Artificial Intelligence and Law
https://doi.org/10.1007/s10506-024-09389-8

ORIGINAL RESEARCH

Code is law: how COMPAS affects the way the judiciary handles the risk of recidivism

Christoph Engel[1] · **Lorenz Linhardt**[2] · **Marcel Schubert**[3]

Accepted: 21 December 2023

Abstract
Judges in multiple US states, such as New York, Pennsylvania, Wisconsin, California, and Florida, receive a prediction of defendants' recidivism risk, generated by the COMPAS algorithm. If judges act on these predictions, they implicitly delegate normative decisions to proprietary software, even beyond the previously documented race and age biases. Using the ProPublica dataset, we demonstrate that COMPAS predictions favor jailing over release. COMPAS is biased against defendants. We show that this bias can largely be removed. Our proposed correction increases overall accuracy, and attenuates anti-black and anti-young bias. However, it also slightly increases the risk that defendants are released who commit a new crime before tried. We argue that this normative decision should not be buried in the code. The tradeoff between the interests of innocent defendants and of future victims should not only be made transparent. The algorithm should be changed such that the legislator and the courts do make this choice.

Keywords Algorithmic decision-aids in the court room · COMPAS · False positives versus false negatives · Anti-defendant bias · Algorithmic correction

1 Introduction

Judges may not refuse to decide. This also holds for the decision to detain a defendant who has been apprehended for purportedly committing a crime. According to Federal US law the defendant is to be detained if "no condition or combination of conditions will reasonably assure [...] the safety of any other person and the community" (18 U.S. Code §3142 (e) (1)). It suffices if this is supported by

✉ Christoph Engel
engel@coll.mpg.de

[1] Max-Planck-Institute for Research on Collective Goods, 53113 Bonn, Germany

[2] Dept. Computer Science, Technical University Berlin, Berlin, Germany

[3] Max-Planck-Institute for Research on Collective Goods, Bonn, Germany

Published online: 09 February 2024

"clear and convincing evidence" (18 U.S. Code §3142 (f) (2) (B)). This intermediate standard of proof is less than certainty, and also less stringent than evidence "beyond a reasonable doubt". It merely requires that the disputed fact is "highly and substantially more likely to be true than untrue" (U.S. Colorado v. New Mexico, 467 U.S. 310 (1984)).

With these rules, the law conditions the judicial decision on a prediction: what is the likelihood that the defendant would commit another (sufficiently serious) crime before tried if they are released? By their nature, predictions are fraught with the risk of being wrong. The fact that a person has been apprehended is a predictive signal. Persons with a criminal history are on average more likely to commit further crimes than those without (Sampson and Laub 1992). Yet, the rules of criminal procedure force the law to strike a balance between two potentially wrong decisions: reacting to the signal with incapacitating an individual who would not have recidivated while waiting for trial, and not reacting to the signal and releasing an individual who reoffends before being tried for the first crime.

For practitioners of predictive modelling, this is a familiar choice: given a Pareto-optimal model based on imperfect data, the incidence of false negative predictions can usually only be reduced when increasing the risk of false positive predictions, and vice versa. Trading the higher risk of false negatives against the lower risk of false positives is a normative decision. In the specific case, this decision has particularly high weight, as judges are deciding over detainment of a defendant. Life, limb, and property of innocent victims are at stake if false positive decisions are minimized. Conversely, if false negative decisions are minimized, innocent defendants risk losing their jobs, families, and being put on a criminal career (Hagan and Dinovitzer 1999; Western et al. 2001, 2004).

The legal system cannot avoid making this choice. It notably also makes a choice if it seeks to maximize accuracy, i.e. if it minimizes the sum of false positive and false negative decisions. The legal system then decides to weigh equally the incidence of false negative decisions, and of false positive decisions, for that matter. Hence, it cannot be a policy question *whether* a weighting should be chosen, but *which*. It would be simple if this decision could be logically derived from first principles—fundamental propositions that are generally agreed upon. However, most societies do not feel comfortable with putting a price tag on life, limb, or fear, nor is there an agreed cost for wrongful conviction or suspicion (Brooks and Simpson 2012). Hence, even if one were to agree on a utilitarian norm, there would be disagreement about parameters. Moreover, it can by no means be taken for granted that the well-being of victims and of potentially innocent defendants should be traded against each other. From a deontological perspective, the freedom of a person from intrusion on their physical well-being should deserve absolute protection, as should the freedom of a person from unjustified sovereign intervention.

As the normatively correct decision cannot be found by deduction, in a democracy, the natural institution for this kind of value judgment is Parliament. Possibly, the constitution prescribes to convey at least some of this authority to the judiciary. As a matter of fact, this happens in the frequent situation of statutory provisions leaving room for interpretation. By contrast, corporations,

such as Northpointe (the developer of the COMPAS algorithm),[1] are no first-order rulemaking bodies, since they lack democratic legitimacy. For pragmatic reasons, legal orders make exceptions. Private ordering is, for instance, frequent in the formulation of technical standards. But at the least, such secondary rulemaking bodies are exposed to scrutiny by institutions with direct democratic legitimacy, like Parliament or regulatory agencies controlled by government.

Already in 1999, scholars working at the intersection of law and computer science alerted the public to an emergent phenomenon: code is law (Lessig 1999). Originally, however, attention was on technical substitutes for traditional private ordering, like the design of a negotiation platform. In this paper we argue, and empirically demonstrate, that normative decisions at the core of constitutionally protected freedoms are now buried in code. This is alarming as these decisions are not at all transparent. To show this, we investigate the "Correctional Offender Management Profiling for Alternative Sanctions" (COMPAS) system. At face value, COMPAS provides judges just with a computer-generated prediction about a fact that the law of criminal procedure cited above determines to be relevant for the choice between bail and jail[2]: is the defendant likely to recidivate?

The company behind COMPAS immunizes itself from criticism by insisting that its software only assesses the risk of recidivism (general or violent), and leaves it to the competent judge to draw their conclusion for the choice between bail and jail. Yet, unless the judge plainly disregards the machine generated prediction, the prediction has the potential to influence their decision (Grgić-Hlača et al. 2019). This influence is normatively highly problematic, as we show that the COMPAS prediction relies upon (normative) considerations encapsulated in the software: we find that COMPAS strongly privileges victims over defendants. This decision may be in line with the preferences of the majority of the legislator in at least some of the US states. Critically, however, these legislative bodies themselves have never made this decision, and the public has never gotten a chance to discuss the choice—it is hidden in the design of the algorithm.

