SOCIAL SCIENCE REVIEW ARCHIVES
https://policyjournalofms.com
ISSN Online: 3006-4708
ISSN Print: 3006-4694

The Role of Artificial Intelligence in Criminal Investigations: Challenges of Bias, Transparency, and Legal Accountability

Syed Abbas Ali Shah

[1] Department of Law, University of Lahore. Email: syedabbasadvocate@gmail.com

DOI: https://doi.org/10.70670/sra.v1i2.2669

Abstract

From facial recognition to predictive policing, automated analysis of digital evidence and probabilistic forensic software to risk scoring, AI has become part of the day-to-day of criminal investigation. They can deliver speed and uniformity, but they're also been responsible for misidentifying suspects, issuing incorrect sentences, and controversy over who's to blame when an algorithm is used to produce an unfair result. This article focuses on three related issues – bias, transparency, and legal accountability. It examines empirical studies, doctrinal writing, and regulatory documents from the United States, the United Kingdom, and the European Union and finds that these issues are not isolated flaws, each of which warrants a remedy in isolation. Bias creeps in at each stage of the investigative process, opacity makes it hard for defendants and courts to identify it, and current legal doctrine assigns blame at the right moments where there are holes in the investigative process. The article proposes a 3-phase accountability model (ex ante authorisation and validation, in operation oversight, ex post disclosure and remedy) and specifies the safeguards needed for each phase. It holds that the legality of the use of AI in investigations is not tied to the accuracy of any one intelligence system, but to maintaining the capacity of people to challenge the product of the intelligence systems.

Keywords: Artificial Intelligence, Criminal Investigation, Algorithmic Bias, Transparency, Explainability, Legal Accountability, Facial Recognition, Predictive Policing

Introduction

A man is arrested at his home in front of his family by police officers in Detroit in January 2020 after his driver's license photo is captured on a blurred surveillance camera of a robbery suspect. He didn't wear the video guy. He was detained for hours without anyone realizing the error and his case was one of the early wrongful arrests that came to be associated with the use of facial recognition in the United States (Hill, 2020). The story is memorable because it is concrete and it's also capturing a structural pattern. The output of an automatic system became a lead, then became grounds for action for human officers; none of the human officers could easily explain why the machine had targeted this man rather than another.

This type of occurrence has been increasing with the implementation of data-driven tools by the police and prosecutors. Facial recognition is just the most obvious use. Police departments have trialed software designed to predict where crime will happen, agencies have tried using statistical instruments to estimate whether a defendant will reoffend, labs use probabilistic software to help decipher mixed up DNA samples, and investigators rely on automated tools to sift through phones,

messages, and financial records that are too vast for a human team to comb (Ferguson, 2017; Brayne, 2020). These systems can help mitigate human bias, reveal trends that investigators would fail to see, and introduce uniformity and consistency to decisions that have previously been inconsistent (Kleinberg et al., 2018). Critics point out that the same systems can also be engineered to enforce historical discrimination in mathematical power and to conceal justifications for trade secrecy (Pasquale, 2015; O'Neil, 2016).

Both assertions are taken seriously in this article. It doesn't say that algorithmic tools are "evil," or worse than human judgment; human investigators are also fallible and subject to human bias. It questions rather than answers what sets apart the changes in the modus operandi of criminal investigation brought by artificial intelligence and why the legislation has not reacted accordingly. The overarching idea is of three problems interacting. Bias is what dictates the fairness of the tool. Transparency determines if anybody can discover it. The issue of legal accountability is whether or not someone is legally responsible to take some sort of action. A system that is biased but transparent can be challenged. An opaqueness and neutrality of a system can be illegitimate in a process that is supposed to provide reasons. The worst case scenario is for bias, opacity, and diffuse responsibility to reinforce each other.

The article makes three contributions. First, it charts the existing tools used in criminal investigations and notes where they get caught in the investigative process. Second, it collates the empirical evidence on bias and performance, provides comparative data from the key audits and demonstrates that there are competing definitions of fairness and they cannot be met simultaneously. Third, it looks at how the laws can address the issue under three different jurisdictions and puts forward a model to link the protection to the stage of the investigation where it can apply.

The following is an outline of the remainder of the article. In Section 2, the conceptual and theoretical framework is developed. The technologies used are surveyed in section 3. The research design is discussed in Section 4. Section 5 analyses bias. Transparency and Explainability are covered in Section 6. Legal accountability is explored in Section 7. The suggested framework is presented in Section 8. Implications are discussed in Section 9 and the text concludes in Section 10.

Conceptual and Theoretical Framework

Defining Artificial Intelligence in the Investigative Setting

There are many techniques that fall under the umbrella of AI, and they don't all seem to be cause for concern. For the purposes of this document, the relevant system is one that relies on statistical learning and/or rule-based inference to produce outputs, including predictions, classifications, rankings, or matches, that are used by investigators or courts to make a decision about a person. The important point is not that it is technically sophisticated but rather what the output does. The reordering of a list of suspects is a decision-influencing system even when its designers represent it as being advisory (Joh, 2016). Mittelstadt et al. (2016) helpfully differentiate concerns regarding the evidence that inform the use of an algorithm, the fairness of its results, and the traceability of responsibility for those results, and here we will distinguish among these three considerations.

Algorithmic Bias

In this context, bias may have a few meanings that need to be differentiated. Statistical bias is systematic error, such as a face-matching algorithm that has a higher rate of false matches for certain demographic groups than other groups. Legal bias – bias as discrimination – is based upon whether there is an unjustified difference of treatment or disparate impact for protected groups (Barocas & Selbst, 2016). Social bias is the process of encoding of historical patterns of inequality

into the data that seem neutral. While these categories overlap, they are not the same. Overall, a model can be well calibrated, while at the same time capturing unequal enforcement, as the model is trained on data that reports on police stops, not crime (Lum & Isaac, 2016; Richardson et al., 2019).

This is captured in a phrase by Mayson (2019): "bias in, bias out". The data used for training a predictive system may be based on some biased decision making process, which will lead the system to learn the pattern and keep on doing it. The ensuing arrests produce new data, which helps confirm a prediction, and thus the process continues on. Ensign et al. (2018) formalize this as a runaway feedback loop in predictive policing, as a small initial discrepancy in crime counts can increase because cops focus their patrol efforts in areas with past unusually high crime counts.

