AI & SOCIETY
https://doi.org/10.1007/s00146-023-01673-6

OPEN FORUM

Machine learning in bail decisions and judges' trustworthiness

Alexis Morin-Martel[1]

Received: 16 November 2022 / Accepted: 11 April 2023
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2023

Abstract

The use of AI algorithms in criminal trials has been the subject of very lively ethical and legal debates recently. While there are concerns over the lack of accuracy and the harmful biases that certain algorithms display, new algorithms seem more promising and might lead to more accurate legal decisions. Algorithms seem especially relevant for bail decisions, because such decisions involve statistical data to which human reasoners struggle to give adequate weight. While getting the right legal outcome is a strong desideratum of criminal trials, advocates of the relational theory of procedural justice give us good reason to think that fairness and perceived fairness of legal procedures have a value that is independent from the outcome. According to this literature, one key aspect of fairness is trustworthiness. In this paper, I argue that using certain algorithms to assist bail decisions could increase three different aspects of judges' trustworthiness: (1) actual trustworthiness, (2) rich trustworthiness, and (3) perceived trustworthiness.

Keywords AI ethics · Machine learning · Procedural justice · Relational theory · Bail decisions · Trustworthiness

1 Introduction

In recent years, judges have started using Machine Learning algorithms (ML) as guiding tools to assess the risk of recidivism for convicted criminals among other things and making legal decisions accordingly (Zerilli et al. 2019, 668). The main reason for using algorithms in such a way is the perceived increased accuracy of the results that they supply (Lin et al. 2020, 1).[1] Of course, getting the right legal outcome is a strong desideratum and some believe that it is in fact the only thing that really matters (Dworkin 2011). However, those who hold a relational view of procedural justice believe that judicial procedures also have intrinsic value, independently of getting the right legal result.[2] They argue that certain basic features of the procedures matter significantly for those facing potential legal consequences, because it affects the degree to which they perceive the legal process as being fair. Some (Meyerson et al. 2020, 1) even argue that the perceived fairness of procedure is sometimes as important for defendants as getting the right legal outcome. For instance, empirical studies have shown that defendants were much more willing to accept a legal decision that went against their interests if they judged that they were treated fairly (Meyerson et al. 2020, 4). This perceived fairness is generally broken down into four more precise features: respect, trustworthiness, neutrality, and voice (Meyerson et al. 2020, 4–5). According to this growing literature, even when we get the right legal result, procedures that are lacking these features can have damaging effects on defendants' capacity for self-respect and are also correlated with reduced compliance with the law, another strong desideratum of legal decisions (Sorial 2020, 153–154).

In this paper I focus on the compatibility of using ML algorithms for bail decisions with the 'trustworthiness' and

[1] This claim is highly contentious, when it comes to current models such COMPAS, as we will see.

[2] Dignitarians such as Waldron (2011) also believe that judicial procedures have intrinsic value. However, this paper focuses solely on relational views of procedural justice.

✉ Alexis Morin-Martel
Alexis.morin-martel@mail.mcgill.ca

[1] Philosophy, McGill University, Montreal, QC, Canada

Published online: 21 April 2023

'voice' criteria. More specifically, I argue that judges could increase their actual and rich trustworthiness using ML as an advisory tool to make bail decisions, as long as they remain in control of the final decisions. I also argue that such a use would not undermine their perceived trustworthiness and that it is compatible with defendants' voice being heard. I focus on bail decisions, because such decisions rely on predictions that should be informed by statistics. Crucially, making predictions that rely partly on statistics is a notoriously hard task for humans (Tversky and Kahneman 1978), and new research shows great promise of improved results using ML (Kleinberg et al. 2017a). In Sect. 3, I first provide a quick overview of what ML is. I then discuss the challenges and limitations that judges face when making bail decisions and address a study from Kleinberg et al. (2017a) suggesting that ML could improve the accuracy of such decisions as well as reduce the effects of racist biases in the decision-making process. In Sect. 4, I present a potential implementation of ML that would still allow defendants to have their voice heard. I then present Jones' (2012) distinction between trustworthiness and rich trustworthiness,[3] and argue that a certain implementation of ML could increase both. Finally, I argue that increasing judges' perceived trustworthiness by use of ML requires choosing algorithmic accuracy over transparency when they come into conflict.

2 Preliminary note

This paper does not address the more fundamental question of whether bail should remain a part of our legal institutions in the first place. Some (Guttman et al. 2020) have put forward good arguments to abolish the monetary aspect of bails for misdemeanors and non-violent crimes in the context of Californian legislation. They argue (correctly in my opinion) that monetary bails for such offenses disproportionately impact marginalized people and that their effectiveness is not supported by the empirical data. The view I put forward in this paper is compatible with bail systems that have a much more limited scope than ours or that rely on different incentives than money.[4] In short, my argument remains relevant unless we argue for complete abolition of the bail system (and not just its reform). From the perspective of nonideal theory, if one believes that the prospects of abolishing bail system are low, then one ought to think about how to make it fairer.

3 The use of ML in bail decisions

3.1 Machine learning algorithms as risk assessment instruments

Judges attempting to statistically assess the risk that a given defendant represents predates the current use of ML. Indeed, it goes back at least to the 1920s when "various factors derived from age, race, prior offense history, employment, grades, and neighborhood background were used to estimate success of parole" (Zeng et al. 2017, 689). Historically, this statistical approach had very limited accuracy and relied on problematic racist and classist assumptions. However, recent developments in ML might allow us to go beyond these limitations.

Traditionally, most Artificial Intelligence technologies were "Expert Systems", which means that they had some logical rules encoded into them on how to process specific data (Maclure 2021, 423). They were, therefore, used as powerful logical reasoners, but the scope of what they could achieve was very limited, because the only way AI could process the information differently was by changing their code (Dreyfus 1978). The ML approach changed that. It aims at making AI algorithms progressively better at a given task over time, emulating human learning abilities to a certain extent. It relies on artificial networks containing multiple layers of interconnected nodes called neural networks (Maclure 2021, 423). Fundamentally, the objective is to feed the algorithm a large amount of information related to a desired task. For instance, to make an algorithm that can assess whether something is a fork or a knife, we would show it multiple images of knives and forks in a format that it can process. We would then provide feedback to the algorithm over time.[5] What would happen next is that the neural network would adjust the weight given to various aspects of the information received until it improves its success rate.