COMPAS has met with considerable criticism. Public and scientific debate has focused on hidden discrimination, by race (Angwin et al. 2013; Fass et al. 2008; Flores et al. 2016; Dieterich et al. 2016; Chouldechova 2017; Agarwal et al. 2020), or by age (Rudin 2019; Jackson and Mendoza 2020; Rudin et al. 2020). It has been pointed out that the accuracy of COMPAS predictions is as low as 68% (Grgić-Hlača et al. 2018; De Miguel Beriain 2018). COMPAS predictions are no better, in terms of accuracy, false positives and false negatives, than untrained human laypersons (Dressel and Farid 2018), at least under comparable circumstances (Jung et al. 2020). Moreover, critics oppose the proprietary nature of the tool (Freeman 2016; Carlson 2017; De Miguel Beriain 2018; Nishi 2019; Rudin 2019), with the ensuing potential for conflicts of interest (Freeman 2016), and the lack of transparency (Rudin 2019).

[1] The company since changed its name to "equivant".

[2] We use the term "jailing" to refer to jailing as well as imprisonment.

All this critique is well taken. Yet it leaves an even deeper normative concern untouched: COMPAS is not only biased against certain groups of defendants, the algorithm is biased against *all* defendants. If the judge takes the COMPAS prediction at face value, they are considerably more likely to jail a defendant who would not have recidivated, rather than releasing a defendant who commits a new crime before trial. Arguably this violates the "clear and convincing evidence" test. This substantive normative concern is exacerbated by a procedural concern. It is not apparent from the way the prediction is communicated to the judge that, and how severely, the prediction is biased to the detriment of defendants. Nor has the legislator struck a balance between the safety of potential victims and the freedom of potentially harmless or even innocent defendants. A decision of high normative relevance is taken by a commercial entity, and it is concealed from public scrutiny. The decision lacks democratic legitimacy.

In our analysis, we not only show that this bias is pronounced, we also introduce a simple procedure for neutralizing the bias, i.e. for maximizing accuracy. This correction also mitigates racial bias and age bias. In line with our call for democratic legitimacy, we do not argue that this corrected version of the algorithm is preferable. We only demonstrate that a correction is feasible, and claim that an option for correction should be provided by COMPAS. The legislator may have normative reasons not to maximize accuracy. It may hold the conviction that victims are more important than defendants, even if defendants are innocent (or vice versa). Our correction should therefore be understood as a proof of concept. In the supplementary material, we make the code available. If the legislator or (if constitutionally empowered to do so) the judiciary want to allow for x% more false positives if this reduces false negatives by y% or more, the predictive tool can be adjusted. Rather than maximizing accuracy, the predictive model then implements a defined ratio of false positives (in the language of machine learning: recall) over precision (the fraction of correct—positive or negative—decisions). But critically this normative decision is no longer concealed. It is transformed into a transparent policy choice for which the legislator (or the judiciary) must assume political responsibility.

2 Anti-defendant bias

2.1 Documenting the bias

At the surface-level, COMPAS leaves potentially contentious normative choices to its judicial users. The user manual says that it is for the user to define which level of predicted risk of recidivism is to be considered problematic, and hence warrants denying release on bail (Northpointe 2015, 5). The manual also explains that the risk scores (decile scores) are calculated in relation to the group of the population to which the defendant belongs (Northpointe 2015, 11). These so-called norm groups may take into account gender, and whether defendants in the training data have been in prison or on parole; in jail; or on probation (Northpointe 2015, 11). Compared to other members of their respective norm group, a defendant with a decile score of

1–4 is characterized by COMPAS as low risk, a defendant with score 5–7 as medium risk, and a defendant with score 8–10 as high risk (Northpointe 2015, 8). For the latter COMPAS suggests "(extended) supervision". Yet, as our analysis shows, the COMPAS algorithm itself is actually pronouncedly normative, to the detriment of defendants.

We are in a position to show this as ProPublica, a nonprofit organization conducting investigative journalism, has exploited freedom of information legislation to compile a dataset containing the relevant COMPAS scores for 5759 defendants in Boward County, Florida. Furthermore, the dataset includes a number of features about the defendant that the COMPAS algorithm uses as input and, additionally, information on whether they committed any new crime within two years since the COMPAS screening event (Angwin et al. 2013; Rudin 2019).[3] For these defendants, we thus know the ground truth: we know whether they recidivated. Consequently, for each COMPAS decile score we also know whether an individual prediction was a "false positive", as COMPAS suggested extended supervision when none was required, or vice versa.

Actually, we can define correct (true) and incorrect (false) predictions conditional on any COMPAS decile score that is used as the decision threshold: If the competent judge takes the COMPAS prediction at face value, given a decile score chosen as threshold, every individual below that threshold would be released, and all other individuals would be kept in prison. Thus, for every decile score threshold, we are able to identify the number of true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs).

In the Appendix, we explain the steps we have taken to clean the data. As we later want to show that it is possible, with a straightforward machine learning method, to remove the bias, we split the complete data into 3239 data points for training, 1080 datapoints for validation and finetuning the algorithm, and 1440 datapoints for testing. Hence the holdout dataset comprises 25% of the entire dataset. And we use 75% of the training data for building the model, and 25% for validation. We enable a direct comparison between the original and the corrected data by using the subset reserved for testing for both purposes.

In Fig. 1, we show the error rates of COMPAS for different thresholds (on the test set, which is comprised of 1440 randomly drawn samples). If a judge were to decide to jail the defendant whenever the score is equal or above 4 (medium risk), 577 (40.1%) defendants would be wrongly classified as too likely to re-offend to be left unsupervised, while only 73 (5.1%) would be wrongly classified as unlikely to re-offend. By contrast, if the judge jailed at or above a score of 7 (high risk in COMPAS language), the incidence of falsely jailed (239 or 16.6%) and of falsely released (275 or 19.1%) would be almost balanced. If the judge aimed to be as accurate as

[3] This time-span does, however, start at different points in time. For defendants who have been released on bail, it starts close to the day when they have been apprehended. For those who have been in jail, the time-span starts once they leave prison, if they have not been convicted, or after they have served time. One has to be aware of this imbalance when interpreting false negatives. The fact that a defendant has not recidivated might result from a deterrent or educative effect of having been in jail.

