# Algorithms and Recidivism: A Multi-disciplinary Systematic Review

**Arul George Scaria[1], Vidya Subramanian[1], Nevin K George[2], Nandana Sengupta[2]**

[1]National Law School of India University, Bengaluru, India
[2]Indian Institute of Technology, Delhi, India
arul.scaria@nls.ac.in, vidya.subramanian@nls.ac.in, nevinkgeorge@gmail.com, nandana@iitd.ac.in

**Abstract**

The adoption of algorithms across different jurisdictions have transformed the workings of the criminal justice system, particularly in predicting recidivism risk for bail, sentencing, and parole decisions. This shift from human decision-making to statistical or algorithmic tool-assisted decision-making has prompted discussions regarding the legitimacy of such adoption. Our paper presents the results of a systematic review of the literature on criminal recidivism, spanning both legal and empirical perspectives. By coalescing different approaches, we highlight the most prominent themes that have garnered the attention of researchers so far and some that warrant further investigation.

## Introduction

The desire for evidence-based sentencing has led to the adoption and use of risk scores in determination of criminal sentencing decisions. Risk scores are generated using different kinds of parameters, including socio-economic and demographic ones (Southerland 2021). Some academics, lawyers and judges see these developments as heralding a new era of progressive reform. Many of the arguments raised by such scholars, including the desire for greater objectivity, promptness, and efficiency in administration of justice, particularly by way of curtailing the incarceration rates, have often persuaded policy makers and courts to employ statistical and algorithmic risk assessment tools (RATs). (Starr 2014). These tools are now used in many areas of the criminal justice system like predictive policing, surveillance, bail related decision making, and sentencing. RATs are particularly used for recidivism prediction, which is the estimation of an individual's risk of reoffending or failing to appear in court (Hamilton 2021).

RATs' promise of objective methods and standards to mitigate misguided decision making is especially pertinent given recent findings which suggest that judges may harbor implicit biases against particular communities (Rachlinski et al. 2008). However, the use of RATs and empirical methods in law is much more contested because law, unlike other domains like medicine or business, lacks a "uniformity of purpose" (Rachlinski n.d.). Therefore, it is difficult to establish singular truths in law, like profit maximization or healthy populations, which complicates the ideas of right or wrong. This also makes it difficult to spell out a normative future and therefore the ability of empirical methods to dispel wrongly held myths in law (Rachlinski n.d.).

RATs pack within them a mixture of legal and empirical knowledge (Eaglin 2017). As one may notice from this paper, both empirical and legal scholars have attempted to unpack some of the opportunities and consequences such tools offer. We bring together both legal and empirical perspectives containing a range of divergent and common themes on the topic. We approach the review of both streams of literature systematically by answering the following three research questions:

1. What are the concerns regarding the deployment of RATs in recidivism decisions?
2. What are the arguments in favor of employing RATs in recidivism decisions?
3. What solutions have been proposed to address concerns regarding RAT use for recidivism prediction?

A majority of papers we reviewed examine various concerns surrounding the use of risk assessment tools. These concerns have resulted in many more works scrutinizing the concerns for their validity. As one may notice from the first and second research questions, our review is structured in a manner reflecting this evolution of the literature. The third question on proposed solutions highlights the current state of knowledge given the concerns. We synthesize findings from literature indexed in five academic databases and the multi-stage review process as detailed in the methodology

section. The literature in the sample is predominantly from law, computer science and economics. To the best of our knowledge, this is the first multidisciplinary systematic review of literature on RATs.

## Methodology

The relevant papers for review were chosen in three stages. Stage I, Identification, involved choosing relevant academic databases and formulation of keywords for the search procedure. We used "Algorithmic bias && Recidivism" and "Algorithm && Recidivism" as the search keywords for identifying legal literature. HeinOnline, Westlaw and JSTOR (US studies alone) databases were used for finding legal literature. Scopus and WoS were used for finding relevant empirical literature and truncated keywords - "Algor* and Recid*" were used to obtain papers that we broadly classify as empirical literature. Truncated keyword search in the legal databases returned over two thousand results, while the extended keywords returned single digit results in the empirical databases, making a uniform approach infeasible. The following table presents keywords and the corresponding total results from each database.

| Keywords | Database | Results |
|---|---|---|
| "Algorithm && Recidivism" | HeinOnline | 1663 |
| | Westlaw | 515 |
| | JSTOR | 479 |
| "Algorithmic bias && Recidivism" | HeinOnline | 147 |
| | Westlaw | 97 |
| | JSTOR | 27 |
| "Algor* and Recid*" | Scopus | 171 |
| | WoS | 51 |

Table 1: Search keywords across databases

At the second stage, Screening I, we filtered papers by scrutinizing their titles and abstracts to check for their appropriateness in answering the three research questions mentioned earlier. We also removed duplicate and irrelevant records across databases at this stage, resulting in a total of 37 papers for the keywords "Algorithmic bias && Recidivism", 79 for "Algorithm && Recidivism", and 63 for "Algor* and Recid*".

In the third and final stage, Screening II, we comprehensively analyzed entire texts of the selected papers for further filtration. 20 papers were selected under the keywords "Algorithmic bias && Recidivism", and 42 under the keywords "Algorithm && Recidivism" for the legal literature. No results were removed at this stage from the empirical list. This methodological approach led to a total of 126 papers.

Figure 1 is a distribution of selected papers published between 2000 and 2022. As one may notice, there is a clear increase in the academic literature on RATs after 2016. Two major events may have contributed to this. In 2016, ProPublica had published an article wherein it accused COMPAS of racial bias against Black defendants. In the same year, the Wisconsin Supreme Court had also delivered the decision in *State v. Loomis*. In this case, the court ruled against the defendant stating that the results of the algorithmic tool which was employed to assist in the determination of sentencing were not the sole determinative factor but were only one amongst the many factors taken into consideration for reaching the final determination. It was also held that the inclusion of algorithmic scores were not violative of the due process rights. Both the events may have triggered interest among scholars to work on diverse issues in the area. In 2022, we also observe a decline in the total number of papers. It is primarily attributable to a decline in the legal papers as there has been a slight increase in the published empirical papers.

The legal and empirical findings relating to each of our three research questions are presented in the next 3 sections. In the legal and empirical subsections, we have thematically grouped and synthesized findings based on similarity of arguments. We also ordered the subsections according to context for reader's ease of understanding.

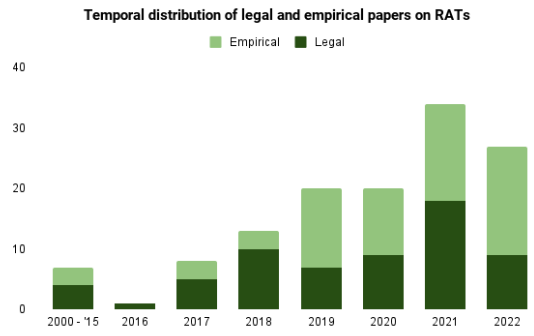


Figure 1: Distribution of selected papers from 2000-2022

## Concerns

RATs and algorithms in general are often portrayed as technological instruments that are objective and value neutral. Since they are generally lacking human frailties, they are considered as superior decision makers. However, empirical evaluations of COMPAS, a RAT used in the US criminal justice system, have challenged these assumptions of unbiasedness and subjected algorithms to deeper empirical and

legal scrutinizes. Legal scholars have also extensively worked on enumerating the concerns associated with the deployment of RATs in the criminal justice system arising out of biased input data and disproportionate outputs along with the legal issues that might follow. The following section discusses the legal concerns and the technical limitations of employing RATs, highlighted in the literature.

## Legal Findings

The concerns highlighted in the legal papers can be classified into three diverse categories - input data bias, primarily resulting from the use of biased data sets; output data bias, discernible from the outcomes generated by the tool; and lastly, the legal concerns, arising out of their employment in decision-making. It needs to be highlighted that 70 percent of the legal literature we reviewed through the outlined methodology focused on the concerns stemming out with the usage of RATs.

### Input Data Bias

Input data bias may arise because of using training data riddled with historic bias, resulting in overrepresentation of persons from certain demographics in the data fed into the model. The application of an offender's immutable traits, the reliance on group aggregate data in place of individualized assessments, and consideration of the offender's criminal history for making recidivism predictions may further contribute in augmenting the inherent bias present in the data fed into the algorithms.