ML allows significant improvements in our daily lives. Among other things, it provides much better traffic prediction, and it also makes automatic translations more accurate (Shaygan et al. 2022; Agrawal et al. 2019, 41). Of course, ML algorithms also show inherent technical limitations. For instance, they are strongly dependent on the quality and the quantity of data available relevant for the task at hand. Biases in data collection can be learned and amplified by ML algorithms, leading to biased predictions or decisions.

[3] The crucial distinction which will be explained later is that displaying actual trustworthiness involves being competent in a specific domain and taking the fact that someone else is counting on you as a reason to act in a certain way (Jones 2012, 70–71). By contrast, rich trustworthiness is the ability accurately to signal to others the domains in which you are trustworthy and those in which you are not (Jones 2012, 76).

[4] See Stevenson and Mayson (2017) for alternatives to the current monetary bail.

[5] This is an extremely simplified presentation of ML algorithm training. There are various ways to provide feedback to an algorithm. Some approaches provide more constant feedback (supervised training) while others are much more hands-off (Semi-supervised and Unsupervised) (Ayodele 2010, 19).

They also are prone to overfitting, which happens when an algorithm learns the training data too well, including the meaningless aspects in the data, which can result in poor performance on new data (Ying 2019). Finally, some algorithms known as "black box" models can have a very high degree of complexity, making their decision-making process almost completely opaque for humans (Petch et al. 2022, 204).

Despite these limitations, ML has found uses in public decision-making. For instance, some police departments have made use of an algorithm called PredPol to predict where criminality is more likely to occur on their territory and to deploy their staff accordingly (Asaro 2019).[6] In recent years, some US judges have also made use of ML such as the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) to assess the risk of recidivism in criminals (Zerilli et al. 2019, 667–668). While COMPAS has been subject to significant criticism due to the racist bias it exhibits as well as its general lack of accuracy (Dressel and Farid 2018), certain new algorithms allow for measured optimism. Indeed, some algorithms such as the one designed by Kleinberg et al. (2017a) show promising results for bail decisions. Kleinberg's algorithm has not been deployed for public use, but I believe that it nevertheless provides an interesting window on the new kinds of algorithms that might very well shape the future of bail decisions. So far, most of the ethical discussions surrounding the use of algorithms in the legal system have focused on obviously flawed and highly biased models, such as COMPAS (with good reasons). Nonetheless, this should not prevent us from discussing the ethical challenges of potentially new and better algorithms that are perhaps just around the corner. The main claim that I will defend in this paper is that using an algorithm such as the one developed by Kleinberg as an assistance system would make judges more trustworthy. However, before discussing Kleinberg's algorithm itself, I will say a little more about how bail decisions usually work.

3.2 Bail decisions

Bail decisions are a part of many criminal trials. They happen after the defendant's arrest, but before the criminal trial officially begins. In most North American jurisdictions, the judge must assess the risk that the defendant will not show up for trial (flight risk) or that they will commit another criminal infraction before the trial (Kleinberg et al. 2017a, 239). If these combined risks are assessed to be low, the judge will usually release the defendant before the trial, with or without a monetary bail as an incentive for the defendant to comply. If the risk appears to be too high, the judge can also choose to detain the defendant until the trial begins. Besides assessing these risks, the judge must also consider the impact of jailing the defendant on carceral resources as well as on the defendant's well-being. As with almost all other decisions pertaining to the criminal system, the judge's decision happens after an adversarial presentation of arguments by the prosecution and the defendant.

It is important to note that even if a bail decision is not a judgment of the defendant's guilt, it can still have two significant impacts on their well-being. First, defendants can remain detained for several months awaiting trial, and this, of course, affects their ability to work and provide for their family (Kleinberg et al. 2017a, 239). Second, bail decisions are harder to contest by defendants than the final verdict of culpability or sentencing. In Canada, it is possible for a defendant to obtain a bail decision review under section 520 (1) of Canada's *Criminal Code* (R.S.C., 1985, c. C-46). However, the conditions set by the jurisprudence in R. v. Singh Garcha (2004 SKQB 92) state that the defendant must either prove that there was a clear legal error in the bail decision or that material changes in the defendant's circumstances make it unjust not to vacate the detaining order. When contesting a bail decision, the burden of proof is on the defendant. Furthermore, it is very demanding, since showing that a legal error took place requires the defendant to prove that the judge did not simply misapprehend the evidence, but that an incorrect legal principle was used to make the decision (*R v. J.M.H* 2011 SCC 45). Since bail decisions have a strong impact on defendants' livelihood and are hard to contest, this gives us particularly good reasons to strive for accurate results.

However, when looking at data on bail decisions, one should note that there is a fundamental asymmetry in our capacity to assess whether a judge was right in their prediction that led to the bail decision. Indeed, if a judge releases a defendant who ends up breaking the law or not showing for trial, it is obvious in retrospect that they were mistaken in their prediction.[7] However, it is nearly impossible to know whether someone who was detained would have complied if they were released. It is important to acknowledge this limitation in the available information and that our statistics on judges' mistakes for individual cases can only be about

[6] However, Lum and Isaac (2016, 18–19) provide strong evidence that the use of PredPol led to the over-policing of predominantly black neighborhoods.

[7] As an anonymous referee rightly points out, this does not necessarily mean that the judge was wrong in their risk assessment. They could have correctly assessed the risk of a defendant's non-compliance to be low and rightly decided to release them before trial. In some cases, the judge's prediction would still turn out to be wrong even if their risk assessment was right. Afterall, low risk doesn't mean no risk. Nevertheless, if what we are interested in are general patterns of discrepancies for bail decisions, I believe that we can safely bracket these cases.

defendants that were released but should instead have been detained.[8]

3.3 Kleinberg's ML algorithm

Keeping this limitation in mind, the study by Kleinberg et al. (2017a) allows us to better understand what often goes wrong when a judge's prediction that a defendant can be released is mistaken. The study was based on 758,027 bail decisions from New York City's judges between 2008 and 2013. 73,6% of these cases were met with a favorable release decision (Kleinberg et al. 2017a, 248–249). For each defendant that was released before trial, the researchers also knew whether the defendant was later found to have broken their bail conditions or not. Kleinberg then trained a ML algorithm through a portion of the cases by providing it with the following data from the defendants: current offense, rap sheet, prior failures to appear to court, age. Furthermore, they did not use any other demographics than age, which means that there was no reference to the defendants' race or ethnic group (Kleinberg et al. 2017a, 247). Their goal was to train the algorithm to assess the probability risk that a defendant would not comply with their release conditions.