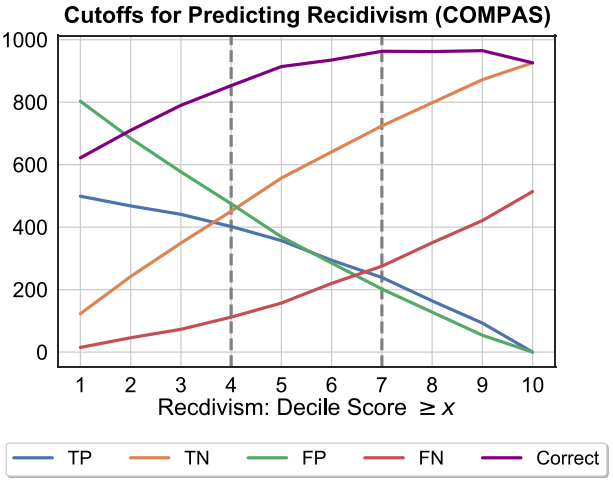

Fig. 1 Anti-defandant bias. On the x-axis, we show the chosen decile score threshold including and above which a judge may consider a defendant for jail. On the y-axis, we show the absolute number of FPs, TPs, FNs, and TNs that would be the outcome of making jail decisions purely based on the respective threshold. The dashed lines are the upper bounds of the "low risk" and "medium risk" ranges

possible, they would have to draw the line at a decile score of 9, i.e. just one decile score below the upper end of the score range. Consequently, if one considers maximum accuracy as the normative standard, whenever the judge chooses a threshold below 9, and relies on the COMPAS prediction, their decision is biased to the detriment of the defendant, as the number of false positives grows faster than the number of false negatives. As Fig. 1 shows, at a threshold of 4, there are even more false positives than true positives: more than half of the defendants have not recidivated in the two years after they have been released. About a third of all defendants would be unnecessarily jailed. Importantly, the anti-defendant bias inherent in choosing a low threshold is not communicated to judges.

2.2 Explaining the bias

One might wonder: isn't anti-defendant bias mechanical? If the judge or the legislator is more cautious and content with a lower risk of recidivism as a reason for detention: does this not inevitably mean that more defendants are incarcerated even if, in retrospect, there was no need for incapacitation? At a very high level, this concern is correct: if a lower predicted risk of recidivism suffices, more defendants are kept in custody unnecessarily. But how many of them are unnecessarily detained, and how the fraction of false positive decisions compares to the fraction of true positive decisions, or to the fraction of false negative decisions, critically depends on the distribution of recidivism risk in the population of those individuals who are apprehended by the police. To show this,

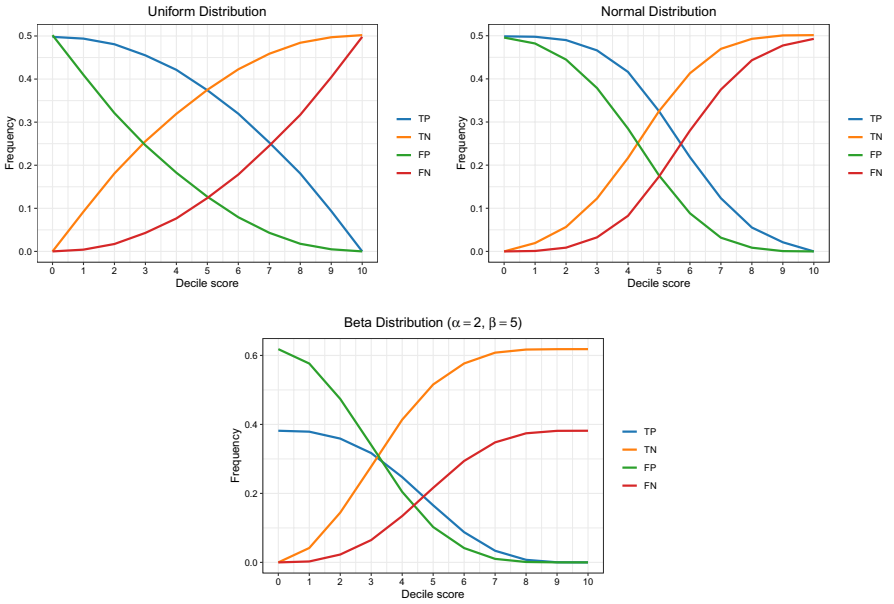

Fig. 2 Simulation of anti-defandant bias. On the x-axis, we show the chosen decile score threshold including and above which a judge may consider a defendant for jail. On the y-axis, we show the fraction of FPs, TPs, FNs, and TNs that would be the outcome of making jail decisions purely based on the respective threshold. The upper left figure assumes that recidivism risk is uniformly distributed in the population of apprehended individuals. The upper right figure assumes that the risk is normally distributed, with mean.5, and standard deviation.2. The lower figure assumes that the risk results from a beta-distribution with $\alpha = 2, \beta = 5$

we revert to simulation. The technical details of the simulation are explained in the Appendix.

Figure 2 shows how strongly the degree of anti-defendant bias depends on the distribution of recidivism risk in the population of apprehended persons, and how much the degree of anti-defendant bias depends on the chosen threshold beyond which the predicted recidivism risk is considered too high. The upper left panel analyses a population in which the recidivism risk is uniformly distributed: the judge sees as many persons whom one predicts to be perfectly unlikely to commit another crime before tried for the crime for which they have been apprehended as the judge sees persons whom one predicts to recidivate with certainty before tried—and any intermediate prediction of recidivism risk is as likely. If every apprehended person is detained (the threshold is at 0), mechanically the risk of false negatives is 0, and there are no true negatives (nobody is released). But since the risk of recidivism is uniformly distributed, only 50% of the detainees would have recidivated (and therefore are true positives), while another 50% would not have recidivated (and therefore are false positives).

The more the threshold moves to the right (the higher the predicted risk of recidivism that is required for detention), the higher the fraction of true negative

decisions (the defendant is released and does not recidivate), and the lower the fraction of true positive decisions (the defendant is detained and would otherwise have committed another crime). As the distribution is uniform, it is perfectly symmetric. Consequently the fraction of true positive and true negative decisions is the same if the threshold is at a predicted risk of recidivism of 50%. At this threshold, the fraction of false positives and false negatives is also identical.

In comparison with Fig. 1, several differences are worth noting. In the perfectly symmetric population of apprehended persons, the fraction of false positive decisions is never higher than the fraction of true positive decisions - while it is much higher in the COMPAS data. The fraction of true negative decisions increases from more lenient threshold to more lenient threshold by the same rate as the fraction of true positive decisions increases from stricter to stricter threshold. By contrast in the COMPAS data, the fraction of true negative decisions increases much faster than the fraction of true positive decisions. Consequently the anti-defendant bias is much more pronounced in the COMPAS data than in a population of apprehended persons with uniformly distributed recidivism risk.

