International Journal of Social Trends
Volume 4, Issue 5 (Sept-Oct; 2026)

Review: Law, Criminology and Public Administration

Algorithmic Justice? A Critical Review of Machine Learning in Criminal Justice, Welfare and Public Decision-Making

Eszter Fischer [1], **Mia Jovanović** [2]

[1] University of Belgrade, SERBIA
[2] Université Félix Houphouët-Boigny, CÔTE D'IVOIRE

ABSTRACT

Machine-learning systems now help decide who is detained before trial, which families are investigated by child-welfare agencies, which patients receive additional care, which neighbourhoods are policed and which job applicants are interviewed. This review essay examines what a decade of research in computer science, law and criminology has learned about these systems. It argues that the debate has moved through three phases: an initial phase in which algorithms were promoted as more objective than human decision-makers; a phase of exposure, in which journalistic and academic audits revealed disparities by race and gender; and a current phase in which attention has shifted from the properties of models to the institutions that build, procure and use them. We review the evidence from landmark cases, including recidivism risk assessment, health-care resource allocation, facial analysis and predictive policing, and explain the formal results showing that common definitions of fairness cannot all be satisfied at once when base rates differ. We then assess accountability mechanisms, from documentation and auditing to due-process rights and the European Union's Artificial Intelligence Act. We conclude that fairness is a property of decision systems rather than of algorithms alone, and that the most consequential choices concern what is predicted, for what purpose, and with what opportunity for contestation.

Keywords: Algorithmic Fairness; Criminal Justice; Risk Assessment; Predictive Policing; Algorithmic Accountability; Due Process; AI Regulation

INTRODUCTION: THE PROMISE AND THE PROBLEM

The case for using algorithms in public decision-making has always rested on a comparison with human judgement. Judges, caseworkers, police officers and hiring managers make high-stakes decisions under time pressure, with incomplete information and in ways that research has repeatedly shown to be inconsistent and biased. Actuarial tools have a long history in criminal justice, and their proponents argue that statistical prediction can be both more accurate and more consistent than unaided judgement. Evidence from economics supports part of this case: a study of pretrial release decisions in New York City estimated that a machine-learning algorithm could reduce crime among released defendants without increasing the jail population, or reduce detention substantially without increasing crime, and could do so while also reducing racial disparities [1].

Yet the same decade produced a sustained critique. Investigative journalists reported that a widely used recidivism risk tool made different kinds of errors for Black and white defendants [2]. Scholars in law and sociology argued that data-driven systems can automate and conceal existing inequalities [3-7]. Computer

scientists proved that several intuitive definitions of fairness are mathematically incompatible when groups differ in their base rates of the predicted outcome [8, 9]. A body of work on algorithmic accountability proposed documentation standards, audits and legal rights [10-13].

This essay reviews that literature from the standpoint of law and criminology, disciplines concerned with legitimacy, due process and the social consequences of state power. We make an argument rather than offering a neutral catalogue. Our claim is that much of the early debate asked the wrong question. Asking whether an algorithm is fair, as if fairness were a property of a model, obscures the decisions that matter most: the choice of what to predict, the data used to measure it, the use to which predictions are put, and the institutional safeguards that allow affected people to understand and contest decisions. The cases reviewed below show that these choices, rather than the sophistication of the learning method, account for most of the harms and benefits observed.

THREE PHASES OF THE DEBATE

Objectivity and the Actuarial Turn

Statistical prediction in criminal justice predates machine learning by most of a century. Critics of the actuarial turn argued that prediction-based policing and sentencing shift the purpose of criminal justice from responding to what people have done to managing what they might do, and that because predictive factors such as prior arrests reflect earlier enforcement decisions, prediction can reinforce the concentration of enforcement on already over-policed groups [16]. These arguments were largely conceptual until the 2010s, when risk assessment instruments spread through pretrial and sentencing systems and data became available to test them.

Advocates of algorithmic tools in this period emphasised consistency and transparency relative to human discretion. A tool applies the same weights to every case, and its inputs can in principle be inspected. The Wisconsin Supreme Court accepted the use of a proprietary risk score at sentencing, provided that judges were warned of its limitations and did not treat it as determinative [17]. The decision illustrates the legal framing of the time: algorithms were acceptable as advice to human judgement, and concerns about secrecy were addressed by caution rather than disclosure.