**Employing Immutable Traits.** Numerous studies (N=24) have revealed how risk-assessment tools can reinforce and exacerbate the inherent societal biases present in the datasets utilized to train the algorithms. These biases may be a consequence of the higher rates of arrests among individuals belonging to certain populations, often including persons belonging to lower socio-economic strata or people living in minority neighborhoods, the data being acquired from arrest records based on policing decisions [(Zavrsnik 2021), (Kehl et al 2017), (Jones 2021), (Mbadiwe 2018), (Hamilton 2021), (Rizer & Watney 2018), (Shi 2022), (Okidegbe 2021), (Villasenor and Foggo 2020), (Thomas and Ponton-Nunez 2022), (O'Brien 2021), (Starr 2014), (Bagaric et al. 2022), (Southerland 2021), (Rankin 2021), (Gravett 2021), (Lewis 2022), (Anderson 2020), (Cyphert 2020), (Wisser 2019), (Dalakian 2019), (Deskus 2018), (Mayson 2019), (Starr 2016)]. Okidegbe (2021) highlights how underrepresentation of racially marginalized communities belonging to lower socio-economic status from the training datasets or during the construction of the algorithm can result in a "group's marginalization in pretrial governance".

**Utilizing Group Data.** Another serious limitation associated with the use of algorithmic tools that has garnered the attention of authors (N=11) is use of generalized aggregate data of groups with similar characteristics as the benchmark in determination of criminal sentencing, as opposed to focusing on the immediate criminal act or an individual's personal conduct for evaluating the recidivism risk [(Eckhouse et al. 2019), (Donohue 2019), (Starr 2014), (Wisser 2019), (Novokmet et al. 2022), (Slobogin 2013), (Monahan and Skeem 2016), (Gravett 2021), (Okidegbe 2021), (Southerland 2021), (Thomas and Ponton-Nunez 2022)]. Starr (2014) & Gravett (2021) have touted the use of group data instead of individualized information as patently unfair. Southerland (2021) also explains how digital profiling can arise when the punishment for a crime committed by a particular individual is determined based on his similarity or affiliation to a certain group. Thomas & Ponton-Nunez (2022) point out how judicial discretion is lost in the quest for minimizing divergence from a pre-defined criteria in the form of group aggregate data severing the possibility of considering the diverse particularities exclusively related to an individual's crime and conditions in the determination of sentence.

**Inclusion of Criminal History.** The static criminal history fed into the algorithmic models may be entrenched with perpetual racial bias affecting the objectivity of the algorithmic tool predicting recidivism. There might be a possibility of arrest data being biased against people of color based on the policing decisions, choosing to ignore the immediate culpability of the individual in question. Anderson (2020), Zavrsnik (2019), Monahan and Skeem (2016), Okidegbe (2022), Jones (2021) and Tonry (2019) highlight how the use of biased and historically inaccurate information fails to tackle the actual crime committed and shroud inequitable outcomes meted out to the marginalized communities. Gravett (2021) and Collins (2018) explain how the failure to focus on the immediate crime in question by resorting to past criminal history of an individual can also have serious ramifications on the determination of the sentencing range. O'Brien (2021) also points to the challenges that a formerly incarcerated person may face, including loss of employment, livelihood and change in family dynamics.

### Output Bias

The pertinence of algorithmic tools has also been called into question based on their output performance. Several authors have elaborated on transparency, accuracy and reliability challenges associated with the application of these tools and it may be worth discussing them separately.

**Transparency.** Lack of transparency in algorithmic decision-making in the criminal justice system is one of the primary concerns indicated by the legal scholars. Several authors (N = 12) have labelled the risk-assessment tools as a 'black box' owing to ambiguities related to their design, weightage assigned to the variables, datasets used and the lack of timely validation [(Donohue 2019), (Seglias 2021),

(Brenner et al 2020), (Deskus 2018), (Wisser 2018), (Cyphert 2020), (Eaglin 2019), (Kehl et al. 2017), (Thomas and Ponton-Nunez 2022), and (Novokmet et al. 2022)].

Intellectual property protection like copyright and trade secrets also present additional hurdles in examining and understanding the working of the proprietary algorithms used in decision-making, especially for the defendants and their respective counsels [(Mbadiwe 2018), (Adler et al. 2020), (Rizer 2021), (Villasenor & Foggo 2020), (Gravett 2021), (Wexler 2018), (Carlson 2017), (Rizer and Watney 2018)]. An example on this point is the invocation of the trade secrets protection argument by the Northpointe to prevent screening of its algorithm during the *Loomis* case.

**Reliability.** Differing standards of reliability or validity in forecasting the behavior of the offender is another concern inviting extensive scrutiny. Data trained on the unique characteristics of a particular population may not be suitable for widespread application across different jurisdictions, as highlighted in numerous studies (N=5). Slobogin (2013), Rankin (2021), Eaglin (2019), Carlson (2017), Hamilton (2021) and Zavrsnik (2019) note how RATs are equipped with the ability to calibrate correlations, but seldom delve into the authenticity of the causations.

The interaction between human judges and ADM tools have also been observed to produce divergent results. There are apprehensions that judges may override the recommendations of the ADM systems altogether ("algorithmic aversion") or rely on the findings of the tools completely without exercising their discretion ("automation bias") [(Rizer & Watney 2018), (Ludwig & Mullainathan 2021), (Thomas & Ponton-Nunez 2022), (Stevenson 2018)]. Interpretation of risk-level categorizations in algorithmic tools also raise questions regarding the reliability of these tools [(Adler et al. 2020), (Jones 2021), (Hamilton 2021)]. Novokmet et al. (2022) explain how the factors and the criteria for the calculation of risk scores vary across different tools. The employment of algorithmic tools has also been observed to contribute to higher rates of mass incarceration (Collins 2018) and nullified subjectivity Zavrsnik (2021).

**Accuracy.** Accuracy concerns in risk-score calibration is another major shortcoming in the adoption of algorithmic tools in criminal justice applications [(Slobogin 2013), (Zavrsnik 2021), (Deskus 2018), (Tonry 2019), (Wisser 2019), (Gravett 2021), (Novokmet et al 2022), (Mbadiwe 2018)].

Eaglin (2017) elucidates how developers make normative decisions while constructing a tool which may not always be in the best interests of the state. Hamilton (2021) highlights how lack of validation studies can result in failure to measure absolute accuracy, necessary to accept a tool's performance. Meanwhile, Villasenor and Foggo (2020) highlight how subjecting an individual to substantially inaccurate information can trigger due process concerns.

**Legal Challenges**

Legal challenges to the employment of RATs may be classified into three categories: those posing a challenge to the equal protection clause, those infringing the due protection clause and those giving rise to fairness concerns.

**Equal Protection.** The equal protection clause in the US Constitution necessitates the law to treat everyone equally. Further, this doctrine also mandates the requirement of a legitimate purpose for employing non-suspect classifications by the state. Resorting to the use of suspect class variables, including race, ethnicity and gender in algorithms used for sentencing is likely to face elevated scrutiny. An equal protection challenge can also lie in the absence of explicit suspect-class classification, provided a defendant can prove that the algorithmic tool produces racially disparate impact accompanied with discriminatory intent. Equal protection analysis also mandates showing that the use of risk assessment tools is narrowly tailored to justify its adoption by the state [(Kehl et al. 2017), (Rizer 2020), (Thomas & Ponton-Nunez 2022), (Brenner et al 2020), (Dalakian 2018), (Wisser 2019), (DiBenedetto 2019), (Krent and Rucker 2021), (Slobogin 2013), (Starr 2014), (Muenster 2022)].

**Due Process.** Due process clause in the U.S Constitution assures the right to a fair trial and requires the defendant not to be sentenced on substantially inaccurate information. The underlying algorithms employed in the risk assessment based on proprietary models do not allow any examination, thereby tactically guarding themselves from due process claims [(Kehl et al. 2017), (Rizer 2020), (DiBenedetto 2019), (Krent and Rucker 2021), (Slobogin 2013)]. Novokmet et al. (2022) also highlights due process concerns that may arise as a result of the usage of RATs, and staunchly opposes their employment in the sentencing stage, as they could lead to denying the right to a fair trial. Some scholars also argue that due process requires the explainability of algorithmic models, which is necessary for interpreting the evidence presented against oneself in a criminal sentencing [(Brenner et al. 2020), (Wisser 2019)].