The difference between the COMPAS data and simulated data from the population of apprehended persons with a *normally* distributed risk of recidivism (right upper panel of Fig. 2) is even starker. In this simulated population, the risk that a person commits another crime before tried when released on bail is most likely 50%. Of course, as this distribution is also symmetric about the midpoint, the fraction of false positive and false negative decisions is the same if the threshold is set at a 50% recidivism risk. At this threshold, the fraction of true positive and true negative decisions is also the same. In such a population, at any lower threshold the fraction of false positive decision quickly grows, and at any higher threshold, the fraction of false negative decisions quickly grows. This is of course due to the fact that a normal distribution has more mass near its midpoint. One may also say that a normal distribution makes prediction easier. One predominantly has to get the midpoint right. Actually normality may not be such a strong assumption in reality. It certainly is not true that the risk of committing crime is normally distributed in the complete population. But the normality assumption may be considerably more plausible for the select population that ends up being apprehended by the police.

The lower panel of Fig. 2 shows why the anti-defendant bias is so pronounced in the COMPAS dataset: the distribution of recidivism risk is pronouncedly right skewed: most persons who have been apprehended are predicted to have a rather low risk of recidivism. The beta distribution defines such a population. As the simulation shows, the more the distribution is skewed to the right, the higher the incidence of false positive decisions, and the bigger the gap between false and true positives if the threshold is set at a low risk of recidivism. The more the distribution is skewed to the right, the more pronounced the anti-defended bias.

2.3 Removing the bias

While the trade-off between false positives and false negatives is seemingly inherent in the data and the method, in this section we show that this is not quite true. With

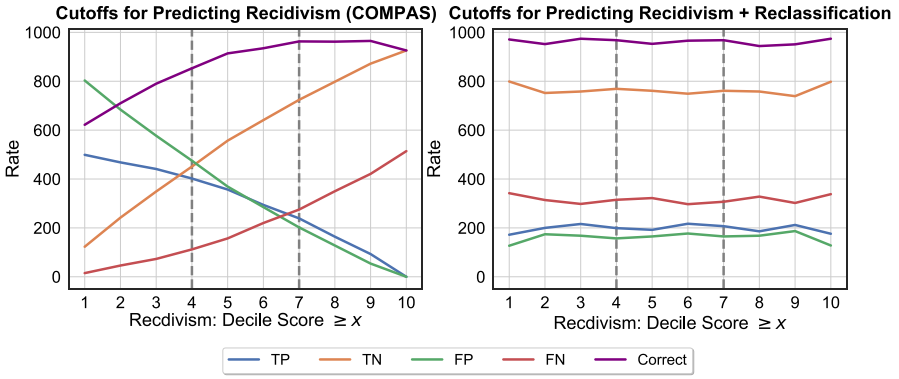

Fig. 3 Anti-defandant bias. On the x-axis, we show the chosen decile score threshold including and above which a judge may consider a defendant for jail. On the y-axis, we show the absolute number of FPs, TPs, FNs, and TNs that would be the outcome of making jail decisions purely based on the respective threshold. The dashed lines are the upper bounds of the "low risk" and "medium risk" ranges

a reasonably simple procedure (using machine learning) the anti-defendant bias can be removed. At a high level, the correction proceeds as follows: For each decile score, we train a neural network model that predicts whether the decision for a certain individual would be incorrect, had the judiciary chosen this threshold. That means we predict whether they would receive bail but will recidivate, or whether they would be jailed but not recidivate. As input, each model gets all the available features in the ProPublica database, except for the ones explicitly indicating ethnicity,[4] as we attempted not to introduce variables not available to the COMPAS algorithm, which would be true for ethnicity, according to the creators of COMPAS (Dieterich et al. 2016). Additionally, the model is fed a binary variable indicating whether the defendant is predicted to recidivate, given the COMPAS score and the threshold. Each model is trained to maximize accuracy, i.e. to minimize the sum of false positives and false negatives, for decisions made based on the given threshold. The COMPAS prediction is replaced by a prediction that maximizes accuracy. Figure 3 contrasts original outcomes based on COMPAS as is (left panel) with the outcomes after applying the proposed correction (right panel). With the correction, all lines are nearly flat. Hence, our correction not only makes predictions more accurate, it also makes the decision for a threshold and the incidence of wrong decisions orthogonal - as it should since accuracy is maximized.

We stress the obvious: the corrected model is not error proof. It just minimizes the risk of committing an error, by retaining an innocent defendant, or by releasing a defendant who seriously risks committing new crime before tried. More importantly even, the proposed correction is also normative - it prioritizes accuracy over preserving an implicit error-rate, seen as acceptable by the judge. With the correction, mostly irrespective of the threshold, more than 2/3 of all decisions are correct.

[4] See the Appendix for detail.

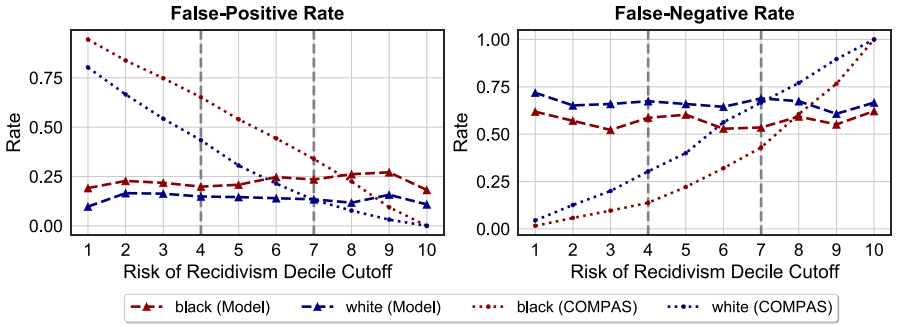

Fig. 4 Racial bias in false positives vs. false negatives. Left panel: rate of defendants jailed although they have not recidivated two years after release. Right panel: rate of defendants released on bail who have recidivated during the next two years. Dotted lines: results when using COMPAS predictions, conditional on threshold chosen by the user (x-axis). Dashed lines: results after our correction. Red: defendants labeled "black", blue: defendants labeled "white". Others are excluded

This is undoubtedly a desirable property. Of the 926 defendants who have actually not recidivated, on average 761, i.e. 82.18%, are released, which is a significant improvement. However, of the 514 defendants who have recidivated, on average only 201 (39.11%) are jailed; considerably more than half of them are released. The judiciary may deem this risk too high. Therefore, we do not argue that the corrected version is normatively superior. But we show that even when only exploiting the relatively small data set published by ProPublica, the costs associated with the trade off between protecting innocent defendants and protecting the public can be changed. If the judiciary decides not to do so, that in itself is a valid choice. However, that decision, in turn, should be made transparent, such that it can be politically discussed and justified.