Exposure: Audits and Disparities

The second phase began with audits. ProPublica obtained risk scores and outcomes for defendants in Broward County, Florida, and reported that the COMPAS tool falsely labelled Black defendants who did not reoffend as higher risk at almost twice the rate of white defendants, while white defendants who did reoffend were more often labelled lower risk [2, 18]. Figure 1 displays the two error rates at the centre of the controversy.

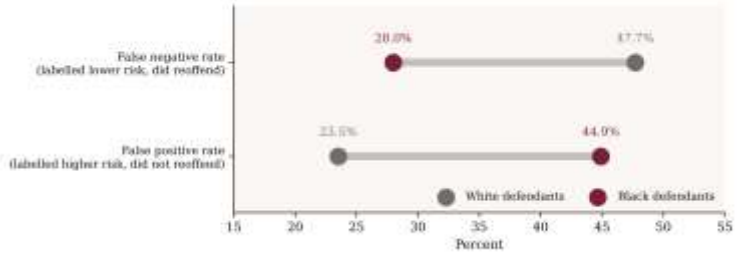

Figure 1: Error rates of the COMPAS recidivism score by race in Broward County, Florida, as reported by ProPublica. Higher-risk labels correspond to scores of 5 to 10. Data source: [18].

The tool's developer responded that COMPAS satisfied a different criterion, predictive parity: among defendants with the same score, Black and white defendants reoffended at similar rates [19], and criminologists published a rejoinder making the same point [20]. Both sides were correct about the criterion they used. The dispute became a turning point because it exposed a structural tension: when groups differ in their measured rates of rearrest, a score that is equally well calibrated for both groups must produce different false-positive and false-negative rates [8, 9]. We return to this result below.

Other audits followed. A study of three commercial gender-classification systems found that they performed far worse on darker-skinned women than on lighter-skinned men, and that two common facial-analysis benchmarks consisted overwhelmingly of lighter-skinned faces (Figure 2) [21]. A follow-up showed that the companies named in the audit reduced these disparities in subsequent releases, suggesting that public, targeted audits can change industry behaviour [22]. In hiring, a widely reported case described a recruiting model abandoned by its developer after it was found to downgrade applications that mentioned women's organisations, because it had learned from historical hiring patterns dominated by men [23].

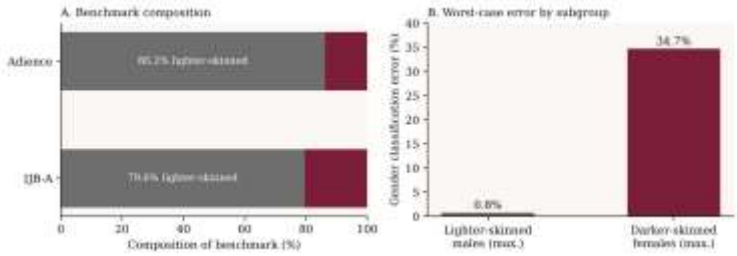

Figure 2: Representation and error in facial analysis. Panel A shows the share of lighter-skinned subjects in two benchmark datasets; Panel B shows the maximum gender-classification error rates for lighter-skinned males and darker-skinned females across three commercial systems. Data source: [21].

The most consequential audit outside criminal justice concerned health care. A commercial algorithm used to identify patients for care-management programmes predicted future health-care costs as a proxy for health needs. Because less money is spent on Black patients than on equally sick white patients, the algorithm systematically underestimated the needs of Black patients: at a given risk score, Black patients were considerably sicker. The authors estimated that remedying this disparity would increase the share of Black patients automatically identified for extra help from 17.7 percent to 46.5 percent (Figure 3) [24].

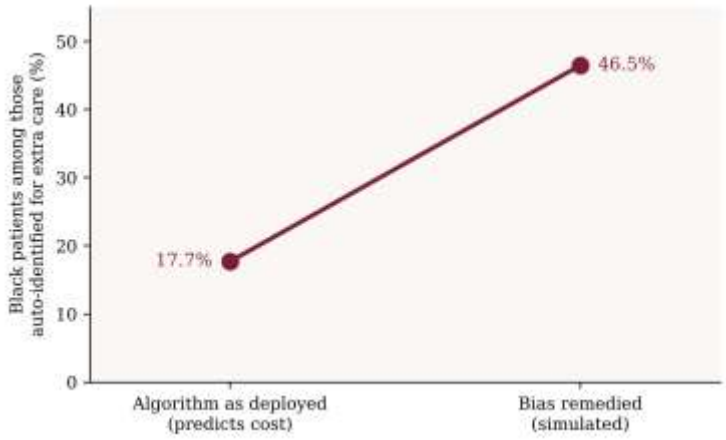

Figure 3: Effect of correcting label bias in a health-care allocation algorithm: share of Black patients among those automatically identified for additional care. Data source: [24].