**Fairness.** Algorithmic tools have the potential to augment and propagate bias that may have permeated during the curation of the datasets employed to train the algorithms. There is also controversy surrounding the competing notions regarding the definition of algorithmic fairness. The most prominent example being the 2016 Propublica study, addressing the prevalent racial bias in the risk assessment tool COMPAS. While Propublica criticized COMPAS by showcasing that people of color were more likely to receive higher risk scores under the predictive parity conception, the tool developer Northpointe exalted the tool's neutrality by relying on equalized odds criteria (Washington, 2019).

Several scholars (N = 6) have examined the implications of employing algorithmic tools for fairness, noting that it

may be impossible for an algorithm to meet different standards of fairness, since they rely on different base rates for gauging recidivism [(Mbadiwe 2018), (Rizer and Watney 2018), (Hill II 2021), (Brenner et al. 2020), and Ludwig and Mullainathan 2021)]. O'Brien (2021) and Zavrsnik (2021) have also worked on the tradeoffs between fairness-accuracy and fairness-equality and concluded that achieving all the conditions simultaneously may be statistically impossible. Succinctly put, there is still an ongoing debate for identifying a universal definition for fairness, as it depends on the clear-cut objectives expected from an algorithm (Kehl et al. 2017).

### Empirical Findings

Notions of fairness, accuracy and bias have mathematical expressions. Mathematical representation makes quantitative estimations of the deviation from defined objectives possible for both humans and algorithms, enabling a comparison of their performance.

Dressel & Farid (2018), in a seminal experimental analysis found that non-experts can make defendant recidivism predictions close to COMPAS' predictions. These results and ProPublica's 2016 findings of COMPAS' higher false positive and lower false negative prediction rates for Black defendants are the empirical arguments most relied on by critics to maintain that algorithms can be biased and perform no better than human decision makers (Angwin et al. 2022). Hamilton (2019) builds on this research concluding that COMPAS systematically over classifies women into high-risk groups.

Dressel and Farid (2018) also report other findings from the experiment which were formative arguments against the use of algorithms. First, the vignette experiments demonstrate that lay persons, with little criminal justice experience, when provided with 7 defendant features, make predictions with similar false negatives and positives for Blacks and whites as COMPAS - racial machine bias (Angwin et al. 2022) therefore, was also found among lay persons (Dressel & Farid 2018).

Second, Dressel & Farid (2018) report that a logistic regression classifier with just 2 features viz., age and number of prior convictions, performs as good as COMPAS, which uses 7 features - conjecturing that COMPAS is no more complex than a simple linear predictor. A more powerful ML algorithm does not improve their results - questioning algorithmic prediction of criminal justice outcomes. These well received findings, emerging during a surge of race discourses in politics, has led to a back-and-forth on the specifics.

Stevenson & Slobogin (2018), through partially reverse engineering COMPAS' VRRS, find that age, not race or others, is the most dominant factor in explaining variation in COMPAS scores. Younger defendants have nearly twice the score of older defendants. Youth, as is encoded in COMPAS, is only treated as a potential criminal risk variable, unlike in a courtroom where a judge can also ideally see a young defendant's diminished culpability. Stewart (2022) argues in a similar vein that multiple sub groups (along identity or seemingly irrelevant traits) can be formed for fairness constraints and that single partitions might be too restrictive. However, no fairness constraint can be satisfied across different partitions of subgroups, making some level of bias inevitable.

T. Greene et al. (2022) draw attention to technical choices in RAT design during their development that can lead to inconsistent predictions. Conceptualizing this as "predictive inconsistency", they propose methods that can identify how specific choices in design can influence a particular prediction. These methods enable questioning RAT predictions while clarifying its political and legal legitimacy (T. Greene et al. 2022).

## Arguments in Favor

The concerns raised in Dressel & Farid (2018), Angwin et al. (2022) and Hamilton (2019) invited a slew of responses. Limitations of Dressel & Farid's (2018) methodology and replications and extensions of the original experiment arrived at different conclusions. With new evidence, these critiques together call for a reanalysis of the original results. In pointing out the limitations, critics of the original experiment also make the case for adoption of algorithms in identified settings. The sub-section on empirical findings first summarizes methodological critiques that probe if the conclusions of Dressel & Farid (2018) are generalizable before addressing black-boxes and racist algorithms. The sub-section on legal findings then discusses legal papers highlighting the benefits achieved with the implementation of RATs for forecasting recidivism, in the form of efficiency, consistency and minimizing judge-made bias.

### Empirical Findings

#### Methodological Critiques of the Dressel & Farid (2018) Experiment

Bansak (2019) observes that there is large variation in the proportion of correct responses made by individual MTurkers (0.38 to 0.80) in the Dresel and Farid (2018) experiment and highlights that this resulted in an analysis based on an average of participant responses. Though participants achieve mean and median accuracy of 0.62 and 0.64, these averages conceal the unreliability of individual estimates. Employing a predetermined single probability cut point in analyzing pooled estimates for multidimensional decision making is also undesirable (Bansak 2019). An example in this regard is a decision space that has more than 2 options - imprisonment, supervised release, or no incapacitation. When probabilistic outputs are used, a comparison

finds that statistical machine learning (ML) algorithms outperform human predictions (Bansak 2019). In a replication study, Lin et al. (2020) found that providing participants with immediate feedback on their predictions improved their accuracy through the experiment as they experienced learning. Additionally, when the defendant pool had a low base rate of rearrest, human participants performed worse than existing risk assessment tools (COMPAS and LSI-R) and logistic regression models without feedback, unlike in Dressel and Farid (2018) where the base rate was 48% (Lin et. al. 2020). Algorithms are also able to outperform humans in ranking accuracy (AUC) and when more predictive information is made available (Lin et al. 2020). Therefore, the results observed in Dressel & Farid (2018) seem like a specific occurrence and less reflective of general and real conditions.

In another modified replication, Biswas et. al (2020) find that using a balanced defendant pool with equal Black and white profiles results in lower accuracy, but relatively fairer responses in terms of false negatives and positives when compared to Dressel & Farid (2018) responses. Here, using a logistic regression model on balanced data results in significantly fairer predictions despite a slightly lower accuracy.

Both the Dressel & Farid (2018) experiment and methodological critiques assume that individuals in vignette experiments are similar to judges making informed judgments. An overreliance on this mode of investigation can prove disadvantageous, if later research finds incongruence between this method and findings.

### Racial Bias in Algorithms

Vincent & Viljoen (2020) conclude that, while scrutiny for bias in RATs is necessary, there exists no evidence that such tools are based against individuals of colour. Specific instruments could be faulty, but generalized calls for banishing algorithms are unfounded. The mathematical principles behind maximising recidivism prediction has revealed the systematic biases in the justice system because social groups scoring higher on an unbiased instrument do so partly because of higher arrest rates for those groups (Vincent & Viljoen 2020).

### Black Box Character of RATs

Black-box RATs are either proprietary models or ML methods that generate uninterpretable outputs - both restrict knowledge of inner mechanisms (Rudin 2019). Black-boxes present legal challenges on explainability, transparency and accountability as racial bias has been found in tools employed in other high-stakes decision making sectors (Obermeyer et al. 2019). While no authors argue in favor of using black-box RATs, many present nuanced arguments on the topic. Most currently employed RATs are interpretable statistical regression models, therefore not black-boxes of the second kind (T. Greene et al. 2022). Rudin (2019) recommends that policy makers must stop accepting unexplainable models and researchers should stop trying to explain such models. Using uninterpretable black-boxes for high-stakes decisions can be unsafe. Interpretable ML models can be used in such cases without common misconceptions like interpretability decreases as prediction accuracy increases (Rudin 2019).