3 Racial and age bias revisited

3.1 Race

Triggered by ProPublica's findings (Angwin et al. 2013), the normative debate has focused on racial discrimination (Angwin et al. 2013; Fass et al. 2008; Flores et al. 2016; Dieterich et al. 2016; Chouldechova 2017; Agarwal et al. 2020). Little more than a third (2079) of the defendants who are labelled as *race_black* in the ProPublica dataset recidivated, 1934 of the defendants labelled as *race_white*, and 886 labelled as being of another ethnicity, which makes it meaningful to revisit racial discrimination. Figure 4 shows for the original COMPAS predictions as well as the corrected model, and for all thresholds, how the rate of false positives and false negatives depends on ethnicity. Irrespective of the threshold, defendants labeled as "black" are substantially more exposed to false positive predictions: in the left panel of Fig. 4, the dotted red line is above the dotted blue

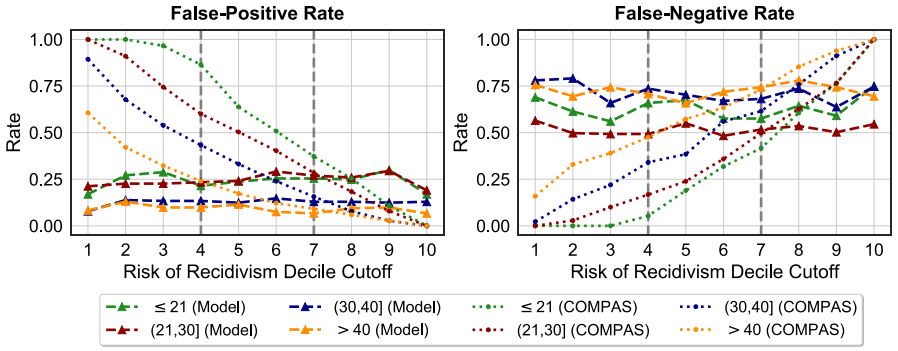

Fig. 5 Age bias in false positives vs. false negatives. Left panel: the rate of defendants jailed that did not recidivate two years after release. Right panel: the rate of defendants released on bail who have recidivated during the next two years. Dotted lines: results when using COMPAS predictions, conditional on threshold chosen by the user (x-axis). Dashed lines: results after our correction. Green: age ≤ 21, red: (21, 30], blue: (30, 40], orange: > 40

line for all possible thresholds. Conversely, the right panel shows that defendants labeled "white" are substantially more likely to benefit from a false negative decision. The correction introduced above is also effective conditional on ethnicity (the dashed lines are largely in parallel). This is worth noting as the correction is tuned towards accuracy and does not directly target ethnicity. As the correction targets accuracy, not race, it does however not remove the gap between defendants labelled as "black" or "white".

3.2 Age

In recent years, academic attention has shifted from racial bias to age bias (Rudin 2019; Jackson and Mendoza 2020; Rudin et al. 2020), chiefly because it has been recognized that both variables are correlated. Defendants labeled as "black" are generally younger. In our dataset individuals labelled as "black" are of an average age of 29.5, whereas defendants labelled "white" are on average 35.2 years old.

As can be seen in Fig. 5, the strong bias against young defendants is reflected in a much higher rate of false positive decisions (the younger the defendant, the higher the dotted line on the left panel). This corresponds to a considerably lower chance for younger defendants of being wrongly released on bail (the younger the defendant, the lower the dotted line on the right panel).

Again, the anti-defendant bias is effectively removed by our correction, also conditional on the respective age bracket (the dashed lines have no trend). For all thresholds below the (fairly generous) threshold of 8, the correction also attenuates the age bias (the dashed lines for the two younger and for the two older age brackets are close to each other, in particular for the false positive rate), but the bias is not fully neutralized (there is a discernible gap between the two younger and the two older age brackets and false positives, to the detriment of younger

defendants; the dashed lines for false negatives still differ by age, although not as intensely as the corresponding dotted lines). This is once more explained by the fact that the correction optimises for accuracy, not for age neutrality.

4 Discussion

In medicine, an evidence-based approach has been widely accepted (Sackett et al. 1996); at least for some diagnostic tasks, dedicated software outperforms human experts. It does not seem far-fetched to draw the analogy to judicial decision making, and to call for evidence-based adjudication (Manski 2020). The more the law wants the decision to rest on a prediction, the more it seems appealing to muster the ever increasing capacity of algorithms, paired with the growing richness of datasets, to make good predictions.

In this paper, we do not argue against algorithmic decision aids in the court room. But the law does not only care about performance. The fact that, on average, in a given domain, one decision maker (the machine) is more accurate than another (the human judge) does not automatically imply that this decision maker should decide. Ultimately, the law does not care about population effects; it cares about individual cases - the opposite of what drives machine learning models. In many contexts, the law must live with imperfection. Facts that are relevant for the decision of the case remain unclear or are contested, and the risk of materially wrong decisions is insurmountable.

Traditionally, this risk is contained procedurally. The most important safeguard is the personality of the judge. They are obliged to weigh the benefits and risks of a decision as best they can, and to be responsible, in person, for the outcome. To the extent that concerns can be generalized, the legislator takes a stand, based on open political debate. The rule of law and democratic legitimacy make the residual imperfection tolerable. In this paper, we show that these safeguards risk being blunted by algorithmic design. The popular COMPAS software does not only perform poorly at the very task for which it has been designed—this has been pointed out before (Grgić-Hlača et al. 2018; De Miguel Beriain 2018)—the software also implicitly assumes the role of a sovereign that, in a democratic country, should be with the courts and the legislator. The software pronouncedly privileges potential victims over potentially innocent defendants. More importantly: this normative decision is concealed. The judge is led to believe that they just decide about the acceptable recidivism risk, while they actually also decide about the risk of unnecessarily jailing a defendant, and about the level of racial and age bias.