Transparency, Explainability, and Contestability

There are at least three levels of transparency. The first is disclosure of some sort that a system was used at all. The second is the technical information regarding its usage such as training data, validation studies and error rates. The third is explanation, the capacity to explain to someone why the system gave this output for this case. Burrell (2016) mentions three causes of opacity: first, the deliberate concealment of information by the owners; second, the general lack of technical expertise on the part of those affected; and third, the complexity of machine learning models working at scales that are beyond human understanding. Each requires a different cure.

What it means to be explainable is open to debate. Some scholars propose post hoc explanation methods which are close to what a complex model does. For decisions with high stakes like those in criminal justice, it is preferable to have models that are inherently interpretable, rather than relying on an explanation of black box models, as the latter can be misleading (Rudin 2019). Wachter et al. (2018) suggest counterfactual explanations that narrate to an affected individual what was different in the past to result in a different outcome, which can be used to contest without allowing proprietary code to be leaked. The thing that is learnt from practice is that explanation is not just a technical characteristic. It's a system, an audience, and a purpose, and what constitutes sufficient will vary between the police supervisor, the judge, and the defendant.

Accountability and Due Process

For accountability to take place, there must be someone identified as accountable for a decision, there must be a need to account for, there must be to account for the reasons, and there must be consequences when a reason is failed. Early, Citron (2008) wrote that automated administrative systems pose a threat to due process because they lack notice, hearing, and reasoned decision. She argued for technological due process, a concept which would be built into the design of automated decision-making systems. Kroll et al. (2017) further advanced this concept, suggesting the development of technical accountability of decision systems by designing the system to have features that enable outsiders to verify that rules used were those intended and that the system was run regularly. Citron and Pasquale (2014) cautioned that the use of scoring systems leads to generating a "scored society" that results in life chances becoming influenced by non-publicly communicated score.

Analytical Approach

Three analytical questions from the literature are used in this article. First, where does bias creep in and where is there evidence of it in the investigative process? Second, why is it impossible for people who are affected by it, and the courts, to find and challenge it? Thirdly, what are the legal principles and policy tools that hold responsible and where are there loopholes? These can be

considered as one single enquiry and the discussion can reveal how the weakness in one dimension can be exacerbated by a weakness in another dimension.

Artificial Intelligence in Criminal Investigations: An Overview

The Tools in Use

The table below presents key categories of AI-powered tools and the potential risks involved with each category. Some of the categories are more closely related to a specific suspect than others, are used earlier in the process than others and allow more discretion for the human decision maker to act than others.

Table 1

Principal AI-Assisted Tools in Criminal Investigation and Associated Risks

Tool	Investigative function	Principal risks	Illustrative sources
Facial recognition	Identifies or verifies persons from images or video	Demographic error disparities; wrongful arrest; mass surveillance	Buolamwini & Gebru (2018); Grother et al. (2019); Garvie et al. (2016)
Predictive policing	Forecasts places or persons at elevated risk of crime	Feedback loops; dirty data; discriminatory deployment	Lum & Isaac (2016); Ensign et al. (2018); Richardson et al. (2019)
Risk assessment instruments	Estimate likelihood of reoffending or failure to appear	Disparate error rates; proprietary secrecy; weak validation	Angwin et al. (2016); Dressel & Farid (2018); Rudin et al. (2020)
Probabilistic forensic software	Interprets complex DNA mixtures and other trace evidence	Undisclosed source code; unvalidated assumptions	National Research Council (2009); President's Council of Advisors on Science and Technology (2016)
Digital evidence analytics	Sorts, clusters, and flags content in seized devices and records	Overbroad searches; opaque ranking; privacy intrusion	Carpenter v. United States (2018); Ferguson (2017)

Note. The table is illustrative rather than exhaustive. Sources are those discussed in the corresponding sections of this article.

The most attention has been on facial recognition. In 2016, Garvie et al. found that the images of just about half of American adults were already available in police databases for law enforcement to search — many of which were compiled from driver's license records, without the individual's knowledge or consent. In contrast, predictive policing systems are based on information from aggregated locations, times, and previous incidents, with some predictive policing programs trying to draw on statistical information to identify individuals who are 'in the mix' for violence. (Brayne, 2020) The typical time to use risk assessment instruments is not at the point of initial investigation but at pretrial and sentencing. They are also all in the same family, in that they use past information about a group to predict information about an individual (Kehl et al., 2017).

Another set of questions is raised by forensic applications. A seminal National Research Council (NRC) review of forensic feature comparison (2009) revealed that the science behind many of these disciplines was not well established and that a recent report by the President's Council of Advisors on Science and Technology (2016) highlighted the importance of empirical evidence of

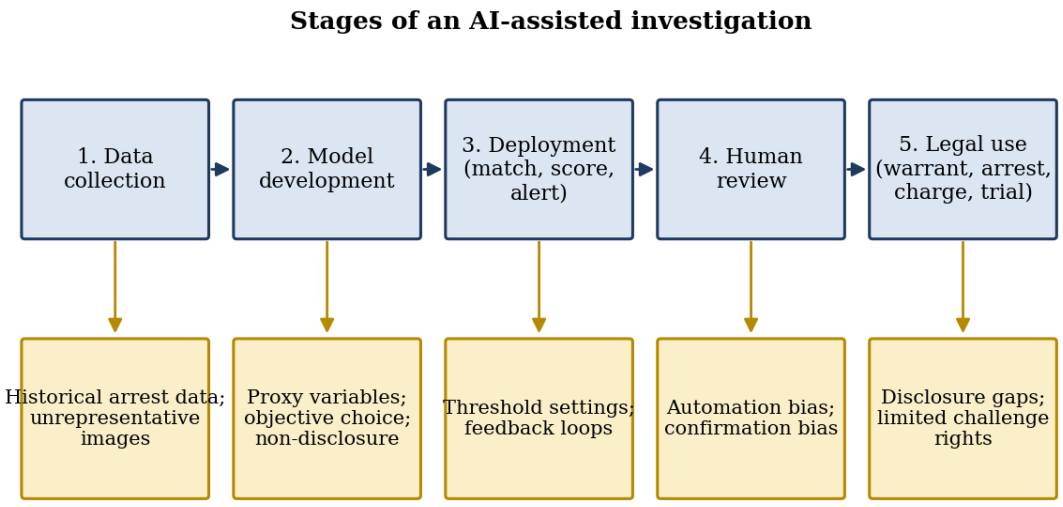
foundational validity and of validity as applied. The same issues apply to software that attempts to interpret complex mixtures; a software program can produce a plausible-seeming likelihood ratio whose validity depends on the assumptions that it makes, which few people outside of the software's authors have the ability to check.

