: There's Software Used across the Country to Predict Future Criminals - And It's Biased Against Blacks. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Apsler, Robert & David O. Sears. 1968. Warning, Personal Involvement, and Attitude Change. 9 *J. Pers. Soc. Psychol.* 162.

Argo, Jennifer J & Kelley J. Main. 2004. Meta-Analyses of the Effectiveness of Warning Labels. 23 *J. Public Policy Mark.* 193–208.

Beriain, Iñigo De Miguel. 2018. Does the Use of Risk Assessments in Sentences Respect the Right to Due Process? A Critical Analysis of the Wisconsin v. Loomis Ruling. 17 *Law Probab. Risk* 45–53.

Berk, Richard. 2017. An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism. 13 *J. Exp. Criminol.* 193–216.

——— 2019. *Machine Learning Risk Assessments in Criminal Justice Settings.* New York: Springer.

Berk, Richard A. & Justin Bleich. 2013. Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment. 12 *Criminol. Public Policy* 513.

Berman, Emily. 2018. A Government of Laws and Not of Machines. 98 *B. U. L. Rev.* 1277.

Blank, Hartmut & Céline Launay. 2014. How to Protect Eyewitness Memory against the Misinformation Effect: A Meta-Analysis of Post-Warning Studies. 3 *J. Appl. Res. Mem. Cogn.* 77–88.

Blomberg, Thomas, William Bales, Karen Mann, Ryan Meldrum & Joe Nedelec. 2010. *Validation of the COMPAS Risk Assessment Classification Instrument.* Tallahassee, FL: College of Criminology and Criminal Justice, Florida State University.

Borland, Ron, Nick Wilson, Geoffrey T. Fong, David Hammond, K. Michael Cummings, Hua-Hie Yong, Warwick Hosking, Gerard Hastings, James Thrasher & Ann McNeill. 2009. Impact of Graphic and Text Warnings on Cigarette Packs: Findings from Four Countries over Five Years. 18 *Tob. Control* 358–364.

Carlson, Alyssa M. 2017. The Need for Transparency in the Age of Predictive Sentencing Algorithms. 103 *Iowa Law Rev.* 303.

Casey, Pamela M., Jennifer K. Elek, Roger W. Warren, Fred Cheesman, Matt Kleiman & Brian Ostrom. 2014. Offender Risk & Needs Assessment Instruments: A Primer for Courts. URL https://nicic.gov/offender-risk-needs-assessment-instruments-primer-courts.

Chouldechova, Alexandra. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. 5 *Big Data* 153–163.

Cino, Jessica Gabel. 2017. Deploying the Secret Police: The Use of Algorithms in the Criminal Justice System. 34 *Ga. State Univ. Law Rev.* 1073.

Collins, Erin. 2018. Punishing Risk. 107 *Ga Law J.* 57.

Deaux, Kay K. 1968. Variations in Warning, Information Preference, and Anticipatory Attitude Change. 9 *J. Pers. Soc. Psychol.* 157.

Deese, James. 1959. On the Prediction of Occurrence of Particular Verbal Intrusions in Immediate Recall. 58 *J. Exp Psychol.* 17.

Deskus, Cassie. 2018. Fifth Amendment Limitations on Criminal Algorithmic Decision-Making. 21 *NYU J. Legislation Public Policy* 237.

Desmarais, Sarah L., Kiersten L. Johnson & Jay P. Singh. 2016. Performance of Recidivism Risk Assessment Instruments in us Correctional Settings. 13 *Psychol. Serv.* 206.

Dieterich, William, Christina Mendoza & Tim Brennan. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. *Northpoint Inc.*

Dietvorst, Berkeley J., Joseph P. Simmons & Massey Cade. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. 144 *J. Exp. Psychol. Gen.* 114.

Donnelly, Grant E., Laura Y. Zatz, Dan Svirsky & John K. Leslie. 2018. The Effect of Graphic Warnings on Sugary-Drink Purchasing. 29 *Psychol Sci* 1321–1333.

Doshi-Velez, Finale & Been Kim. 2017. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv*:1702.08608.

Dressel, Julia & Farid Hany. 2018. The Accuracy, Fairness, and Limits of Predicting Recidivism. 4 *Sci. Adv.* eaao5580.

DRI. 2012. DRI *2012 National Poll on the Civil Justice System*. URL https://www.dri.org/advocacy/center-for-law-and-public-policy/poll/2012-poll/.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold & Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Eaglin, Jessica M. 2017. Constructing Recidivism Risk. 67 *Emory Law J.* 59–122.

Echterhoff, Gerald, William Hirst & Walter. Hussy 2005. How Eyewitnesses Resist Misinformation: Social Postwarnings and the Monitoring of Memory Characteristics. 33 *Mem. Cogn.* 770–782.