Before discussing the results from Kleinberg's experiment, I will say a few words about the actual algorithm that they used, which is known as a 'gradient boosted decisions trees.' The algorithm works as "an average of multiple decision trees that are built sequentially on the training data, with each subsequent iteration up-weighting the observations that have been predicted most poorly by the sequence of trees up to that point" (Kleinberg et al. 2017a, 252). This means that, depending on the feedback that the algorithm receives, it will readjust the importance given to some of the decisions trees. This is what is meant here by 'weight.' The average result is then turned into a prediction of the degree of risk that a given defendant represents.

One thing to note here is that gradient boosted decisions trees can become extremely complex for humans to analyze depending on their depth and can be qualified as "black box" models, because they are "sufficiently complex that they are not straightforwardly interpretable to humans" (Petch et al. 2022, 204). While we know the data that was given to the algorithm for the study in question, the actual weight that the system ended up giving to discrete part of the data can be highly difficult to decipher.[9] Furthermore, the use of a black box model rather than a model that is interpretable by humans was not mere caprice. Indeed, the researchers compared the results that they could obtain with an interpretable model (a logic regression) and there were significant gains to using the black box model (Kleinberg et al. 2017a, 260). The fact that Kleinberg's algorithm is a black box model is especially relevant, because some (Demirdag and Shu 2020, 398) argue that such opacity significantly reduces the algorithm's perceived trustworthiness, as we will see in Sect. 4.4.[10]

When Kleinberg and his team judged that their algorithm had enough training, they could test how well the algorithm's predictions fared compared to the judges' decisions in 203,338 cases that were not used for training the algorithm. To compare the algorithm with the judges, they had to determine the percentage of risk that would serve as a threshold to detain or release a defendant. The algorithm was given two different tasks. The first was to match the judges' detaining rate by releasing every defendant under the relevant risk threshold. The second was to attempt to match the judge's number of cases of non-compliance while releasing as many defendants as possible. With both metrics, the results were overwhelmingly favorable to the algorithm:

> The algorithmic rule, at the same jailing rate as the judges, could reduce crime by no less than 14.4% and up to 24.7%; or without any increase in crime, the algorithmic rule could reduce jail rates by no less than 18.5% and up to 41.9%. (Kleinberg et al. 2017a, 241)

What is particularly interesting with this study is that it gives us indications as to when judges went particularly wrong in their evaluations. The result revealed that the judges' most egregious mistakes involved mis-ranking the riskiest defendants. Indeed, "the riskiest 1% of defendants, when released, fail to appear for court at a 56.3% rate and are rearrested at a 62.7% rate. Yet judges release 48.5% of them" (Kleinberg et al. 2017a, 240). It is also important to note that this mis-ranking of the riskiest defendants was not a simple matter of judges being more lenient than the algorithm in general. The study also revealed that "stricter judges do not jail the riskiest defendants first; instead, they appear to draw additional detainees from throughout the predicted risk distribution" (Kleinberg et al. 2017a, 240). This means that while we might have expected stricter judges to detain a higher proportion of the riskiest defendants, this did not occur; rather, they detained both larger numbers of lower risk defendants and larger numbers or higher risk ones.

Obviously, the algorithm did not have as much information about the defendants as the judges did. For instance, the algorithm had no access to the defendants' appearance,

[8] Of course, that does not prevent us from suspecting more general racist patterns, when we see for instance that black people are statistically much more likely to be jailed while awaiting their trials than white people are.

[9] In Kleinberg et al.'s (2017b, 13) unpublished manuscript, they refer to their own algorithm as being a black box and there is no indication in the published version that the algorithm is interpretable by humans.

[10] Robbins (2019, 498) went even further in arguing that black box models should not be used for any decisions with a moral significance because such decisions necessarily require an explanation.

general demeanor, etc. However, this additional information accessible to human judges might partly explain why they were making worse predictions. Studies have reliably shown that human reasoners struggle to make accurate predictions that rely on both statistical correlation and qualitative data. Humans tend not to give enough weight to statistical data and to give too much weight to salient qualitative data. This is a bias known as "base rate neglect" (Lin et al. 2020, 2). Furthermore, recent studies show that judges do not fare much better than ordinary citizens when it comes to this bias (Goel et al. 2021, 10). This is particularly detrimental to accurate bail decisions, since there is a strong statistical correlation between certain statistical data about the defendant and the likelihood of their compliance with the bail conditions. However, even when they know about the relevance of these statistical data, judges still tend not to give it an adequate weight (Kleinberg et al. 2017a, 242). The salient aspects of the case, the defendant's reluctance to make eye contact, their appearance, the judge's mood as well as all other information unavailable to the algorithm did provide the right information (also called "signal") to improve the judge's decision. It instead created a distraction (also called "noise") that prevented rather than aided accurate prediction-making (Kleinberg et al. 2017a, 242). In addition, Danziger et al.'s (2011) research suggests that judges may also be distracted by other factors, such as their level of hunger. Their study of Israeli judges found that prisoners were significantly less likely to be granted parole before lunch than after lunch (Danziger et al. 2011, 6889).[11]

3.4 Algorithms as a tool to mitigate racist biases

At this point, one might worry that the algorithm achieves its superior results by neglecting relevant social considerations that human judges consider. This is known as the "omitted payoffs bias" (Kleinberg et al. 2017a, 242). For instance, if marginalized populations were systematically deemed to present a higher non-compliance risk by the algorithm, it could contribute to their stigmatization and to social inequities.[12] To prevent this kind of scenario, human judges are supposed to avoid relying on discriminatory predictive factors for their judgment. In the New York City context, there was a reasonable fear that racism might play a role in the algorithm's predictions. As we noted earlier, the only demographic characteristic of defendants that was given to the algorithm was age. However, researchers could not exclude that proxy discrimination took place, which is the "possibility that the algorithm winds up using [race or ethnicity] inadvertently—if other predictors are correlated with race or ethnicity" (Kleinberg et al. 2017a, 278).

Fortunately, the New York City database allowed them to test this hypothesis in practice, because the defendants were grouped into three categories: whites, blacks, and Hispanics.

