The limits of human predictions of recidivism

Zhiyuan "Jerry" Lin,[1] Jongbin Jung,[1] Sharad Goel,[1] Jennifer Skeem[2]

[1]Stanford University
[2]University of California—Berkeley

Dressel and Farid (*1*) recently found that laypeople were as accurate as statistical algorithms in predicting whether a defendant would reoffend, casting doubt on the value of risk assessment tools in the criminal justice system. We report the results of a replication and extension of Dressel and Farid's experiment. Under conditions similar to the original study, we found nearly identical results, with humans and algorithms performing comparably. However, algorithms beat humans in the three other datasets we examined. The performance gap between humans and models was particularly pronounced when—in a departure from the original study—participants were not provided with immediate feedback on the accuracy of their responses. Models also outperformed humans when the information provided for predictions included an enriched (vs. restricted) set of risk factors. These results suggest that statistical models can outperform human predictions of recidivism in ecologically valid settings.

Introduction

Algorithms and predictive analytics inform decisions in almost every sector of public policy, including criminal justice. When judges, correctional authorities, and parole boards make decisions regarding incarceration, supervision, and release, they now routinely turn to *risk assessment instruments* (RAIs)—checklists that summarize "risk factors" for estimating a person's likelihood of future reoffending. The chief rationale is a belief that RAIs outperform unaided human judgment in predicting recidivism (*2, 3*).

The validity of this rationale, however, has been questioned.[1] In a recent high-profile study, Dressel and Farid (*1*) found that a widely used RAI called COMPAS "is no more accurate ... than predictions made by people with little or no criminal justice expertise." The authors recruited 400 online workers through Amazon's Mechanical Turk platform to participate in an experiment on predicting crime. They showed each participant 50 short descriptions of real defendants drawn from a publicly available COMPAS dataset, and asked participants to indicate whether they thought each defendant would commit another crime within two years. Averaging across these responses, participants' overall accuracy was 62%, comparable to the accuracy of algorithmic COMPAS predictions (65%).

But a closer look suggests that the study elicited laypeople's predictions in a manner that may not best represent unaided human judgment—particularly the kind that judges, probation officers, and other professionals must exercise when predicting reoffending in the real world. In particular, the study design focused people's attention on the most predictive factors and promoted learning over the course of the experiment, perhaps boosting accuracy rates as a result. In a new series of experiments, we test the impact of three conditions on the relative accuracy of human judgment and RAIs in predicting reoffense. Collectively, these experiments are de-

[1] Aside from their effectiveness, some have also questioned the equity of risk assessment instruments. See, for example, Angwin et al. (*4*) and responses to those critiques (*5–9*).

signed to illuminate both the situations in which humans can predict recidivism as accurately as algorithms, and settings in which algorithms can provide better estimates than humans.

First, we test the impact of providing streamlined versus enriched information for prediction. Dressel and Farid provided people with brief vignettes that listed five risk factors for recidivism per case in narrative form: the defendant's sex, age, current charge, and number of prior adult and juvenile offenses. This format mimics structured checklists of selective risk factors that have been shown to increase professionals' ability to make accurate predictions (*10*). However, the information available in justice settings is far less constrained. Presentence investigation reports, attorney and victim impact statements, and the defendant's demeanor all add complex, inconsistent, risk irrelevant, and potentially biasing information. We hypothesize that algorithms predict better than humans when both are provided with more complex or "noisy" risk information.

We test this hypothesis by manipulating whether streamlined information (Dressel and Farid's five risk factors) or enriched information (those five factors plus ten more) is provided. We ensure that all information is consistent and risk-relevant. Given that the COMPAS dataset lacked these additional risk factors, we used similar datasets on an RAI called the "Level of Service Inventory – Revised", or LSI-R (*11*). The LSI-R includes, for example, information on one's criminal history, employment status, and substance use. For each of the 10 risk factors assessed by the LSI-R, we wrote phrases to describe each score on that factor (e.g., "has a serious drinking problem that interferes with work" for a substance abuse score of 3). These phrases were combined to create enriched vignettes that described many aspects of real offenders, as a contrast to streamlined vignettes that only described the most predictive risk factors for the same offenders.

Second, we test the impact of providing people with feedback on their accuracy across a series of trials. In each of the 50 rounds of Dressel and Farid's study, participants made a

prediction, were informed whether the prediction was correct (and their cumulative accuracy), and then moved on to the next vignette. In other words, prediction events were experienced sequentially, with immediate feedback on accuracy. This created a "kind" environment—one shown to be ideal for humans to intuitively learn the probabilities of specific outcomes, even when the rules are not transparent (*12*). Kind environments can promote accuracy, unlike the "wicked" learning environments that characterize most justice settings, where outcomes cannot be observed immediately or are never observed at all (*10, 13*). In the absence of such feedback, we hypothesize that algorithms predict better than humans. We test this hypothesis by manipulating whether people are provided with feedback on their accuracy, using both the COMPAS dataset and our LSI-R datasets.

Third, we test the impact of base rates—a group's probability of reoffending—on the relative predictive accuracy of algorithms and humans. Base rates vary substantially across contexts. For example, in the COMPAS data used by Dressel and Farid the base rate of re-arrest for any type of crime was 48%, whereas the base rate of re-arrest for violent crime in the same dataset was only 11%. Even when people are explicitly told base rates, they often fail to update their prior beliefs—a phenomenon called "base rate neglect" (*14–17*). Statistical algorithms, in contrast, are designed to incorporate this information accurately and consistently. For this reason, we expect the accuracy of human predictions, but not the accuracy of algorithmic decisions, to be particularly sensitive to base rates.

