Journal of Economic Perspectives—Volume 35, Number 4—Fall 2021—Pages 71–96

Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System

Jens Ludwig and Sendhil Mullainathan

The criminal justice system routinely fails at its central mission: delivering justice. Empirical studies reveal a system that is inconsistent in its judgments, mistaken in its predictions, and disparate in its impacts. The same type of defendant handled by different judges is treated very differently, and the same judge treats cases differently from day to day—what behavioral scientists call "noise" (Kahneman, Sibony, and Sunstein 2021). Decisions are often systematically mistaken in ways that could have been better identified in advance: for example, in pre-trial decisions, judges release many high-risk people while simultaneously jailing low-risk people (Kleinberg et al. 2018a). Moreover, certain groups (often disadvantaged in other ways) disproportionately bear the brunt of these problems and receive worse treatment, to the point that one could credibly argue the system is discriminatory against them.

Algorithms have a long history in criminal justice as a potential solution to these problems. Statistical models date back to the 1920s. Explicit guidelines for judges were used even earlier and are themselves primitive algorithms: that is, they are explicit rules for how a judge should decide based on case and defendant characteristics. *In principle*, carefully formed rules provide a way to reduce inconsistency,

■ *Jens Ludwig is the Edwin A. and Betty L. Bergman Distinguished Service Professor, University of Chicago Harris School of Public Policy, Chicago, Illinois. Sendhil Mullainathan is the Roman Family University Professor of Computation and Behavioral Science, University of Chicago Booth School of Business, Chicago, Illinois. Both authors are Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are jludwig@uchicago.edu and sendhil@chicagobooth.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.4.71.

error, and (if constructed with that aim) racial bias (Milgram et al. 2014). It is no surprise, then, that new tools from machine learning have drawn a great deal of interest in criminal justice (Berk 2018). They offer a superior version of what already appealed to many in a very crude form: algorithms trained on large datasets can extract greater predictive signal, and can also rely on inputs that could not have entered simple statistical models or guidelines, such as speech, text, or video. The result is the growing proliferation of algorithms across a wide range of criminal justice applications, as shown in Table 1.

But the optimism for machine learning in criminal justice did not last long. *In practice*, algorithms often proved less helpful than anticipated.[1] In many cases, they were even actively harmful. Some algorithms proved to be no more accurate than the judges whose prediction errors they were purported to correct. Reports emerged of algorithms that were themselves discriminatory, producing racially disparate outcomes at a high enough rate that the phrase "algorithmic bias" has entered the lexicon.[2] The algorithms also introduced new problems of their own, such as a lack of transparency—defendants unable to access the "black boxes" that dictated their fates—and concerns that the system is being depersonalized in a way that compromises due process. The best that could be said, it sometimes seemed, is that at least algorithms are consistent—if inscrutable—in their mistakes.

Why were hopes dashed? One common critique points to features of machine learning itself. According to this argument, the data used to train algorithms are too noisy and biased. The complexity of criminal justice objectives cannot be quantified. These decisions are too important to cede control to black boxes. Consequently, the introduction of algorithms into criminal justice is increasingly viewed as an inherently flawed enterprise. We argue that each of these problems follow from a deeper one. Algorithms fail because of shoddy construction: human decisions about how to build and deploy them is the root cause of problems. Machine learning algorithms in criminal justice are not doomed to fail, but algorithms are fragile: if crucial design choices are made poorly, the end result can be (and is often) disastrous.

One reason for the fragility of algorithms comes from important econometric problems that are often overlooked in building them. Decades of empirical work by economists shows that in almost every data application, the data is incomplete, not fully representing either the objectives or the information that decision-makers possess. For example, judges rely on much more information than is available to algorithms, and judges' goals are often not well-represented by the outcomes provided to algorithms. These problems, familiar to economists, riddle every case where

[1] Throughout the paper we use the term "algorithms" not only in the general sense but also to refer to the more specific end-product of work in the artificial intelligence field of machine learning. We use "artificial intelligence" and "machine learning" interchangeably in what follows and make clear from the context which definition of the term "algorithm" we mean.

[2] "Bias" or "violations of fairness" in social science and legal scholarship usually refers to some combination of disparate treatment, disparate impact, or the principle of fair representation (some also call this "statistical parity"). Computer science, as we discuss below, adds a number of additional definitions. We use the term broadly for most of the paper, but where relevant note which specific definition we mean.

Table 1
Illustrative Applications of Artificial Intelligence in Criminal Justice

Type of application	*Examples*
Investigative/forensic uses	Facial recognition to match closed-circuit television images to mugshots
	Social media image searches to find defendant alibis
	Uses of images for investigations (for example: using backgrounds in image to link suspect to image in child abuse case)
	License plate readers
	Auditing police body-worn camera footage
Detection/monitoring/surveillance	Facial recognition to find lost children, other missing persons
	Gunshot detection
	Chatbots to combat grooming and "sex tourism"
	Closed-circuit television to help airport security decide whom to investigate further
Decision aids	Risk tools for pre-trial release
	Risk tools for diversion decisions
	Risk tools for sentencing
	Risk tools for parole decisions
	Predictive policing (places and times)
	Predictive policing (people)

algorithms are being applied. Another reason is that in criminal justice settings, the algorithm is not the final "decider"—a human is. Building good algorithms requires understanding how human decisions respond to algorithmic predictions. Algorithm builders too often fail to address these types of technical challenges because they haven't had to. Existing regulations provide weak incentives for those building or buying algorithms, and little ability to police these choices.

Economists and other social scientists have a key role to play in building and studying algorithms, because such efforts require econometric, regulatory, and behavioral expertise. The return to such efforts is high: if designed well, algorithms have a chance to undo human fallibility. Algorithms have another benefit—when regulated well, their problems are easier to diagnose and more straightforward to fix than are the problems of human psychology (Kleinberg et. al. 2018c). It is easier to improve fragile algorithms than fallible decision-makers.

We illustrate these ideas for the case of algorithmic bias: why racial disparities arise in algorithms and what can be done about it. We illustrate how poorly built algorithms can exacerbate bias. At the same time, *well-built* algorithms can *reduce* bias. They can, in fact, be a force for social justice. So there is room for cautious optimism: algorithms can still do some good in criminal justice, but only if great care is taken.

The problems and opportunities we highlight for algorithms in criminal justice apply more broadly. Algorithms are increasingly used in a range of areas of interest to economists including the labor market, education, credit, and health care. The issues we raise have equal importance there, and in several cases, have already started to make an appearance. Anyone interested in the effect of algorithms on society has lessons to learn from the criminal justice experience.

Inconsistency, Error, and Discrimination in Judicial Decision-Making

America's criminal justice system is clearly broken. An overwhelming majority of people see the problems and think they must be fixed.[3] The incarceration rate has exploded in a way that has no historical or international precedent (as discussed by Bruce Western in this issue). Nor is the current system, with all of its social costs, providing public safety: America's murder rate far exceeds that of any other high-income nation. Meanwhile, the burden of both crime and incarceration falls disproportionately on minority communities: for example, 70 percent of Black male high school dropouts spend time in prison by their mid-30s.

In this paper, we will focus on inconsistency, error, and discrimination in the criminal justice system. These problems pervade almost every part of the system, ranging from law enforcement to how cases are adjudicated innocent or guilty (plea-bargaining, trials, and other steps) to how people are supervised out in the community on probation or parole. We will focus here on three types of criminal justice decisions that are representative of the broader challenges and substantively important in their own right: pre-trial detention, sentencing, and parole decisions. Given the vast literature, we focus here on selected examples.

The *pre-trial detention* decision occurs soon after an arrest. The defendant must appear in front of a judge within 24 to 48 hours. The judge typically has several choices: release the defendant under their own recognizance (a promise to return for trial); set release with certain conditions, like wearing a location-monitoring device; requiring cash collateral (bail) for release to ensure return to court; or refusal to release the defendant before trial at all. In general, this decision is supposed to depend on the judge's assessment of the defendant's risk to public safety and/or the likelihood that the defendant will appear in court for trial.

The *sentencing decision* occurs when a defendant has been found guilty. This decision will depend on the crime for which the person was convicted but also on the likelihood of future re-offending, as well on other factors like the defendant's remorse and society's sense of just deserts. Depending on the criminal charge, sentencing options could include a fine, probation (the defendant goes free but must report to a probation officer), or detention time either in jail (more common for a misdemeanor charge) or prison (more common for a felony charge).

[3]For an example of such polling data, see Benenson Strategy Group (2017), a survey done for the American Civil Liberties Union.

The *parole decision* arises because historically most defendants sentenced to prison would receive an indeterminate sentence; for example, it might be from four to seven years. After the inmate had served the minimum term, a parole board would then decide when exactly an inmate would go free. Inmates out on parole would typically be required to report periodically to a parole officer. Criteria for parole decisions are similar to those for sentencing but can make use of information about the defendant's behavior in prison as well. The role of parole boards declined in the 1970s with the shift towards determinate sentencing (Kuziemko 2013).

We refer to these three decisions as "judicial decisions" for convenience, recognizing that in practice other criminal justice actors also play a role. Prosecutors, for instance, make recommendations to judges as part of both pre-trial release and sentencing decisions. Prosecutors play a particularly important role for sentencing, given that 90–95 percent of all convictions result from a plea bargain (Devers 2011). The quality of legal representation that defendants receive can vary enormously. For the cases that do make it to trial, a jury may play a role in sentencing. Also, as noted above, while judges often set sentences over a certain range, parole decisions are usually made by a parole board staffed by people who are typically not judges.