> The algorithm can achieve the same crime rate as the judges but by jailing 40.8% fewer minorities, including 38.8% fewer blacks and 44.6% fewer Hispanics. As a whole, these results suggest our gains are not coming from the hidden costs of increasing racial inequity. If anything, we see that the algorithm can reduce racial inequity. (Kleinberg et al. 2017a, 278)

As we can see, far from detaining more Hispanic or black defendants, the algorithm managed to keep the same level of non-complying released defendants while significantly reducing the percentage of detained defendants from those groups. While Kleinberg does not make that case, it seems very plausible that implicit (or explicit) racist biases could have distracted human judges and partly explains that discrepancy with the algorithm's predictions. Indeed, the data from their study indicates that, from 2008 to 2013, human judges detained Hispanic and black defendants at higher rates than they did white defendants. Furthermore, the claim that judges might suffer from racist biases in their bail decisions is corroborated by a wider study across the US judiciary system. Research shows a substantial bias against black defendants when it comes to bail decisions and that these "results cannot be explained by black–white differences in the probability of being arrested for certain types of crimes" (Arnold et al. 2018, 1929). With those results in mind, it seems that the use of ML could not only be seen as more efficient than judges making such decisions, but it could also help mitigate the effect of racist biases on marginalized groups.[13] Of course, the training set of Kleinberg's algorithm only used data from New York City, and we should be prudent in our assessment on how well it would perform in other jurisdictions. As previously stated, ML algorithms are susceptible to overfitting. To assess its performance in different populations, we must broaden the scope of the research.

[11] However, Chatziathanasiou (2022) gives us some reasons to be more skeptical of the causal relationship at play here.

[12] COMPAS for instance has been rightfully attacked because of the racist biases on which it seems to operate (Dodge et al. 2019, 277).

[13] Another potential conclusion that we could draw from this discrepancy is that we need to modify the judges' environment so that their decision is not impacted as much by potential biases and noise. It could be that human judges would perform as well as algorithms if they were given only the impoverished information that the algorithm used. However, showing this would require further empirical research. I am grateful for an anonymous reviewer pointing this out to me.

4 Use of ML in bail decisions and the relational theory of procedural justice

Kleinberg's study gives strong indication that future algorithms could help make better bail decisions than human judges can achieve without them and that they could help mitigate the negative impacts of racist biases against certain minorities. If we believe that getting the right legal result is all that matters to procedural justice, those facts alone could justify a strong shift towards algorithmic bail decisions. However, I do not claim that we should abandon the adversarial procedure in which the prosecution and the defendant provide arguments. Indeed, even if we achieved more accurate predictions without a bail hearing, there would still be good reasons to believe that procedures are intrinsically important, independently from the fact that they allow reaching correct verdicts. To argue for the intrinsic value of judicial procedures, I turn to the relational theory of procedural justice. According to this theory, good judicial procedures are essential for people to conceive themselves as meaningful and respected members of a community (Meyerson et al. 2020, 4).

A very interesting aspect of the relational theory of procedural justice is that it relies on psychological research to assess individual perception of fairness of people who face legal procedures of various types. One key finding of this type of empirical research is that people who went through a legal process did not only value the final decision regarding their case, but also gave as much importance to the perceived fairness of the procedure. This perception of fairness is generally broken down into four more precise criteria: respect, trustworthiness, neutrality and voice (Meyerson et al. 2020, 4–5). Fostering a positive perception of procedures based on these criteria is valuable for many reasons, but I will focus on two. First, studies show that proper procedures are important for people to conceive themselves as meaningful and respected members of a community, which can promote their sense of self-worth (Meyerson et al. 2020, 4). Second, proper procedures have an impact on future compliance of the individual within the system (Tyler 2006, 379). I believe that these facts alone give us strong reasons to pay close attention to the potential impact that introducing ML in bail decisions could have on perceived fairness.

The aim of Sect. 4 is to provide a potential implementation of ML that could be in line with the requirements of the relational theory of procedural justice. For the sake of brevity, I focus almost exclusively on the 'trustworthiness' component. However, I also include a short but important discussion of the 'voice' component at the beginning of this section. Of course, this means that additional work would be needed to assess whether the 'respect' and 'neutrality' criteria could be satisfied with my suggested implementation.

4.1 Finding the right balance: 'judge assistance system'

Starting with the 'voice' component, I follow Meyerson who defines it as "people's perceptions that authorities have provided them with opportunities to participate in the decision-making process and to voice their points of view." (Meyerson et al. 2020, 5) I believe that getting rid of the adversarial aspect of bail hearings altogether would critically undermine this component of the relational theory. Indeed, it would be hard to argue that defendants have any chance of expressing themselves if the bail decision does not even require a single statement by them or their legal representative. Similarly, if the bail decision hearing was held normally but the final decision gave very limited weight to the adversarial process compared to the algorithm's prediction, it remains very dubious that the 'voice' requirement could be meaningfully fulfilled. Therefore, to acknowledge this dimension of the relational theory of procedural justice, ML would have to remain an advisory tool for judges rather than being the main contributor for the bail decision. How exactly to implement this in practice is an empirical question that falls beyond the scope of this paper. Nonetheless, I would like to sketch a potential avenue. One way to use ML that is consistent with empirical findings on the struggle of human judges is to conceive of ML as similar to a driver assistance system:

> The resulting tool would be analogous to a 'driver assist' system that provides a warning whenever the car begins to do something the driver might not really wish to do, like drift into another lane. In the bail context, the warning might trigger whenever the judge is about to release a high-risk defendant … (Kleinberg et al. 2017b (unpublished manuscript), 24)

In this case, a judge releasing a defendant that is deemed to constitute a very high risk by the algorithm is presented as analogous to a driver veering off the highway at a very high speed. They might have very good reasons to do so, but it still makes sense for the algorithm to flag the issue. Here, I believe that we should resist seeing the two situations as being analogous, since they present fundamental differences. For instance, when a driver receives a warning from their car, they can quickly check if the situation is truly hazardous, which judges cannot do. The comparison also does not mean that implementing such a system in court settings is just as innocuous as driver assistant systems.[14] Nevertheless, comparing the two systems can still be helpful in highlighting the restricted range of application for the

[14] I am grateful for an anonymous reviewer pointing this out to me.

Judge Assistance System (JAS) that I am proposing. Similar to how most driver assistance systems only intervene in unusual situations, a JAS would only provide outputs when a judge's decision significantly conflicts with its own risk assessment. Of course, limiting JAS influence to cases where a high-risk defendant is being released would only ensure that more people are detained, which might not be the most productive way of implementing such a system. I believe that the algorithm should also flag very low risk defendants being detained. This could be compared to a driver assistance system telling you that you are being overly prudent by going at 60 km/h when the speed limit is 90 km/h.