Noticing the hidden normative dimension of the algorithm is not an argument against algorithmic decision aids in the court room. However, it is only due to ProPublica's efforts that COMPAS predictions can be compared to ground truth. With this analysis, to the best of our knowledge, we are the first to show the hidden anti-defendant bias in the design of the COMPAS software. Legislators and

(constitutional) courts should only clear the use of prediction software in the courtroom if such decisions have been made transparent. It would then be for democratically legitimate authorities to make the inevitable normative decisions.

As a proof of concept, we also offer an algorithmic cure to the algorithmic problem. We readily acknowledge two limitations: the correction requires data. We were only able to provide the correction since we had access to the data collected by ProPublica. But this limitation is not severe. Even the providers of COMPAS stress that their predictive model should be validated in the jurisdiction that decides to use it in the courtroom (equivant 2019). Moreover if the legislator wanted to privilege victims over defendants (or defendants over victims), the correction algorithm would have to be adjusted. But this would only imply a change in parameters, not the design of a new algorithm.

We note that the ProPublica dataset is only a sample, coming from one jurisdiction, and that we also did not have access to the complete feature set that COMPAS uses for prediction. Hence, in the evaluation of the proposed correction, we could not exploit the full richness of the data. As the correction performs well regardless, this does not seem to be an important limitation. Moreover, we are also forced to rely on the correctness of the ProPublica data, e.g., that a repeat offender was indeed rearrested. These limitations, however, apply to all studies on the COMPAS algorithm.

We have a message that transcends the case of COMPAS: Before algorithmic aids are introduced into the court room, they must be probed for inherent normative choices. As our analysis shows, the normative content may not be immediately visible. Algorithmic decision aids do in particular make it necessary for the legal users to understand the normative choice between false positive and false negative decisions. Once they have made this decision, it must be implemented algorithmically. Our correction algorithm shows that this can be done. The legislator may not hide behind the computer screen: code is law.

Appendix

Raw data

The subset of the ProPublica COMPAS dataset used in this analysis contains outputs (scores) of the COMPAS algorithm. We use data for 5759 defendants who have been tried in Boward County, Florida and for whom ProPublica, using freedom of information legislation, has retrieved the criminal record (Angwin et al. 2013). However, ProPublica's processed dataset is often criticized: While they explain how they calculate individual variables, the supplementary data needed for the calculations is not publicly available. For that reason, Rudin et al. (2020) have gone to tremendous lengths, to recollect the necessary supplementary data. They made that data available to us upon request. Moreover, they publish the code for generating the final data on github.

Table 1 Overview over available input variables—"History of Violence"-subscale items

Feature name	Type	Values	Explanation
p_juv_fel_count	Integer	Count	Prior number of felonies committed by person while the individual was still juvenile
p_felprop_violarrest	Integer	Count	Prior violent felony property offense arrests
p_murder_arrest	Integer	Count	Prior voluntary manslaughter/murder arrests
p_felassault_arrest	Integer	Count	Prior felony assault offense arrests (excluding murder, sex, or domestic violence)
p_misdemassault_arrest	Integer	Count	Prior misdemeanor assault offense arrests (excluding murder, sex, domestic violence)
p_famviol_arrest	Integer	Count	Prior family violence arrests
p_sex_arrest	Integer	Count	Prior misdemeanor assault offense arrests (excluding murder, sex, domestic violence)
p_famviol_arrest	Integer	Count	Prior family violence arrests
p_weapons_arrest	Integer	Count	Prior weapons offense arrest

Table 2 Overview over available input variables—"History of Criminal Involvement"-subscale items

Feature Name	Type	Values	Explanation
p_charge	Integer	Count	Prior number of charges
p_arrest	Integer	Count	Prior number of arrests
p_jail30	Integer	Count	Prior number of times sentenced to jail 30 days or more
p_prison30	Integer	Count	Prior number of times sentenced to prison 30 days or more
p_prison	Integer	Count	Prior number of times sentenced to prison
p_probation	Integer	Count	Prior number of times sentenced to probation as an adult
is_misdem	Integer	[0,1]	If all charges connected to the current offenses are only misdemeanors = 1, otherwise 0 (i.e. at least one charge is in regards to a felony)

Combining both sources, we know for each defendant whether they have previously been jailed or released on bail; whether they have been charged for any other crime during two years after release,[5] as well as 32 more directly observable characteristics, most of which are demographic or concerning the defendant's criminal record. We do not have access to the remaining features COMPAS receives as input, which are answers to questionnaires on personal, vocational, and educational information conducted during the screening process (Northpointe 2015). We slightly adapted the data, to fit our needs, e.g. we constructed the "married"-variable and, unlike Rudin et al. (2020), did not drop all original variables from the input data. The exact build of our final dataset, before we applied any preprocessing to it, may be found in Tables 1, 2, 3, 4, and 5

Constructing decile scores

We want to estimate the frequency of false positive and false negative predictions, conditional on the decile score selected by the judicial user of COMPAS. The decile score is contained in the dataset. We know that scores result from partitioning the raw risk scores into 10 equally sized bins, conditional on the population that was used for normalization. We also know that norm groups are stratified by a combination of the following: gender, prison, jail, parole,

Table 3 Overview over available input variables—"History of Noncompliance"-subscale items

Feature Name	Type	Values	Explanation
p_n_on_probation	Integer	Count	Prior number of offenses while on probation
p_current_on_probation	Boolean	[0,1]	Current offense committed while on probation
p_prob_revoke	Integer	Count	Number of times probation terms were violeted or probation was revoked

[5] Except if the crime is either unobserved or if the new charge was a traffic ticket or a minor municipal ordinance violation, failure to appear in court, or a later charge with a crime that had occurred before the COMPAS screening (Larson et al. 2016). For defendants who have been in jail, the time until recidivism is measured from the day of release from prison onward. This imbalance is inherent in the data.