The literature has identified three problematic aspects of how decisions are made: for a selective review of some prominent studies, see Table 2. One long-standing concern with these judicial decisions is misprediction: not simply that there are inevitable errors, but that predictions made by judges are systematically mistaken. For example, in the case of sentencing, Gottfredson (1999) asked judges in Essex County, New Jersey in 1977–78 to record their subjective predictions about the recidivism risk of 960 defendants. The correlation between judge predictions and actual recidivism outcomes 20 years later is very modest, on the order of 0.2. These low levels of predictive accuracy also jibe with data from pre-trial release (for example, Jung et al. 2017; Kleinberg et al. 2018a). Concerns with the accuracy of recidivism predictions for parole, which historically have often been made by psychiatrists, dates back at least to the 1940s (Jenkins et al. 1942; Schuessler 1954). More recently, Kuziemko (2013) finds some positive correlation between predictions made by parole boards and recidivism, but Berk (2017) shows there is also substantial misprediction.

Second, judicial decisions are inconsistent in several ways. One way is that they are inconsistent with each other. For example, some judges tend to be "tough" and others "lenient." This discretion was long justified on the basis that judges could then account for the circumstances of each case (Alschuler 1991). But the data shows that even with randomly assigned caseloads, the average level of leniency varies dramatically. Kling (2006), among others, documents this for sentencing, while for pre-trial release decisions, the difference in pre-trial release rates between the most- and least-lenient quintile of judges in New York City was nearly 25 percentage points (Kleinberg et al. 2018a). As one judge complained, "[I]t is obviously repugnant to one's sense of justice that the judgment meted out to an offender should depend in large part on a purely fortuitous circumstance; namely, the personality of the particular judge before whom the case happens to come for disposition" (Diamond and

Table 2
Selected Studies of Inconsistency, Error, and Discrimination in the Criminal Justice System

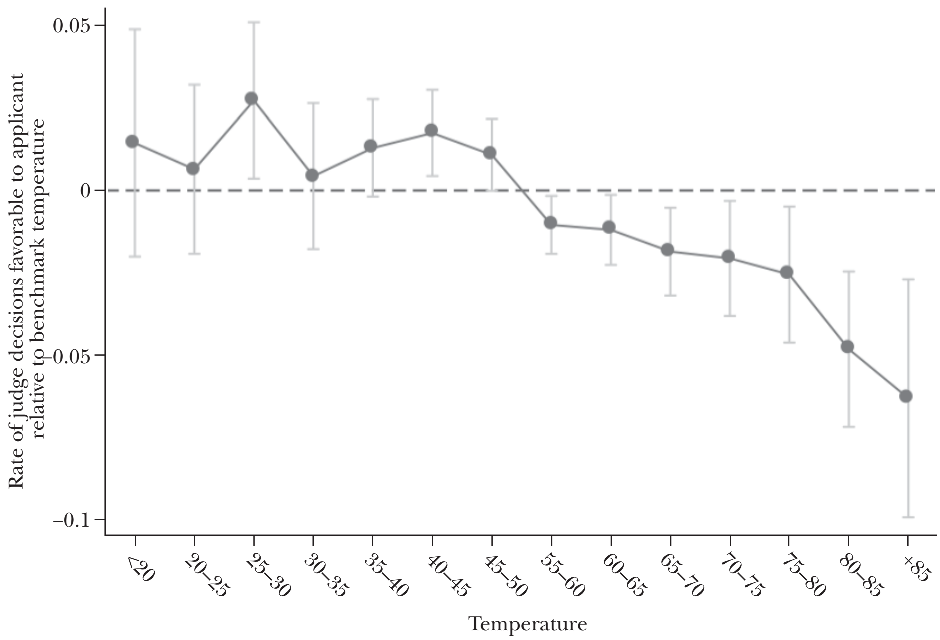
	Error (misprediction)	*Inconsistency*	*Discrimination*
Pre-trial decisions	Judge predictions (implicit in their release decisions) disagree with algorithmic predictions in ways that suggest misprediction (Jung et al. 2017; Kleinberg 2018a)	Judges seeing similar caseloads differ widely in pre-trial decisions, and also deviate from their central tendencies in ways that reduce decision quality (Kleinberg et al. 2018a)	Judges discriminate against minority defendants in an "outcome test" analysis (Arnold, Dobbie, and Yang 2018; Arnold, Dobbie and Hull 2020)
Sentencing decisions	Correlation of judge recidivism predictions with observed recidivism outcomes 20 years later is modest (Gottfredson 1999)	Judges seeing similar case-loads differ widely in prison sentencing (Kling 2006), and are also influenced by irrelevant factors like recent sports-team losses (Eren and Mocan 2018), heat (Heyes and Saberian 2019), or the features of recent cases (Chen, Moskowitz, and Shue 2016)	Judges assign longer sentences to observationally similar minority defendants compared to Whites (Kennedy 2001; Loury 2008; Blumstein 2015)
Parole decisions	Psychiatrist assessment of parolee risk adds little signal beyond structured data (Jenkins et al. 1942); parole predictions disagree with algorithmic predictions in ways that suggest misprediction (Berk 2017)	Parole members pay attention to very different sources of information in their decisions (Ruhland 2020)	Parole may be one of the few parts of the system without substantial racial bias (Anwar and Fang 2015; Mechouan and Sahuguet 2015)

Zeisel 1975, p. 111). Sentencing guidelines were introduced in the 1970s partly to address this problem. But they may have simply shifted discretion to the decisions made by prosecutors about what specific crimes will be charged and what plea-bargain deals will be made (Davis 2005), and of course sentencing guidelines have no effect on pre-trial decisions. For parole decisions, Ruhland (2020) shows parole board members pay attention to very different types of information about a case.[4]

Judges not only differ from each other; they also differ from themselves: the same judge can decide differently on the same case from day to day. For example, Eren and Mocan (2018) show how irrelevant circumstances can skew decisions: Upset losses by the Louisiana State University football team increase the sentences Louisiana judges hand out by about 6 percent—and the effect is larger for judges who are LSU alumni. Heyes and Saberian (2019) show that a 10-degree increase in outdoor temperatures reduces the likelihood an immigration judge rules in favor of

[4] As Kahneman, Sibony, and Sunstein (2021) point out, a more subtle version of across-person inconsistency is when some judges are relatively more lenient on cases of type A and more harsh on cases of type B, while other judges have the reverse pattern.

Figure 1
Variation in Immigration Judge Decisions Favorable to Defendant by Outdoor Temperature



Source: Heyes and Saberian (2019).

an applicant by 7 percent, as shown in Figure 1. Chen, Moskowitz, and Shue (2016) show that judge decisions in a case depend on the features of other cases the judge just heard. Kleinberg et al. (2018a) present evidence for inconsistency in judicial decisions around pre-trial release. For a defendant, what will happen to you depends on the happenstance of which judge you see and when you happen to see them.

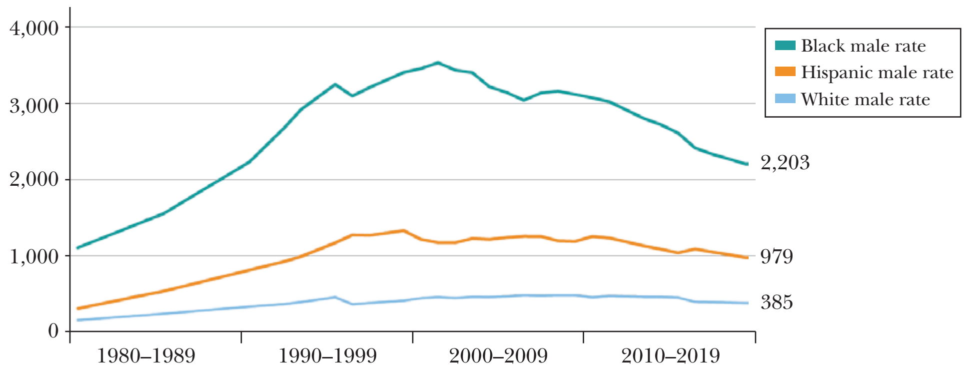
Finally, there are striking racial disparities. For example, African Americans make up 13 percent of the US population, but 26 percent of those who get arrested and 33 percent of those in state prisons.[5] Although disparities in imprisonment have been declining in recent years, they remain substantial (as shown in Figure 2). While disentangling exactly how much of the overall disparity is due to discrimination by the criminal justice system itself is a challenging task, there is little question that some of it is.

As one example of this evidence, Arnold, Dobbie, and Yang (2018) capitalize on the fact that cases are as good as randomly assigned to judges and that judges have systematically different propensities to release defendants pre-trial. They conduct an

[5] Data on the general population is here: https://www.census.gov/quickfacts/fact/table/US/PST045219. Data on state and federal prison inmates is here: https://www.bjs.gov/content/pub/pdf/p19.pdf.

Figure 2

Trends in Incarceration Rates per 100,000 People, by Race and Ethnicity, United States



Source: Bureau of Justice Statistics data analyzed by Statista. https://www.statista.com/chart/18376/us-incarceration-rates-by-sex-and-race-ethnic-origin.

"outcomes test" for marginal defendants. If judges were unbiased, we would expect to see similar re-arrest rates for White and Black defendants with similar probabilities of release. Yet re-arrest rates are lower for Black than White defendants, consistent with judges holding Black defendants to a higher standard. Arnold, Dobbie, and Hull (2020) suggest that around two-thirds of the Black-White disparity in release rates appears to be due to racial discrimination, with statistical discrimination also playing a role. Parole, in contrast, may be one of the few parts of the system where we do not consistently see evidence of substantial racial discrimination (Anwar and Fang 2015; Mechoulan and Sahuguet 2015). For reviews of the larger literature on discrimination in sentencing and many other parts of the justice system, useful starting points include Kennedy (2001), Loury (2008), and Blumstein (2015).

In short, a considerable body of evidence suggests that the criminal justice system is often inconsistent, error-prone, and discriminatory.