4.2 JAS and actual trustworthiness

Even if ML is merely used as a JAS, some might argue that its use would still undermine the actual trustworthiness of judges. The argument would go as follows: a judge using a JAS partly relies on a tool which they do not really understand to make their decision, which, therefore, make their decision process untrustworthy. I think that this argument fails for two reasons. First, judges already use various tools to reach their verdicts which they do not totally understand, and I believe that they are justified to do so. One could think here of the various technical instruments involved in DNA evidence or fingerprint analysis, among others. To this, some might object that a JAS is qualitatively different, because it gives reasons for a judge to change their mind without allowing them to really understand the underlying decision process of the algorithm. Even if we grant that a JAS is more akin to a human decider that we cannot understand than to a simple measuring instrument, judges routinely accept evidence from expert witnesses despite not always having the ability to completely understand the underlying reasoning. Of course, expert witnesses can be subject to cross-examination in a way that is hardly achievable for algorithms. However, this cross-examination will also require the expertise of other expert witnesses if it pertains to very technical knowledge and judges must often decide which expert to give more credit to in cases of conflicting expert testimonies. In those cases, the judge is still expected to adjudicate based on incomplete understanding. Crucially, in all these cases, it is reasonable to assume that relying on expertise that they do not totally understand still provides judges with better information to make a final decision.[15]

Second, I argue that the use of JAS for bail decisions increases judges' actual trustworthiness by raising their level of competence. To make my case, I turn to Karen Jones' account of trustworthiness, which I believe rightfully explains what being trustworthy entails in judicial settings. Jones' particular understanding of trustworthiness goes as follows:

> Three-place trustworthiness: B is trustworthy with respect to A in domain of interaction D, if and only if she is competent with respect to that domain, and she would take the fact that A is counting on her, were A to do so in this domain, to be a compelling reason for acting as counted on. (Jones 2012, 70–71)

For the sake of my argument, I take B to be the judge, A to be the defendant and the domain of interaction D to be the bail decision. The first thing that we need to determine is what it means in that context for the defendant to 'count' on the judge. I take it to mean that the defendant has a legitimate expectation that the judge will fairly assess their case and not order them to be detained pending trial without good reasons. As Jones puts it, being trustworthy towards someone implies taking the fact that they count on us as a compelling reason to act in a certain way, but not as an overriding reason (Jones 2012, 71). In that sense, it would be unreasonable for a defendant to count on the judge to unconditionally release them, since the defendant is aware that the judge also needs to be trustworthy with respect to the rest of the population. An unconditional release might be at odds with this other objective. However, a fair assessment of the defendant's case also requires the judge to go beyond the merely quantitative non-compliance probability in their analysis. They also need to take into consideration the significant impact that their decision will have on the defendant's well-being. The upshot of this is that a judge relying *only* on a JAS to make their decision would probably not fulfill the role that the defendant counts on them to fulfill. Indeed, a JAS only assess statistical probability, but does not account for the well-being impact of the decision on the defendant. However, the JAS implementation that I suggested earlier would still allow judges to account for well-being in their final decisions and, therefore, is compatible with what defendants count on them to do.

Another notable aspect of Jones' definition is the notion of 'competence' in the domain. I take competence here to involve a certain level of consistency from the judge when acting in the way that the defendant counts on them to act. Of course, this consistency needs to be tied to a very

[15] As an anonymous referee aptly remarks, there are safeguard mechanisms for expert witnesses that are not available for ML algorithms. For instance, if experts fail to live up to certain ethical and legal standards in their testimonies, they can potentially be held legally responsible in a way an algorithm cannot. They can also provide explanation of their reasoning and debate with other experts, which current black box algorithms cannot do. I take this to show that the use of algorithms will require different safeguard mechanisms, involving frequent audits within human institutions that are held accountable for the algorithm's discrepancies, as we will see in Sect. 4.4.

specific kind of activity: applying the right legal standards to the defendant's case. If a judge was to release a defendant based on random dice results, even a positive outcome for the defendant would not mean that the judge is trustworthy. Indeed, a judge reaching the desired outcome based on luck does not show enough consistency for our general understanding of competence. Furthermore, they also do not apply any relevant legal standards to the case, which is the action that they are expected to do. I believe that a JAS would really bolster judges' competence, since it would act as a heuristic tool that should sometimes lead them to reassess the weight that they gave to information that is not available to the algorithm. Broadly speaking, it would act as a reminder for judges to engage in metacognition and make sure that their decision is based on actual signal, and not noise or biases. As human reasoners, judges are prone to biases and bail decisions trigger such biases, because they rely on the qualitative aspects of a case and on statistical base rate. In that context, ML would serve as a tool to help judges to critically reassess whether they have given enough weight to statistical base rate and whether they might have given too much importance to qualitative and very salient but poorly predictive aspects of the case, such as the defendant's appearance or lack of eye contact. Being competent for a judge certainly involves a reasonable use of the statistical tools at their disposal for the risk evaluation and I believe to have shown that using a JAS could help achieve that goal.

4.3 JAS and rich trustworthiness

Crucially, I also believe that the recognition by judges that they need a JAS would also be a way for them to exhibit what Karen Jones calls 'rich trustworthiness.' Rich trustworthiness involves the ability for B to signal to A the domains in which they are competent and can be counted on (Jones 2012, 74). As Jones puts it:

> Rich trustworthiness requires not only competence in a domain, but also competence in assessing my own competence, so that I neither signal competences I do not have, nor "hide my light under a bushel." I need to engage in ongoing reflective self-monitoring of my own competences so that I know them and their limits. (Jones 2012, 76)

Rich trustworthiness can be seen as a kind of 'meta-trustworthiness', because it is not constrained to one domain but requires B to have a kind of reflective understanding of the scope of their own competence across various domains. It also requires the ability to communicate accurately the domains in which B is reliable to others. For instance, if my neighbor asks for my help in reviewing grammatical mistakes in their resume written in Romanian, showing rich trustworthiness could involve telling them that my own understanding of Romanian grammar is shaky at best, and that they cannot rely on me to detect potential mistakes. However, I would not be showing rich trustworthiness if I always signaled my inability to be counted on to my neighbor whenever they asked for my help across multiple domains.