There is one common exception to this expectation. People often do take base rates into account when the probabilities are "directly experienced through trial-by-trial outcome feedback" (*18*). Dressel and Farid's feedback protocol creates this kind of intuitive learning environment. When feedback is provided, we thus expect the accuracy of humans to be substantially less sensitive to base rates. We test these twin hypotheses by varying the base rate of recidivism

in both the COMPAS dataset (using any vs. violent re-arrest) and the LSI-R datasets (where base rates differ by location), across feedback conditions.

Design

Following Dressel and Farid, we recruited participants on Amazon's Mechanical Turk platform to estimate the likelihood that defendants would be rearrested within two years of release, based on brief descriptions of the individuals. In the original study, participants were simply asked for binary yes/no predictions of recidivism. We altered this design to instead elicit predictions on a 30-point probability scale. To do so, as shown in Figure **1**, we first asked participants to select one of six risk buckets, ranging from "almost certainly NOT arrested (1-16%)" to "almost certainly arrested (84-99%)". Based on this initial response, we then asked individuals to select one of five subcategories to obtain more specific probability estimates. For example, in the lowest risk bucket, the subcategories were 2%, 5%, 8%, 12%, and 15%. Participants could not indicate exactly 50% likelihood of rearrest, and so reported probabilities could unambiguously be converted to binary predictions based on a 50% probability threshold.

We extended the original study in three additional ways. First, whereas Dressel and Farid focused on a single dataset, we repeated our experiments on four: (1) COMPAS balanced base rate assessments of any recidivism in Broward County, Florida—the dataset used by Dressel and Farid; (2) COMPAS low base rate assessments of *violent* recidivism, also in Broward County; (3) LSI-R balanced base rate assessments of recidivism in North Dakota; and (4) LSI-R low base rates assessments of recidivism in Texas.[2] A summary of these four datasets is presented in Table 1. Second, we examined the effects of immediate feedback on human predictions. In the original study, participants were told after each prediction whether a defendant was indeed rear-

[2]In the first three datasets, "recidivism" means rearrest; in the fourth, recidivism means reincarceration. See the Materials and Methods for more detail.

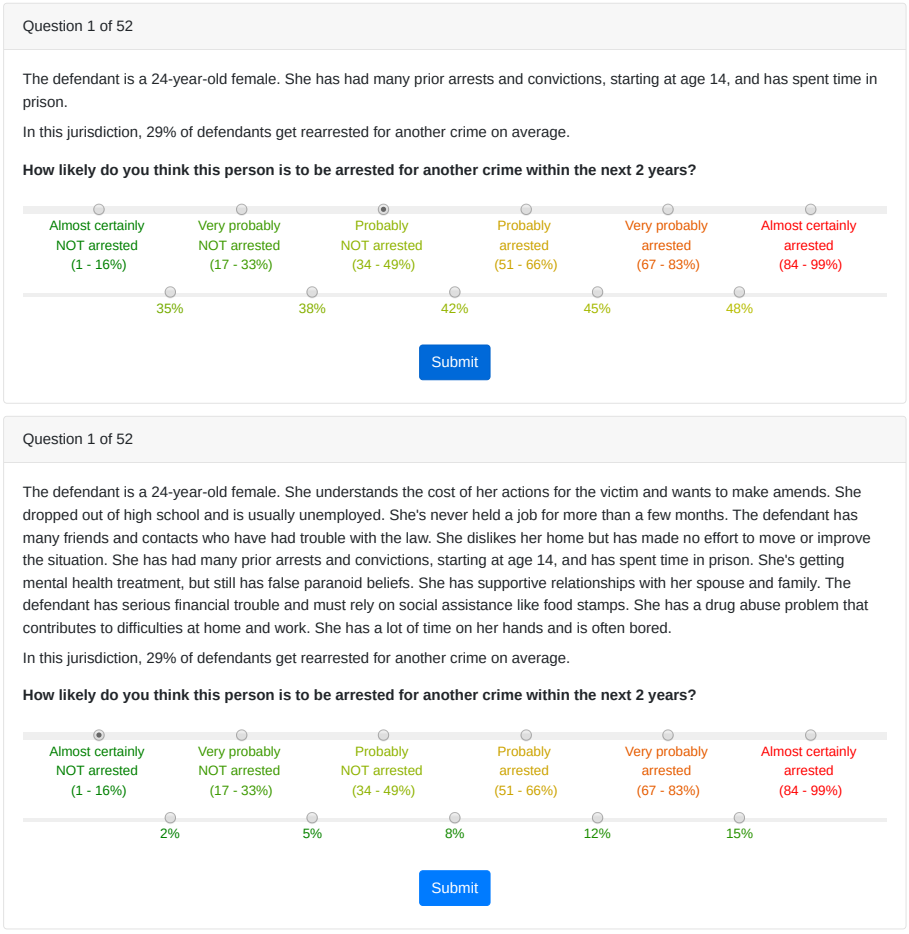

Fig. 1: *Example survey questions for the "streamlined" (top) and "enriched" (bottom) conditions. Participants were asked to assess the likelihood of rearrest on a 30-point scale.*

rested. We instead randomly assigned participants to either receive or not to receive feedback. Finally, we investigated the effects of information richness on predictive accuracy. In the two COMPAS datasets—including the dataset used in the original study—relatively little information is available about individuals, and that which is available (e.g., age, gender, and number of past arrests) is strongly associated with recidivism risk. Vignettes based on COMPAS datasets are necessarily "streamlined", i.e., restricted to the five risk factors available. However, in the two LSI-R datasets, we have more complete information on each individual, including ten addi-