Where Bias and Opacity Enter the Pipeline

It is helpful to consider AI-assisted investigation as a five-step process: Data collection, Model development, Deployment, Human review, and Legal use. These stages are presented in figure 1, with the points when problems typically occur. There are four ways bias can creep in: at the first, by inputting unrepresentative or historically biased data; at the second, by selecting an objective and by selecting proxy variables; at the third, by setting thresholds and by feeding back; at the fourth, by giving too much credence to outputs of the machine. Opacity has an impact at all stages, but at the fifth stage, when evidence or recommendations emerging from the system are used in a charging decision, a hearing, or a trial, it has a legal impact.

Figure 1

The Investigative Pipeline and Points of Vulnerability



Note. Conceptual model developed by the author from the literature reviewed in Sections 2 and 3.

The human review stage is worthy of special mention since it is often touted as "the catch all stage that makes the whole thing palatable. In practice, a human in the loop might provide some protection. Automated suspicion does not replace discretion, but rather complicates it, writes Joh (2016), and that officers who are alerted by an algorithm might take it as gospel. Ferguson (2012) demonstrates that the output of a predictive information model can interact with the legal concept of reasonable suspicion in a manner with little understanding in the courts, as it is possible for its output to be treated as an objective fact and not as a probabilistic estimate based on facts that may be disputed.

Research Design

This study is in the form of a structured narrative review, which is supplemented by a doctrinal analysis. The review is based on literature, including the peer-reviewed publications in the field of law, computer science, criminology and science and technology studies, as well as government and institutional reports. Sources were chosen because they appeared to have direct relevance to one of the three problems studied and because they would influence future scholarship and policy. The empirical evidence reported in Section 5 is based on three widely cited published audits: an intersectional audit of commercial gender classification systems (Buolamwini & Gebru, 2018); an audit of a commercial recidivism risk instrument (Angwin et al., 2016); and an experimental comparison between the recidivism risk instrument and human judgment and simple statistical models (Dressel & Farid, 2018). Numbers shown here are not new calculations, but are taken from those studies.

The doctrinal component covers some of the most significant decisions and instruments of legislation of the United States, the United Kingdom and the European Union. The jurisdictions selected are typical of three different regulatory approaches: prescriptive approach (European Union), common law approach (European Union) based on human rights and data protection principles, and a fragmented, case-driven approach (United States). The analysis does not aim at a comprehensive survey of the law and practice in this field, and it is aware that there is a rapid evolution of the law and practice in this area.

There are three restrictions to be noted. First, the audits reviewed are audits of certain systems at certain times, and the performance of existing systems could be different. Second, despite all the use of investigative AI, much of it takes place in practice without being reported. This means that what is published is likely to be an under-estimate of its use, and less of a failure. Third, the article does not try for estimation of a general effect of AI on the crime control, because it would require other research design. The purpose is to study the issues in terms of problem and the effectiveness of the solutions provided by the law.

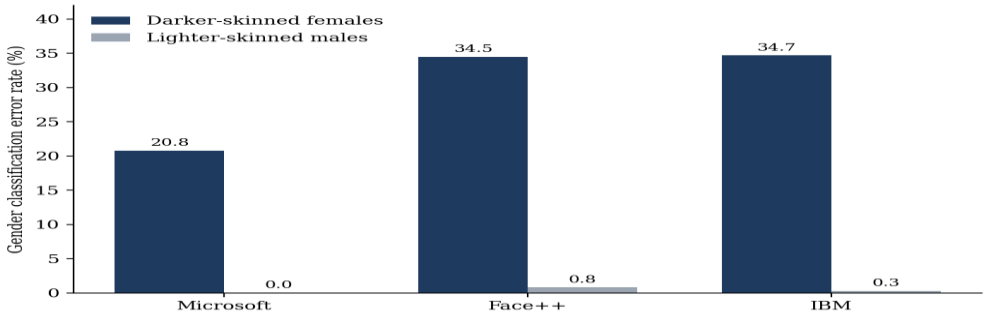
Bias in AI-Assisted Investigation

Disparities in Facial Analysis

Buolamwini and Gebru (2018) conducted the most impactful empirical study on the subject of demographic differences in commercial face analysis. The researchers developed a representative dataset of images that were balanced for skin colour and gender and evaluated three commercial gender classification systems. They discovered all three did much worse with darker skinned females than with lighter skinned males. The error rates are reported in Graph 1. The systems made less than one percent errors when applied to lighter-skinned men, and between 21 percent and nearly 35 percent errors when applied to darker-skinned women.

Graph 1

Gender Classification Error Rates by Intersectional Subgroup in Three Commercial Systems



Note. Data from Buolamwini and Gebru (2018). The task was gender classification rather than identity matching.

Two warnings should be given. The task examined was gender classification, rather than the one-to-many identity matching that police do, and so the numbers don't reflect the risk of arrest. The study revealed, however, a tendency for subgroup differences that would not have been apparent in one overall accuracy figure, and that the composition of the benchmarks and training data can yield very large differences in accuracy. In a one-to-one matching context, the National Institute of Standards and Technology found false positive rates to be significantly different across demographic groups, with for some groups' rates ten to hundred times larger than for other groups, though the magnitude of the effect differed widely between the algorithms (Grother et al. 2019). That variation matters. It implies that disparity is not intrinsic to the technology but exists in certain systems and data, and that procurement and validation choices can mitigate disparity.

In the policing context, a false positive is not a theoretical construct. A false match could lead officers to the door of an innocent person, as in the case of Detroit (Hill, 2020). Error rates based on the well-lit, whole image or from a favourable angle can underestimate the risk in the operational scenario where the images are poorly lit, partial, or from an unfavourable angle. Garvie et al. (2016) reported that only a few agencies had detailed testing protocols, and that many had a policy that allowed for testing in any crime.