Engel, Christoph & Nina Grgić-Hlača. 2020a. Machine Advice with a Warning about Machine Limitations: Experimentally Testing the Solution Mandated by the Wisconsin Supreme Court. URL https://osf.io/gewns?view_only=60363e5bc128410ab4214fa245dc925f.

Engel, Christoph & Nina. Grgić-Hlača 2020b. Machine Advice with a Warning about Machine Limitations: Experimentally Testing the Solution Mandated by the Wisconsin Supreme Court. Supplementary Treatments. URL https://osf.io/5xua2?view_only=7081f7833878416893fe640b6b5234b0.

——— 2020c. Machine Advice with a Warning about Machine Limitations: Experimentally Testing the Solution Mandated by the Wisconsin Supreme Court. Supplementary Treatments 2. URL https://osf.io/hy7fs?view_only=89e6521aeb9d4fc49620daea209994da.

Engel, Christoph & Lilia Zhurakhovska. 2017. You Are in Charge: Experimentally Testing the Motivating Power of Holding a Judicial Office. 46 *J. Legal Stud.* 1–50.

Farabee, David. 2005. *Rethinking Rehabilitation: Why Can't We Reform Our Criminals?* Washington, DC: AEI Press.

Farabee, David, & Zhang Sheldon. 2007. *COMPAS Validation Study: First Annual Report*. Los Angeles, CA: Department of Corrections and Rehabilitation.

Fass, Tracy L., Kirk Heilbrun, David DeMatteo & Fretz Ralph. 2008. The LSI-R and the COMPAS: Validation Data on Two Risk-Needs Tools. 35 *Crim. Justice Behav.* 1095–1108.

Freeman, Katherine. 2016. Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis. 18 *N. C. J. Law Technol.* 75.

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

Grgić-Hlača, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi & Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 51-60.

Groves, Robert M., Floyd J. Fowler Jr, Mick P. Couper, James M. Lepkowski, Eleanor Singer & Roger Tourangeau. 2011. *Survey Methodology*, vol. 561. Hoboken, NJ: John Wiley & Sons.

Hamilton, Melissa. 2019a. The Biased Algorithm: Evidence of Disparate Impact on Hispanics. 56 *Am. Crim. L. Rev.* 1553.

——— 2019b. The Sexist Algorithm. 37 *Behav. Sci Law* 145–157.

Hammond, David, Geoffrey T. Fong, Ron Borland, K. Michael Cummings, Ann McNeill & Pete. Driezen 2007. Text and Graphic Warnings on Cigarette Packages: Findings from the International Tobacco Control Four Country Study. 32 *Am. J. Prevent. Med.* 202–209.

Harvard LR. 2016. Criminal Law - Sentencing Guidelines - Wisconsin Supreme Court Requires Warning before Use of Algorithmic Risk Assessments in Sentencing - State v. Loomis. 130 *Harv. Law Rev.* 1530.

Holsinger, Alexander M., Christopher T. Lowenkamp, Edward Latessa, Ralph Serin, Thomas H. Cohen, Charles R. Robinson, Anthony W. Flores & Scott W. VanBenschoten. 2018. A Rejoinder to Dressel and Farid: New Study Finds Computer Algorithm Is More Accurate than Humans at Predicting Arrest and as Good as a Group of 20 Lay Experts. 82 *Fed. Prob.*50.

Huq, Aziz Z. 2019. Racial Equity in Algorithmic Criminal Justice. 68 *Duke Law J.* 1043.

Israni, Ellora. 2017. Algorithmic Due Process: Mistaken Accountability and Attribution in State v. Loomis. URL https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in-state-v-loomis-1.

Kaskutas, Lee A. 1993. Changes in Public Attitudes toward Alcohol Control Policies since the Warning Label Mandate of 1988. 12 *J. Public Policy Market.* 30–37.

Kleinberg, Jon, Lakkaraju Himabindu, Leskovec Jure, Jens Ludwig & Sendhil. Mullainathan 2017. Human Decisions and Machine Predictions. 133 *Q. J. Econ.* 237–293.

Lansing, Sharon. 2012. New York State COMPAS-Probation Risk and Need Assessment Study: Examining the Recidivism Scale's Effectiveness and Predictive Accuracy. *Criminal Justice Research Report.*

Larson, Jeff, Mattu Surya, Lauren Kirchner & Julia. Angwin 2016. How We Analyzed the COMPAS Recidivism Algorithm. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Lin, Zhiyuan, Jung Jongbin, *et al.* 2020. The Limits of Human Predictions of Recidivism. 6 *Sci. Adv.* eaaz0652.

Liu, Han-Wei, Ching-Fu Lin & Yu-Jie. Chen 2019. Beyond State v Loomis: Artificial Intelligence, Government Algorithmization and Accountability. 27 *Int. J. Law Inf. Technol.* 122–141.