Institutions: From Models to Systems

The Obermeyer study marks the transition to the third phase. Its central finding is not that the model was inaccurate. It predicted costs well. The problem was that the institution asked the wrong question: cost is a biased measure of need [24]. A parallel argument had been made in law, where scholars showed that discrimination can enter data-driven systems through the definition of the target variable, the labelling of training examples, the collection of data, the selection of features and the use of proxies, all of which are institutional choices rather than properties of learning algorithms [3]. Work on sociotechnical systems identified a set of "abstraction traps" into which designers fall when they treat fairness as a property of a model detached from the social context in which it is used [15].

Research on how people use algorithmic advice reinforces the point. In an experiment in which participants predicted risk with and without an algorithm's help, people did not reliably improve on the algorithm and deviated from it in ways that introduced new racial disparities [25]. A study of the implementation of risk assessment in Kentucky found that the effects of the tool depended heavily on how judges used it, and that its benefits eroded over time as judges returned to prior practices [26]. In child welfare, a detailed case study of the Allegheny County Family Screening Tool documented how the agency combined the score with caseworker judgement and examined the tool's fairness properties in detail [27], while an ethnographic account of the same system and others emphasised the experience of the poor families who are subject to them [4]. These studies show that a model's performance in isolation tells us little about the justice of the decisions it informs.

WHAT FAIRNESS CAN AND CANNOT MEAN

Formal definitions of fairness have multiplied, but most fall into a small number of families. Table 1 summarises them.

Table 1: Families of formal fairness criteria.

Criterion	Informal definition	Typical justification	Main objection
Fairness through unawareness	Do not use protected attributes as inputs	Formal equality before the law	Proxies reproduce group differences anyway
Demographic parity	Equal rates of positive decisions across groups	Remedying historical exclusion	Ignores real differences in need or risk
Equalised odds / equal opportunity	Equal false-positive and/or true-positive rates across groups	Equal burden of error	Conflicts with calibration when base rates differ
Calibration / predictive parity	Same score means same probability for all groups	Scores mean the same thing for everyone	Permits unequal error rates
Individual fairness	Similar individuals receive similar decisions	Treating like cases alike	Requires an agreed similarity metric
Counterfactual fairness	Decision unchanged if the protected attribute had been different	Removes causal effects of group membership	Depends on contested causal models

Note. Definitions follow the formulations in [28], [29], [8] and [9].

The Impossibility Results

Two papers published at the height of the COMPAS controversy proved that, except in degenerate cases, a risk score cannot simultaneously be calibrated within groups and equalise both false-positive and false-negative rates across groups when the groups have different base rates of the outcome [8, 9]. The results explain why ProPublica and Northpointe could both be right: they were measuring incompatible criteria. They also imply that choosing a fairness criterion is a normative decision about whose errors matter and how, not a technical detail. Equalising false-positive rates protects people who would not reoffend from being wrongly detained; calibration ensures that a given score carries the same meaning for everyone. Which matters more depends on the purpose of the decision and the harms of each kind of error.

A further analysis framed algorithmic fairness as a constrained optimisation problem and showed that, for pretrial release, the decision rule that maximises public safety holds all defendants to a single risk threshold, whereas satisfying common fairness constraints requires different thresholds for different groups and therefore some loss of safety [30]. Legal scholars have drawn a different lesson: because base rates are themselves products of unequal enforcement, any predictive tool trained on arrest data projects the past onto the future. On this view, the problem is not the choice of fairness metric but prediction itself, and the remedy lies in changing what is predicted or whether prediction is used at all [31].

Measurement: The Hidden Choice

The impossibility results take the outcome variable as given. In practice, the outcome is a measurement, and its validity is often the decisive issue. Recidivism tools predict rearrest or reconviction, not offending, and arrests depend on where police patrol and whom they stop. The health-care algorithm predicted cost rather than need [24]. The hiring model learned from past hires rather than from job performance [23]. In each case the target variable embedded the very inequalities the tool was meant to avoid. This is why we argue that fairness audits should begin with the label, not the model.