Predictive Policing and Feedback Loops

One other way of bias is through predictive policing. Lum and Isaac (2016) tested the commercial predictive policing algorithm on drug crimes in Oakland with police data. They compared the pattern the algorithm would predict with estimates of actual drug use based on health survey data, and found that the software would focus drug driving patrols disproportionately on neighbourhoods with higher Black population numbers despite the fact drug use is more evenly distributed. The algorithm was training itself on crime locations on police records, not on actual crime locations.

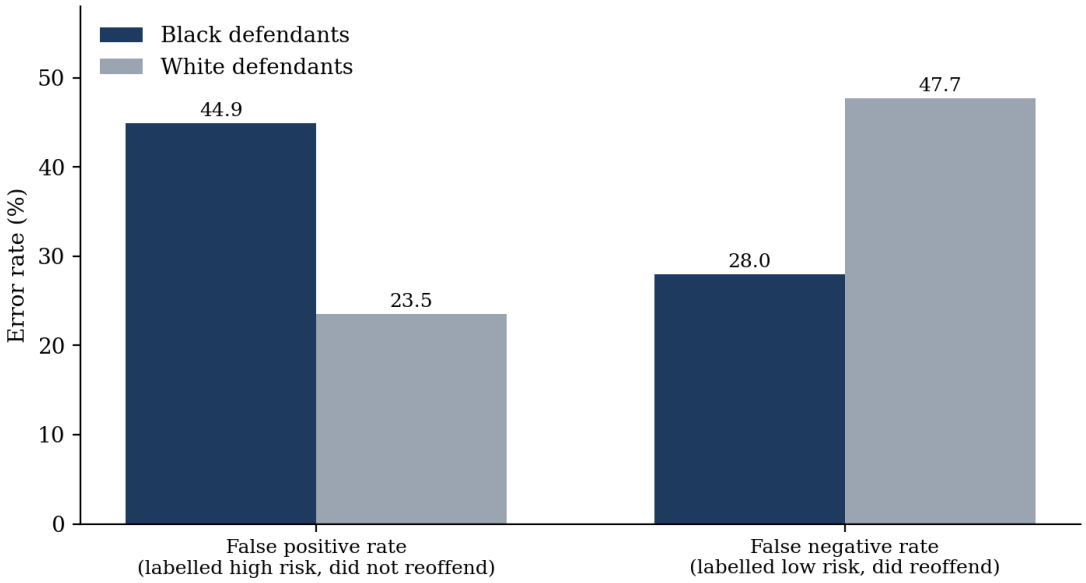
Ensign et al. (2018) analysed the dynamics theoretically. Where patrols are placed based on historic recorded incidents, and where patrols go effects the number of recorded incidents, this can cause a small difference to grow into a large one over time. Richardson et al. (2019) made an important institutional discovery. They used data from jurisdictions whose police departments were previously discovered to have committed illegal activities, and found that predictive systems were trained on data collected during times of documented police bad behavior, known as dirty data. It is as if the issue of data quality is not simply a technical one but one of institutional integrity.

Risk Assessment and the Problem of Competing Fairness Criteria

Perhaps the most popular debate on algorithmic fairness is that around the danger scoring of recidivism. Angwin et al. (2016) studied the scores given defendants by a popular evaluative tool in the Florida County of Broward and matched them with future criminal records. They found blacks defendants who had not reoffended were nearly two times more likely to be classified as higher risk than similar white defendants, and whites who did reoffend were more likely to be classified as lower risk. The figures are represented in graph 2.

Graph 2

Error Rates of a Recidivism Risk Instrument by Race (Broward County Analysis)



Note. Data from Angwin et al. (2016).

The developer said that the test had been calibrated and thus a score would represent a similarly high rate of reoffending in both groups. These two points may be true simultaneously. Chouldechova (2017) proved formally that if the base rates of recorded reoffending vary across groups, an imperfect instrument cannot meet calibration and equal error rates at the same time. Kleinberg et al. (2017) arrived at a similar impossibility result. The discovery doesn't solve the issue. It converts it from a technical question into a normative question of the choice of the criteria of fairness that can't be resolved by a statistician, and the choice should be made explicitly by decision makers whose role it is to be accountable. This also implies that the base rates themselves (reflecting the pattern of enforcement) are part of the moral question (Mayson, 2019).

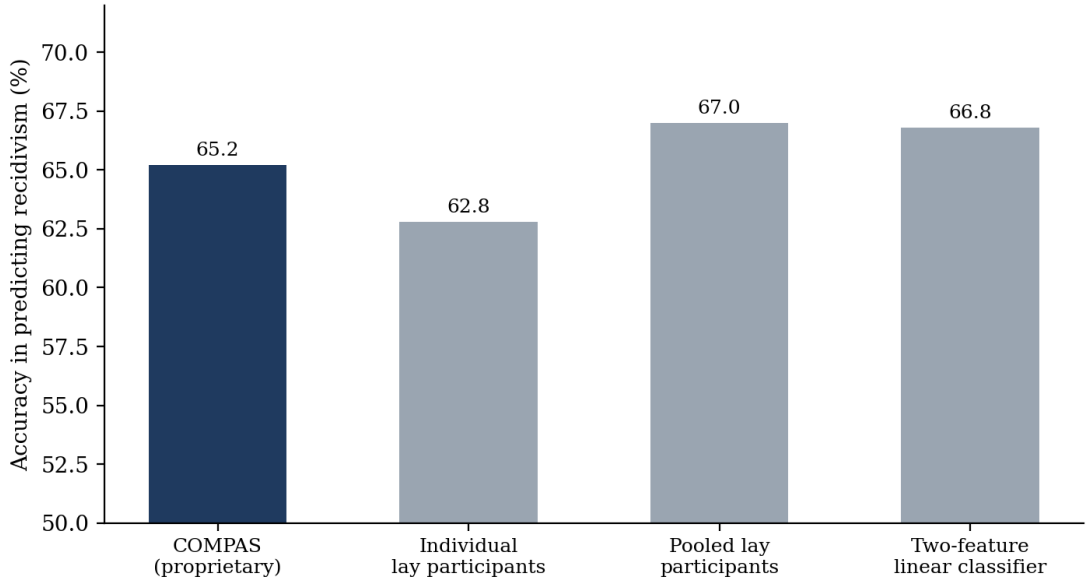
Does the Algorithm Outperform Humans?