	COMPAS balanced BR	COMPAS low BR	LSI-R balanced BR	LSI-R low BR
Number of cases	1,000	1,000	311	1,959
Base rate of recidivism	48%	11%	29%	8%
Features	Streamlined	Streamlined	Streamlined/Enriched	Streamlined/Enriched
# Responses (No feedback)	2,700	2,400	2,850 / 2,500	2,850 / 2,900
# Responses (Feedback)	2,400	3,000	2,700 / 2,400	3,000 / 2,550

Table 1: *Characteristics of datasets (Note: BR = base rate).*

tional risk factors like education and employment problems. Vignettes based on LSI-R datasets can be streamlined like those presented in the original study, or "enriched" with descriptions containing more detailed information.

In summary, we carried out four separate experiments, one for each of the four datasets we consider. In the two COMPAS experiments, participants were randomly assigned to receive or not to receive feedback. In the two LSI-R experiments, participants were randomly assigned to one of the two feedback conditions, and independently assigned to see streamlined or enriched vignettes, in a 2×2 design. In all cases, participants provided 50 predictions and received financial compensation for accuracy, in line with the original study. In aggregate, across all the experiments, we collected 32,250 responses from 645 participants.

Results

As detailed below, we compared human predictions of recidivism to those from existing algorithms (COMPAS and LSI-R); we also fit our own statistical models to the data as another point

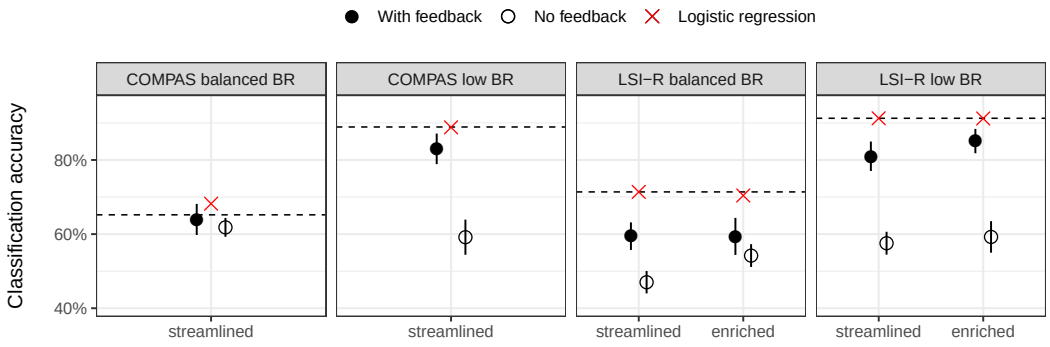

Fig. 2: *Classification accuracy of human predictions, with and without immediate feedback, and a logistic regression model that uses the same information provided to study participants. Dashed lines represent performance of the existing tools, COMPAS or LSI-R. For participants in the feedback condition, only the last 10 responses for each participant are used, to account for the effects of learning. Error bars represent 95% confidence intervals.*

of comparison. Algorithmic predictions were restricted to providing responses on the same 30-point probability scale available to study participants. To quantify performance, Dressel and Farid focused on binary classification accuracy, which we simply call "classification accuracy" here. We expanded on that analysis by considering not only classification accuracy but also ranking accuracy, as measured in part by the area under the receiver operating characteristic curve (commonly called "area under the curve", or AUC).

Classification accuracy

Figure **2** shows the classification accuracy of study participants and models. In particular, the solid circle in the left-most panel corresponds to the setting considered in Dressel and Farid: the COMPAS dataset, with immediate feedback provided to participants. In line with that study, we found participants performed on par with the COMPAS algorithm (64% accuracy for participants vs. 65% for COMPAS, indicated by the dashed line). Further, even without immediate feedback (open circle, 62%), the study participants did reasonably well. Our own statistical model, which was trained on the local data, did somewhat better (red **x**, 68%), but the gap is

relatively modest.[3] Thus, despite some experiment design differences (e.g., we elicit probability judgments rather than binary predictions), we were largely able to replicate the main result reported by Dressel and Farid.

However, we saw qualitatively different patterns in the three other datasets we consider. In those datasets, as shown in Figure **2**, study participants performed consistently worse than both the existing risk assessment instruments (COMPAS and LSI-R) and our own statistical models. The gap between participants and algorithms was particularly large when feedback was not provided. For example, in the COMPAS low base rate dataset, COMPAS and our model both achieve 89% classification accuracy, but participants attain only 83% accuracy when provided with feedback, and just 60% accuracy without feedback. Figure **A2** in the Appendix shows that classification accuracy improved over time when feedback was provided, but not fast enough for participants to be competitive with the statistical models.[4]

The effect of feedback on participants' performance appears most pronounced when base rates for recidivism are relatively low. In the COMPAS balanced base rate data used by Dressel and Farid, 48% of defendants recidivated. In comparison, base rates were 11%, 8% and 29% in the COMPAS low base rate, LSI-R low base rate, and LSI-R balanced base rate datasets, respectively. In these latter three datasets, participants consistently over-estimated risk, hurting their classification accuracy (cf. Figure **A3** in the Appendix)—despite the fact that study participants were explicitly and repeatedly informed of the lower base rates, as shown in Figure **1**.