Table 4 Overview over available input variables—"Characteristics"

Feature name	Type	Values	Explanation
uid	String	–	Unique identifier; Concatination of id and screening date
first_offense_date	String	–	Date of first offense commited
current_offense_date	String	–	Date of the current offense in question for which COMPAS screening took place
offenses_within_30	Integer	Count	Count all offenses that occured up unitl 30 days prior to screening date
p_felony_count_person	Integer	Count	Prior number of felonies committed by person
p_misdem_count_person	Integer	Count	Prior number of misdemeanours committed by person
p_charge_violent	Integer	Count	Number of charges against individual falling under violent crimes/offenses
p_current_age	Integer	Age	Age in years of the individual when committing the offense
p_age_first_offense	Integer	Age	Age when committing the first offense (static)
is_married	Boolean	[0,1]	Baseline is "single"
is_divorced	Boolean	[0,1]	Baseline is "single"
is_widowed	Boolean	[0,1]	Baseline is "single"
is_separated	Boolean	[0,1]	Baseline is "single"
is_sig_other	Boolean	[0,1]	Baseline is "single"
is_marit_unknown	Boolean	[0,1]	Baseline is "single"
sex	string	[Female, Male]	Gender
race_black	Integer	[0,1]	Individual is black = 1 (baseline is race_other)
race_white	Integer	[0,1]	Individual is white = 1 (baseline is race_other)
race_hispanic	Integer	[0,1]	Individual is hispanic = 1 (baseline is race_other)
race_asian	Integer	[0,1]	Individual is asian = 1 (baseline is race_other)
race_native	Integer	[0,1]	Individual is native = 1 (baseline is race_other)
crim_inv_arrest	Integer	Count	"Criminal Involvement"-scale calculated from features (using arrests) as outlined. Scale is a simple sum of count-based-features. Uses p_charge
crim_inv_charge	Integer	Count	"Criminal Involvement"-scale calculated from features (using charges) as outlined. Scale is a simple sum of Count-based features. Uses p_arrest

Table 4 (continued)

Feature name	Type	Values	Explanation
vio_hist	Integer	Count	"History of Violence"-scale calculated from features as outlined. Scale is simple sum of count-based features
history_noncomp	Integer	Count	"History of Noncompliance"-scale calculated from features as outlined. Scale is simple sum of count-based features

Table 5 Overview over target characteristics

Feature Name	Type	Values	Explanation
Risk of Failure to Appear_score_text	String	low, medium, high	Formed from decile score. Medium and high necessitate special consideration for incarceration decision
Risk of Failure to Appear_decile_score	Integer	1–10	Normed raw score. Normed by underlying data we do not have but could approximate. The normalization is done within the county and within age and gender as well as race groups
Risk of Failure to Appear_raw_score	Integer	11–48	COMPAS score Failure to appear
Risk of Recidivism_score_text	String	low, medium, high	Formed from decile score. Medium and high necessitate special consideration for incarceration decision
Risk of Recidivism_decile_score	Integer	1–10	Normed raw score. Normed by underlying data we do not have but could approximate. The normalization is done within the county and within age and gender as well as race groups
Risk of Recidivism_raw_score	Double	(−3) – (2.36)	COMPAS score “Risk of Recidivism”
Risk of Violence_score_text	String	low, medium, high	Formed from decile score. Medium and high necessitate special consideration for incarceration decision
Risk of Violence_decile_score	Integer	1–10	Normed raw score. Normed by underlying data we do not have but could approximate. The normalization is done within the county and within age and gender
as well as race groups. Risk of Violence_raw_score	Double	(−4.63) - (0.5)	COMPAS score “Risk of Violence”
recid	Integer	[0,1]	Individual is recidivist within two years after screening = 1
recid_violent	Integer	[0,1]	Individual is violent recidivist within two years after screening = 1
recid_proPub	Integer	[0,1]	Individual is recidivist within two years after screening = 1 as calculated by ProPublica
recid_violent_proPub	Integer	[0,1]	Individual is violent recidivist within two years after screening = 1 as calculated by ProPublica

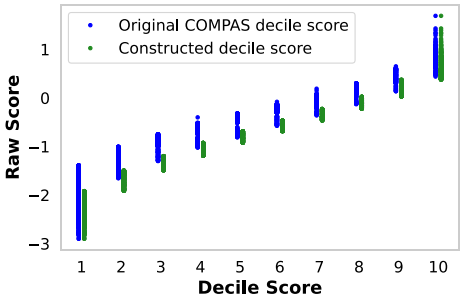

Fig. 6 Original mapping of raw scores onto decile scores vs. our constructed mapping. The constructed mapping creates ten bins with equal numbers of defendants

and probation (Northpointe 2015, p. 11), potentially leading to different decile scores for the same defendant. However, we do not know which norm groups were applied to arrive at the decile scores of individual datapoints. Knowing this reference group is necessary for estimating the correction term. Actually, as may be seen in Fig. 6, the underlying raw scores are not exclusive to one decile score, but rather overlapping at the boundaries—sometimes even the boundaries of three consecutive decile scores (as it is the case for decile scores 3, 4, and 5). To prevent the application of norm groups from biasing results, we reconstruct decile scores, working with raw scores and transforming them into decile scores, the exact same way as described in the COMPAS practitioners guide Northpointe (2015): We distribute the raw scores into 10 equally sized bins, each of which corresponds to a decile score. Consequently, our norm group is the complete training dataset.

To avoid assigning individuals with the same raw score to different bins, we take the uppermost raw score of each decile (see Table 1) as the upper bound of the respective decile bracket. In a second step, we then re-allocate all individuals such that each individual in the next bracket, which has a raw score on the upper bound of the decile bracket below, will be resorted into the decile bracket below. That enables us to have non-overlapping decile brackets. As a consequence, the brackets are not perfectly uniform in size anymore, though the effect is negligible in terms of impact on the absolute bracket sizes.

Simulations

In all three simulations, we generate data for $N = 10000$ persons. The three simulations differ by the assumptions about the distribution of recidivism risk in the population of persons that are apprehended. We either assume this risk to be uniformly distributed on the unit interval $\mathcal{U}[0, 1]$, or normally distributed $\mathcal{N}(\mu = .5, \sigma = .2), [0, 1]$, or use a beta distribution $\mathcal{B}(\alpha = 2, \beta = 5), [0, 1]$. With these parameters, very few observations of the normal or the beta distribution are outside the unit interval. These observations are mapped to the left or right border of the interval. We use a binomial distribution, with the simulated probability of recidivism, to simulate ground truth. We allow for thresholds from 0 to 100%, in steps of 10%. For every threshold, we assume that a person with a

Table 6 Raw score and risk cutoffs for each constructed decile score

Decile score	Raw score upper bound	Risk upper bound (%)
1	−1.92	12.8
2	−1.5	18.2
3	−1.2	23.1
4	−0.93	28.3
5	−0.69	33.4
6	−0.47	38.5
7	−0.23	44.3
8	0.02	50.5
9	0.37	59.1
10	1.69	84.4

Risk, in percent, is calculated as the sigmoid transformation of the raw score times 100

predicted recidivism risk at or higher than this threshold is detained. Using the simulated realizations we classify the observation as a true or false positive if the predicted probability is at or above the threshold, and as a true or false negative if the predicted probability is below the threshold.