In her account of rich trustworthiness, Jones focuses on interpersonal relations, but Alfano and Hjuits extend this notion to institutional contexts with the notion of 'global rich trustworthiness' which "measures not just how B is disposed towards some particular person but how B is disposed towards other people more broadly" (Alfano and Hjuits 2020, 258). In the context of this discussion, the relationship would be between judges and the defendants for whom they must make bail decisions. What does it mean for a judge to exhibit rich trustworthiness in that setting? Contrary to individuals who can often simply signal their incompetence relative to a domain like in my resume reviewing example, judges cannot always do this. Of course, rich trustworthiness for judicial institutions can involve refusing to hear cases when they fall outside of one's jurisdiction. By doing so, a court sends the signal that it is beyond its mandate or competence to settle such cases. Similarly, individual judges can signal that they cannot be counted on when they are in a situation of potential or real conflict of interest. However, when it comes to bail decisions, some judge will have to hear the case at some point and to adjudicate. I believe that using JAS can be a way for judges to exhibit rich trustworthiness. As I have argued earlier, humans are prone to various biases such as base rate neglect and struggle to make accurate enough predictions for bail decisions. Since they must nevertheless adjudicate, using JAS would be a way for judges to acknowledge the limits of their competence as human thinkers and to offset these limits in a particular domain (e.g., bail decisions). I believe that it would not be very different from the many other domains in which we now expect professionals to make use of technology to improve their accuracy. For instance, nobody would trust a doctor to perform eye surgery without relying on a technology that minimizes the potential for human mistakes. Under normal circumstances, I believe that an eye surgeon who tells their clients they can perform surgery without such a technology would probably not be displaying rich trustworthiness.

As we can see, there is a strong connection between this claim about rich trustworthiness and the notion of competence which Jones recognizes as an essential feature of actual trustworthiness. I believe the interaction goes as follow. Judges first need to display rich trustworthiness by acknowledging that relying on their judgment alone for bail decisions makes them untrustworthy in that domain. This can then lead to the use of a JAS, which increases their competence, and, therefore, makes them more trustworthy to defendants in that domain.

4.4 JAS and perceived trustworthiness

Even if I am right and JAS could increase judges' actual and rich trustworthiness, relational theories of procedural justice also care deeply about perceived trustworthiness. It is not enough for JAS to improve judges' actual trustworthiness; it is also crucial for defendants who interact with the judge to be convinced of this. Here, algorithms seem to face an uphill battle. For instance, Wang's (2018, 20) extensive survey's conclusion is that the public has "an intrinsic distrust of algorithms" when it comes to their predictions and decisions. Another study shows that this extends to the use of algorithms in managerial decisions (Höddinghaus et al. 2021), while a study by Dietvorst et al. (2015) concludes that "people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake." When it comes to perceived trustworthiness of predictive algorithms, a common line of reasoning is that it is the opacity of algorithms that make them seem untrustworthy and that we, therefore, need more interpretable models to improve perceived trustworthiness (Ribeiro et al. 2016, 1144). In this last subsection, I argue that the best way to increase judges' perceived trustworthiness when they use a JAS is not to avoid black box models, but to use models that are the most accurate (transparent or not). I also argue that perceived trustworthiness requires these algorithms' accuracy to be closely monitored within legal institutions.

At first glance, having stronger algorithmic transparency seems like a promising avenue to increase perceived trustworthiness.[16] As noted in Sect. 3.3, Kleinberg's algorithm is a black box model because of the great complexity of the operations that it involves. Currently, the best explanation of their algorithm's decisions would most likely come from a second (post-hoc) algorithm that would highlight salient factors that seem to generate shifts in the first algorithm's predictions.[17] If JAS is also a black box algorithm, judges would only be able to approximate what salient statistical parameters of the case led to the algorithm's result by referring to such a post-hoc explanation. In the JAS implementation that I have in mind, this explanation would then serve as a heuristic tool for judges to critically reassess if they have given enough weight to the statistical data and whether to change their initial decision or not. Some (Rudin 2019, 206) worry that a post-hoc explanation might not be a true explanation at all, but more akin to a statistical guess. Consequently, there is a strong desire among AI ethicists for intrinsically interpretable models, i.e., models that are simple enough for humans to understand. Rudin (2019, 206), a leading expert in the field, even claims that no algorithms useful in high stake public decisions require the level of complexity that would make them black boxes.

However, as I explained earlier, Kleinberg et al. (2017a, 259–260) employed a black box model (gradient boosted decisions trees) specifically, because it allowed a predictive accuracy that was unmatched with more interpretable models (such as a logistic regression). In addition, a recent study gives us reasons to doubt the intuitive idea that intrinsically interpretable algorithms really allow human reasoners to better understand and predict algorithmic decisions than a black box models with post-hoc explanation. It seems that complex intrinsically explainable model can end up confusing human users, causing them to fare worse in their ability to explain what the algorithm is doing than for some black box models (Bell et al. 2022, 248). One possible interpretation of their results is that we might be misguided altogether in believing that we need to give up on more accurate black box algorithms in favor of more interpretable (but less accurate) algorithms. Since our goal with JAS would be to help judges critically reassess their decisions, this seems to cast doubt on the idea that this goal would be better achieved using intrinsically interpretable algorithms instead of a black box algorithm with a post-hoc explanation.

Furthermore, I argue that there also are reasons to doubt that moving away from black box models in favor of less accurate but interpretable models would be the most efficient path to increase the perceived trustworthiness of judges who use them. Of course, it is well-known that there is a general distrust of algorithms among the population (Wang 2018, 20). However, in a recent survey, a strong majority of respondents (65.6 to 69.7%) preferred algorithmic decisions to human decisions when the algorithm was more accurate (Wang 2018, 13). Furthermore, when made to choose between a transparent and a non-transparent but more accurate algorithm, 63.4% of the respondents preferred the non-transparent to the transparent one (Wang 2018, 16). Of course, to have an algorithm that is more accurate and transparent would be perceived as even more desirable by the public. Nevertheless, this suggests using black box algorithms such as the one deployed by Kleinberg could lead to

[16] Algorithms can lack transparency for many reasons. Some algorithms (such as COMPAS) are opaque not in virtue of their intrinsic complexity but because they are owned by companies who do not want to share the data used to preserve trade secrets or otherwise to allow others to benefit from their unique knowledge of the algorithm (Rudin 2019, 209). In my opinion, there are strong ethical reasons to refuse using algorithms that do not share their data such as COMPAS for public decisions. However, a proper defense of this argument would go beyond the scope of this paper.