One of the common arguments for algorithmic tools is that they are more reliable than humans. Dressel and Farid (2018) directly tested this. They invited lay participants to be recruited online to predict recidivism from a brief description of each defendant, and compared the accuracy of their predictions with the accuracy of the commercial instrument and with simple statistical models. As can be seen from Graph 3, the accuracy of the commercial instrument was not significantly different from that of an untrained individual, and using a linear classifier with just two of the features yielded comparable accuracy.

Graph 3

Accuracy of Recidivism Prediction: Commercial Instrument, Lay Participants, and a Simple

Model



Note. Data from Dressel and Farid (2018); figures rounded to one decimal place.

The result doesn't imply that algorithms are of no use, and for one instrument, and for one task. In other contexts, however, well-designed algorithms can lead to less discrimination than human decision makers, as an algorithm can be audited and modified in ways that human intuition cannot be. (Kleinberg et al., 2018). The less ambitious statement is that accuracy claims must be tested, not assumed, and that a proprietary tool must make a claim that it is adding value to simpler, more transparent alternatives (Rudin et al., 2020). If it's a simple model that works the same, there's not much justification for the opacity.

Human Factors: Automation Bias

Models are not the only thing that has bias. It also occurs in the reactions of people to them. When officers are given a ranked list or a score, they can anchor on that and not be as thorough in looking for disconfirming evidence. As described by Ferguson (2017), big data policing can lead to the use of risk scores as a form of short-cut suspicion. The risk is greater when the output comes with an aura of scientific neutrality, which Pasquale (2015) identifies as a core feature of black box governance. As such, a mandate for a human to review an algorithmic output is not adequate. The reviewer has to be equipped with information, time, training and authority to differ.

Transparency and Explainability

The Opacity Problem

There are multiple causes for opacity in the criminal justice system. Source code, training data, and validation studies are common items that are claimed as trade secrets. Government agencies may have an interest in not disclosing due to concerns for operational security. Some models, especially deep neural networks, are hard to interpret even by the model designers (Burrell, 2016). In the criminal justice system, the use of proprietary tools has created an era of opacity and injustice in which people whose liberty is at stake are unable to consider the instruments used to judge it. (Rudin et al. 2020)

The Wisconsin Supreme Court, in State v. Loomis (2016), addressed a challenge by a defendant to the use of a proprietary risk score in sentencing. The defendant claimed that he was denied due process because the instrument could not be explained to him because he could not examine how

it was scored, and that the inclusion of gender in the instrument was inappropriate. The court confirmed the use of the instrument, but it placed restrictions on the instrument. It was believed that this score should not be utilized to decide whether a defendant will be incarcerated or how harshly they will be punished, and that sentencing reports should include written cautions regarding the instrument's proprietary status, the use of group data and questions about the instrument's accuracy across groups. The decision is noteworthy. It allowed for some opacity, but attempted to control it through warnings, an approach which many commentators believe to be a substandard replacement for being able to scrutinize and question the tool (Kehl et al., 2017; Rudin et al., 2020).

Technical Remedies and Their Limits

There are a number of answers to the question of technical research. Interpretable models allow the reasoning to be made explicit by design. Rudin (2019) suggests that the purported compromise between accuracy and interpretability often is exaggerated in the case of structured data and that in high-stakes domains it's the developers' responsibility to demonstrate why an uninterpretable model is warranted. One approach to providing affected individuals with actionable reasons without disclosing the model is counterfactual explanations (Wachter et al., 2018). Kroll et al. (2017) suggest several methods of accountability, including cryptographic commitments and auditable logs, that enable an outside entity to ensure that a decision was made in accordance with the declared procedure.

These responses have merit but are not technical and cannot be achieved without the help of the teacher. Interpretable models will not be useful unless agencies are forced to use them. Consumer scoring might be appropriate for counterfactual explanations, but the same is not true for evidence: A court is interested in the reliability of a method, but not what would alter the result. Auditable logs are helpful if they are reviewed by a person with authority and expertise. The key issue here is that transparency should be regulated and required by law and linked with the rights of people concerned.

Transparency as a Condition of a Fair Trial

Fighting the test of evidence is at the heart of adversarial criminal procedure. In the United States, the Daubert v. Merrell Dow Pharmaceuticals, Inc. (1993) holding imposes upon trial judges the duty to serve as gatekeepers to determine if expert testimony is founded on a reliable basis, such as whether the expert method has been tested, whether the method has an established error rate, and whether the method is accepted by the relevant community. It is hard to apply this to an algorithm of which the code is not available. The President's Council of Advisors on Science and Technology (2016) emphasized that the reliability of feature-comparison methods needs to be demonstrated using empirical studies and that claims of validity without the use of such studies should not be accepted as science. The same reasoning applies to software in a similar fashion. There are worries surrounding machine-generated evidence as well. An algorithm does not produce a witness to testify to the end of the world, but it may have testifying value. A defendant has no access to the system design and validation, and he may not know all of the specifics of what the software did when the human expert testifies. The transparency consequently impacts the trial's design. That means, disclosure requirements should be considered as part of the right to a fair hearing and not as courtesy to the defence (Citron, 2008).

Public and Institutional Transparency

Transparency is also a collective aspect. If the technology doesn't exist than citizens and legislators can't discuss its legitimacy. Numerous police departments are using surveillance systems without prior announcement and approval from the public and it is only after they are in use that they become controversial and only then investigated by the media. There would be public registers of systems deployed, published impact assessments, and regular reports on usage and error rates for democratic oversight. These are minor expenses, given the trust they can inspire, and lends to a larger concern about the potential for algorithmic governance to replace democratic deliberation with technical administration (Eubanks, 2018; Citron & Pasquale, 2014).

Legal Accountability

The Accountability Gap

Legal responsibility for harm caused through an AI assisted investigation is hard to pinpoint. A number of actors have played their role in a situation: the vendor who constructed the system; the agency who purchased the system; the analyst who operated it; the officer who relied upon it; the prosecutor who acted upon it; the court which introduced it. Current precedent has a tendency to look at each case individually. There are many reasons why no one should be held responsible, including qualified immunity, prosecutorial discretion, and rules on expert evidence. Selbst (2017) suggests that the legal framework that supports individual decisions around criminal procedure is not well suited to decisions based on data aggregations and statistical inferences, and may require new tools like algorithmic impact assessments.