## Legal Findings

We could identify eight legal studies that discuss the benefits realized with the application of RATs in criminal justice settings. Desmarais et al. (2021) argue that the performance of RATs must be scrutinized with reference to the status quo, and not an ideal scenario that may have never existed. They argue that RATs offer an efficient and reasonable basis for decision-making in comparison to the current judge-made decisions, wherein cognitive assumptions and inherent implicit biases of judges may affect the decision-making process. Desmarais (2021), Rizer (2021), Hunter et al. (2020) and Dalakian (2018) expound how RATs are relatively more reliable than human-made decisions, as they are devoid of any generalizations or extraneous considerations. Bagaric and Wolf (2018) also highlight how judge-made decisions have been found to be biased against marginalized communities, a concern that can be resolved using algorithmic tools which are uninfluenced by instinctive bias of any sort.

Consistency of results observable in relation to similar crimes is another factor that can propel their further use in sentencing applications. Bagaric and Wolf (2018) endorse the employment of RATs by stressing how they can uphold the rule of law by providing a mechanism for doing away with arbitrary and opaque sentencing decisions by human judges. They further highlight how the application of algorithmic tools can minimize the decision-making time as computer models can process large swathes of data quicker in comparison to human judges. Hunter et al. (2020) draws attention to the speed and the efficiency with which a sentencing determination can be fetched by an algorithmic tool.

Pre-trial detention in the absence of bail hearings has been touted to be one of the prominent drivers of mass incarceration. Dalakian (2018), Hamilton (2021), Lowden (2018), Rizer (2021) and Slobogin (2021) elucidate how the application of RATs seems to have had a significant effect in curbing the growing numbers of prison population by bringing down the duration of incarceration period for offenders during the pre-trial phase, without affecting public safety.

Bagaric & Wolf (2018) and Dalakian (2018) also try to illustrate how the use of RATs can significantly curtail government revenues directed towards servicing prisons due to mass incarceration. They argue that application of RATs can further contribute in selecting the combination of factors and

variables deserving higher weightage in the determination of recidivism. Lowden (2018) also cites the example of Mecklenburg County, wherein the Public Safety Assessment (PSA) tool was introduced and it emphasizes the benefits attained in cases of money bail, especially for indigent defendants having limited means to secure their release before trial.

## Proposed Solutions

Legal papers have proposed numerous solutions for addressing and mitigating the observable concerns of RAT assisted judicial decision-making. These range from the need for public participation in the construction of tools to providing training to decision-makers, and increased transparency. In the technical literature, there is increasing realisation that mere exclusion of defendants' race is no solution for algorithmic bias concerns, as race is correlated with many other variables (Dressel & Farid 2018). Technical solutions to algorithmic problems primarily focus on the algorithm to improve performance metrics. These methods can be broadly classified into pre-processing, in-processing and other methods. The following subsections elaborate on these proposals.

### Legal Findings

#### Mitigation of Input Bias

Ludwig and Mullainathan (2019) attribute the fallacy of the algorithmic tools on the humans (or the designers) responsible for constructing it and propose making appropriate design choices for the former to work as a viable option. Rizer & Watney (2018), Adler et al. (2020), Washington (2019) and Garrett & Monahan (2020) suggest tool developers to supply design information to the public including the input factors considered, the criteria for the inclusion of variables and the weightage assigned to the factors in the algorithmic model. Further, these papers direct attention to the notion of assigning risk level labels and the inconsistency that might arise in the absence of a well-defined framework for the allocation of risk scores across the different Risk Assessment Tools. Southerland (2021) proposes a radicalized idea of accounting for race in the inputs and the outputs of the RATs by opting to be race-conscious in order to eradicate racial discrimination in the predictions. Humerick (2020) also endorses affirmative action in this regard by acknowledging and explicitly incorporating race as a "plus-factor" in the RATs.

Okidegbe (2021) and Garrett & Monahan (2020) propose public participation as one of the solutions to mend input bias problems. They stress how adequate representation combined with deliberation from members of all communities during the tool creation (including ensuring diversity in the training data) and implementation stages can ensure democratic inclusion in the algorithmic governance framework. Hill (2021), in addition to lending support for the above point, proposes minimizing the role of developers and government officials in making policy decisions concerning the range and levels of risk categories. Okidegbe (2022) offers a solution to the persisting social discrimination in the pretrial system by replacing carceral knowledge sources (such as, police arrest data) in the input datasets with knowledge about the most impacted groups in the criminal justice framework, also termed as 'community knowledge sources'.

Mayson (2018) acknowledges the associated concerns arising out of judges' cognitive bias, while highlighting the futility of algorithmic tools exhibiting racial disparities and pushes for measures that can aid in eliminating the social inequality pervading throughout the functioning of the criminal justice system. Tonry (2019) suggests doing away with predictive sentencing, or minimizing its use due to lack of any empirical or moral proof of its effectiveness.

Against the background of the first ever legislation on AI regulation, the EU's AI Act and its consequent adaptation in the realm of risk-assessment instruments, Van Dijck (2022) underscores the significance of human oversight for correcting biases resulting from the employment of RATs.

#### Mitigation of Output Bias

Several papers [(Kehl et al. 2017), (Villasenor & Foggo 2020), (Bagaric et al. 2022), (Dibenedetto 2019), (Slobogin 2021), (Wisser 2019), (Gravett 2021) and (O'Brien 2021)] have advocated for increased transparency to address the systemic biases ascribed to the usage of algorithmic tools. Eaglin (2017) enumerates three levels of opacity plaguing the algorithmic tools and proposes measures to counter them by ensuring transparency for tool developers and criminal justice actors; providing accessibility regarding the construction choices by establishing democratic accountability; and incorporating interpretability measures in statistical modeling.

Quite a few papers [(Rizer 2021), (Rizer & Watney 2018) and (Carlson 2017)] have also contextualized the daunting challenge posed by trade secret laws in preventing access to the algorithms employed in the RATs. Villasenor & Foggo (2020) subscribe to the idea of granting access to the proprietary algorithms with proper safeguards while also ensuring trade secret protection. Carlson (2017) stresses how carving out an exception for trade secrets under the Freedom of Information Act (1967) has resulted in frustrating the vision of the legislation and recommends an amendment to the legislation for necessitating disclosure from private companies developing RATs. On the other hand, Wisser (2019) proposes moving away from privately-owned algorithms to government-owned algorithms to address the impediments posed by the trade secrets. Further, the author also suggests demanding public explanation in the written format from the algorithm designers about the working of the algorithm.

Jones (2021) posits scrutinizing the data for ensuring accuracy and reliability of the datasets employed to train the algorithm. Anderson (2020), Wisser (2019), O'Brien (2021), and Garrett & Monahan (2020) highlight how providing additional training to judges to get them acquainted with the inner workings of algorithmic tools can lead to improved results, thereby addressing reliability concerns. Guaranteeing the involvement of judges in the decision-making process and stipulating explanation on deviating from the tool recommendations can also aid in bolstering acceptance by the public (Rizer & Watney 2018). Further, they also argue that presenting the risk information in a comprehensible manner along with a structured decision-making framework can aid in effectively regulating their use.

Conducting validation studies to infuse accountability into the decision-making process, and facilitating auditing and impact studies by third parties can also contribute to rigorous evaluation of the tools as proposed by some of the papers [(Kehl et al. 2017), (Villasenor & Foggo 2020), (Bagaric et al. 2022), (Rankin 2021), and (Gravett 2021)]. Villasenor and Foggo (2020) recommends preservation and storage of the data regarding the defendants and the algorithms used for the assessments for allowing future analysis whenever necessary.

Periodic evaluation and review (Hunter et al. 2020), using RATs as cognitive assistants in determining sentences (Donohue 2021), opting for a multidisciplinary approach combining legal and empirical sciences (Hamilton 2021), and ensuring consistency in the outputs by minimizing the interference of proxies (Villasenor and Foggo 2020) are some of the other suggestions articulated to tackle the output bias concerns posed by the RATs.

## Empirical Findings

The technical solutions identified here, importantly, do not solve the problem of structural discrimination that could be responsible for the significantly different between-group base rates (Vincent & Viljoen 2020). It also fails to address the concerns of poor convertibility of social science and legal concepts to clearly defined data science variables on which algorithms work (C. Greene 2022). On the other hand, these solutions implicitly acknowledge the limitations of algorithmic methods and propose computational or other technical approaches to align technological advancements with moral values like fairness or unbiasedness. This is achieved in the literature through debiasing or bias mitigating strategies or by facilitating better interpretability.