PREDICTION, POLICING AND FEEDBACK

Predictive policing illustrates how measurement problems compound over time. Place-based systems forecast where crimes will be recorded, and police are sent to those places. Because crimes are recorded partly where police are present, the forecasts influence the data from which future forecasts are made. A simulation using drug-arrest data from Oakland, California, showed that a predictive algorithm trained on police records would have concentrated enforcement in predominantly Black and low-income neighbourhoods, although public-health survey data suggested that drug use was spread much more evenly across the city [32]. A formal analysis showed that such systems can produce "runaway feedback loops" in which an area receives ever more attention regardless of its true crime rate, and proposed corrections that account for how the data are generated [33].

Fieldwork with the Los Angeles Police Department documented how big-data tools changed surveillance in practice: they lowered the threshold for inclusion in police databases, extended surveillance to people with no direct police contact, and integrated data from institutions outside criminal justice [34]. For criminology, the lesson is that predictive tools are not neutral observers of crime but participants in its production as a social and administrative category. Evaluations that measure accuracy against police records cannot detect this, because the records themselves are affected by the tool.

ARE ALGORITHMS BETTER THAN PEOPLE?

The comparative question that motivated the adoption of algorithms remains contested. The pretrial study described above suggests that, in principle, well-designed algorithms can improve on judges on several dimensions at once [1]. A study of COMPAS reached a sobering conclusion of a different kind: people recruited online, given only seven pieces of information about each defendant, predicted two-year recidivism about as accurately as COMPAS, which uses 137 features, and a simple linear model using only age and number of prior convictions matched the tool's accuracy (Figure 4) [35].

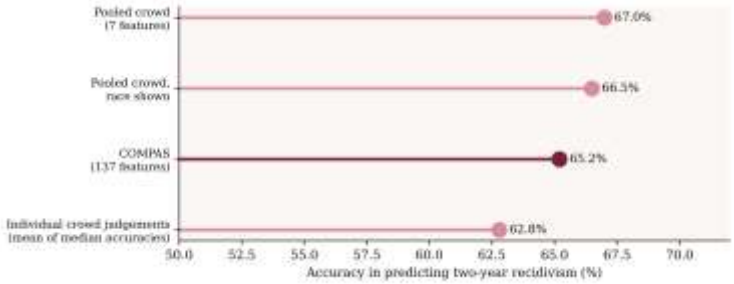

Figure 4: Accuracy of COMPAS and of untrained human participants in predicting two-year recidivism for 1,000 Broward County defendants. Pooled predictions aggregate the judgements of 20 participants per defendant. Data source: [35].

The two findings are not contradictory. The pretrial study concerns the potential of algorithms trained on large administrative datasets and evaluated with careful attention to selection bias in observed outcomes [1]; the COMPAS study concerns the actual performance of a particular commercial tool [35]. Together they suggest that the accuracy gains available from prediction in criminal justice are real but modest, that simple and transparent models often perform as well as complex proprietary ones, and that the case for opaque tools in high-stakes settings is correspondingly weak [36]. Where a transparent model achieves comparable accuracy, there is little justification for using a model whose reasoning cannot be inspected by the people it affects.

Table 2: Landmark cases of algorithmic decision-making in public and quasi-public settings.

Case	Domain	What was predicted	Key finding	Institutional lesson
COMPAS (Broward County)	Pretrial and sentencing	Two-year rearrest	Unequal false-positive and false-negative rates by race	Choice of fairness criterion is normative
Pretrial algorithm (New York City)	Pretrial release	Crime and failure to appear	Potential to reduce both crime and detention	Gains depend on how judges use predictions
Care-management algorithm	Health care	Future health-care costs	Black patients underestimated at same score	Proxy labels can encode inequality
Commercial gender classifiers	Facial analysis	Perceived gender	Error up to 34.7% for darker-skinned women	Unrepresentative benchmarks hide failures

Case	Domain	What was predicted	Key finding	Institutional lesson
Predictive policing simulation	Policing	Recorded drug crime	Enforcement concentrated in Black neighbourhoods	Feedback loops distort the data
Allegheny Family Screening Tool	Child welfare	Future placement or re-referral	Complex trade-offs among screening outcomes	Human-algorithm interaction shapes outcomes
Recruiting model	Employment	Hiring decisions in historical data	Downgraded applications associated with women	Historical decisions are not ground truth

Note. Sources in row order: [2]; [1]; [24]; [21]; [32]; [27]; [23].

ACCOUNTABILITY: TOOLS AND THEIR LIMITS

If fairness is a property of systems, accountability must be too. Researchers and regulators have proposed a range of mechanisms, summarised in Table 3.

Table 3: Mechanisms for algorithmic accountability.