United States

The United States lacks a comprehensive federal statute governing AI in criminal investigation. Constitutional protections do the work in a piecemeal fashion. In Carpenter v. United States (2018), the Supreme Court determined that the government's collection of the historical cell-site location records of a cell phone user typically requires a warrant, finding that the interests in privacy of extensive digital records of an individual's movements are not ones that were foreseen by prior information-sharing with third parties precedents. The move is important for AI because numerous investigative tools rely on the processing of vast quantities of seemingly innocuous data. For example, the aggregation could be a search, according to Carpenter, though the courts have just started to consider this with predictive systems and facial recognition.

The case of State v. Loomis (2016) and Daubert (1993) above are examples of the case-by-case adjudication and the evidentiary gatekeeping role of the trial courts. The requirements of disclosure differ from jurisdiction to jurisdiction and some courts have refused to require the disclosure of source code. Consequently, the protections afforded to a defendant are very sensitive to the forum and on the resources available to challenge an expert, further compounding the inequities of access to justice.

United Kingdom

The Court of Appeal in R (Bridges) v Chief Constable of South Wales Police (2020) was called upon to determine the lawfulness of a police force's use of live automated facial recognition in public. The court's ruling was based on three claims that the deployment was unlawful. The law left too much discretion to individual officers in deciding who to place on a watchlist and how the technology could be used, causing an interference with privacy that wasn't in line with law. The data protection impact assessment was inadequate. Nor had the force taken reasonable steps to carry out an investigation to determine whether there was a racist or sexist bias in the software, and therefore failed to meet the public sector equality duty. The case is significant in that it

demonstrates how existing legal obligations, such as privacy, data protection and equality, can be used to limit the use of AI; and because it was not waited for harm to become apparent but instead the force had to look out for bias before it was deployed.

European Union

The EU has been the most prescriptive. The Law Enforcement Directive (European Union, 2016) lays down requirements for the processing of personal data by competent authorities and limits any decision taken on the sole basis of automated processing which has an adverse legal effect, to the extent that it is not made in accordance with the law with suitable safeguards. The Artificial Intelligence Act (European Union, 2024) extends this and adopts a risk-based classification. It prohibits certain practices and subjects other uses to strict requirements. In principle, real-time, remote biometric identification in public areas for police use is banned – applicable in extremely narrowly defined exceptions. Various other law enforcement uses, such as some of those used to determine risk of reoffending and the reliability of evidence, are deemed to be at 'high risk' and have obligations relating to risk management, data governance, documentation, human oversight, accuracy and robustness. The obligations are gradually being implemented, and the effect of the obligations will be contingent on enforcement and on the interpretation of exceptions.

Završnik's analysis on human rights and AI in criminal justice highlights that risk-based regulation should be grounded in fundamental rights as they are not simply risks to be managed, but fundamental rights of individuals, including the presumption of innocence, the right to a fair trial, and the prohibition on discrimination (Završnik, 2020). That's a very helpful disinhibition from the desire to confuse technical requirements with legality.

Comparison

Table 2 outlines the key instruments and decisions discussed, and notes their relevance to AI supported investigation.

Table 2

Selected Legal Instruments and Decisions Relevant to AI in Criminal Investigation

Instrument or decision	Jurisdiction	Year	Significance
Daubert v. Merrell Dow Pharmaceuticals, Inc.	United States	1993	Judicial gatekeeping of expert evidence for reliability, including error rates
State v. Loomis	United States (Wisconsin)	2016	Permitted proprietary risk score at sentencing with cautions and limits on use
Carpenter v. United States	United States	2018	Warrant generally required for historical cell-site data; aggregation raises privacy concerns
R (Bridges) v Chief Constable of South Wales Police	United Kingdom	2020	Live facial recognition unlawful given weak legal framework, deficient assessment, and unmet equality duty
Law Enforcement Directive (EU) 2016/680	European Union	2016	Data protection rules for policing; limits on solely automated decisions
Artificial Intelligence Act (EU) 2024/1689	European Union	2024	Risk-based regime; restrictions on biometric identification; high-risk obligations for law enforcement AI

Note. Compiled by the author from the primary sources cited in the text.

The comparison of the two tells one about the different regulatory philosophies. In the American model, the onus falls on individual defendants to litigate, in the British model, the general public law obligations are extended to new technology, and the European model tries to regulate ahead of time. They all have their pros and cons. Ex ante regulation can avert harm, but can be rigid and exceptions may be found. Case-by-case adjudication is flexible, but reactive and not equally available. None of the three systems offer a full solution to the question of who's accountable when a system is responsible for an unjust result.