The Promise of Artificial Intelligence for Criminal Justice

The limits of human cognition have motivated interest in statistical methods of prediction; in the criminal justice system, "supervised learning" algorithms have become the dominant form of artificial intelligence used. Though the details of building these algorithms can be arcane, they are in essence quite simple. The problem they solve is simple and familiar: given *x*, predict *y* (called the "label"). The goal is to look at previous data and form a rule that can be deployed to new situations where *x* is known, but *y* is not. Forming those predictions, though, requires large datasets of so-called "labelled observations," where both x and y are available. It is worth noting that every machine learning algorithm is actually two algorithms. The "prediction algorithm" takes as input x and predicts y. It is produced by the "training algorithm," which takes as inputs an entire dataset of (x, y) pairs. In addition, the training algorithms needs an exact objective function specified: more specifically, what is the loss in predicting y incorrectly?

Stated this way, it is clear that familiar economic tools can be viewed as forms of "machine learning." For example, linear regression is one way to predict y from x. The least squares fitting algorithm is, in this case, the "training algorithm" and the "predictor algorithm" is the code which takes the inputs x and multiplies each input by the estimated coefficient. One thing that is new about the current machine learning tools is that they can work with far more complex functional forms and inputs: methodologies like random forests, gradient boosted trees, and neural networks are all examples of non-parametric functional forms which the training algorithm "learns" from the data. Importantly, these tools can also take as a result very novel forms of input: x can be images, audio files, or even video. In this journal, Mullainathan and Spiess (2017) provide an introduction to how machine learning fits in the econometric toolbox.

Importantly, machine learning algorithms fit these complex functions without pre-specification of a functional form by the analyst and while avoiding "over-fitting." A function that fits a specific given dataset as well as possible will inevitably learn more than the general relationship between x and y: it will also be based on statistical noise that is idiosyncratic to that dataset (the overfitting problem), which will, in turn, lead the prediction function to perform poorly on new data. To avoid this problem, these algorithms use sample-splitting techniques where one partition of the data is used for training and model-selection and another for evaluation, ensuring that whatever function is found works well *out*-of-sample.

A well-developed framework in computer science has emerged for building and applying supervised learning algorithms. This framework has enabled breakthroughs in areas like web search, manufacturing, robotics, customer service, automobile safety, and translation. The potential of statistical prediction has only increased over time with the growing availability of "big data" and development of new tools from the artificial intelligence field of machine learning. For excellent reviews at different levels of technical detail, see Athey and Imbens (2019), Berk (2008, 2018), Hastie, Tibshirani, and Friedman (2009), and Jordan and Mitchell (2015), as well as Varian (2014) in this journal.

Building these algorithms requires making key decisions: what outcome to predict, what candidate predictors to make available to the algorithm, and what objective function to provide. For pre-trial release, the relevant outcome is usually guided by state law, usually public safety risk (measured by re-arrest, or measured by re-arrest for violence specifically) and/or flight risk (skipping a required future court case). Typical algorithms used for sentencing and parole instead focus more narrowly on some sort of re-arrest or recidivism risk. Most of these algorithms then use as candidate predictors some combination of the criminal charge for which the person is currently in the justice system, prior criminal record, and a narrow set of demographic factors (usually age, which is legally allowed and predictive of risk given the strong age patterning of criminal behavior, and sometimes gender). Some algorithms can also include factors like employment or some proxy for "community ties" like duration of residence in the area.

These tools differ in important ways in terms of the construction of the functional form that relates the candidate predictors to the outcome of interest. For example, the COMPAS tool that is used for predicting risk of recidivism—and which was the focus of a widely read *Pro Publica* article (Angwin et al. 2016)—is billed as an "evidence-based software product" (http://www.northpointeinc.com/files/downloads/Risk-Needs-Assessment.pdf). But COMPAS is not actually a machine learning tool at all; it seems to be driven instead, as Rudin, Wang, and Coker (2020) note, in large part by human judgments, "a product of years of painstaking theoretical and empirical sociological study" (p. 5). The widely used Public Safety Assessment (PSA) developed by Arnold Ventures for pre-trial release decisions uses a logistic regression to determine the coefficients (weights) that each predictor should get. The tool that the current paper's authors helped develop for use in New York City estimated the relationship between the predictors and the outcome with machine learning, but presents the predictor algorithm to the user as a linear weighted average of predictors to help with interpretability (see also Rudin et al. 2021).

The final ingredient of any algorithm deployed in the criminal justice system is how the results are presented to end-users. Most algorithms map the predictions from the algorithm into recommendations for the final (human) decider. This mapping, typically known as a "decision-making framework," requires making some normative policy judgments about where the right risk thresholds should be to recommend one outcome versus another (like the choice of release versus detain in the pre-trial setting). In practice, those judgments are sometimes made by the algorithm builders alone, sometimes by government agencies, and sometimes through a collaboration. Another question is whether to give the end-user just the recommendations or also the underlying risk predictions, which in principle could help humans learn the algorithm's "confidence" in the recommendation of its decision-making framework (for example, whether a defendant's risk is far from or close to a decision threshold).

These supervised learning algorithms have the potential to improve on human prediction by, for starters, being more accurate. Decades of psychology research show that statistical models, on average, predict more accurately than human beings can in a range of applications (Meehl 1954; Dawes et al. 1989; Grove et al. 2000; Salzinger 2005). That advantage might be even greater today in criminal justice given new supervised learning methods, which allow for increasingly accurate prediction together with the growing availability of larger and larger datasets, which allow for the construction of increasingly accurate algorithms.

Because the predictor algorithm is mechanical, it is necessarily consistent (in the plain English sense of the term). Inputting a given set of predictor-variable values into a predictor algorithm always outputs the same predicted risk. If the human judges and other relevant actors pay attention to the algorithm, there is the potential for an overall reduction in inconsistency (that is, variability in decisions for similar cases) within the justice system.

Finally, statistical models, unlike humans, themselves do not have intrinsic "in-group" preferences—although they can readily acquire such patterns in the training process. What the statistical models learn is a consequence of the training process. As we discuss below, depending on how they are built, their predictions can either mirror historical patterns of discrimination or can undo them.

The prevalence of algorithms within the justice system is hard to determine precisely. This is because data collection and reporting about *anything* in America's justice system is mostly voluntary—even about basic crime statistics—leading to a very underdeveloped criminal justice data infrastructure (Bach and Travis 2021). For pre-trial release, some specific algorithm providers like Arnold Ventures voluntarily share information about use of their tools (at https://advancingpretrial.org/psa/psa-sites). The Arnold tool is used statewide in Arizona, Kentucky, New Jersey, and Utah, in cities like Chicago, Cleveland, Houston, New Orleans, and Pittsburgh, and in a number of suburban and rural counties as well, jurisdictions that together are home to over 66 million people. For sentencing, Stevenson and Doleac (2021) report that algorithms are used in sentencing decisions in a politically, geographically, and demographically diverse set of 28 states; another seven states have at least one county using a risk tool for sentencing. Most states seem to have adopted decision guidelines for parole decisions, if not formal machine learning algorithms, by the mid-1980s (Glaser 1985).

The Disappointing Record of Artificial Intelligence in US Criminal Justice

Against a clear track record of human fallibility and error in the existing criminal justice system, algorithms may seem to offer some hope of improvement. Things have not turned out that way. Supervised learning algorithms and other statistical models in the US criminal justice system have often not only failed to redress problems, they've often created new ones.

The literature analyzing algorithms has focused heavily on documenting racial bias. For example, the widely read *Pro Publica* analysis of the COMPAS risk tool found the tool has a higher false-positive rate in predicting recidivism for Black than White defendants (Angwin et al. 2016). While subsequent research noted the limitations of that specific measure of algorithmic bias,[6] we see examples of algorithms

[6] It is not possible to have both calibration and similar false positive rates with any prediction method (human or algorithmic) in a situation where two groups have different "base rates" for the underlying outcome, unless the prediction method predicts perfectly (Kleinberg, Mullainathan, and Raghavan 2016; Chouldechova 2017).

violating other common definitions of algorithmic fairness as well. For example, *calibration* refers to whether the actual outcomes people experience differ for majority versus protected group members, conditional on the algorithm's risk predictions. This test is complicated by the fact that, for example, we don't observe outcomes for pre-trial defendants who get detained (a point we return to below). With that caveat in mind, we see gender bias in the COMPAS tool (Hamilton 2019). We also see miscalibration by race in the Arnold Ventures Public Safety Assessment, in states like Kentucky, in ways that sometimes create advantages for White defendants (higher crime rate for White than Black defendants at a given risk prediction) and sometimes for Black defendants, as shown in Figure 3. These findings are consistent with evidence of bias in other parts of the justice system that shape the data used by the algorithm, such as police decisions (Fryer 2020; Goncalves and Mello 2021; Hoekstra and Sloan 2020) and jury decisions (Anwar, Bayer, and Hjalmarsson 2012), and consistent with evidence for algorithmic bias in other domains like health (Obermeyer et al. 2019).

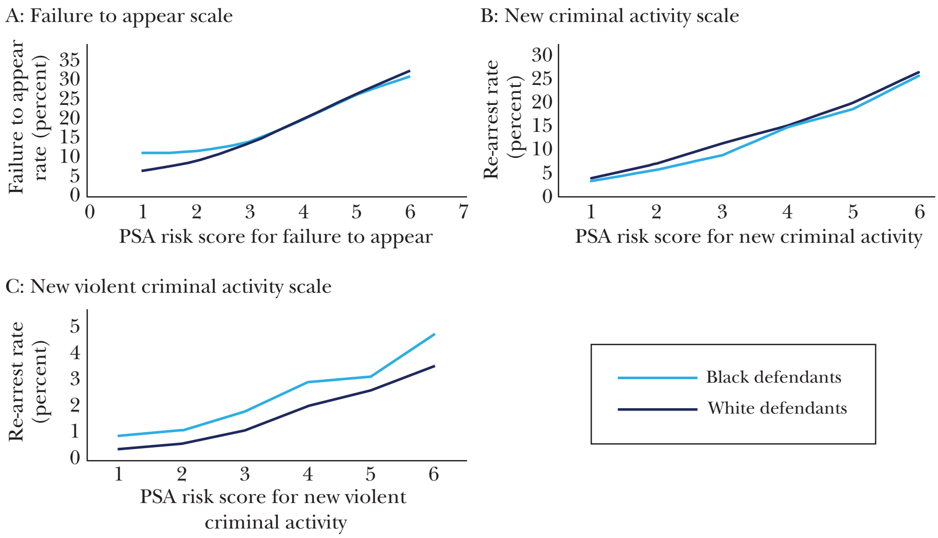
Moreover, many of the algorithms that are deployed are either no more accurate than humans or simply have no effect on actual criminal justice outcomes. One review of 19 risk tools used in correctional facilities found them "moderate at best in terms of predictive validity" (Desmarais and Singh 2013; see also Berk 2019). We also see examples where within a few years of adopting algorithms, decisions revert back to the same patterns as before (Stevenson 2018) or fail to meet the objectives policymakers had initially laid out (Stevenson and Doleac 2019) like reduced pre-trial detention.