Mechanism	What it requires	Strength	Limitation
Model cards	Standard reporting of intended use and performance by subgroup	Makes limitations visible to users	Voluntary; depends on honest reporting
Datasheets for datasets	Documentation of how training data were collected and labelled	Exposes label and sampling choices	Rarely applied to administrative data
Internal audits	End-to-end review across the development life cycle	Catches problems before deployment	Lacks independence
External audits	Independent testing by researchers, journalists or regulators	Demonstrated effect on vendor behaviour	Needs access to systems and data
Explanation rights	Information about the logic of automated decisions	Supports contestation	Explanations may not reveal what matters
Technological due process	Notice, hearing and reasons for automated decisions	Grounds safeguards in constitutional values	Courts slow to apply to algorithms
Risk-based regulation	Obligations scaled to the risk of the application	Covers whole sectors at once	Implementation and enforcement still developing

Note. Mechanisms are discussed with sources in the text.

Documentation and Audit

Documentation standards ask developers to report what a model is for, how it was evaluated and how it performs across groups [10], and to record how training data were collected, labelled and maintained [11]. A framework for internal auditing places such documents within a structured process running from scoping to post-deployment review [12]. External audits have a distinctive record of impact: after the public release of the gender-classification audit, the named companies reduced their error disparities [22]. Audits face practical limits, however. Proprietary systems, trade-secret claims and the absence of outcome data often prevent outsiders from testing the systems that matter most [37].

Law: Explanation, Due Process and Regulation

Legal responses operate at several levels. In the United States, scholars argued early that automated decision-making in public administration threatens procedural due process, because it can deny people notice of the rules applied to them and an opportunity to be heard, and proposed "technological due process" safeguards such as audit trails and public testing of software [13]. In Europe, the General Data Protection Regulation gives individuals rights concerning decisions based solely on automated processing that significantly affect them [38]. Because the regulation's right to an explanation is limited and contested, some scholars proposed counterfactual explanations, which tell people what would have had to be different for a decision to change, as a way to make decisions contestable without disclosing the model [39].

More recent regulation addresses whole categories of systems. The European Union's Artificial Intelligence Act classifies uses of AI by risk and imposes obligations on high-risk systems, which include systems used in law enforcement, migration, access to essential public services, employment and the administration of justice. Obligations cover risk management, data governance, documentation, human oversight, accuracy and robustness, and certain practices are prohibited outright [40]. At the local level, New York City requires employers that use automated employment decision tools to commission bias audits and notify candidates [41]. These developments move accountability from voluntary practice to legal obligation, but their effect will depend on enforcement, on the capacity of regulators and on whether audits measure what matters.

Table 4: Selected legal frameworks governing algorithmic decision-making.

Instrument	Jurisdiction	Scope	Core obligations or rights	Open questions
GDPR, Article 22	European Union	Decisions based solely on automated processing with legal or similarly significant effects	Right not to be subject to such decisions, subject to exceptions; safeguards including human intervention and contestation	Meaning of "solely" automated; extent of explanation rights

Instrument	Jurisdiction	Scope	Core obligations or rights	Open questions
Artificial Intelligence Act (Regulation (EU) 2024/1689)	European Union	AI systems classified by risk; high-risk uses include law enforcement, justice, employment and public services	Risk management, data governance, documentation, human oversight, accuracy; some practices prohibited	Enforcement capacity; standards for bias testing
Local Law 144 of 2021	New York City	Automated employment decision tools	Independent bias audit; publication of results; notice to candidates	Narrow scope; audit metrics limited to selection rates
State v. Loomis (2016)	Wisconsin, United States	Use of a proprietary risk score at sentencing	Use permitted with written warnings about the tool's limitations	Defendants cannot inspect the proprietary model
Procedural due process	United States (constitutional)	Government decisions depriving people of liberty or property	Notice and an opportunity to be heard	Application to automated systems largely untested

Note. Compiled from [38], [40], [41], [17] and [13]. The table summarises; it is not legal advice.

Table 4 shows how unevenly the law has developed. European instruments regulate broadly and in advance, defining categories of systems and the obligations attached to them. American law, by contrast, has developed largely through litigation over particular deployments and through local legislation on specific uses. Neither approach has yet produced a reliable answer to the question that matters most to an affected person: how can I find out why the system treated me as it did, and how can I challenge it? The gap between formal rights and practical contestability is likely to define legal scholarship on algorithms in the coming decade.