[17] This general project is known as explainable AI (XAI), but discussing it is beyond the scope of this paper. See Dodge et al. (2019) for various kinds of post-hoc explanations available for judicial decisions and Wachter et al. (2017) on the topic of counterfactual explanations.

a larger increase of perceived trustworthiness than using a transparent but less accurate algorithm.[18]

Here, one might object that I have not shown that people perceive non-transparent algorithms as being trustworthy. I have merely shown that accuracy considerations outweighed their distrust. In other words, people would accept to forfeit some perceived trustworthiness in exchange for better outcomes. To this, I argue that while actual trustworthiness and outcomes are very distinct considerations, it is not so clear to me that perceived trustworthiness and outcomes are so distinct in the case of algorithms' use.[19] For instance, recent studies have shown that people trust algorithms over experts for many tasks such as estimating someone's weight based on a photo, recommending romantic matches, and predicting the popularity of songs (Logg et al. 2019, 90). By contrast, older studies ranging from the 1950's and 1970's showed much stronger distrust of algorithmic decisions compared to human decisions (Wang 2018, fn. 4). Furthermore, studies show a strong correlation between how much better algorithms get at their task and how much trust the public puts in them (Logg et al. 2019, Wang 2018, Yin et al. 2019). It, therefore, seems that the perception that algorithms are effective at what they are doing (securing the right outcomes) can increase perceived trustworthiness.

One could object that JAS' statistical risk assessment is very different in nature from algorithms estimating someone's weight. Indeed, while it seems relatively easy to assess whether an algorithm estimates weight accurately, JAS would have to assess the non-compliance risk of countless defendants whose cases always are unique. How could we ever know whether JAS truly is accurate? Here, I believe that we need to remember that JAS would be deployed within legal institutions, where they would undergo frequent audit processes. It seems to me that courts would accumulate quite a lot of data on the accuracy of JAS by tracking bail decisions results over time. To do so, they could compare defendants' non-compliance rate over time for judges' decisions that triggered a JAS red flag. They could assess the comparative accuracy of judges' final decisions when they modified their decision based on JAS' assessment as opposed to when they did not.[20] Comparisons using other metrics than non-compliance rate should also be made, such as detention before trial rate, as well as detention rate of minorities. Ultimately, the assessment of JAS' accuracy would be made within legal institutions based on quantifiable data that could lead us to discard it or improve it over time.[21] If JAS shows strong results across all these metrics, as Kleinberg et al. (2017a) let us hope it could, this should go a long way to increase perceived judges' perceived trustworthiness once the public is made aware of such metrics.

5 Conclusion

In this paper, I have argued that future use of Judge Assistance System (JAS) as a support to bail decisions could allow judges to display more actual and rich trustworthiness towards defendants. I also argued that focusing on the JAS' effectiveness rather than on its transparency when they come into conflict is more likely to help preserve the public perceived trustworthiness of judges who will use them. If this is right, the use of JAS can be made consistent with at least one parameter of the relational theory of procedural justice. Nevertheless, to show that JAS really is truly compatible with the relational theory of procedural justice, much further work would need to be done.

Indeed, while this paper briefly discussed the 'voice' requirement, nothing has been said of the 'neutrality' and 'respect' requirements. Ultimately, I believe that neutrality criterion could very probably be met. Indeed, algorithms can be designed in a way that makes them more consistent in their predictions than humans. It may also be possible to run counterfactual scenarios to assess the potential biases of an algorithm, something that we cannot do for human judgment. When it comes to the 'respect' criterion, it seems harder to assess whether allowing an algorithm to guide judges' decisions could be done in a respectful way. When facing very persuasive defendants, JAS could lead judges to give less weight to their testimony, which could be seen as disrespectful towards them. However, in cases where the judges' negative assessment of a defendant rests

[18] This is not to say that we should not value algorithms' transparency for other reasons. I merely argue that accuracy should be favored over transparency when it comes to algorithms' perceived trustworthiness.

[19] While my conflation of perceived trustworthiness and outcomes for algorithms might seem at odds with core principles of the relational theory of procedural justice, I do not think that it is. For instance, Bottoms and Tankebe (2020, 101) argue that citizen's assessment of procedural fairness also include perception about getting the right outcomes, especially in high stake contexts.

[20] I believe my suggested JAS implementation to be compatible with Maclure's (2021, 433–434) institutional approach, which involves keeping human institutions in the loop when using black box algorithms.

[21] As an anonymous referee remarks, another reason why an algorithm such as JAS should be frequently audited is that human behavior may change over time. This could lead the algorithm's accuracy to decrease after its implementation. As evidenced during the COVID-19 pandemic, numerous algorithms experienced difficulties due to significant changes in people's behavior (Heaven 2020).

on disrespectful biases[22] towards the defendant, JAS could be a way to limit the consequences of such disrespect.

Acknowledgements An earlier version of this paper was presented to the philosophy department at McGill University. I wish to thank Pr. Natalie Stoljar, Pr. Jocelyn Maclure, Philippe-Antoine Hoyeck, Alexander Carty, Christophe Facal, and Keven Bisson for their help.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends intechnology, arts, science and society, commenting on issues of concernto the research community and wider society. Whilst the drivefor super-human intelligence promotes potential benefits to widersociety, it also raises deep concerns of existential risk, therebyhighlighting the need for an ongoing conversation between technologyand society. At the core of Curmudgeon concern is the question:What is it to be human in the age of the AI machine? -Editor.

Funding This research project is supported in part by funding from the Canadian Social Sciences and Humanities Research Council (SSHRC).

Data Availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The author certify that he has no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Agrawal A, Gans JS, Goldfarb A (2019) Artificial intelligence: the ambiguous labor market impact of automating prediction. J Econ Perspect 33(2):31–50. https://doi.org/10.1257/jep.33.2.31

Alfano M, Huijts N (2020) Trust in institutions and governance. In The Routledge handbook of trust and philosophy. Routledge, pp 256–270

Arnold D, Dobbie W, Yang CS (2018) Racial bias in bail decisions. Q J Econ 133(4):1885–1932. https://doi.org/10.1093/qje/qjy012

Asaro PM (2019) AI ethics in predictive policing: from models of threat to an ethics of care. IEEE Technol Soc Mag 38(2):40–53. https://doi.org/10.1109/MTS.2019.2915154

Ayodele T (2010) Types of machine learning algorithms. In: Zhang Y (eds) New advances in machine learning. InTech, pp 19–48. https://doi.org/10.5772/9385

Bell A, Solano-Kamaiko I, Nov O, Stoyanovich J (2022) It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. 2022 ACM Conference on Fairness, Accountability, and Transparency, pp 248–266. https://doi.org/10.1145/3531146.3533090