Finally, the adoption of algorithms has also introduced new problems into the criminal justice system, such as limited transparency and concerns about due process. A core value of the American constitutional system, enshrined in the Sixth Amendment to the US Constitution, is the defendant's right to face and confront one's accuser to probe and debate the veracity of the accusations. But many algorithms are not made public, so the defense is deprived of this ability. The Sixth Amendment's "confrontation clause," which was designed reasonably well for the eighteenth century, is severely stretched in the twenty-first. The inability to understand what is happening and why also raises natural concerns about whether the system is treating people in a depersonalized way that compromises Fifth and Fourteenth Amendment due process protections.

Why is Artificial Intelligence Problematic in Practice?

Why have risk tools in the criminal justice system been so disappointing in practice relative to the hoped-for initial promise? The problem is frequently viewed as intrinsic to the machine learning enterprise. Surveys regularly show that the public has a dim view of not just current algorithms, but about their potential to *ever* be useful. For example, one Pew survey found that 58 percent of American adults believe algorithms will inevitably be biased (Smith 2018). Of course many people

Figure 3
Evidence of Mis-calibration in the Public Safety Assessment in Kentucky



Source: DeMichele et al. (2018).

recognize that the alternative to the algorithm—human judgment—can also be biased. So it is revealing that 56 percent of people said in the same survey that they find it "unacceptable" to use algorithms for criminal justice applications like parole. (Majorities also oppose use of algorithms for applications like hiring or credit scoring.) This view is common among experts, too. For example, as one researcher put it: "There's no way to square the circle there, taking the bias out of the system by using data generated by a system shot through with racial bias" (as quoted in Schwartzapfel 2019). Harcourt (2015, p. 237) argued risk tools will "unquestionably aggravate the already intolerable racial imbalance in our prison populations." Similar concerns show up for other problems like accuracy. For example, the belief that reality is easily approximated by a simple combination of one or two factors leads Dressel and Farid (2018) to "cast significant doubt on the entire effort of algorithmic recidivism prediction."

While there are legitimate arguments here, we argue that the overarching reason algorithms perform poorly in practice in the criminal justice system lies elsewhere: many algorithms have been poorly built. Algorithm design, as noted, requires a set of choices and the outcome is highly sensitive to these choices. This creates a fragile process: mistakes in design can lead to consequential errors of the kind we have seen.

The mistakes are, perhaps first and foremost, basic technical ones that arise in working with messy data generated by past human decisions. If econometrics = statistics + human agency, most algorithms are built not through an econometric approach but through a statistical one that ignores key aspects of this messiness. The development of these tools also too often ignores a key sociological challenge: Algorithms don't make decisions, people do. Regulatory failures provide the underlying reason for the persistence of both of these types of technical failures: No safeguards are in place currently to stop inadequately built algorithms from being deployed.

Badly Built Algorithms

Many algorithms cause harm because they have not been constructed to solve two types of econometric challenges. The first is the potential for misalignment between algorithmic objectives and human decision-maker objectives, a problem that is rampant in criminal justice and also shows up in many other areas as well such as child welfare screening (Coston, Chouldechova, and Kennedy 2020). The second is that the data we have are filtered by past decisions of humans who may see things about cases that are not captured in the data.

Nearly all machine learning algorithms simply predict outcomes. But a judicial decision often depends on more—sometimes much more—than the prediction of a single outcome. By assuming that all that matters is the outcome being predicted, an algorithm can wind up leaving out many of the elements the decision-maker cares about. We call that problem *omitted payoff bias* (Kleinberg et al. 2018a).

To see the problem, note that artificial intelligence tools are regularly built for all three judicial decisions we study here: pre-trial, sentencing, and parole. An implicit assumption is that prediction of an outcome like re-arrest or recidivism is equally useful in each case, but in fact, the role that prediction plays in the decision is quite different. For sentencing, for example, countless examples make clear that the objective function of real-world judges is richer than this; it can also include defendant circumstances, personal culpability, remorse, and society's sense of just deserts. Thus, decision-makers are receiving predictions only for a subset of what matters for their decision, creating risk of distorting the decision outcome. In contrast, pre-trial release decisions are supposed to depend on a narrower set of criteria: the judge's prediction of the defendant's flight or public safety risk. A recidivism predictor is better suited for pre-trial decisions than for sentencing because what the algorithm is specifically predicting is better aligned with the judge's objectives. This difference helps to explain why so much recent work on algorithms in the criminal justice system has been focused on the pre-trial release decision.[7] For an economist, an obvious way to address this problem is to inform the algorithm design with a model of the human decision-maker's actual objective function.

A related danger lies in mistakenly concluding the algorithm improves upon human-only decisions because it is better on one dimension, even if it ignores other dimensions the human decision-makers may care about. For example, in the case of pre-trial release tools, Kleinberg et al. (2018a) build an algorithm to predict failure

[7] For example, see Anderson et al. (2019); Angelino et al. (2017); Berk, Sorenson, and Barnes (2016); Corbett-Davies et al. (2017); Cowgill and Tucker (2017); Jung et al. (2017); Kleinberg et al. (2018a); Stevenson (2018); Jung, Goel, and Skeem (2020); and Wang et al. (2020).

to appear in court, the outcome which local law in New York state says should be the focus of risk considerations in pre-trial decisions. But they then show that the algorithm also dominates judge decisions on *other* outcomes that could potentially enter the judge's objective function in practice like re-arrest risk, risk of any violence, or serious violence specifically or, as discussed further below, racial disparities. However, this sort of comprehensive assessment of multiple objectives is not yet standard practice in the field.

The second econometric complication that arises in constructing algorithms also stems from the basic fact that there is a human in the loop in criminal justice applications, namely: we are selectively missing some of the data we need to evaluate the algorithms. Conceptually, the problem is to compare status quo decision-making with the decisions that would happen with an algorithm in the decision loop. However, the data available to estimate that counterfactual are generated by past judicial decisions. If the algorithm recommends pre-trial detention of a defendant that past judges had detained pre-trial, constructing counterfactual arrests is easy because the number of crimes committed before trial is, by construction, zero in both cases. But if the algorithm recommends release of a defendant that judges detained, what do we do? We are missing a measure of the defendant's behavior if released in that case.[8]

We cannot simply impute the missing outcomes for these defendants by looking at the outcomes of released defendants who appear similar on measured variables. There may be cases where judges make inconsistent decisions, as we discussed earlier, but in addition, the judge may have access to information not captured by the algorithm. The presence of *unobserved* variables means two observations we *think* are comparable based on our data may not actually be comparable. We call this the *selective labels problem* (Kleinberg et al. 2018a).

Econometrics has tools to address this issue. For example, economists studying pre-trial judicial decisions have used the fact that judges vary in their leniency rates, and that sometimes defendants are as good as randomly assigned to judges (Kleinberg et al. 2018a; Arnold, Dobbie, and Hull 2020, 2021; Rambachan and Roth 2020; Rambachan 2021). For example, we can take the caseload of more lenient judges, use the algorithm to select the marginal defendants to detain to get down to the detention rate of stricter judges, then compare the observed crime under the simulated lenient judges plus algorithm detention rule compared to the observed stricter judges' crime rate. This allows us to evaluate the algorithm's performance, focusing only on the part of the counterfactual estimation problem (contracting or shrinking the released set) where the selective labels problem is not binding. This sort of exercise confirms that algorithms are indeed able to predict risk much more accurately than can human judges.

[8] There is another problem here from missing outcome data for defendants the judge detains, which is that we can only build an algorithm using data from released defendants. That problem could reduce accuracy of the algorithm when applied to the full set of defendants who come in for pre-trial or sentencing hearings in the future. That possibility just further reinforces the importance of being able to solve the evaluation challenge mentioned above to determine whether the built algorithm really could predict more accurately than do judges.

This selective-labels problem shows up in any situation where the human decision affects the availability of the label, such as hiring where we only observe performance on the job for the employees that a firm decided to hire (for example, Hoffman, Kahn, and Li 2018). In criminal justice applications, it shows up repeatedly and sometimes in different forms. For example, in predictive policing the crime data we have for evaluating the potential performance of any new algorithm are generated by past policing decisions about where to deploy resources. In this approach, the outcome values are contaminated by the treatment effects caused by past decisions (Mullainathan and Obermeyer 2017). The evaluation tools provided by econometrics have not yet diffused into standard operating practice for algorithms before they are deployed at scale.

Human Plus Machine Challenges

A second type of technical challenge that real-world algorithms often fail to address adequately stems from the fact that the *algorithm* does not decide. Humans remain in the loop as the ultimate decision-makers. Thus, any successful algorithmically informed system will need to not only design the algorithm correctly, but also understand and allow for how humans use these algorithms in practice.

A common approach is to assume that because algorithmic predictions can be more accurate than those of humans *on average*, the goal should just be to get the human to follow the algorithm's recommendations as often as possible. The assumption that the algorithm is (almost) always right is reflected in the increasingly common term "algorithm aversion"—the behavioral science description for people's reluctance to always follow the recommendation of a prediction tool (Dietvorst, Simmons, and Massey 2015). Similarly, when economists and others have focused on evaluating deployments of artificial intelligence in criminal justice, they often focus on the "problem" of the human not following the algorithm enough.