### Pre-Processing Methods

Duwe (2019) suggests pre-processing techniques like modifying input variables to create single predictors from multiple variables that are independently not significantly predictive. Using multiple measures beyond the AUC to estimate fairness and testing multiple classifiers before employing a particular one is also suggested here.

Johndrow and Lum (2019) and Calmon et al. (2017) propose algorithmic solutions to pre-processing. The former, using a probabilistic approach, suggests an algorithm to make variables independent of sensitive attributes and the latter presents an algorithm that transforms original data and reduces group discrimination when classifiers such as logistic regression and random forests are trained on this new data. Both methods don't impair accuracy metrics significantly.

Skeem and Lowenkamp (2020) test multiple algorithms inputting racial information at varying degrees. They find that providing information regarding race improves both algorithmic accuracy metrics as well as fairness metrics. They argue that there are inherent trade-offs between competing values (like crime prevention and racial justice) in algorithmic (and human) decision making and algorithms make such trade-offs apparent. Therefore, they argue that policymakers should choose an algorithm based on the trade-off they find most acceptable. Ma et al. (2022) proposes a new algorithm responding to the need for interpretable systems in high-stakes decision making. They develop a new algorithm which is a single decision tree ML algorithm structure instead of the more complex (and less interpretable) decision tree structures more prevalent in the literature. This new algorithm achieves fairness and accuracy metrics similar to more complex ML algorithms, and provides higher interpretability (Ma et al. 2022).

### In-Processing Methods

In-processing methods address fairness during model creation. Biswas & Mukherjee (2021) address a common violation of an assumption that enables such techniques - prior probability shifts. Prior probability shifts occur when structural distributional changes are present in training and test data, especially within population subgroups. The authors develop an algorithm (CAPE) which accounts for such shifts and are able to ensure fair classification.

In another bias mitigation strategy to improve fairness of ML models, a combination of parity in generalized false positive and negative rates while maintaining sub-group calibration is proposed by Karimi-Haghighi & Castillo (2021). Their results show that this procedure can substantially reduce generalized false positive rate disparities across multiple groups, but the decline in bias is achieved at the expense of increased inequalities in other metrics.

In Zhang & Ramesh (2020), instead of a regularization method to avoid bias, they develop Fair-A3SL, an algorithm which directly induces the structure of the ML model and optimizes for multiple fairness measures. They demonstrate that the approach reduces structural model bias and gives interpretable fairer predictions for COMPAS and other datasets.

**Other Methods**

Methods other than pre-process and in-process techniques are a diverse set of technical approaches for equally diverse goals ranging from improving interpretability of ML models to improving fairness metrics of these models.

Meyer et al. (2022) propose an innovative solution by flipping the script on risk assessment. They develop models to estimate the risk the US criminal justice system poses to defendants by predicting the likelihood that the system delivers a lengthy sentence based on factors that may be legally immaterial. Such a tool enables defendants to bring to the judge's knowledge the unjust treatment by an RAT. Through the limitations and value choices inherent in developing such a model, the authors inform the discourse of the same limitations and choices existing for RATs currently used on individuals.

Another approach to technically improving fairness in ML models borrows from developments in causal inference techniques that use potential outcomes instead of observed outcomes used in RATs. This approach problematizes the influence of RAT predictions on ensuing outcomes (Mishler et al. 2021). Coston et al. (2020) suggest a novel method for evaluating RATs that takes into account the effects of interventions while projecting outcomes based on past data. To make sure that projections represent the risk under historical policy rather than the multiple decision options available, they propose using causal inference techniques. Building on this research, Mishler et al. (2021) develops approximate counterfactual equalized odds (approximate cEO), a fairness criterion which proposes an alternative to the widely used criteria based on observed outcomes. Since risk assessment predictions can themselves influence downstream outcomes, cEO is a novel method that seeks to reduce the misleading effects of predictions on outcomes. cEO allows users to negotiate between fairness and performance, employing equalized odds as the best suited fairness criteria to reduce discriminatory disparate impact. To apply such a criterion to existing RA tools, making counterfactual fairness easily applicable to most extant RA tools, the authors also propose a post-processed predictor that displays favorable convergence properties.

In Van Berkel et al. (2019), the authors propose a scalable method to identify fair predictors that are socially acceptable by majority vote. They propose this because of the lack of ground truth in predictor selection which makes it a subjective task in the construction of equitable algorithms. They suggest that the role of selecting predictors should be with the society instead of private parties or other stakeholders.

Additionally, there are methods in the algorithmic fairness literature like Cabrera et al.'s (2019) FAIRVIS. FAIRVIS is a novel tool for auditing ML models. It addresses performance disparities within populations which is often overlooked in ML models trained for overall accuracy. The tool contains an integrated subgroup discovery mechanism and visual analytics system. In a simulation on COMPAS data, the authors confirm that the Black male subgroup has a very high base rate of recidivism and the highest False Positive Rate relative to other subgroups. The accuracy metrics here also matched that of white males. Such comparisons inform users to rework their models to produce more equitable results.

## Discussion

Algorithms were adopted in criminal justice systems with an expectation of neutrality and today judges in eleven US states and an additional 178 countries use some form of risk assessment. Neutrality of such algorithms however were probed across disciplines partly due to the desired sanctity of judicial systems in society. However, most of such examinations have been limited by their methods of investigation and/ or disciplinary boundaries. We recognise this isolation in two simultaneously evolving strands of scholarship and one of the primary aims of this review is to map this disjunction.

Systematic review of the legal papers makes it evident that the literature on the legal side focuses on diverse concerns regarding use of algorithmic tools in the criminal justice system. They primarily discuss concerns relating to reinforcement of the inherent existing societal biases in the input data employed to train the algorithms, issues observable in the output generated, and legal issues inviting challenges before a court of law. Legal papers attempt to provide some solutions and workarounds to address some of the highlighted concerns. The importance of public representation in the construction and implementation of algorithmic tools and providing training to judges on the working of RATs are some of the other important workarounds suggested. A few of the legal studies also highlight the benefits of adopting algorithmic tools in the criminal legal system. Speed, efficiency, consistency, and addressing inherent human biases are some of the positive dimensions highlighted in the legal literature.

As noted, the empirical literature was spurred by Dressel and Farid's (2018) experiment and ProPublica's findings on COMPAS. The literature does not prescribe imprudent employment of algorithms. However, there is a disproportionate anchoring (and use of diverse methods) to investigate technical features and fairness implications. This also implies that there is limited empirical research on topics like comparative predictive accuracy of algorithms and humans or the effects of algorithm-human interaction. Future empirical research should shift away from the excessive focus on pushing up RAT prediction metrics. Evidence from human-automation studies suggest that optimizing an automated aid's performance will not always lead to best performance of the "human-automated team" (Sorkin & Woods 1985). Judges have also been found to misinterpret RAT outputs

(Yacoby et al. 2022) and machines to slightly bias individuals towards predicting no recidivism (Grgic-Hlaca et al. 2019). Algorithmic and human predictions also specifically depend on the characteristics of the data. Future studies on the goodness of either, should be wary of making broad claims from one exercise. Research employing diverse methods and data for each concern may reveal different perspectives.

RATs place excessive focus on individual psychology for criminal behavior, diverting attention from structural factors and societal responsibility. The fixation of the literature on whether algorithms are racially unfair sidelines research into other input features that make them unfair. Technical research in recent years has progressed to provide algorithmic solutions to algorithm generated problems, but second order algorithmic solutions risk the creation of problems anew. With this characterization, we may ask, if algorithmization of judicial systems is inevitable (Volokh, n.d.) then how can we ensure that using them does not mechanically edge society towards what's undesirable. As fast paced technical advancements already make promising developments in prediction beyond ML models, comprehensive evaluations of such systems need to keep up to avoid the perils of techno-solutionism (Shih et al. 2019).