Human Oversight as a Safeguard

Many legal frameworks rely on human oversight: a human must review algorithmic recommendations before decisions are made. The empirical evidence reviewed above suggests caution. Human decision-makers may defer to algorithms when they should not, override them in biased ways, or drift back to prior practices [25, 26]. Oversight is meaningful only when overseers have the time, information, training and authority to disagree with a model, and when their decisions are themselves monitored. Treating a nominal

human in the loop as a guarantee of fairness repeats the abstraction trap of treating any single component as the locus of justice [15].

BEYOND THE COURTROOM: WELFARE, CHILD PROTECTION AND EMPLOYMENT

Although criminal justice has dominated public debate, algorithmic systems arguably affect more people through welfare administration, child protection and employment. These domains share features that make them especially consequential. Those affected are often poor, have limited access to legal representation and interact with the state in conditions of dependency. Decisions are made at scale, frequently without the procedural formalities of a courtroom. And the data available to the system are shaped by prior contact with public services, so that people who rely on those services are more visible to the algorithm than people who can afford private alternatives [4].

Child-protection screening illustrates the tensions. Hotline screeners must decide quickly which reports warrant investigation, and agencies have adopted predictive tools to help prioritise. The detailed evaluation of the Allegheny County tool examined how the score related to outcomes across racial groups and how screeners combined the score with their own judgement, finding that the interaction between human and algorithm shaped disparities in ways neither component would predict alone [27]. Critics have argued that such systems confuse poverty with risk, because families who use public benefits, public health care and public housing generate more administrative records than families who do not [4]. Both perspectives agree that the choice of outcome to predict, such as later placement in foster care, embeds earlier institutional decisions in the label.

Employment raises a different configuration of risks. Automated tools are used to screen résumés, analyse interviews and rank candidates, usually by private firms rather than the state. Here the historical-data problem is acute: a model trained to reproduce past hiring decisions will reproduce whatever patterns those decisions contained [3, 23]. Because applicants rarely know that an automated tool was used, discrimination is hard to detect and harder to challenge. New York City's bias-audit requirement is one of the first attempts to address this through disclosure [41], but its focus on selection rates captures only one of the fairness criteria discussed earlier.

Search and information systems form a final, more diffuse category. Research on search engines documented how ranking and autocompletion reproduced racist and sexist stereotypes, shaping the information environment in which people form judgements about others [6]. Although such systems do not make formal decisions about individuals, they show that algorithmic harms can be representational as well as allocative: they can affect how groups are portrayed as well as what resources individuals receive [7].

EVALUATING ALGORITHMIC SYSTEMS: METHODOLOGICAL PROBLEMS

Empirical claims about algorithmic fairness and accuracy are only as good as the methods used to establish them. Several problems recur across the literature and deserve explicit attention from researchers in law and criminology who use or evaluate these studies.

Selective Labels and Counterfactual Outcomes

In many settings the outcome is observed only for people who received a particular decision. We observe whether released defendants are rearrested, but not whether detained defendants would have been; we observe loan repayment only for applicants who received loans. This "selective labels" problem means that a naïve comparison of algorithmic and human decisions can be badly biased, because the algorithm is evaluated on a population that human decision-makers have already filtered [42]. The pretrial study addressed the problem by exploiting the quasi-random assignment of cases to judges who differ in leniency, allowing the authors to estimate what would have happened under alternative release rules [1]. Few deployed systems are evaluated with comparable care.

Measurement Validity of Outcomes

As discussed above, outcomes such as arrest, re-referral or cost are proxies for the quantities institutions care about. Their validity varies across groups because enforcement and service use vary across groups [24, 32]. Evaluations should therefore report not only predictive accuracy against the recorded outcome but also evidence about the relationship between that outcome and the underlying construct. Where independent measures are available, such as survey data on drug use or clinical measures of health, comparing them with administrative outcomes can reveal the extent of label bias.

Access, Reproducibility and the Politics of Data

Many landmark studies were possible only because a public-records request, a vendor partnership or an unusual data release made the necessary data available [2, 24]. Most deployed systems have never been independently evaluated. Proprietary claims, privacy rules and institutional reluctance all limit access [37]. Researchers should be explicit about how their data were obtained and what that implies for generalisation, and journals and funders should support the publication of data and code where privacy permits. The Broward County data released with the ProPublica analysis became a standard benchmark precisely because so few alternatives exist [18], which itself risks narrowing the field's understanding to a single jurisdiction.