Finally—and in contrast to the feedback condition—we find that providing enriched information for predictions had minimal effect on classification accuracy. As shown in Figure **2** for the two LSI-R datasets, both human participants and our own statistical models show little to

[3]Reported results for our own statistical models are based on out-of-sample predictions, as described in the Materials and Methods.

[4]To adjust for these learning gains, in the feedback condition we report accuracy on the final 10 of the 50 questions answered by a participant.

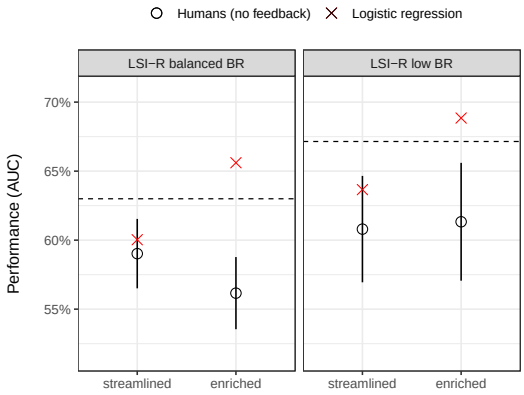

Fig. 3: *AUC performance of human predictions compared to a logistic regression model that uses the same information provided to study participants. Dashed lines represent performance of the existing LSI-R tools. Error bars represent 95% confidence intervals.*

no improvement when given more information on which to base judgments, in the "enriched" condition.

Ranking accuracy

Classification accuracy is a useful starting point, but it can also be a problematic measure of performance in unbalanced datasets. For example, in the COMPAS low base rate dataset (with 11% recidivism), COMPAS and our own statistical model have about the same accuracy as a naive classifier that predicts no one recidivates—though human participants performed considerably worse than even this simple classifier. We thus next gauge performance in terms of AUC, a popular measure that mitigates this issue of class imbalance. Loosely, AUC measures the extent to which predictions correctly rank individuals by risk, ignoring the absolute stated risk level.[5]

[5]AUC is formally defined as follows. In a given dataset, suppose X_1 is a randomly selected defendant who ultimately recidivates, and X_0 is a randomly selected defendant who ultimately does not. Then, AUC is $\Pr(r(X_1) \geq r(X_0))$, where $r(x)$ is the reported probability x recidivates.

Figure **3** shows performance of the human and machine predictions as measured by AUC for the two LSI-R datasets, where the "enriched" condition is available.[6] We find that the study participants performed worse than the LSI-R risk assessment instrument.[7] Further, as was the case for classification accuracy, human AUC performance does not improve with enriched information. Machine performance, however, does improve when provided with more information.

Our findings suggest one possibility to explain these results: humans perform well when a very limited amount of information is sufficient to create accurate risk rankings. For example, when age and number of past offenses is all one needs to assess risk, it is reasonable that appropriately motivated humans can compete with statistical models. Indeed, in the cases we consider, human participants generally performed on par with statistical models that were based on relatively little information, in the "streamlined" condition. But when more information was useful, the models appropriately incorporated that information while the human participants often did not, as seen in the "enriched" condition.[8] In both panels of Figure 3, the gap between human and machine performance is wider in the enriched condition.

Finally, we took a complementary, cost-benefit approach to assess ranking performance. Suppose a policymaker aims to allocate limited resources (e.g., community supervision) to those individuals deemed most likely to recidivate. To assess the performance of different ranking strategies, one can compute the proportion of recidivists that are listed in the top p-percent of candidates in each strategy. (This measure is also known as "recall-at-k" in the machine-learning literature.) Figure **4** traces out the corresponding curves for humans and algorithms in

[6]To compute AUC for human predictions, we first calculated the AUC for each study participant, and then averaged across all participants in a given dataset and treatment condition.

[7]We find similar findings for the other conditions and datasets, shown in Figure **A1** in the Appendix.

[8]For classification accuracy, in Figure **2**, model performance is similar in both the streamlined and enriched conditions for the two LSI-R datasets. In those datasets, very few individuals have estimated likelihood of recidivism that exceeds 50%, and so the optimal binary prediction is "no recidivism" in almost every case, even when provided with the enriched information. Ranking accuracy, in contrast, allows for more nuanced distinctions between individuals, which in turn lets us see the gap between models based on streamlined and enriched predictors.