Both legal and empirical scholarship discuss transparency concerns. Lack of transparency around the model design, variables, and training data have raised questions and recommendations from the legal scholars. The technical literature has discussed different methods to make black-box algorithms interpretable and the inefficiencies of these methods. Another common theme emerging from both strands of literature are reliability concerns associated with the use of algorithmic tools. Lack of a standard risk level categorization and human-machine interaction have been highlighted by legal scholars as giving rise to confusing standards of reliability.

A key issue for future research likely to be of interest to legal academics, and of enormous practical significance, is the roles of lawyers via-a-vis in the proliferation of RATs. Research on gauging legal practitioners' perspective when it comes to incorporation of algorithmic decision-making in judicial functions can also immensely contribute in extending the literature on this topic. Future work could also consider exploring tangible solutions to address the use of a single composite score for determination of different outcomes (for instance, failure to appear and risk of reoffending) by RATs. Finally, issues of variable choice, inability to draw clear causal models and inconsistencies in prediction are some of the open issues which need to be tackled by the technical literature.

## Limitations

While we are able to answer our broad research questions, we acknowledge the constraints resulting from our methodological choices. Here we discuss specific limitations of our approach and note formative studies that could not be included due to methodological constraints.

Our search keywords have been framed to focus solely on acquiring papers related to algorithmic RATs, overlooking other methodological dimensions of risk assessment. Some RATs predate algorithmic and statistical methods. While it is imperative to study and actively derive from earlier scholarship on such tools, our paper has a more narrow agenda. Our principal concern is to review the interdisciplinary scholarship that has risen after ProPublica published *Machine Bias* and the literature existing shortly preceding it.

There have been quite a few empirical studies focussing on the interactions between human judges and risk assessment tools, which couldn't be included. For example, Pruss (2023) conducted an interview-based impact evaluation of RATs after implementation in Pennsylvania and observed that judges frequently ignore the recommendations offered owing to organizational factors, including systemic issues related to information dissemination by the tools. Experimental evidence from Green and Chen (2019) and Fogliato et. al. (2021) also raises concerns regarding fairness and efficacy of RATs. Both studies argue that employing RATs and achieving objective judgements will require deeper evaluations of RAT-human interaction beyond maximizing algorithm accuracy due to disparate human behavior while interacting with RATs. Brayne & Christin (2020) conducted an ethnographic fieldwork to assess the effects of deployment in a criminal court and police department at varying stages of the criminal justice system. Their findings showed instances of professional resistance arising out of concerns, including fears of managerial surveillance, devaluation of professional experience and deskilling arising out of technocratic oversight and reliance on big data analytics. However, these findings could not be included in our results due to the methodological constraints.

## References


Adler, J.; Picard, S.; and Flood. C. 2019. Arguing the Algorithm: Pretrial Risk Assessment and the Zealous Defender. *Cardozo Journal of Conflict Resolution* 21:581-596.

Anderson, C. 2020. Risk Assessment Instruments Are Inappropriate for Sentence Reform: Real Solutions for Reform Address Racial Stratification. *Georgetown Journal of Law & Model Critical Race Perspectives* 12(2):187-206.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. *Ethics of data and analytics* 254-264.

Bagaric, M.; Svilar, J.; Bull, M; Hunter, D.; and Stobbs, N. 2022. The Solution to the Pervasive Bias and Discrimination in the Criminal Justice System: Transparent and Fair Artificial Intelligence. *American Criminal Law Review* 59(1):95-148.

Bagaric, M.; and Wolf, G. 2017. Sentencing by Computer: Enhancing Sentencing Transparency and Predictability, and (possibly) Bridging the Gap Between Sentencing Knowledge and Practice. *Georgetown Mason Law Review* 25(4):653-703.

Bansak, K. 2019. Can Nonexperts Really Emulate Statistical Learning Methods? A Comment on "The Accuracy, Fairness, and Limits of Predicting Recidivism. *Political Analysis 27*(3): 370-380. https://doi.org/10.1017/pan.2018.55

Beyond Intent: Establishing Discriminatory Purpose in Algorithmic Risk Assessment. 2021. *Harvard Law Review* 134(5):1760-1781.

Biswas, A.; Kolczynska, M.; Rantanen, S.; and Rozenshtein, P. 2020. The Role of In-group Bias and Balanced data: A Comparison of Human and Machine Recidivism Risk Predictions. In Proceedings of the 3rd Association for Computing Machinery SIGCAS Conference on Computing and Sustainable Societies 97-104. https://doi.org/10.1145/3378393.3402507.

Biswas, A.; Mukherjee, S. 2021. Ensuring Fairness Under Prior Probability Shifts. In Proceedings of the Association for the Advancement of Artificial Intelligence/Association for Computing Machinery Conference on AI, Ethics, and Society. https://doi.org/10.1145/3461702.3462596.

Brayne, C.; Christin, A. 2020. Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts, *Social Problems* 68(3): 608–624. https://doi.org/10.1093/socpro/spaa004.

Cabrera A. Á.; Epperson, W.; Hohman, F.; Kahng, M.; Morgenstern, J.; and Chau, D. H. 2019. FairVis: Visual Analytics for Discovering Intersectional Bias in Machine Learning. *In Institute of Electrical and Electronics Engineers Conference on Visual Analytics Science and Technology (VAST)* 46-56. https://doi.org/10.1109/VAST47406.2019.8986948.

Calmon, F.; Wei, D.; Ramamurthy, K. R.; and Varshney, K. R. 2017. Optimized Pre-Processing for Discrimination Prevention. *In Advances in neural information processing systems*.

Carlson, M. A. 2017. The Need for Transparency in the Age of Predictive Sentencing Algorithms. *Iowa Law Review* 103(1):303-330.

Coglianese, C.; and Ben Dor, M. L. 2021. AI in Adjudication and Administration. *Brook. Law Review* 86(3):791-838.

[15] Collins, E. 2018. Punishing Risk. *Georgetown Law Journal* 107:57-108.

Coston, A.; Mishler, A.; Kennedy H. E.; and Chouldechova, A. 2020. Counterfactual Risk Assessments, Evaluation, and Fairness. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 582-593.

Cyphert, B. A. 2020. Reprogramming Recidivism: The First Step Act and Algorithmic Prediction of Risk, *Seton Hall Law Review* 51(2):331-382.

Dalakian II, J. G. 2018. Open the Jail Cell Doors, Hal: A Guarded Embrace of Pretrial Risk Assessment Instruments. *Fordham Law Review* 87(1):325-370.

Deskus, C. 2018. Fifth Amendment Limitations on Criminal Algorithmic Decision-Making. *New York University Journal Legislation & Public Policy* 21(1): 237-286.

Desmarais, L. S.; Monahan, J.; and Austin, J. 2021. The Empirical Case for Pretrial Risk Assessment Instruments. *Criminal Justice & Behavior* 49(6):807-816. https://doi.org/10.1177/00938548211041651

DiBenedetto, R. 2019. Reducing Recidivism or Misclassifying Offenders? How Implementing Risk and Needs Assessment in the Federal Prison Will Perpetuate Racial Bias. *Journal of Law & Policy* 27(2):414-452.

Dijck, V.G.V. 2022. Predicting Recidivism Risk Meets AI Act. *European Journal on Criminal Policy & Research.* 28 (3): 407-424. https://doi.org/10.1007/s10610-022-09516-8

Donohue, E. M. 2019. A Replacement for Justitia's scales? Machine learning's role in sentencing. *Harvard Journal of Law & Technology* 32(2):657-678.

Dressel, J.; and Farid, H. 2018. The Accuracy, Fairness, and Limits of Predicting Recidivism. *Science advances* 4, 1 https://doi.org/10.1126/sciadv.aao5580.

Duwe, G. 2019. Better Practices in the Development and Validation of Recidivism Risk Assessments: The Minnesota Sex Offender Screening Tool–4. *Criminal Justice Policy Review* 30(4):538-564. https://doi.org/10.1177/0887403417718608.

Duwe, G.; and Kim, K. 2017. Out With the Old and in With the New? An Empirical Comparison of Supervised Learning Algorithms to Predict Recidivism. *Criminal Justice Policy Review* 28(6): 570-600. https://doi.org/10.1177/0887403415604899

Eaglin, M. J. 2017. Constructing Recidivism Risk. *Emory Law Journal* 67(1):59-122.