From Disparity to Discrimination

Finally, a statistical disparity is not by itself proof of unjust discrimination. Disparities may reflect differences in the measured outcome, legitimate differences in relevant characteristics, or bias in data and decisions. Distinguishing among these requires causal reasoning and normative argument, not only statistics [14, 31]. Conversely, the absence of a measured disparity does not establish fairness, since the chosen metric may not capture the harm that matters. Legal doctrine on disparate impact offers one framework for this reasoning, asking whether a practice with disparate effects is justified by legitimate need and whether less discriminatory alternatives exist [3]. Applying that framework to algorithms is a promising direction for interdisciplinary work.

THE POLITICAL ECONOMY OF PROCUREMENT

Public agencies rarely build their own algorithmic systems; they buy them. Procurement decisions therefore shape fairness and accountability before any model is trained. Contracts determine whether agencies can inspect a model, obtain its training data, commission independent audits or modify the system when

problems are found. Trade-secret protections, invoked by vendors in litigation such as the challenge to COMPAS at sentencing [17], can prevent courts and defendants from examining the tools used against them [37].

Agencies also face incentives that favour adoption over evaluation. Algorithmic tools promise efficiency at a time of constrained budgets, and their adoption can signal modernisation. Evaluation is costly and may produce unwelcome findings. The result is that systems are often deployed with limited pilot testing and without the outcome data needed to assess them afterwards. Requirements in the European Union's AI Act for documentation, risk management and post-market monitoring of high-risk systems can be read partly as a response to this imbalance [40]. Elsewhere, procurement rules that require transparency, audit rights and interpretable models where they perform comparably [36] would be a practical lever for change that does not depend on new legislation.

Table 5: A research agenda for algorithmic justice in law and criminology.

Research question	Why it matters	Suitable methods
How do decision-makers actually use algorithmic recommendations over time?	Effects depend on human uptake, which can drift	Longitudinal administrative data; ethnography; natural experiments
How valid are administrative outcomes as measures of risk or need?	Label bias drives many observed disparities	Linkage to surveys or independent measures; measurement models
Do explanation and contestation rights change outcomes for affected people?	Formal rights may not be usable in practice	Legal case analysis; field experiments; interviews
What procurement terms produce accountable systems?	Contracts determine access and auditability	Comparative document analysis; case studies of agencies
How do algorithmic systems spread outside the United States and Europe?	Evidence base is geographically narrow	Comparative and regional studies; policy mapping
Can prediction be replaced by policy change?	Some harms stem from the decision to predict at all	Policy evaluation; simulation of alternative regimes

Note. Agenda derived from the gaps identified in this review.

TOWARD A SYSTEMS VIEW OF ALGORITHMIC JUSTICE

The literature reviewed here supports a reorientation of research and policy around five questions, which we present as a framework for evaluating any algorithmic decision system in public life.

First, what is being predicted, and is it the right target? The health-care and policing cases show that the choice of label is the single most consequential design decision [24, 32]. Audits should begin by asking whether the outcome measure is a valid and unbiased proxy for the quantity the institution cares about.

Second, what is the decision for? Predictions used to allocate support, such as care management or social services, raise different concerns from predictions used to impose burdens, such as detention or investigation. A tool that triggers help can tolerate false positives that would be intolerable in a tool that triggers punishment. Fairness criteria should be chosen with the purpose in view [14].

Third, is prediction necessary? Some critics argue that risk prediction in criminal justice should be abandoned or radically limited because it inevitably reproduces past inequality [16, 31]. Others argue that well-designed tools can reduce harms relative to current practice [1]. Both positions agree that prediction must be justified against alternatives, including changes in policy that reduce the need for individual risk classification.

Fourth, can affected people understand and contest decisions? Transparency, explanation and due process are not optional extras. Where simple, interpretable models perform as well as complex ones, as they often do in criminal justice [35, 36], there is a strong case for using them.

Fifth, who audits the system over time? Feedback loops, drift and changes in human use mean that a system evaluated once may behave differently later [26, 33]. Ongoing, independent and published monitoring is essential, and regulation increasingly requires it [40, 41].

IMPLICATIONS FOR CRIMINOLOGY AND LEGAL SCHOLARSHIP

For criminology, algorithmic systems are both objects of study and research instruments. As objects, they invite the same questions criminologists ask of any institution of social control: whose behaviour is defined as risky, how classifications are produced, and how they shape the life chances of those classified. Ethnographic work on policing shows that big-data tools reorganise surveillance rather than simply making it more accurate [34], and work on welfare shows how automated eligibility and risk systems change the experience of being poor [4]. As instruments, machine-learning methods can help criminologists study discretion itself, for example by revealing how judges' decisions depart from predicted risk [1]. Both kinds of work require access to data that institutions often withhold.