But simply getting the human to mindlessly follow the algorithm as often as possible is not the right goal, not only because few humans will love the idea of effectively being replaced like this, but also because it need not be the social welfare-maximizing approach. While an algorithm does indeed have an advantage over humans in being able to access a large number of administrative data (a "longer" dataset) to form predictions, humans often have access to data the algorithm does not (a "wider" dataset). This raises the possibility that at least in some cases the human can have an advantage over the algorithm (for example, De-Arteaga, Fogliato, and Chouldechova 2020). Determining when the human should follow the algorithm's prediction, or not, is what we call the *override problem.*

Consider a situation with two sources of information for making a decision about pre-trial release: information observable to both the algorithm and the judge, and information unobservable by the algorithm but observable by the judge. In this setting, consider two possible scenarios that might arise. In the first scenario, the judge using the additional information always estimates more accurately, which in some cases leads to correcting errors that would have been made by the algorithm.

That is, when the algorithm and the judge disagree, the judge is correct to override the algorithm—if the algorithm had the additional information, it would agree with the judge's decision. In the second scenario, the judge uses the additional information in a way that always leads to an incorrect decision: that is, if the algorithm had full information on not just its usually observed data but also the unobserved information usually seen just by the judge, it would still disagree with the judge. In this scenario, when the algorithm and the judge disagree, the algorithm is correct even based on limited information—because the judge draws the wrong inference from fuller information.

Solving the override problem raises new frontier-science challenges that the omitted payoffs and selective labels problems typically do not. The deep problem that has not yet been fully figured out is to understand the contexts in which humans and machines working together might do better than either alone (for example, Salzinger 2005; Jussupow, Benbasat, and Heinzl 2020). Solving the override problem requires not just helping judges use their information as well as possible, but also helping them learn where they have comparative advantage over the algorithm and vice versa. That, in turn, requires figuring out ways of helping judges better understand the algorithm, a focus of computer science work on interpretable algorithms.

It's worth noting that what it even means for something to be interpretable as "an explanation" is unclear. Psychology shows that people find even vacuous explanations acceptable if they simply begin with the word "because." For example, Langer, Blank, and Chanowitz (1978) show that study subjects are more likely to let someone cut in line in front of them at the photocopy machine when the person offers a reason ("because I'm in a hurry") than when they don't. But they're equally likely to let someone cut in line with a real reason as with the vacuous veneer of a reason ("because I need to make copies"). Identifying ways of communicating the process and recommendations of artificial intelligence to humans is as much about understanding the human as it is about the algorithm. More fundamentally, given the importance of due process, solving this problem is essential: when a person is detained or imprisoned based in part on an algorithm's recommendations, "it's a complicated black box" is not an acceptable answer for why.

The fact that algorithms often fail in criminal justice because of the behavior of the human users rather than the artificial intelligence technology itself means that social science will inevitably have an important role to play in solving these problems. Progress on these issues will require creativity in data collection of the sort at which applied economists have become adept, combined with the ability of artificial intelligence methods to make use of unstructured data sources that may help capture the sources of the judge's private signal such as text (courtroom transcripts) or images (perhaps use of video from the courtroom).

Evidence that progress on these human plus machine challenges is possible comes from the progress that fields other than criminal justice have made. For example, to help radiologists detect breast cancer from mammograms Jiang et al. (1999) not only built an algorithm but designed a user interface that presented the doctors with the information in ways they are used to seeing, which in turn, improved diagnostic outcomes. Tschandl et al. (2020) tested multiple user interfaces for the algorithm and came up with an algorithm-human combination that leads to better diagnosis than either the algorithm or the doctor alone (also, see the review in Doi 2007). The fact that this type of progress shows up in medicine, but not in criminal justice, is no accident—as we discuss next.

Inadequate Procurement and Regulation

Why have so many real-world algorithms failed to deal with problems like omitted payoff bias, the selective labels problem, and the override problem? The answer, in short, is that they have not had to. The parties involved in building and deploying algorithms lack either the information or motivation needed to solve those problems, and there are no corrective mechanisms to prevent the flawed algorithms that result from being deployed widely.

Part of the problem is that algorithms used by criminal justice agencies are often not built by those agencies. Vendors can often have asymmetric information with regard to buyers, as well as potentially divergent interests—ideas that are very familiar to economists. With algorithms in the criminal justice system specifically, the vendors often have incorrectly specified the problem to be solved. For example, the allocation of social programs for those in the criminal justice system is often guided by algorithms that predict risk of crime involvement (a standard predictive-inference problem) rather than by the predicted benefit from intervention (a causal-inference problem). Even if the problem is correctly specified, the algorithm's ability to achieve that goal is unclear because few algorithms are properly evaluated prior to deployment. But the buyers don't have the ability to tell, and the result will be a system that does not perform as hoped.

We often rely on regulation to deal with underinformed consumers, but (as is often the case with new technologies) the law and larger regulatory apparatus is still catching up to the ways in which artificial intelligence can cause harm. For example, in health, the Food and Drug Administration requires new medicines or medical devices to be rigorously evaluated through a series of randomized controlled trials before they are deployed. No similar requirement currently exists for algorithms.

The limitations of current algorithmic regulations are not limited to procurement. For example, current discrimination laws are designed to deal with human bias, but fail to deal with how algorithms discriminate (Kleinberg et al. 2018c). Discrimination law for humans focuses on ensuring that people don't pay attention to protected group characteristics. The human brain is the ultimate black box, so we can't tell when a person would use such characteristics to enhance versus detract from accepted societal goals. In contrast, as we discuss further below, for algorithms the use of protected group characteristics can actually help undo bias (Dwork et al. 2012; Kleinberg et al. 2018b; Goel et al. 2021). Discrimination law built for humans is silent on what we outside observers need to monitor algorithms for bias, such as access to data and the predictor algorithm for "fairness audits" and improved transparency in general (Rudin, Wang, Coker 2020).

Algorithmic Bias

Much of the public debate around algorithms explicitly or implicitly assumes that their problems are intrinsic to the underlying technology. Our argument instead is that the problems with algorithms stem not from something intrinsic to artificial intelligence but instead from human decisions about how to construct, evaluate, deploy, and regulate these tools, as shown in Table 3. Indeed, we argue that there are principled ways to address the problems with these underlying human decisions. To illustrate that argument, we consider the case of algorithmic bias.

Under our framework, algorithmic bias is largely an example of omitted payoff bias. Society has a strong social preference for fairness (as well as predictive accuracy). Yet the algorithm builder may ignore this preference and focus only on predictive accuracy. As a result, the wrong data can be used (for example, an algorithm that predicts an outcome like arrests for low-level offenses where officer discretion is high, hence risk of bias is high); tools are evaluated using the wrong outcome criteria (for example, by accuracy alone versus a comparison along multiple dimensions that includes fairness as one); or how the algorithm's output is presented to the judge (for example, if many other factors matter to the judge, providing recommendations rather than the specific narrow predictions can be misleading).

In contrast, once fairness objectives are recognized, they can be incorporated. Concerns about bias in data can lead the algorithm builder to focus on using data on more serious rather than less serious offenses, if discretion (and hence bias) is attenuated with the former, or focusing on convictions over arrests. Different machine learning models can have similar rates of overall predictive accuracy but differ in their predictions for specific cases (the so-called "Rashomon effect"), and so can lead to different implications for fairness objectives (for example, Coston, Rambachan, and Chouldechova 2021).

There are also additional design choices that could be made to improve algorithmic fairness, even if some of them are currently prohibited by laws designed to deal with how humans rather than algorithms discriminate (Kleinberg et al. 2018c; Goel et al. 2021). For example, allowing a properly built algorithm to access information about protected-group membership can help *undo* the effects of bias in the underlying data (Dwork et al. 2012; Kleinberg et al. 2018b). As an example, imagine that in some city, half of all arrests of minority residents are false arrests (the person did not actually commit a crime), while none of the arrests of White residents are. In that case, an algorithm blinded to group membership has no choice but to treat each arrest as equally informative about risk of flight or re-arrest. In contrast, an algorithm that knows a defendant's race or ethnicity has the potential to learn that arrests to minority residents contain less "signal" about future outcomes than do arrests to White residents and could estimate a different arrest-to-risk relationship for each group and so undo some of the bias baked into the underlying arrest data. A similar approach would involve setting different risk thresholds for release for different groups.

Not only is fairness too often ignored, the variability of fairness preferences are also ignored. After all, the most widely used risk tools were built for use in multiple

Table 3
Common Concerns with Algorithms as Explained by Our Framework

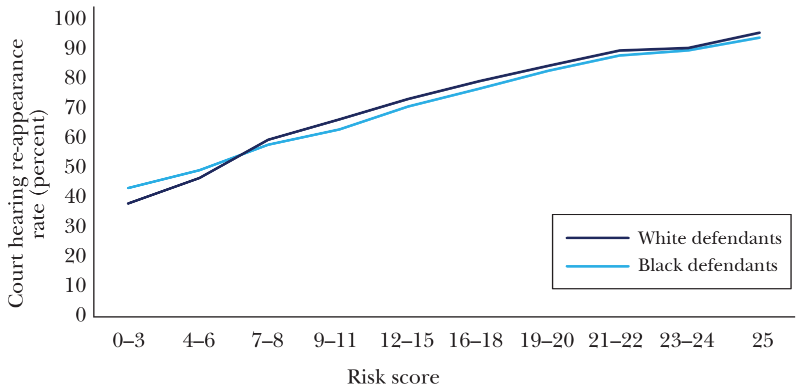
Concern	*Example of failure to solve technocratic problem (omitted payoffs, selective labels, override)*	*Example of regulation/procurement problem*
Ineffectiveness	Inaccurate algorithm mistakenly evaluated to be effective because of failure to deal with selective labels	Algorithm not required to be adequately evaluated before deployment
Transparency		Algorithm not made public because buyer and regulations did not require it
Due process/ depersonalization	Judges may override highly accurate algorithms in ways that reduce differentiation across defendants	Algorithms with low predictive accuracy fail to adequately distinguish among defendants in pre-trial release decisions
Fairness	Algorithm built without adequate attention to human decision-maker's equity objectives	Procurement of biased algorithms when unbiased algorithms for same purpose are available

jurisdictions; they were not designed to reflect the specific equity or other preferences of any particular place. Put differently, algorithms (unlike humans) come with "equity knobs"—the ability to make adjustments in response to the specific equity objectives of a given policymaker.