Correction model

For our correction, we train neural networks on the training set, optimize hyperparameters based on validation set error, and report the results on the test set, which we left untouched until the final analysis. Specifically, for each decile score threshold, one model was trained to predict errors in the COMPAS predictions. Each model is a feed-forward neural network consisting of four fully-connected ReLU layers, of 28 neurons. We consider each individual in the dataset as one sample and normalize all features contained in the dataset to $[-1, 1]$. Additionally, the network receives information $p \in \{0, 1\}$ about whether the sample would be assigned a positive or negative prediction in terms of recidivism, dependent on the COMPAS decile score $d \in \{1, 2, 3, \ldots, 10\}$ and the threshold $t \in \{1, 2, 3, \ldots, 10\}$. Hence, p is defined as follows:

$$p(d, t) = \begin{cases} 0 & d < t \\ 1 & otherwise \end{cases}$$

For each threshold, a model m_t is trained to predict the errors of COMPAS. In order to construct these errors, we use the ground truth information g on whether an individual recidivated within two years after release, either immediately when released on bail, or after having been in jail or in prison, with $g = 1$ if they did recidivate and $g = 0$ otherwise. Consequently, for any sample, the COMPAS error e is defined as follows:

$$e(d,t,g) = \begin{cases} 0 & p(d,t) = g \\ 1 & otherwise \end{cases}$$

Finally, our correction model is specified on the input features x, each of which comes from Tables 1, Table 2, 3, and 4, excluding the five features pertaining to ethnicity. Each model's objective is to predict, on a per-defendant basis, whether COMPAS' assessment is erroneous:

$$\hat{e}(x,d,t) := m_t(x, p(d,t)) \tag{1}$$

Each model is trained independently, using the mean squared error loss overall N training samples:

$$MSE_t = \sum_{i=1}^{N} ||\hat{e}_i(x_i, d_i, t) - e_i(d_i, t, g_i))||^2 \tag{2}$$

The models are optimized over 250 epochs with a batch size of 500, using the Adam optimizer (Kingma and Ba 2015) with a learning rate of 10^{-3}. We optimized the number of epochs as well as the number of neurons in the hidden layers, making use of a validation set (Tables 5, 6).

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Agarwal R, Frosst N, Zhang X, Caruana R, Hinton GE (2020) Neural additive models: interpretable machine learning with neural nets. arXiv preprint arXiv:2004.13912

Angwin J, Larson J, Mattu S, Kirchner L (2013) Machine bias. There's software used across the country to predict future criminals. and it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

De Miguel Beriain Iñigo (2018) Does the use of risk assessments in sentences respect the right to due process? A critical analysis of the Wisconsin v. Loomis ruling. Law Probab Risk 17(1):45–53

Brooks J, Simpson A (2012) Find the cost of freedom: The state of wrongful conviction compensation statutes across the country and the strange legal odyssey of Timothy Atkins. San Diego L Rev 49:627–668

Carlson A (2017) The need for transparency in the age of predictive sentencing algorithms. Iowa Law Review 103:303–330

Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 5(2):153–163

Dieterich W, Mendoza C, Brennan T (2016) Compas risk scales: demonstrating accuracy equity and predictive parity. Northpoint Inc 7(7.4):1

Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv 4(1):eaao5580
equivant. Practitioners Guide to COMPAS, (2019). URL https://github.com/mhschubert/compas-analysis/blob/master/additional_documents/practitioners-guide_2019.pdf. Accessed 2021 April 30
Fass TL, Heilbrun K, DeMatteo D, Fretz R (2008) The lsi-r and the COMPAS: validation data on two risk-needs tools. Crim Justice Behav 35(9):1095–1108
Flores AW, Bechtel K, Lowenkamp CT (2016) False positives, false negatives, and false analyses: a rejoinder to machine bias: there's software used across the country to predict future criminals and it's biased against blacks. Fed Probat 80:38
Freeman K (2016) Algorithmic injustice: How the Wisconsin Supreme Court failed to protect due process rights in State v. Loomis. North Carolina J Law Technol 18(5):75–106
Grgić-Hlača N, Zafar MB, Gummadi KP, Weller A (2018) Beyond distributive fairness in algorithmic decision making: feature selection for procedurally fair learning. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
Grgić-Hlača N, Engel C, Gummadi KP (2019) Human decision making with machine assistance: an experiment on bailing and jailing. Proc ACM Hum-Comput Interaction 3(CSCW):1–25
Hagan J, Dinovitzer R (1999) Collateral consequences of imprisonment for children, communities, and prisoners. Crime Justice 26:121–162
Jackson E, Mendoza C (2020) Setting the record straight: What the COMPAS core risk and need assessment is and is not. Harvard Data Sci Rev 2(1):1–15
Jung J, Goel S, Skeem J et al (2020) The limits of human predictions of recidivism. Sci Adv 6(7):eaaz0652
Kingma DiP, Ba J (2015) Adam: a method for stochastic optimization. In: Bengio Y, LeCun Y (eds) 3rd International Conference for Learning Representations (ICLR). http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#KingmaB14
Larson J, Mattu S, Kirchner L, Julia A (2016) How we analyzed the COMPAS recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed 2021 July 11
Lessig L (1999) Code: and other laws of cyberspace. ReadHowYouWant.com
Manski CF et al (2020) Judicial and clinical decision-making under uncertainty. J Inst Theor Econ (JITE) 176(1):33–43
Nishi A (2019) Privatizing sentencing. Columbia Law Rev 119(6):1671–1710
Northpointe. Practitioners Guide to COMPAS (2015). https://github.com/mhschubert/compas-analysis/blob/master/additional_documents/practitioners-guide_2015.pdf. Accessed 2021 April 30
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215 arxiv: 1811.10154
Rudin C, Wang C, Coker B (2020) Broader issues surrounding model transparency in criminal justice risk scoring. Harvard Data Sci Rev 2(1):1–16
Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS (1996) Evidence based medicine: what it is and what it isn't
Sampson RJ, Laub JH (1992) Crime and deviance in the life course. Ann Rev Sociol 18(1):63–84
Western B, Kling JR, Weiman DF (2001) The labor market consequences of incarceration. Crime Delinquency 47(3):410–427
Western B, Lopoo L, McLanahan S (2004) Incarceration and the bonds among parents in fragile families. Imprisoning America: The social effects of mass incarceration, pp 21–45