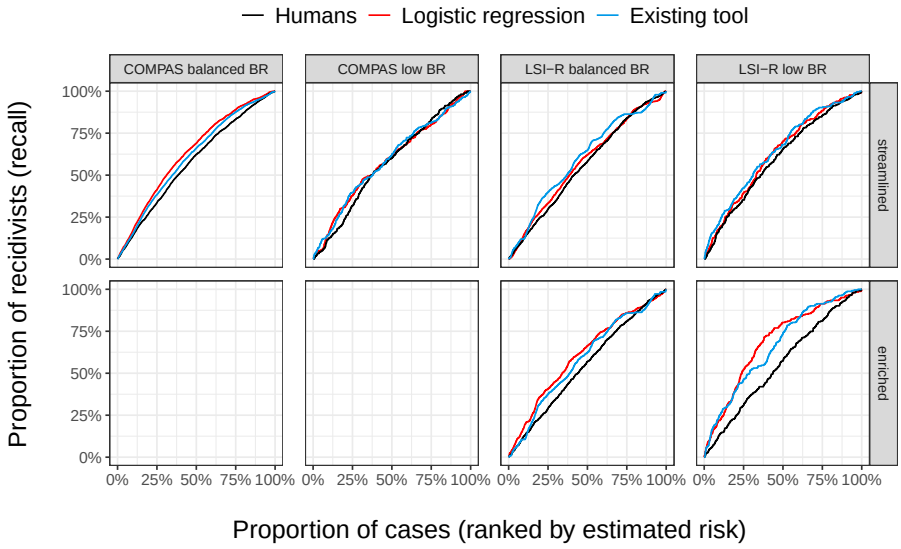

Fig. 4: *Proportion of people who recidivated that were identified when ranking by the risk assessments of humans (in the no-feedback condition), a logistic regression model, and existing tools (COMPAS or LSI-R). For each value p on the horizontal axis, the vertical axis shows the proportion of all recidivists that are included among the p-percent of the population deemed riskiest. Human performance is generally comparable to algorithmic tools in the "streamlined" condition (top panels), but algorithmic tools outperform humans when more information was made available ("enriched" condition, in the bottom panels).*

our experiments, for all values of p from 0% to 100% on the horizontal axis. For simplicity, we only show the curves for study participants who did not receive feedback. In line with our results above, we found that study participants generally performed on par with statistical models in the "streamlined" condition, but that algorithmic tools outperformed humans when more information was available. For example, in the LSI low base rate dataset in the "enriched" condition, the top 50% of defendants deemed riskiest by study participants in reality contain 58% of recidivists in the dataset, just slightly better than random. In comparison, the top 50% deemed riskiest by LSI-R and our own statistical model contain 74% and 80% of recidivists, respectively.[9]

[9]In the LSI balanced base rate dataset with "enriched" condition, the top 50% deemed riskiest by study participants contain 57% of recidivists, while LSI-R and our own statistical model contain 62% and 66%, respectively.

Discussion

Risk assessment is "the engine that drives" a federal prison reform bill recently signed into law (*19*) and a component of many jurisdictions' efforts to reduce incarceration rates without compromising public safety (*20*). When risk is a legally relevant factor, judges, correctional authorities, and other professionals have been advised to consider risk assessment instruments when making decisions. The assumption is that RAIs predict reoffending better than unaided human judgment.

Dressel and Farid's findings challenge this assumption in a setting where risk information is constrained, feedback on accuracy is provided across many trials, and base rates of recidivism are balanced. In the present series of experiments, we examined the robustness of that result by manipulating these three features. We replicated Dressel and Farid's finding that people perform as well as algorithms under the conditions they investigate. However, we also found that algorithms tended to outperform humans in settings where decision makers have access to extensive information, do not receive immediate feedback, and base rates are far from balanced—features of many real-world scenarios.

On the whole, our findings are consistent with much of the past research comparing human and algorithmic decisions (*21–23*); for crime-specific reviews, see (*24, 25*). For example, based on a meta-analysis of 41 studies, Ægisdóttir et al. (*21*) found that statistical methods were reliably superior to humans in predicting a range of outcomes. For predicting violence and other criminal behavior specifically, they note that algorithms were "clearly superior to the clinical [human] approach". Similarly, several studies conducted with judges and correctional officers indicate that algorithms and RAIs outperform their professional judgment in predicting recidivism (cf. Goel et al. (*26*) for an overview).

Against this backdrop, Dressel and Farid's finding was unusual. Their work, however, helps to provide hints about the conditions under which humans may perform as accurately as algorithmic RAIs. Although we could not examine every possibility in the present set of experiments, our results point toward two sets of conditions that influence the relative accuracy of humans. First, when base rates are unbalanced, our results suggest that providing people with feedback can improve their classification accuracy to rival that of algorithms. We explicitly informed all participants in our experiments about the base rate of recidivism, but classification accuracy improved only among the subset of participants who also received trial-by-trial outcome feedback. Across trials, people who received feedback became less likely to guess that the defendant would reoffend, compared to people who did not receive feedback. But, while feedback corrected participants' tendencies to overpredict recidivism, it did not improve their ranking accuracy, highlighting the limitations of feedback.

Second, our results suggest that people can predict recidivism as well as algorithms if only a few simple predictive factors are specified as inputs—as was the case in Dressel and Farid's study. In this context of "streamlined" inputs, the accuracy of algorithms and humans (without feedback) was largely interchangeable. In contrast, when inputs were "enriched" with additional predictive factors, algorithms outperformed human judgment. This is not because the additional risk information compromised human judgment (in fact, people's performance did not differ much in "streamlined" vs. "enriched" conditions). Instead, it is because algorithms made better use of the additional information than did humans.