Proof-of-concept of what is possible from accounting for equity preferences comes from an algorithm to inform pre-trial release decisions in New York City that one of our research centers (the University of Chicago Crime Lab) helped construct. New York was one of the first places in the United States to implement a pre-trial risk tool back in the 1960s, as part of the Vera Institute of Justice's Manhattan Bail Project. The new tool that our team worked to develop with Luminosity and New York's Criminal Justice Agency was implemented in November 2019. The previous tool that had been in use since 2003 (!) showed signs of miscalibration by race. In contrast, the new tool that our team built meets the calibration test, as seen in Figure 4.

Perhaps even more important than the algorithm's statistical properties are its effects on decision outcomes, as shown in Table 4. The older tool recommended release for 32 percent of Black defendants and 41 percent of White defendants. New York City government set the new release thresholds based on estimates for how much higher the release rate could go without increasing failure-to-appear rates, where the possibility of increasing release without increasing the risk of failure-to-appear for a future court proceeding comes from better prioritizing the truly high risk for detention. As shown in Table 4, the new tool our team helped build recommends for release 83.9 percent of Black defendants and 83.5 percent for White defendants—a large absolute gain in release rates for both groups, and a reduction in the racial gap from nine percentage points down to effectively zero. That is, our new tool meets not only the calibration definition for algorithmic fairness, but even the more stringent (and more controversial) definition of "statistical

Figure 4
Calibration Test of New York City's New Release Assessment—Reappearance Rates by Predicted Risk Bin, by Race/Ethnicity



Source: Luminosity and University of Chicago Crime Lab (2020).

Table 4
Results from Algorithm for Pre-trial Release Decisions in New York City

	Release recommendations under old tool	*Release recommendations under new tool*	*Judge release decisions under new tool (2019–20 data)*
Black defendants	31.7%	83.9%	69.4%
White defendants	41.1%	83.5%	72.0%
Black-White gap	9.4 percentage points	0.4 percentage points	2.6 percentage points

Source: Peterson (2020). The new algorithmic tool was built by the University of Chicago Crime Lab in partnership with Luminosity and the NYC Criminal Justice Agency.

parity" (Hertweck, Heitz, and Loi 2021). To underscore the point that, at the end of the day, the justice system is more about the humans than the technology, Table 4 also shows what ultimately happened in practice when the tool was deployed: the human judges took release recommendations that were similar across race groups and turned them into a three-point gap in favor of Whites (Peterson 2020).

Our key point is that with the right motivations for the human algorithm builders and deployers, algorithms have the *potential* not only to avoid bias, but even to be a force for social justice. We see other examples in policing, for instance, where incorporating fairness objectives changed algorithmic outcomes for hiring decisions by the Los Angeles Police Department (Ridgeway 2013) and, according to evidence from a randomized trial, led to a predictive policing tool that helped reduce crime without increasing overall arrests or the racial composition of those arrested (Mohler et al. 2015; Brantingham, Valasik, and Mohler 2018). Examples of how to incorporate fairness objectives into algorithms, and examples of how doing so can lead to gains relative to the status quo, show up in many other domains of interest to economists as well, such as hiring (Li, Raymond, and Bergman 2020), lending (Bartlett et al. 2019), housing (Ross and Yinger 2002), and health (Obermeyer et al. 2019).

Racial bias provides a useful contrast between human and algorithmic decision-making. Discrimination by people is hard to discover (Charles and Guryan, 2011). Once found, it is hard to fix. As an example, intricate hiring audits are needed to uncover bias in resume screening, and even despite the widespread dissemination of those findings, little has changed over the last two decades (Bertrand and Mullainathan 2004; Kline, Rose, and Walters 2021). Algorithms can, with the right transparency measures, be more straightforwardly audited and adjusted. With the right motivations and regulations in place, algorithmic bias can be easier to find and fix than human bias (Kleinberg et. al. 2018c).

Conclusion

Very often the discussion of algorithms happens in a vacuum. For many social systems, including but not limited to criminal justice, we cannot understand the algorithms without understanding the human beings. Humans set the benchmark for algorithms through their existing decisions. Humans produce the data that the algorithm uses. Humans build and deploy the algorithm. Viewed this way, we can see that algorithms cannot be expected to be an automatic panacea for all the problems of our criminal justice system. Algorithms can be, and too often in practice are, deeply problematic.

But they need not be. Designed correctly, they offer a potential remedy for human fallibility. The challenge to overcome is that algorithms themselves are fragile, extremely sensitive to design choices. Those choices are made and the resulting algorithms are built, deployed, and procured by a social system riddled with the very problems we seek to address, a system that has been designed and implemented by fallible humans. These problems are complex but not hopeless. Economists and other social scientists have an important role to play in ensuring that algorithms do no harm and even do social good.

■ *Thanks to Amanda Agan, Charles Brown, Alexandra Chouldechova, Philip Cook, Amanda Coston, Dylan Fitzpatrick, Barry Friedman, Jonathan Guryan, Peter Hull, Erik Hurst, Sean Malinowski, Nina Pavcnik, Steve Ross, Cynthia Rudin, Greg Stoddard, Cass Sunstein, Timothy Taylor, Heidi Williams, and Morgan Williams for valuable comments, and to Kristen Bechtel, Megan Cordes, Ellen Dunn, Rowan Gledhill, and Elizabeth Rasich for their assistance. All opinions and any errors are of course our own. The authors gratefully acknowledge support for the construction of the New York City algorithm discussed in the paper from the non-profit Criminal Justice Agency to the University of Chicago Crime Lab, and for support for the writing of this paper from the Sloan Foundation and the Center for Applied Artificial Intelligence at the University of Chicago Booth School of Business.*

References

Alschuler, Albert W. 1991. "The Failure of Sentencing Guidelines: A Plea for Less Aggregation." *University of Chicago Law Review* 58 (3): 901–51.

Anderson, Chloe, Cindy Redcross, Erin Valentine, and Luke Miratrix. 2019. *Evaluation of Pretrial Justice System Reforms that Use the Public Safety Assessment: Effects of New Jersey's Criminal Justice Reform*. New York City: MDRC.

Angelino, Elaine, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. 2017. "Learning Certifiably Optimal Rule Lists for Categorical Data." Paper presented at KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, August 13–17.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias". *ProPublica*, May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Anwar, Shamena, Patrick Bayer, and Randi Hjalmarsson. 2012. "The Impact of Jury Race in Criminal Trials." *The Quarterly Journal of Economics* 127 (2): 1017–1055.

Anwar, Shamena, and Hanming Fang. 2015. "Testing for Racial Prejudice in the Parole Board Release Process: Theory and Evidence." *Journal of Legal Studies* 44 (1).

Arnold, David, Will S. Dobbie, and Peter Hull. 2020. "Measuring Racial Discrimination in Bail Decisions." NBER Working Paper 26999.

Arnold, David, Will Dobbie, and Peter Hull. 2021. "Measuring Racial Discrimination in Algorithms." *AEA Papers and Proceedings* 111: 49–54.

Arnold, David, Will Dobbie, and Crystal S. Yang. 2018. "Racial Bias in Bail Decisions." *Quarterly Journal of Economics* 133 (4):1885–1932.

Athey, Susan, and Guido W. Imbens. 2019. "Machine Learning Methods that Economists Should Know About." *Annual Review of Economics* 11: 685–725.

Bach, Amy, and Jeremy Travis. 2021. "Don't Ignore the Infrastructure of Criminal Justice." *The Hill*, August 16. https://thehill.com/blogs/congress-blog/politics/568063-dont-ignore-the-infrastructure-of-criminal-justice.

Barlett, Robert, Adair Morse, Richard Stanton and Nancy Wallace. 2019. "Consumer-Lending Discrimination in the Fintech Era." NBER Working Paper 25943.

Benenson Strategy Group. 2017. "2017. ACLU National Survey." https://www.aclu.org/report/smart-justice-campaign-polling-americans-attitudes-criminal-justice.

Berk, Richard A. 2008. *Statistical Learning from a Regression Perspective*, Vol. 4. New York: Springer.

Berk, Richard A. 2017. "An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism." *Journal of Experimental Criminology* 13: 193–216.

Berk, Richard A. 2018. *Machine Learning Risk Assessments in Criminal Justice Settings*. New York: Springer.

Berk, Richard A. 2019. "Accuracy and Fairness for Juvenile Justice Risk Assessments." *Journal of Empirical Legal Studies* 16 (1): 175–94.

Berk, Richard A., Susan B. Sorenson, and Geoffrey Barnes. 2016. "Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions." *Journal of Empirical Legal Studies* 13 (1): 94–115.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94 (4): 991–1013.

Blumstein, Alfred. 2015. "Racial Disproportionality in Prison." In *Race and Social Problems*, edited by Ralph Bangs and Larry E. Davis, 187–93. New York: Springer.

Brantingham, P. Jeffrey, Matthew Valasik, and George O. Mohler 2018. "Does Predictive Policing Lead to Biased Arrests? Results from a Randomized Controlled Trial." *Statistics and Public Policy* 5 (1).

Charles, Kerwin Kofi, and Jonathan Guryan. 2011. "Studying Discrimination: Fundamental Challenges and Recent Progress." *Annual Review of Economics* 3: 479–511.