Eaglin, M. J. 2019. Technologically Distorted Conceptions of Punishment Technologically Distorted Conceptions of Punishment. *Washington University Law Review* 97:483-543.

Eckhouse, L.; Lum, K; Conti-Cook, C.; and Ciccolini, J. 2018. A Unified Approach for Understanding Problems With Risk assessment. *Criminal Justice and Behavior* 20 (10):1-25. 10.1177/0093854818811379

Elyounes, A. D. 2020. Bail Or Jail? Judicial Versus Algorithmic Decision-Making in the Pretrial System. *Columbia Science & Technology Law Review* 21(2):376-446.

Garrett, L. B.; and Monahan, J. 2020. Judging Risk. *California Law Review* 108:439-493. DOI: https://doi.org/10.15779/Z38B56D515.

Ghasemi, M.; Anvari, D.; Atapour, M.; Wormith, J. S.; Stockdale, K. C.; and Spiteri, R. J. 2021. The Application of Machine Learning to a General Risk–Need Assessment Instrument in the Prediction of Criminal Recidivism. *Criminal Justice and Behavior* 48, 4, 518-538. https://doi.org/10.1177/0093854820969753

Gravett, W. 2021. Jailed by a "Black Box": The Impact of Opaque Algorithms on the Right to a Fair Trial in the United States of America. *Tydskrif vir hedendaagse Romeins-Hollandse*. 84(3): 299-317.

Gravett, W. 2021. Sentenced by an Algorithm-Bias and lack of Accuracy in Risk Assessment Software in the United States Criminal Justice System. *Journal of Criminal Justice* 34(1):31-54. https://doi.org/10.47348/SACJ/v34/i1a2.

Greene, C. 2022. AI and the Social Sciences: Why All Variables are Not Created Equal. *Res Publica* 1-17. https://doi.org/10.1007/s11158-022-09544-5.

Greene, T.; Shmueli, G.; Fell, J.; Lin, C.; and Liu, H. 2022. Forks over knives: Predictive Inconsistency in Criminal Justice Algorith-

mic Risk Assessment Tools. *Journal of the Royal Statistical Society Series A: Statistics in Society* 185:2 https://doi.org/10.1111/rssa.12966.

Grgić-Hlača, N.; Engel, C.; and Gummadi, K. P. 2019. Human decision making with machine assistance: An Experiment on Bailing and Jailing. In Proceedings of the Association for Computing Machinery on Human-Computer Interaction 3, 1-25. https://doi.org/10.2139/ssrn.3465622.

Hamilton, M. 2019. The Sexist Algorithm. *Behavioral sciences & the law* 37(2): 145-157. https://doi.org/10.1002/bsl.2406.

Hamilton, M. 2021. Algorithmic Risk Assessment: A Progressive Policy in Pretrial Release. *Idaho Law Review* 57 (3):615-634.

Hill, II A. S. 2021. Bail Reform and the (False) Racial Promise of Algorithmic Risk Assessment. *UCLA Law Review* 68(4):910-987.

Humerick, D. J. 2020. Reprogramming Fairness: Affirmative Action in Algorithmic Criminal Sentencing. *Columbia Human Rights Law Review* 4(2):213-245.

Hunter, D.; Bagaric, M.; and Stobbs, N. 2019. A Framework for the Efficient and Ethical Use of Artificial Intelligence in the Criminal Justice System. *Florida State University Law Review* 47: 749-800.

Johndrow, J. E.; and Lum, K. 2019. An Algorithm for Removing Sensitive Information. *The Annals of Applied Statistics* 13(1): 189-220.

Jones, N. C. 2020. A Broken PATTERN: A Look at the Flawed Risk and Needs Assessment Tool of the First Step Act. *Howard Human & Civil Rights Law Review* 5:185.

Karimi-Haghighi, M.; and Castillo, C. 2021. Enhancing a Recidivism Prediction Tool with Machine Learning: Effectiveness and Algorithmic Fairness. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, 210-214. https://doi.org/10.1145/3462757.3466150.

Kehl, D.; Guo, P.; and Kessler, S. 2017. Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. *Responsive Communities Initiative, Berkman Klein Center for Internet & Society*, Harvard Law School. 1-36.

Kigerl, A.; Hamilton, Z.; Kowalski, M.; and Mein, X. 2022. The great methods bake-off: Comparing Performance of Machine Learning Algorithms. *Journal of Criminal Justice* 82. https://doi.org/10.1016/j.jcrimjus.2022.101946.

Krent, J. H.; and Rucker, R. 2022. The First Step Act - Constitutionalizing Prison Release Policies. *Rutgers University Law Review* 74:631-680.

Leng, J.; Xu, W.; Tonghong, L.; Chen, L.; and Xu, M. 2022. A Prediction Model of Recidivism of Specific Populations Based on Big Data. *Wireless Communications and Mobile Computing.* https://doi.org/10.1155/2022/9167590.

Lewis, C. 2022. Risk-based Sentencing and the Principles of Punishment. *Journal of Criminal Law & Criminology* 112(2):213-264.

Li, N.; Goel, N.; and Ash, E. 2022. Data-centric Factors in Algorithmic Fairness. In Proceedings of the Association for the Advancement of Artificial Intelligence/ Association for Computing Machinery Conference on AI, Ethics, and Society, 396-410. https://doi.org/10.1145/3514094.3534147.

Lin, Z. J.; Jung, J.; Goel, S.; and Skeem, J. 2020, The Limits of Human Predictions of Recidivism. *Science advances* 6. https://doi.org/10.1126/sciadv.aaz0652.

Lowden, F. R. 2018. Risk Assessment Algorithms: The Answer to an Inequitable Bail System? *North Carolina Journal of Law & Technology* 19(4):221-251.

Ludwig, J.; and Mullainathan, S. 2021. Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System. *The Journal of Economic Perspectives* 35(4):71-96.

Ma, Y.; Nakamura, K.; Lee, E.; and Bhattacharyya, S. S. 2022 EADTC: An Approach to Interpretable and Accurate Crime Prediction. In Institute of Electrical and Electronics Engineers International Conference on Systems, Man, and Cybernetics 170-177. https://doi.org/10.1109/SMC53654.2022.9945130

Marius, M.; Tolan, S.; Gómez, E.; and Castillo, C. 2021. Evaluating Causes of Algorithmic Bias in Juvenile Criminal Recidivism. *Artificial Intelligence and Law* 29(2):111-147. https://doi.org/10.1007/s10506-020-09268-y

Mayson, G. S. 2019. Bias In, Bias Out. *Yale Law Journal* 128(8): 2122-2473.

Mbadiwe, T. 2018. Algorithmic injustice. *The New Atlantis* 54. 3-28.

Meyer, M.; Horowitz, A.; Marshall, E.; and Lum, K. 2022. Flipping the Script on Criminal Justice Risk Assessment: An Actuarial Model for Assessing the Risk the Federal Sentencing System Poses to Defendants. In Association for Computing Machinery Conference on Fairness, Accountability, and Transparency, 366-378. https://doi.org/10.1145/3531146.3533104

Michael, B.; Gersen, S. G.; Haley, M.; Lin, M.; Merchant, A.; Millett, R.J.; Sarkar, S. K.; and Wagner, D. 2020. Constitutional Dimensions of Predictive Algorithms in Criminal Justice. *Harvard Civil Rights - Civil Liberties Law Review* 55:267-310.

Mishler, A.; Kennedy, E. H.; and Chouldechova, A. 2021. Fairness in Risk Assessment Instruments: Post-Processing to achieve Counterfactual Equalized Odds. In Proceedings of the 2021 Association for Computing Machinery Conference on Fairness, Accountability, and Transparency, 386-400.

Monahan, J.; and Skeem, L. J. 2016. Risk Assessment in Criminal Sentencing. *Annual Review Clinical Psychology* 12:489-513.

Morgan, Andrew.; and Pass, R. 2019. Paradoxes in Fair Computer-Aided Decision Making. In Proceedings of the 2019 Association for the Advancement of Artificial Intelligence/ Association for Computing Machinery Conference on AI, Ethics, and Society 85-90. https://doi.org/10.1145/3306618.3314242

Muenster, E. 2022. The First Step Act Took One Step Forward and Two Steps Back. *Houston Law Review* 60(1):135-174.