We note, however, that even in the "enriched" condition, the additional information we provided was still relevant for recidivism prediction, as it was included in the LSI-R. Like Dressel and Farid's study, then, our experiments compare the accuracy of algorithms and RAIs with that of *structured* human judgment—which has been found to consisently outperform unstructured judgment in predicting violence and other recidivism (cf. Goel et al. (*26*) for a summary). To

better represent human judgment in justice settings, we hope future studies provide even more realistic and complete inputs for prediction, including irrelevant or potentially distracting information (*27*). Still—together with past work—our results support the claim that algorithmic risk assessments can often outperform human predictions of reoffending.

Acknowledgements: The authors thank Alex Chohlas-Wood and Ravi Shroff for helpful feedback. **Author Contributions:** The authors contributed equally to all aspects of this work and manuscript preparation. **Competing Interests** The authors declare that they have no competing financial interests. **Data and materials availability:** All data and code associated with this research will be made public.

References

1. J. Dressel, H. Farid, The accuracy, fairness, and limits of predicting recidivism. *Science Advances* **4**, 1–6 (2018).

2. J. Monahan, J. L. Skeem, Risk assessment in criminal sentencing. *Annual review of clinical psychology* **12**, 489–513 (2016).

3. A. Neufeld, In defense of risk assessment tools. .

4. J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica* **23** (2016).

5. J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair determination of risk scores. *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* **67** (2017).

6. A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* **5**, 153–163 (2017).

7. S. Corbett-Davies, S. Goel, The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

8. S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 797–806 (2017).

9. Z. Lipton, J. McAuley, A. Chouldechova, Does mitigating ML's impact disparity require treatment disparity? *Advances in Neural Information Processing Systems* pp. 8125–8135 (2018).

10. C. Guthrie, J. J. Rachlinski, A. J. Wistrich, Blinking on the bench: How judges decide cases. *Cornell Law Review* **93**, 1–44 (2007).

11. D. A. Andrews, J. Bonta, *The level of service inventory-revised* (Multi-Health Systems Toronto, Ontario, Canada, 2000).

12. R. M. Hogarth, E. Soyer, Sequentially simulated outcomes: Kind experience versus non-transparent description. *Journal of Experimental Psychology: General* **140**, 434–463 (2011).

13. R. M. Hogarth, T. Lejarraga, E. Soyer, The two settings of kind and wicked learning environments. *Current Directions in Psychological Science* **24**, 379–385 (2015).

14. D. Kahneman, A. Tversky, On the psychology of prediction. *Psychological review* **80**, 237–251 (1973).

15. J. S. Carroll, Judgments of recidivism risk: The use of base-rate information in parole decisions. *New directions in psycholegal research* pp. 68–86 (1980).

16. C. K. Cannon, V. L. Quinsey, The likelihood of violent behaviour: Predictions, postdictions, and hindsight bias. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* **27**, 92 (1995).

17. G. D. Walters, D. G. Kroner, D. DeMatteo, B. R. Locklair, The impact of base rate utilization and clinical experience on the accuracy of judgments made with the hcr-20. *Journal of Forensic Psychology Practice* **14**, 288–301 (2014).

18. J. J. Koehler, The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and brain sciences* **19**, 1–17 (1996).

19. B. L. Garrett, The prison reform bills implementation will be tricky: Heres how to ensure its a success (2018). *Slate*, `https://slate.com/news-and-politics/2018/12/prison-reform-bill-success.html` [Accessed: 2019-06-19].

20. Justice reinvestment: State resources (2017). *National Conference of State Legislatures*, `http://www.ncsl.org/research/civil-and-criminal-justice/justicereinvestment.aspx` [Accessed: 2019-06-19].

21. S. Ægisdóttir, M. J. White, P. M. Spengler, A. S. Maugherman, L. A. Anderson, R. S. Cook, C. N. Nichols, G. K. Lampropoulos, B. S. Walker, G. Cohen, J. D. Rush, The Meta-Analysis of clinical judgment project: Fifty-Six years of accumulated research on clinical versus statistical prediction. *Couns. Psychol.* **34**, 341–382 (2006).

22. W. M. Grove, D. H. Zald, B. S. Lebow, B. E. Snitz, C. Nelson, Clinical versus mechanical prediction: a meta-analysis. *Psychol. Assess.* **12**, 19–30 (2000).

23. R. M. Dawes, D. Faust, P. E. Meehl, Clinical versus actuarial judgment. *Science* **243**, 1668–1674 (1989).

24. D. A. Andrews, J. Bonta, J. S. Wormith, The recent past and near future of risk and/or need assessment. *Crime & Delinquency* **52**, 7–27 (2006).

25. R. K. Hanson, K. E. Morton-Bourgon, The accuracy of recidivism risk assessments for sexual offenders: a meta-analysis of 118 prediction studies. *Psychol. Assess.* **21**, 1–21 (2009).

26. S. Goel, R. Shroff, J. L. Skeem, C. Slobogin, The accuracy, equity, and jurisprudence of criminal risk assessment (2018).

27. M. Stevenson, Assessing risk assessment in action. *Minn. L. Rev.* **103**, 303 (2018).

28. C. T. Lowenkamp, E. J. Latessa, *Evaluation of Ohio's community based correctional facilities and halfway house programs* (University of Cincinnati, Division of Criminal Justice, Center for Criminal , 2002).

29. A. W. Flores, C. T. Lowenkamp, P. Smith, E. J. Latessa, Validating the level of service inventory-revised on a sample of federal probationers. *Fed. Probation* **70**, 44 (2006).