Bottoms A, Tankebe J (2020) Procedural justice, legitimacy, and social contexts. In Procedural justice and relational theory. Routledge, pp 85–110

Chatziathanasiou K (2022) Beware the lure of narratives: "hungry judges" should not motivate the use of "artificial intelligence" in law. German Law J 23(4):452–464. https://doi.org/10.1017/glj.2022.32

Danziger S, Levav J, Avnaim-Pesso L (2011) Extraneous factors in judicial decisions. Proc Natl Acad Sci 108(17):6889–6892. https://doi.org/10.1073/pnas.1018033108

Demirdag I, Shu S (2020) Insights into the black box: input explainability of algorithmic decisions drives consumer satisfaction in the digital world. NA Adv Consum Res 48:297–298

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 144(1):114–126. https://doi.org/10.1037/xge0000033

Dodge J, Liao QV, Zhang Y, Bellamy RKE, Dugan C (2019) Explaining models: an empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp 275–285. https://doi.org/10.1145/3301275.3302310

Dreyfus HL (1978) What computers can't do: the limits of artificial intelligence. Harper Collins

Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv 4(1):eaao5580. https://doi.org/10.1126/sciadv.aao5580

Dworkin R (2011) Justice for hedgehogs. Belknap Press of Harvard University Press

Goel S, Shroff R, Skeem J, Slobogin C (2021) The accuracy, equity, and jurisprudence of criminal risk assessment. Research handbook on big data law, pp 9–28

Guttman Z, Hebner Y, Mori K, Balk J (2020) Beyond cash bail: public health, risk assessment, and california senate bill 10. J Sci Pol Govern. https://doi.org/10.38126/JSPG170107

Heaven WD (2020) Our weird behavior during the pandemic is messing with AI models. MIT Technology Review). https://Www.Technologyreview.Com/2020/05/11/1001563/Covid-Pandemic-Broken-Ai-Machine-Learning-Amazon-Retail-Fraud-Humans-in-the-Loop/. Accessed 15 June, 2020

Höddinghaus M, Sondern D, Hertel G (2021) The automation of leadership functions: would people trust decision algorithms? Comput Hum Behav 116:106635. https://doi.org/10.1016/j.chb.2020.106635

Jones K (2012) Trustworthiness. Ethics 123(1):61–85. https://doi.org/10.1086/667838

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2017a) Human decisions and machine predictions. Q J Econ. https://doi.org/10.1093/qje/qjx032

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2017b). Human decisions and machine predictions. Unpublished manuscript, Working Paper 23180

Lin Z, Jung J, Goel S, Skeem J (2020) The limits of human predictions of recidivism. Sci Adv 6(7):eaaz0652. https://doi.org/10.1126/sciadv.aaz0652

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: people prefer algorithmic to human judgment. Organ Behav Hum Decis Process 151:90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lum K, Isaac W (2016) To predict and serve? Significance 13(5):14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Maclure J (2021) AI, explainability and public reason: the argument from the limitations of the human mind. Mind Mach 31(3):421–438. https://doi.org/10.1007/s11023-021-09570-x

Meyerson D, Mackenzie C, MacDermott T (2020) Procedural justice in law, psychology, and philosophy. Proc Just Relation Theory. https://doi.org/10.4324/9780429317248

Petch J, Di S, Nelson W (2022) Opening the black box: the promise and limitations of explainable machine learning in cardiology. Can J Cardiol 38(2):204–213. https://doi.org/10.1016/j.cjca.2021.09.004

R. v. J.M.H., 3 SCR 197 (Canada (Federal › 2011). https://canlii.ca/t/fnbb2

[22] See Stoljar (2020) on whether implicit biases can be disrespectful.

R. v. Singh Garcha, No. 54 (Saskatchewan › March 9, 2004). https://canlii.ca/t/1gvqp

Ribeiro MT, Singh S, Guestrin C (2016) "Why Should I Trust You?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 1135–1144. https://doi.org/10.1145/2939672.2939778

Robbins S (2019) A misdirected principle with a catch: explicability for AI. Mind Mach 29(4):495–514. https://doi.org/10.1007/s11023-019-09509-3

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. https://doi.org/10.1038/s42256-019-0048-x

Shaygan M, Meese C, Li W, Zhao XG, Nejad M (2022) Traffic prediction using artificial intelligence: review of recent advances and emerging opportunities. Transport Res Part c: Emerg Technol 145:103921. https://doi.org/10.1016/j.trc.2022.103921

Sorial S (2020) Legal legitimacy and the relevance of participatory procedures. Proced Just Relation Theory. https://doi.org/10.4324/9780429317248-9

Stevenson M, Mayson SG (2017) Bail reform: new directions for pretrial detention and release. In: Academy for justice. A Report on Scholarship and Criminal Justice Reform

Stoljar N (2020) Racial profiling as pejorative discrimination. Proced Just Relation Theory. https://doi.org/10.4324/9780429317248-15

Tversky A, Kahneman D (1978) Judgment under uncertainty: heuristics and biases: biases in judgments reveal some heuristics of thinking under uncertainty. Uncert Econ 17–34

Tyler TR (2006) Psychological perspectives on legitimacy and legitimation. Annu Rev Psychol 57(1):375–400. https://doi.org/10.1146/annurev.psych.57.102904.190038

Wachter S, Mittelstadt B, Russell C (2017) Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv JL & Tech 31:841

Waldron J (2011) The rule of law and the importance of procedure. NOMOS: Am Soc Pol Legal Phil 50:3.

Wang AJ (2018) procedural justice and risk-assessment algorithms. SSRN Electron J. https://doi.org/10.2139/ssrn.3170136

Yin M, Wortman Vaughan J, Wallach H (2019) Understanding the effect of accuracy on trust in machine learning models. In: Proceedings of the 2019 CHI conference on human factors in computing systems, pp 1–12. https://doi.org/10.1145/3290605.3300509

Ying X (2019) An overview of overfitting and its solutions. J Phys Confer Ser 1168(2):022022. https://doi.org/10.1088/1742-6596/1168/2/022022

Zeng J, Ustun B, Rudin C (2017) Interpretable classification models for recidivism prediction. J R Stat Soc A Stat Soc 180(3):689–722. https://doi.org/10.1111/rssa.12227

Zerilli J, Knott A, Maclaurin J, Gavaghan C (2019) Transparency in algorithmic and human decision-making: is there a double standard? Philos Technol 32(4):661–683. https://doi.org/10.1007/s13347-018-0330-6

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.