McCabe, David P. & Anderson D. Smith. 2002. The Effect of Warnings on False Memories in Young and Older Adults. 30 *Memo. Cogn.* 1065–1077.

Nishi, Andrea. 2019. Privatizing Sentencing. 119 *Colum. Law Rev.* 1671–1710.

Oeberst, Aileen & Hartmut. Blank 2012. Undoing Suggestive Influence on Memory: The Reversibility of the Eyewitness Misinformation Effect. 125 *Cognition* 141–159.

Palan, Stefan & Christian. Schitter 2018. Prolific.ac—a Subject Pool for Online Experiments. 17 *J. Behav. Exp. Finan.* 22–27.

Pew Research Center. 2016. *2016 Party Identification Detailed Tables. URL https://www.pewresearch.org/politics/2016/09/13/2016-party-identification-detailed-tables/.*

Raghu, Maithra, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer & Sendhil Mullainathan. 2019. The Algorithmic Automation Problem: Prediction, Triage, and Human Effort. *arXiv preprint arXiv:1903.12220.*

Redmiles, Elissa M., Yasemin Acar, Sascha Fahl & Michelle L. Mazurek. 2017. A Summary of Survey Methodology Best Practices for Security and Privacy Researchers. Technical Report.

Rizer, Arthur & Caleb. Watney 2018. Artificial Intelligence Can Make Our Jail System More Efficient, Equitable, and Just. 23 *Tex. Rev. Law Policy* 181.

Roediger, Henry L. & Kathleen B. McDermott. 1995. Creating False Memories: Remembering Words Not Presented in Lists. 21 *J. Exp. Psychol Learn. Mem. Cogn.* 803.

Schul, Yaacov. 1993. When Warning Succeeds: The Effect of Warning on Success in Ignoring Invalid Information. 29 *J. Exp. Soc. Psychol.* 42–62.

Simmons, Ric. 2017. Big Data and Procedural Justice: Legitimizing Algorithms in the Criminal Justice System. 15 *Ohio State J. Crim. Law* 573.

Skeem, J & J. Eno Louden, 2007. Assessment of Evidence on the Quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). Unpublished report prepared for the California Department of Corrections and Rehabilitation. https://webfiles.uci.edu/skeem/Downloads.html.

Spielkamp, Matthias. 2017. Inspecting Algorithms for Bias. 120 *MIT Technol. Rev.* 96–98.

Starr, Sonja B. 2014. Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. 66 *Stanford Law Rev.*803.

Stevenson, Megan. 2018. Assessing Risk Assessment in Action. 103 *Minn Law Rev.*303.

Stevenson, Megan T. & Christopher. Slobogin 2018. Algorithmic Risk Assessments and the Double-Edged Sword of Youth. 36 *Behav. Sci. Law* 638–656.

Tan, Sarah, Julius Adebayo, Kori Inkpen & Ece. Kamar 2018. Investigating Human+ Machine Complementarity for Recidivism Predictions. *arXiv preprint arXiv:1808.09123.*

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

Verwijmeren, Thijs, Johan C. Karremans, Stefan F. Bernritter, Wolfgang Stroebe & Daniël H.J. Wigboldus 2013. Warning: You Are Being Primed! The Effect of a Warning on the Impact of Subliminal Ads. 49 *J. Exp. Soc. Psychol.* 1124–1129.

Washington, Anne L. 2019. How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate. 17 *Colo. Technol. Law J. 131-160.*

Wilder, Bryan, Eric Horvitz & Ece. Kamar 2020. Learning to Complement Humans. *arXiv preprint arXiv:2005.00582.*

Woelfel, Tiffany. 2016. Go Directly to Jail, Do Not Pass Go, Do Not Collect $200: Improving Wisconsin's Pretrial Release Statute. 2016 *Wisconsin Law Rev.* 207–237.

Wogalter, Michael S., Scott T. Allison & Nancy A. McKenna. 1989. Effects of Cost and Social Influence on Warning Compliance. 31 *Hum. Factors* 133–140.

Wogalter, Michael S., Sandra S. Godfrey, Gail A. Fontenelle, David R. Desaulniers, Pamela R. Rothstein & Kenneth R. Laughery. 1987. Effectiveness of Warnings. 29 *Hum. Factors* 599–612.

Wright, Daniel B. 1993. Misinformation and Warnings in Eyewitness Testimony: A New Testing Procedure to Differentiate Explanations. 1 *Memory* 153–166.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rogriguez & Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Artificial Intelligence and Statistics*, 962–970.

Zhang, Sheldon X., Robert E.L. Roberts & David. Farabee 2014. An Analysis of Prisoner Reentry and Parole Risk Using COMPAS and Traditional Criminal History Measures. 60 *Crime Delinq.* 167–192.