Chen, Daniel L., Tobias J. Moskowitz, and Kelly Shue. 2016. "Decision Making under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires." *Quarterly Journal of Economics* 131 (3): 1181–1242.

Chouldechova, Alexandra. 2017. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *Big Data* 5 (2): 153–63.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. "Algorithmic Decision

Making and the Cost of Fairness." Paper presented at the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, August 13–17.

Coston, Amanda, Alexandra Chouldechova, and Edward H. Kennedy. 2020. "Counterfactual Risk Assessments, Evaluation, and Fairness." Paper presented at 2020 Conference on Fairness, Accountability and Transparency, Barcelona, January 27–30.

Coston, Amanda, Ashesh Rambachan, and Alexandra Chouldechova. 2021. "Characterizing Fairness over the Set of Good Models under Selective Labels." arXiv pre-print.

Cowgill, Bo, and Catherine Tucker. 2017. "Algorithmic Bias: A Counterfactual Perspective." https://bitlab.cas.msu.edu/trustworthy-algorithms/whitepapers/Bo%20Cowgill.pdf.

Davis, Angela J. 2005. "The Power and Discretion of the American Prosecutor." *Droit et Cultures* 49: 55–66.

Dawes, Robyn M., David Faust, and Paul E. Meehl. 1989. "Clinical versus Actuarial Judgement." *Science* 243 (4899): 1668–74.

De-Arteaga, Maria, Riccardo Fogliato, and Alexandra Chouldechova. 2020. "A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores." Paper presented at ACM Conference on Human Factors in Computing Systems, Honolulu, April 25–30.

Desmarais, Sarah L., and Jay P. Singh. 2013. "Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States." https://csgjusticecenter.org/wp-content/uploads/2020/02/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf.

Devers, Lindsey. 2011. *Plea and Change Bargaining*. Washington, DC: Bureau of Justice Assistance, US Department of Justice.

Diamond, Shari Seidman, and Hans Zeisel. 1975. "Sentencing Councils: A Study of Sentence Disparity and Its Reduction." *University of Chicago Law Review* 43 (1): 109–49.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology: General* 144 (1): 114–26.

DeMichele, Matthew, Peter Baumgartner, Michael Wenger, Kelle Barrick, and Megan Comfort, and Shilpi Misra. 2018. "The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky." http://dx.doi.org/10.2139/ssrn.3168452.

Doi, Kunio. 2007. "Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential." Computerized Medical Imaging and Graphics 31 (4–5): 198–211.

Dressel, Julia, and Hany Farid. 2018. "The Accuracy, Fairness, and Limits of Predicting Recidivism." Science Advances 4 (1).

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold and Richard Zemel. 2012. "Fairness through Awareness." Paper presented at 3rd Innovations in Theoretical Computer Science Conference, Cambridge, MA, January 8–10.

Eren, Ozkan, and Naci Mocan. 2018. "Emotional Judges and Unlucky Juveniles." *American Economic Journal: Applied Economics* 10 (3): 171–205.

Fryer, Roland G., Jr. 2020. "An Empirical Analysis of Racial Differences in Police Use of Force: A Response." *Journal of Political Economy* 128 (10): 4003–4008.

Glaser, Daniel. 1985. "Who Gets Probation and Parole: Case Study versus Actuarial Decision Making." *Crime and Delinquency* 31 (3): 367–78.

Goel, Sharad, Ravi Shroff, Jennifer Skeem, and Christopher Slobogin. 2021. "The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment." In *Research Handbook on Big Data Law*, edited by Roland Vogl. Cheltenham: Edward Elgar Publishing.

Goncalves, Felipe, and Steven Mello. 2021. "A Few Bad Apples? Racial Bias in Policing." *American Economic Review* 111 (5): 1406–41.

Gottfredson, Don M. 1999. *Choosing Punishments: Crime Control Effects of Sentences.* Washington, DC: US Department of Justice.

Grove, William, David Zald, Boyd Lebow, Beth Snitz, and Chad Nelson. 2000. "Clinical versus Mechanical Prediction: A Meta-Analysis." *Psychological Assessment* 12 (1): 19–30.

Hamilton, Melissa. 2019. "The Sexist Algorithm." *Behavioral Sciences & The Law* 37 (2): 145–57.

Harcourt, Bernard E. 2015. "Risk as a Proxy for Race: The Dangers of Risk Assessment." *Federal Sentencing Reporter* 27 (4): 237–43.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Hertweck, Corinna, Christoph Heitz, and Michele Loi. 2021. "On the Moral Justification of Statistical Parity." Paper presented at FAccT '21, Virtual Event, Canada, March 3–10.

Heyes, Anthony, and Soodeh Saberian. 2019. "Temperature and Decisions: Evidence from 207,000 Court Cases." *American Economic Journal: Applied Economics* 11 (2): 238–65.

Hoekstra, Mark, and CarlyWill Sloan. 2020. "Does Race Matter for Police Use of Force? Evidence from 911 Calls." NBER Working Paper 26774.

Hoffman, Mitchell, Lisa B. Kahn, and Danielle Li. 2018. "Discretion in Hiring." *Quarterly Journal of Economics* 133 (2): 765–800.

Jenkins, R.L., Henry Harper Hart, Philip I. Sperling, and Sidney Axelrad. 1942. "Prediction of Parole Success: Inclusion of Psychiatric Criteria." *Journal of Criminal Law and Criminology* 33: 38–46.

Jiang, Yulei, Robert Nishikawa, Robert Schmidt, Charles Metz, Maryellen Giger, and Kunio Doi. 1999. "Improving Breast Cancer Diagnosis with Computer-Aided Diagnosis." *Academic Radiology* 6 (1): 22–33.

Jordan, Michael I., and Tom M. Mitchell. 2015. "Machine learning: Trends, Perspectives, and Prospects." *Science* 349 (6245): 255–60.

Jung, Jongbin, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. 2017. "Simple Rules for Complex Decisions." *arXiv preprint arXiv:1702.04690.*

Lin, Zhiyuan, Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. "The Limits of Human Predictions of Recidivism." *Science Advances* 6 (7).

Jussupow, Ekaterina, Izak Benbasat, and Armin Heinzl. 2020. "Why Are We Averse towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion." Paper presented at the 28th European Conference on Information Systems, virtual event, June.

Kahneman, Daniel, Oliver Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgement.* New York City: Little, Brown, and Company.

Kennedy, Randall. 2001. "Racial Trends in the Administration of Criminal Justice." *America Becoming: Racial Trends and Their Consequences* 2: 1–20.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018a. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133 (1): 237–93.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018b. "Algorithmic Fairness." *AEA Papers and Proceedings* 108: 22–27.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. 2018c. "Discrimination in the Age of Algorithms." *Journal of Legal Analysis* 10: 113–74.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *arXiv preprint arXiv:1609.05807.*

Kline, Patrick M., Evan K. Rose, and Christopher R. Walters. 2021. "Systemic Discrimination among Large US Employers." NBER Working Paper 29053.

Kling, Jeffrey R. 2006. "Incarceration Length, Employment, and Earnings." *American Economic Review* 96 (3): 863–76.

Kuziemko, Ilyana. 2013. "How Should Inmates Be Released from Prison? An Assessment of Parole versus Fixed Sentence Regimes" *Quarterly Journal of Economics* 128 (1): 371–424.

Langer, Ellen J., Arthur Blank, and Benzion Chanowitz. 1978. "The Mindless of Ostensibly Thoughtful Action: The Role of 'Placebic' Information in Interpersonal Interaction." *Journal of Personality and Social Psychology* 36 (6): 635–42.

Li, Danielle, Lindsey R. Raymond, and Peter Bergman. 2020. "Hiring as Exploration." NBER Working Paper 27736.

Loury, Glenn C. 2008. *Race, Incarceration, and American Values.* Cambridge, MA: MIT Press.

Luminosity and University of Chicago Crime Lab. 2020. *Updating the New York City Criminal Justice Agency Release Assessment.* New York: Criminal Justice Agency.

Mechoulan, Stéphane, and Nicolas Sahuguet. 2015. "Assessing Racial Disparities in Parole Release." *Journal of Legal Studies* 44 (1).

Meehl, Paul E. 1954. *Clinical versus Statistical Prediction: A Theoretical Analysis and Review of the Evidence.* Minneapolis, MN: University of Minnesota Press.

Milgram, Anne, Alexander M. Holsinger, Marie Vannostrand, Matthew W. Alsdorf. 2014. "Pretrial Risk Assessment: Improving Public Safety and Fairness in Pretrial Decision Making." *Federal Sentencing Reporter* 27 (4): 216–21.

Mohler, George O., Martin B. Short, Sean Malinowski, Mark Johnson, George E. Tita, Andrea L. Bertozzi, and P. Jeffrey Brantingham. 2015. "Randomized Controlled Field Trials of Predictive

Policing." *Journal of the American Statistical Association* 110 (512): 1399–1411.

Mullainathan, Sendhil, and Ziad Obermeyer. 2017. "Does Machine Learning Automate Moral Hazard and Error?" *American Economic Review* 107 (5): 476–80.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–53.

Peterson, Richard R. 2020. *Brief No. 46: CJA's Updated Release Assessment*. New York: Criminal Justice Agency.

Rambachan, Ashesh, and Jonathan Roth. 2020. "Bias In, Bias Out? Evaluating the Folk Wisdom." Paper presented at 1st Symposium on the Foundations of Responsible Computing (FORC 2020), LIPIc, June 1–3.

Rambachan, Ashesh. 2021. "Identifying Prediction Mistakes in Observational Data." https://asheshrambachan.github.io/assets/files/asheshr_identifyingPredictionMistakes_Main.pdf.

Ridgeway, Greg. 2013. "The Pitfalls of Prediction." *NIJ Journal* 271: 34–40.