30. B. Green, Y. Chen, Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. *Proceedings of the Conference on Fairness, Accountability, and Transparency* pp. 90–99 (2019).

31. G. W. Brier, Verification of forecasts expressed in terms of probability. *Monthey Weather Review* **78**, 1–3 (1950).

32. K. Rufibach, Use of brier score to assess binary predictions. *Journal of clinical epidemiology* **63**, 938–939 (2010).

33. J. Hernández-Orallo, P. Flach, C. Ferri, A unified view of performance metrics: translating threshold choice into expected classification loss. *Journal of Machine Learning Research* **13**, 2813–2869 (2012).

34. K. Bansak, Can nonexperts really emulate statistical learning methods? A comment on the accuracy, fairness, and limits of predicting recidivism. *Political Analysis* **27**, 370–380 (2019).

Materials and Methods

Datasets. Our two COMPAS experiments are based on a single dataset that is comprised of 7,214 defendants from Broward County, Florida (*4*) who were scored with the COMPAS for both risk of recidivism and risk of violent recidivism. While the Dressel and Farid (*1*) study only considered human predictions of overall recidivism, we additionally considered predictions of violent recidivism. The COMPAS dataset contains: individual-level demographic information (age and gender); criminal history (current charge and number of past arrests); and whether or not each defendant was arrested for a new crime or, separately, a new violent crime within two years of COMPAS scoring, excluding any initial detention period. We restrict to the same randomly selected subset of 1,000 defendants used by Dressel and Farid.

We also make use of two LSI-R (*11*) datasets: one containing 311 individuals under correctional supervision in North Dakota, drawn from Lowenkamp & Latessa (*28*); and another containing 1,959 individuals on probation in Texas, drawn from Flores, Lowenkamp, Smith & Latessa (*29*). These datasets include numerical scores for 10 risk factors or subscales of the LSI-R (e.g., criminal history, antisocial peers, substance abuse). All individuals in these datasets were followed for a minimum of one year after the LSI-R was administered to assess recidivism—which was for any arrest in North Dakota and any reincarceration (a low base rate phenomenon) in Texas.

Vignettes. For the COMPAS data, we generated short vignettes to show study participants following the method of Dressel and Farid. In particular, participants were given a brief description of each individual's age, gender, current criminal charge, and number of past arrests. For the LSI-R data, we first created several one-sentence descriptions for each of the 10 LSI-R factors and risk levels, discretized to "low", "medium", or "high". For example, one such sentence for "low" risk on the criminal history scale is: "He has never been convicted of a

prior offense". These sentences were created by consulting the LSI-R scoring manuals for each jurisdiction. Participants in the "streamlined" condition were presented with the individual's age, gender, and number of past arrests, as in the COMPAS experiments. (Current charge was not available in the LSI-R datasets.) Participants assigned to the "enriched" condition were shown a brief paragraph describing an individual's age and gender, followed by one-sentence descriptions for each of the 10 LSI-R risk factors, displayed in random order.

Participants. Study participants were recruited through Amazon's Mechanical Turk. The study was advertised as follows: "You are invited to participate in a research study on predicting criminal behavior. You will be presented with a series of descriptions that requires a classification decision (e.g., assessing an individuals risk of recidivism). We will not ask or record any personal information." Each participant was asked to assess the recidivism risk for 50 individuals randomly selected from the corresponding dataset. As in Dressel and Farid (*1*), we additionally asked participants to answer two attention checks at different points in the experiment, and we only include in our analysis responses from the 71% of participants who passed both attention checks. Our four experiments were conducted in close succession over the course of several weeks, and participants were only allowed to complete one experiment.

Each participant received \$1 for completing the study, and a bonus of up to \$5 based on performance. Following previous work (*30*), we measured performance via Brier scoring, an incentive-compatible payment scheme for eliciting probabilities (*31–33*). For each question i, the Brier score is $1 - (Y_i - p_i)^2$, where p_i is the probability of recidivism reported by the participant, $Y_i = 1$ if the individual described was indeed arrested for a new (violent) crime within two years of release, and $Y_i = 0$ otherwise. A participant's final score is computed by summing the Brier scores earned for the 50 substantive questions, excluding the two attention checks.

Statistical models. For comparison, we fit a logistic regression on the same data that are made available to human participants. In the case of COMPAS, we use a separate training set of 6,214 cases—the remainder after excluding the 1,000 cases used as test data. For the LSI-R data, we use leave-one-out evaluation over all available data. Probabilistic predictions from the models are then rounded to the nearest value of the 30-point scale presented to participants in our experiment, to ensure that performance of the statistical models are evaluated under the same technical constraints.