8. A Three-Phase Accountability Framework

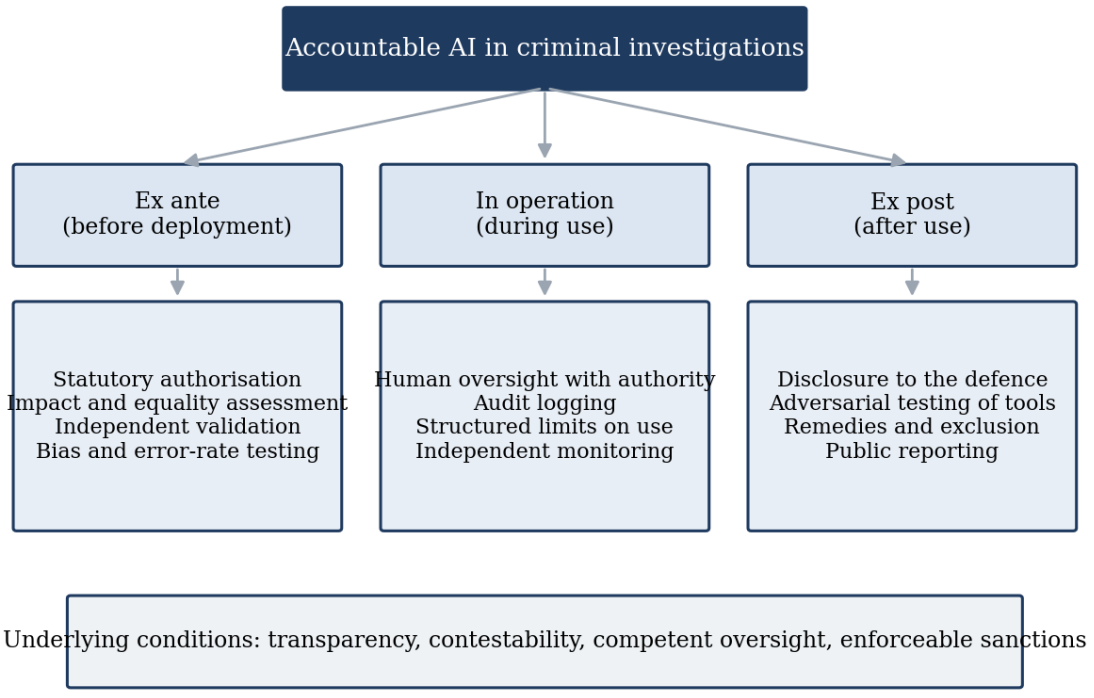
Rationale

The ideas that have been developed to date indicate that none of the reforms can solve the issues of bias, opacity, and diffuse responsibility. If one technology is banned, another is not, if there's a right to an explanation, there must be an explanation that makes sense, if there's a need for human review, there must be a human with the authority to review. What is required is a framework which spreads safeguarding throughout the life cycle of an AI-assisted investigation (AI), with control mechanisms in each phase and responsible actors for each phase. It is based on the notion of technological due process (Citron, 2008) and the literature on technical accountability (Kroll et al., 2017), which is adapted to the context of criminal investigation.

The framework is shown in figure 2. It comprises of three dimensions: ex ante, in operation and ex post, grounded in transparency, contestability, competent oversight, and enforceable sanctions.

Figure 2

A Three-Phase Framework for Accountable AI in Criminal Investigation



Note. Framework developed by the author.

Phase One: Ex Ante Authorisation and Validation

There are three safeguards that should be put in place before an agency deploys a system. First, the use should be authorized by a legislative or equivalent democratic process that includes purposes of the use, categories of offences and conditions for the use of the tool. This is relevant to the finding in R (Bridges) v Chief Constable of South Wales Police (2020), that an inadequate legal framework gave too much leeway to individual officers to decide how to act. Second, the agency should publish an impact assessment which considers bias, privacy and equality impacts, as it does in that case in respect of the public sector equality duty. Thirdly, the tool should be independently validated and tested on actual populations and image or data conditions similar to its operational use. The President's Council of Advisors on Science and Technology (2016) and the National Research Council (2009) both made this point—reliability claims must be based on empirical evidence of performance—and this would be the case with AI tools.

Procurement is a critical lever in this phase. Agencies should not buy systems whose vendors refuse to allow independent testing or to disclose the information necessary for meaningful evaluation. Kroll et al. (2017) note that accountability is easier to achieve when it is designed into a system from the start than when it is attempted afterwards. Contracts should therefore require access to documentation, performance data disaggregated by relevant groups, and logging capability.

Phase Two: Oversight in Operation

During use, the central safeguard is meaningful human oversight. Because of the risk of automation bias discussed above, oversight should be structured rather than nominal. This means that the reviewing officer should have the training to understand the tool's limits, access to the information needed to evaluate an output, the time to do so, and the institutional authority to reject it without penalty. An output from a facial recognition system, for example, should be treated as an investigative lead requiring independent corroboration and should not be treated as probable cause on its own. Policies should specify what corroboration is needed before an arrest or search.

Audit logging is the second safeguard. Each use of the system should be recorded, including the input, the output, the identity of the user, and the action taken, so that errors can be traced and patterns of misuse identified. An independent body, such as an inspector, a data protection authority, or a court-appointed monitor, should have the power to review these logs and the data on which systems are trained. Richardson et al. (2019) show that data produced under documented misconduct can contaminate predictive systems, which supports periodic review of data provenance.

Phase Three: Ex Post Disclosure and Remedy

After an investigation yields charges or other consequences, the key safeguards are disclosure and remedy. A defendant should be told whenever an AI-assisted tool contributed to the identification, arrest, charging decision, or evidence against them, and should receive the information necessary to challenge it, including error rates, validation studies, and, subject to protective orders where necessary, access to the code or to an independent expert who can examine it. The trade secret interest of the vendor should not defeat the defendant's right to test the evidence. Where such access is refused, courts should be prepared to exclude the evidence or to give it reduced weight (Daubert v. Merrell Dow Pharmaceuticals, Inc., 1993; Rudin et al., 2020).

Remedies should also exist for those wrongly affected. These may include exclusion of tainted evidence, civil damages for unlawful reliance on unvalidated systems, and administrative sanctions against agencies that deploy systems in violation of policy. Public reporting of error rates and

incidents would provide the information needed for democratic oversight and for continuous improvement.

Allocating Responsibility

Table 3 links the principal accountability gaps identified in this article with proposed safeguards and the actors responsible for them.

Table 3

Accountability Gaps, Safeguards, and Responsible Actors

Accountability gap	Proposed safeguard	Primary responsible actor
Deployment without clear legal basis	Statutory authorisation specifying purposes, offences, and limits	Legislature
Undetected demographic error disparities	Independent validation and disaggregated testing before and during use	Agency, independent evaluator
Biased or contaminated training data	Data provenance review and periodic audits	Agency, oversight body
Automation bias in human review	Structured oversight; corroboration requirement before arrest or search	Agency leadership, supervisors
Proprietary secrecy	Procurement terms requiring testability; disclosure to the defence under protective order	Procuring agency, courts
Diffuse responsibility for harm	Audit logs; defined duty holders; civil and administrative remedies	Legislature, courts, regulators
Lack of public knowledge	Public registers, published impact assessments, error-rate reporting	Agency, legislature

Note. Developed by the author from the analysis in Sections 5 to 7.

The table makes a simple point. Responsibility must be assigned in advance and in specific terms. A general expectation that someone will ensure fairness leaves the matter to whoever happens to be nearest when harm occurs, which is typically a junior officer or the defendant. Assigning duties to identifiable actors at each phase creates the conditions for accountability in the meaning set out in Section 2.4.