Novokmet, A.; Tomicic, Z.; and Vinkovic, Z. 2022. Pretrial Risk Assessment Instruments in the US Criminal Justice System—What Lessons can be Learned for the European Union. *International Journal of Law and Information Technology* 30(1):1-22. https://doi.org/10.1093/ijlit/eaac006

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science* 366 (6464): 447-453. https://doi.org/10.1126/science.aax2342

O'Brien, T. 2021. Compounding Injustice: The Cascading Effect of Algorithmic Bias in Risk Assessments. *Georgetown Journal of Law & Model Critical Race Perspectives* 13(1):39-84.

Okidegbe, N. 2022. Discredited Data. *Cornell Law Review* 107: 2007-2066.

Okidegbe, N. 2022. The Democratizing Potential of Algorithms. *Connecticut Law Review* 53:739-789.

Oleson, J.C. 2011. Risk in Sentencing: Constitutionally Suspect Variables and Evidence-based Sentencing. *South Methodist University Law Rev* 64:1329-1402.

Pruss, D. 2023. Ghosting the Machine: Judicial Resistance to a Recidivism Risk Assessment Instrument. In Proceedings of the 2023 Association for Computing Machinery Conference on Fairness, Accountability, and Transparency 312–323. https://doi.org/10.1145/3593013.3593999

Rachlinski, J. J. 2010. Evidence-based law. *Cornell Law Review* 96: 901.

Rachlinski, J. J.; Johnson, S.; Wistrich, J. A.; and Guthrie, C. 2008. Does Unconscious Racial Bias Affect Trial Judges? *Notre Dame Law Review* 84:1195.

Rankin, G. M. S. 2021. Technological Tethereds: Potential Impact of Untrustworthy Artificial Intelligence in Criminal Justice Risk Assessment Instruments. *Washington & Lee Law Review* 78(2):647-724.

Rizer, A. 2021. Artificial Intelligence and Risk Assessment Tools: Problems and Solutions. *Washburn Law Journal* 60(3):495-510.

Rizer, A.; and Watney, C. 2018. Artificial Intelligence Can Make Our Jail System More Efficient, Equitable, and Just. *Texas Review of Law & Policy* 23(1):181-228.

Rudin, C. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and use Interpretable Models Instead. *Nature machine intelligence* 5(1):206-215. https://doi.org/10.1038/s42256-019-0048-x

Sandburg, J. I.; Skardhamar, T.; Kristoffersen, R.; and Friestad, C. 2021. Testing the Static-99R as a Global Screen for Risk of Sex Crime Recidivism in a Norwegian Routine Rample. *Sexual Abuse* 33(6):725-742. https://doi.org/10.1177/1079063220951194

Seglias, G. 2021. Bias and Discrimination in Opaque Automated Individual Risk Assessment Systems: Challenges for Judicial Review under the Equality Act 2010. *Oxford University Undergraduate Law Journal* 53-79.

Shi, J. 2022. Artificial Intelligence, Algorithms and Sentencing in Chinese Criminal Justice: Problems and Solutions. *Criminal Law Forum* 121-148. https://doi.org/10.1007/s10609-022-09437-5

Shih, P.; Chiu, C.; and Chou, C. 2019. Using Dynamic Adjusting NGHS-ANN for Predicting the Recidivism Rate of Commuted Prisoners. *Mathematics* 7(12), 1187. https://doi.org/10.3390/math7121187.

Skeem, J.; and Lowenkamp, C. 2020. Using Algorithms to Address Trade-offs Inherent in Predicting Recidivism. *Behavioral Sciences & the Law* 38(3):259-278. https://doi.org/10.1002/bsl.2465.

Slobogin, C. 2013. Risk Assessment and Risk Management in Juvenile Justice. *Criminal Justice* 27(4):10-25.

Slobogin, C. 2018. Principles of Risk Assessment: Sentencing and Policing. *Ohio State Journal of Criminal Law* 15(2):583-596.

Slobogin, C. 2021. Preventive Justice: How Algorithms Parole Boards, and Limiting Retributivism Could End Mass Incarceration Retributivism Could End Mass Incarceration. *Wake Forest Law Review* 56(1):97-168.

Sorkin, R. D.; and Woods, D. D. 1985. Systems with Human Monitors: A Signal Detection Analysis. *Human-computer interaction* 1(1):49-75. https://doi.org/10.1207/s15327051hci0101_2

Southerland, M. V. 2021. The Intersection of Race and Algorithmic Tools in the Criminal Legal System. *Maryland. L. Rev.* 80 (3): 487-564.

Starr, B. S. 2014. Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Review* 66(4):803-871.

Starr, S. 2016. Actuarial Risk Prediction and the Criminal Justice System. *The Odds of Justice.* 29(1):49-51. https://doi.org/10.1080/09332480.2016.1156368

Stevenson, M. 2018. Assessing Risk Assessment. Action. *Minnesota Law Review* 103(1):303-384

Stevenson, T. M.; and Slobogin, C. 2022. Algorithmic Risk Assessments and the Double-edged Sword of Youth. *Behavioral sciences & the law* 36:638-656.

Stewart, R. T. 2022. Identity and the Limits of Fair Assessment. *Journal of Theoretical Politics* 34(3):415-442. https://doi.org/10.1177/09516298221102972.

Thomas, C.; and Pontón-Núñez, A. 2022. Automating Judicial Discretion: How Algorithmic Risk Assessments in Pretrial Adjudications Violate Equal Protection Rights on the Basis of Race. *Minnesota Journal of Law & Inequality* 40(2): 370-407.

Ting Ming, H.; Chu, C. M.; Zeng, G.; Li, D.; and Chng, G. S. 2018. Predicting Recidivism Among Youth Offenders: Augmenting Professional Judgment with Machine Learning Algorithms. *Journal of Social Work* 18(6), 631-649. https://doi.org/10.1177/1468017317743137.

Tolan, S.; Miron, M.; Gómez, E.; and Castillo, C. 2019. Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia. In Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, 83-92. https://doi.org/10.1145/3322640.3326705

Tonry, M. 2019. Predictions of Dangerousness in Sentencing: Déjà Vu All Over Again. *Criminal and Justice —A Review of Research* 48:439-482. https://doi.org/10.1086/701895

Van Berkel, N.; Goncalves, J.; Hettiachchi, D.; Wijenayake, S.; Kelly, R. M.; and Kostakos, V. 2019. Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study. In Proceedings of the Association for Computing Machinery on Human-Computer Interaction 3, 1-21.

Villasenor, J.; and Foggo, V. 2020. Artificial Intelligence, Due Process and Criminal Sentencing. *Michigan State Law Review* 2020(2):295-354.

Vincent, M. G.; and Viljoen, L. J. 2020. Racist Algorithms or Systemic Problems? Risk Assessments and Racial Disparities. *Criminal Justice and Behavior* 47(12):1576–1584. https://doi.org/10.1177/0093854820954501

Volokh, E. 2018. Chief justice robots. *Duke Law Journal* 68:1135.

Washington, L. A. 2019. How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate. *Colorado Technology Law Journal* 17(1):131-161.

Wexler, R. 2018. Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System. *Stanford Law Review* 70:1343-1429.

Wisser, L. 2019. Pandora's Algorithmic Black Box: The Challenges of Using Algorithmic Risk Assessments in Sentencing. *American Criminal Law Review* 56(4):1811-1832.

Yacoby, Y.; Green, B.; Griffin Jr, C. L.; and Doshi-Velez, F. 2022. "If it Didn't Happen, Why would I Change my Decision?": How Judges Respond to Counterfactual Explanations for the Public Safety Assessment. In Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Human Computation and Crowdsourcing, 10(1):219-230 https://doi.org/10.1609/hcomp.v10i1.22001.

Završnik, A. 2021. Algorithmic justice: Algorithms and Big Data in Criminal Justice Settings. *European Journal of criminology* 18(5):623-642.

Zhang, Y.; and Ramesh, A. Learning Fairness-aware Relational Structures. 2020. In European Conference on Artificial Intelligence.