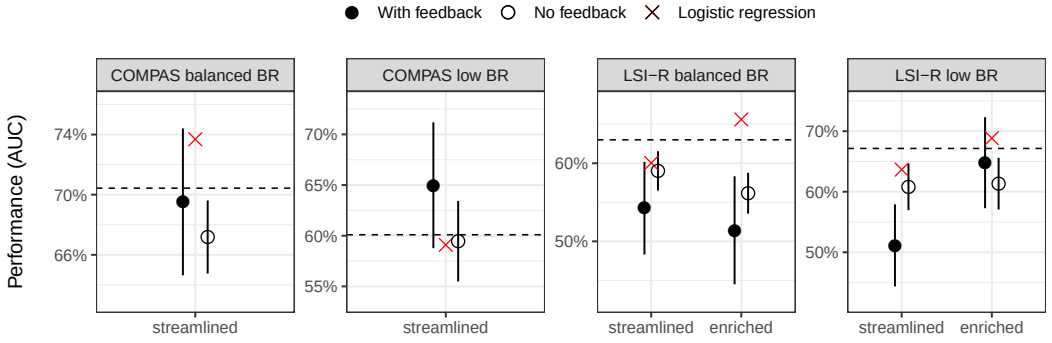

Fig. A1: *AUC performance of human predictions, with and without immediate feedback, and a logistic regression model that uses the same information provided to study participants. Dashed lines represent performance of the existing tools, COMPAS or LSI-R. For humans in the feedback condition, only the last 10 responses for each participant are used to account for the effects of learning. Error bars represent 95% confidence intervals.*

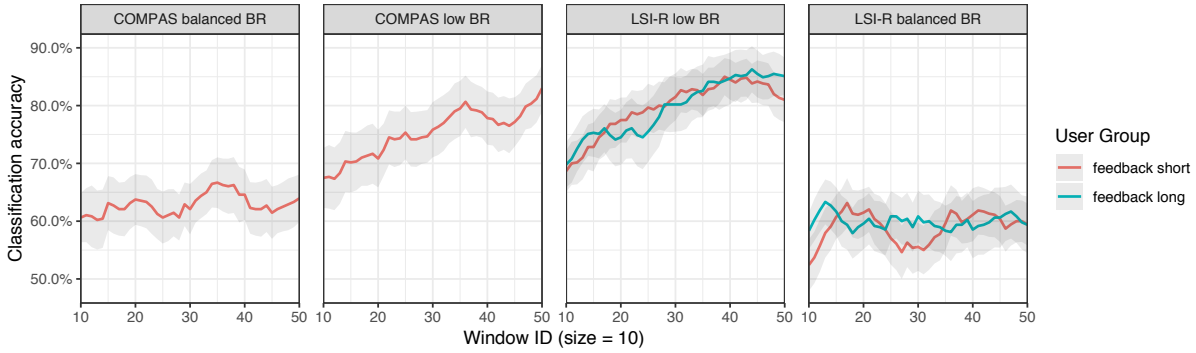

Fig. A2: *Average classification accuracy over time for a sliding window of 10 questions. The window ID indicates the last question of that window. As a result of feedback, humans recalibrate, and we observe accuracy increasing. The largest improvements occur for groups with low base rates. Grey bands indicates 95% confidence intervals of average accuracy.*

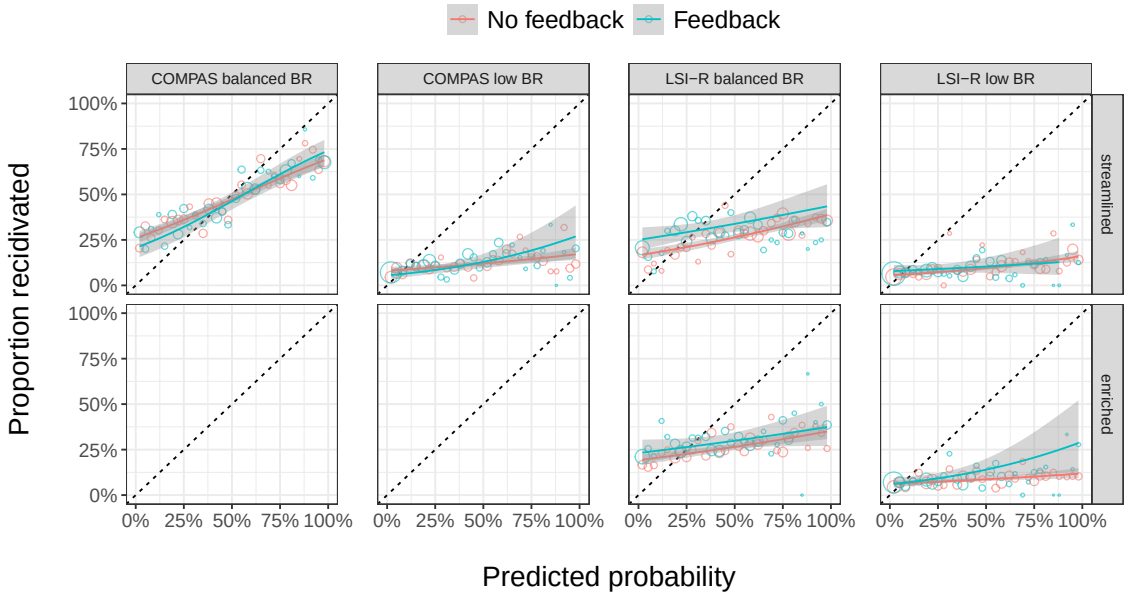

Fig. A3: *Calibration plot for human responses. Circles show the proportion of defendants who actually recidivated, binned by the probability estimate given by human participants and sized according to the number of responses. Lines show a logistic regression of participants' estimated probabilities against the actual outcomes, and indicate that human predictions suffer from poor calibration. In a re-analysis of Dressel and Farid's data, Bansak (34) likewise found evidence of poor calibration in human predictions of recidivism.*