Discussion

Principal Findings

The analysis reveals four key conclusions. First of all, bias is not one fault but a number of faults that can be introduced at various points. The empirical audits analysed here reveal differences in error rates by demographic group in face analysis, predictive policing feedback effects, and risk scoring profiles of errors. Second, in cases of differing base rates, it is impossible to satisfy all competing notions of fairness – so the selection of any one to the exclusion of any others is a normative decision which must be explicit and public (Chouldechova, 2017; Kleinberg et al., 2017). Third, the accuracy of other claims must be tested and in one important comparison, a proprietary instrument was not found to be more accurate than untrained individuals and a simple model (Dressel & Farid, 2018). The results of our four analyses indicate that legal accountability

is fragmented among the three jurisdictions studied and none has yet provided a fully developed scheme, connecting the dots from validation, oversight, disclosure, to remedy.

Implications for Policy

The findings suggest to legislators that there should be specific statutory authorisation, mandatory impact assessment and rules as to disclosure to the defence. The European Union's regime is a model for ex ante regulation, but it will only have an impact if it is enforced and the exceptions must be monitored (European Union, 2024). The findings indicate that courts should be more skeptical of algorithmic evidence and of claims of trade secret, extending the reliability logic of Daubert (1993) and the due process concerns raised in State v. Loomis (2016). The results indicate that procurement requirements, oversight, and corroboration measures can help mitigate risks to law enforcement in a relatively cost-free manner. Opacity is increasingly unsustainable for vendors in a context where freedom of choice is on the line.

Counterarguments

There are a couple of objections that warrant answers. The first is that regulation will slow down the adoption of tools which could increase safety. This ignores the waste and injustice involved when tools are not reliable, and the higher the level of credible validation, the more likely the tools are to be effective and trustworthy. The second is a bias in human decision making, which should not be expected of algorithmic tools. The answer is that algorithms work at scale and give the impression of objectivity, which amplifies the errors and that algorithms are auditable and provide a window for laws to examine them in ways that laws cannot examine human intuition (Kleinberg et al., 2018). A third is that releasing data might make it easier for criminals to avoid getting caught. In some situations this is a concern, and the framework provides for protective orders and independent expert review as alternatives to open publication.

Limitations and Directions for Future Research

The limitations that were mentioned above apply to this study. As the empirical evidence is based on published audits of specific systems, it is not possible to draw conclusions about the performance of current systems in all jurisdictions. Further research is needed to systematically gather metrics on the use of AI in investigations, like the percentage of arrests that result from AI-provided leads and the percentage of wrongful arrests, and to assess the impact of procedural protections like a requirement to corroborate leads. Comparative studies could be used to consider how access to source code is dealt with in courts in the UK and how the gradual introduction of the European regime will impact police practice. The psychology of automated bias in investigative contexts would also aid in establishing if there is effective human oversight.

Conclusion

AI has gotten ahead of the regulators that are meant to keep it in check when it comes to criminal investigations. The evidence cited in this article demonstrates that risk is real and specific: errors that are subject to different rates in different demographic groups, feedback loops that reinforce established patterns of enforcement, proprietary systems whose rationale may not be scrutinized, and legal doctrines that diffuse responsibility to the point where no one's to blame. The same evidence demonstrates that these risks cannot be taken for granted. There are significant differences in systems, simple and transparent systems can sometimes do as well as opaque ones, and existing legal obligations could be used to require testing prior to deployment.

In the article, it has been claimed that legality of AI-assisted investigation relies on the ability of humans to challenge what AI produces. It needs to be validated prior to use, supervised in use and

disclosed and remedied after use. It also demands someone to be specifically responsible at every step. If AI is to play a legitimate role in investigations, the law must ensure that no algorithmic output, however confident its presentation, can displace the human obligation to justify the exercise of coercive power.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732.

Brayne, S. (2020). *Predict and surveil: Data, discretion, and the future of policing*. Oxford University Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 77–91.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12. https://doi.org/10.1177/2053951715622512

Carpenter v. United States, 585 U.S. 296 (2018).

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Citron, D. K. (2008). Technological due process. *Washington University Law Review, 85*(6), 1249–1313.

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review, 89*(1), 1–33.

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

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

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. *Proceedings of Machine Learning Research, 81*, 160–171.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

European Union. (2016). Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences. *Official Journal of the European Union, L 119*, 89–131.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.

Ferguson, A. G. (2012). Predictive policing and reasonable suspicion. *Emory Law Journal, 62*(2), 259–325.

Ferguson, A. G. (2017). *The rise of big data policing: Surveillance, race, and the future of law enforcement*. New York University Press.

Garvie, C., Bedoya, A., & Frankle, J. (2016). *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law Center on Privacy & Technology.

Grother, P., Ngan, M., & Hanaoka, K. (2019). *Face recognition vendor test (FRVT) part 3: Demographic effects* (NISTIR 8280). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8280

Hill, K. (2020, June 24). Wrongfully accused by an algorithm. *The New York Times*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Joh, E. E. (2016). The new surveillance discretion: Automated suspicion, big data, and policing. *Harvard Law & Policy Review, 10*(1), 15–42.

Kehl, D., Guo, P., & 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.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis, 10*, 113–174. https://doi.org/10.1093/jla/laz001

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou (Ed.), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (Article 43). Schloss Dagstuhl.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review, 165*(3), 633–705.

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

Mayson, S. G. (2019). Bias in, bias out. *Yale Law Journal, 128*(8), 2218–2300.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1–21. https://doi.org/10.1177/2053951716679679

National Research Council. (2009). *Strengthening forensic science in the United States: A path forward*. National Academies Press.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

President's Council of Advisors on Science and Technology. (2016). *Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods*. Executive Office of the President.

R (Bridges) v Chief Constable of South Wales Police, [2020] EWCA Civ 1058.

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online, 94*, 192–233.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review, 2*(1). https://doi.org/10.1162/99608f92.6ed64b30

Selbst, A. D. (2017). Disparate impact in big data policing. *Georgia Law Review, 52*(1), 109–195.

State v. Loomis, 881 N.W.2d 749 (Wis. 2016).

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.

Završnik, A. (2020). Criminal justice, artificial intelligence systems, and human rights. *ERA Forum, 20*(4), 567–583. https://doi.org/10.1007/s12027-020-00602-0