Ross, Stephen L., and John Yinger. 2002. *The Color of Credit: Mortgage Discrimination, Research Methodology, and Fair-Lending Enforcement*. Cambridge, MA: MIT Press.

Rudin, Cynthia, Caroline Wang, and Beau Coker. 2020. "The Age of Secrecy and Unfairness in Recidivism Prediction." *Harvard Data Science Review* 2 (1).

Rudin, Cynthia, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2021. "Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges." arXiv:2103.11251.

Ruhland, Ebony L. 2020. "Philosophies and Decision Making in Parole Board Members." *Prison Journal* 100 (5): 641–61.

Salzinger, Kurt. 2005. "Clinical, Statistical, and Broken-Leg Predictions." *Behavior and Philosophy* 33: 91–99.

Schuessler, Karl F. 1954. "Parole Prediction: Its History and Status." *Journal of Criminal Law, Criminology, and Police Science* 45 (4): 425–31.

Schwartzapfel, Beth. 2019. "Can Racist Algorithms Be Fixed?" *The Marshall Project*, July 1. https://www.themarshallproject.org/2019/07/01/can-racist-algorithms-be-fixed.

Smith, Aaron. 2018. "Public Attitudes Toward Computer Algorithms." *Pew Research Center*, November 16. https://www.pewresearch.org/internet/2018/11/16/attitudes-toward-algorithmic-decision-making.

Stevenson, Megan. 2018. "Assessing Risk Assessment in Action." *Minn. Law Review* 103 (303).

Stevenson, Megan T., and Jennifer L. Doleac. 2021. "Algorithmic Risk Assessment in the Hands of Humans." https://ssrn.com/abstract=3489440.

Tschandl, Philipp, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda et al.2020. "Human–Computer Collaboration for Skin Cancer Recognition." *Nature Medicine* 26: 1229–34.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2): 3–28.

Wang, Caroline, Bin Han, Bhrij Patel, Feroze Mohideen, and Cynthia Rudin. 2020. "In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction." *arXiv preprint arXiv:2005.04176.*

This article has been cited by:

1. Ziheng Song, Shafiq Ur Rehman, Chun PingNg, Yuan Zhou, Patick Washington, Ricardo Verschueren. 2024. Do FinTech algorithms reduce gender inequality in banks loans? A quantitative study from the USA. *Journal of Applied Economics* **27**:1. . [Crossref]
2. Eleonora Cresto. 2024. Knowledge, algorithmic predictions, and action. *Asian Journal of Philosophy* **3**:2. . [Crossref]
3. Fritz Schiltz, Kelsey J MacKay, Philippe Vandekerckhove. 2024. Measuring the Size of the "Third Pillar": A Global Dataset. *The World Bank Economic Review* **38**:4, 861-873. [Crossref]
4. Brendon McConnell, Kegon Teng Kok Tan, Mariyana Zapryanova. 2024. How do parole boards respond to large, societal shocks? Evidence from the 9/11 terrorist attacks. *Journal of Public Economics* **238**, 105206. [Crossref]
5. Ana Valdivia, Cari Hyde-Vaamonde, Julián García Marcos. 2024. Judging the algorithm. *AI & SOCIETY* **16**. . [Crossref]
6. Jens Ludwig, Sendhil Mullainathan, Ashesh Rambachan. 2024. The Unreasonable Effectiveness of Algorithms. *AEA Papers and Proceedings* **114**, 623-627. [Abstract] [View PDF article] [PDF with links]
7. A Patrick Behrer, Valentin Bolotnyy. 2024. Heat and law enforcement. *PNAS Nexus* **3**:5. . [Crossref]
8. Daniel Garcia, Juha Tolvanen, Alexander K. Wagner. 2024. Strategic Responses to Algorithmic Recommendations: Evidence from Hotel Pricing. *Management Science* **133**. . [Crossref]
9. Hyunjin Kim, Edward L. Glaeser, Andrew Hillis, Scott Duke Kominers, Michael Luca. 2024. Decision authority and the returns to algorithms. *Strategic Management Journal* **45**:4, 619-648. [Crossref]
10. Luca Grilli, Sergio Mariotti, Riccardo Marzano. 2024. Artificial intelligence and shapeshifting capitalism. *Journal of Evolutionary Economics* **34**:2, 303-318. [Crossref]
11. Brendan O'Flaherty, Rajiv Sethi, Morgan Williams. 2024. The nature, detection, and avoidance of harmful discrimination in criminal justice. *Journal of Policy Analysis and Management* **43**:1, 289-320. [Crossref]
12. Eirini Manga, Nikitas Karanikolas, Catherine Marinagi. Building a Smarter Government Using Machine Learning Applications: Benefits and Challenges 77-98. [Crossref]
13. Reena Roy, Anil Kumar Dixit, S. K. Chaturvedi, Sagar Saxena, Gagandeep Kaur, Ravi Kant. A Socio-Legal Perspective on Gender-Based Discrimination in Machine Learning Algorithm 435-446. [Crossref]
14. Kai Hao Yang, Nathan Yoder, Alexander Zentefis. 2024. Explaining Models. *SSRN Electronic Journal* **110**. . [Crossref]
15. Gregory Weitzner. 2024. Reputational Algorithm Aversion. *SSRN Electronic Journal* **40**. . [Crossref]
16. Arna Woemmel, Hendrik Hüning, Lydia Mechtenberg. 2024. Public Attitudes Toward Algorithmic Risk Assessments In Courts: A Deliberation Experiment. *SSRN Electronic Journal* **23**. . [Crossref]
17. Disa Hynsjo, Luca Perdoni. 2024. Mapping Out Institutional Discrimination: The Economic Effects of Federal "Redlining". *SSRN Electronic Journal* **86**. . [Crossref]
18. Cara Gillingham, Jessica Morley, Luciano Floridi. 2024. The Effects of AI on Street-Level Bureaucracy: A Scoping Review. *SSRN Electronic Journal* **13**. . [Crossref]
19. Greg Stoddard, Dylan Fitzpatrick, Jens Ludwig. 2024. Predicting Police Misconduct. *SSRN Electronic Journal* **31**. . [Crossref]

20. Greg Stoddard, Dylan Fitzpatrick, Jens Ludwig. 2024. Predicting Police Misconduct. *SSRN Electronic Journal* **31**. . [Crossref]

21. Julie Brun Bjørkheim, Odd E. Nygård. Gender Differences in Tax Evasion: Evidence from Norwegian Administrative Data **1**, . [Crossref]

22. Terry-Ann Craigie, Vis Taraz, Mariyana Zapryanova. 2023. Temperature and convictions: evidence from India. *Environment and Development Economics* **28**:6, 538-558. [Crossref]

23. Shaojun (Marco) Qin, Nan Jia, Xueming Luo, Chengcheng Liao, Ziyao Huang. 2023. Perceived Fairness of Human Managers Compared with Artificial Intelligence in Employee Performance Evaluation. *Journal of Management Information Systems* **40**:4, 1039-1070. [Crossref]

24. S. Mills, S. Costa, C. R. Sunstein. 2023. AI, Behavioural Science, and Consumer Welfare. *Journal of Consumer Policy* **46**:3, 387-400. [Crossref]

25. Jens Ludwig, Jon Kleinberg, Sendhil Mullainathan. 2023. Cohort bias in predictive risk assessments of future criminal justice system involvement. *Proceedings of the National Academy of Sciences* **120**:24. . [Crossref]

26. Erika Montana, Daniel S. Nagin, Roland Neil, Robert J. Sampson. 2023. Cohort bias in predictive risk assessments of future criminal justice system involvement. *Proceedings of the National Academy of Sciences* **120**:23. . [Crossref]

27. Caroline Wang, Bin Han, Bhrij Patel, Cynthia Rudin. 2023. In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction. *Journal of Quantitative Criminology* **39**:2, 519-581. [Crossref]

28. Oren Asman, Amir Tal, Yechiel Michael Barilan. 2023. Conversational Artificial Intelligence—Patient Alliance Turing Test and the Search for Authenticity. *The American Journal of Bioethics* **23**:5, 62-64. [Crossref]

29. Adel Ismail Al-Alawi, Albandri Mohammed A-Lmansouri. Artificial Intelligence in the Judiciary System of Saudi Arabia: A Literature Review 83-87. [Crossref]

30. Norbert Jastroch. Sustainable Artificial Intelligence: In Search of Technological Resilience 317-326. [Crossref]

31. Douglas Araujo. 2023. Gingado: A Machine Learning Library Focused on Economics and Finance. *SSRN Electronic Journal* **59**. . [Crossref]

32. Stuart Mills, Samuel Costa, Cass R. Sunstein. 2023. The Opportunities and Costs of AI in Behavioural Science. *SSRN Electronic Journal* **46**. . [Crossref]

33. Sara Wakefield. 2022. Criminal Justice Reform and Inequality. *American Journal of Criminal Justice* **47**:6, 1186-1203. [Crossref]

34. Xin Ouyang, Zhen Sun, Xinzhen Xu. 2022. Patent system in the digital era - Opportunities and new challenges. *Journal of Digital Economy* **1**:3, 166-179. [Crossref]

35. Thomas A. Loughran, Megan Augustyn, Mauri Matsuda, Kimberly L. Henry. 2022. Prognosticating Offending in Early Adulthood: How Early Can We Predict?. *Journal of Quantitative Criminology* **55**. . [Crossref]

36. Nitasha Kaul. 2022. 3Es for AI: Economics, Explanation, Epistemology. *Frontiers in Artificial Intelligence* **5**. . [Crossref]

37. Brendan O'Flaherty, Rajiv Sethi, Morgan Williams Jr. 2022. The Nature, Detection, and Avoidance of Harmful Discrimination in Criminal Justice. *SSRN Electronic Journal* **111**. . [Crossref]

38. Arthur Dyevre. 2022. Adopting the Gold Standard: Field Experimentation for Evidence-Based Judicial Lawmaking. *SSRN Electronic Journal* **150**. . [Crossref]