For legal scholarship, the central task is to translate principles of equality and due process into requirements that can be applied to sociotechnical systems. Anti-discrimination law developed around intentional discrimination and disparate impact in human decision-making [3]; its application to systems in which no individual intends discrimination, and in which disparities emerge from data, remains unsettled. Scholars have argued that existing doctrine is poorly matched to these harms and that new procedural rights are needed [13, 37]. The European regulatory approach offers one model of such rights, but its effectiveness is not yet known.

Interdisciplinary collaboration is the clearest route forward. Computer scientists have supplied formal tools for measuring disparities and proving trade-offs [8, 9, 28, 29]; legal scholars and criminologists supply the normative and institutional analysis needed to decide which trade-offs are acceptable and who should make them [14, 15, 31]. The most influential studies of the past decade combined the two, often with the investigative persistence of journalists [2, 21, 24].

CONCLUSION

A decade of research has dispelled the idea that algorithms are inherently objective, and it has also complicated the opposite view that they are inherently discriminatory. What the evidence shows is that algorithms reproduce the choices embedded in their design and use: what counts as the outcome, whose data are collected, which errors are tolerated, and how predictions enter decisions. The impossibility results make clear that some of these choices cannot be avoided and must be made openly [8, 9]. The audits show that unexamined choices can produce large disparities [2, 21, 24]. The institutional studies show that human oversight, often presented as a safeguard, does not reliably correct them [25, 26]. Justice in algorithmic decision-making therefore depends less on finding the right fairness metric than on building institutions in which the purpose of prediction is justified, labels are valid, systems are transparent and contestable, and performance is monitored by people with the power to stop a system that causes harm. Law and criminology are well equipped to contribute to that work. Their engagement with computer science over the past decade has already changed the questions being asked; the next decade should change the institutions that answer them.

REFERENECS

[1] Potharaju, B. (2026). Infrastructure-level intelligence: Embedding AI into data movement and validation layers. *International Journal of Computational and Experimental Science and Engineering*.

[2] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics, 133*(1), 237–293.

[3] Potharaju, B., & Ravipudi, J. (2023). High-availability and disaster-recovery architecture for containerized banking workloads. *International Journal of Data Engineering and Intelligent Computing*, 6(1), 1–32.

[4] Potharaju, A. (2026). Embedded compliance architecture for real-time financial systems modernization. *Journal of Information Systems Engineering and Management*, 11(2s), 1147–1158.

[5] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*.

[6] Potharaju, A. (2026). Human-guided artificial intelligence for adaptive compliance automation in financial institutions. *Journal of Computational Analysis and Applications*, 35(4), 57–68.

[7] Potharaju, A. (2026). Microservices-based modernization of financial regulatory systems. *International Journal of Computer Science and Engineering Innovations*, 2(1), 68–85.

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

[9] Movva, C. K., & Manohar, V. (2021). OCF protocol implementation for smart home IoT applications: An Android SDK development case study. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 2(1), 89–96.

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

[11] Aluri, G. N. C., & Mupparapu, H. K. (2021). Spring Batch framework patterns for large-scale data processing in cloud-native microservices: A GCP deployment study. *International Journal of AI, BigData, Computational and Management Studies*, 2(4), 130–142.

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

[13] Mallempati, A., & Jaladi, D. S. (2023). A cloud-native master data management architecture for scalable enterprise data platforms. *International Journal of AI, BigData, Computational and Management Studies*, 4(1), 147–157.

[14] Potharaju, B., & Sajja, R. R. (2024). Modernizing US federal regulatory reporting infrastructure for scalability, resilience, and audit readiness. *International Journal of Data Engineering and Intelligent Computing*, 7(1), 1–36.

[15] Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

[16] Mallempati, A. (2024). Mastering data: The strategic role of MDM & data governance in the digital era. *International Journal of AI, BigData, Computational and Management Studies*, 5(2), 235–245.

[17] Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Polity Press.

[18] Yuvaraj, N., Kumar, M. S., & Manoj, M. S. (2026). Unified agentic AI architecture for data lineage, governance, and predictive workflow automation in enterprise systems. In *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 1817–1828). IEEE.

[19] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference* (Article 43). Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

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

[21] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM.

[22] Shammah, R. S., Rahman, I., Guria, Z. M., Kowser, S., Faruq, O., & Morshed, N. (2025, December). Scalable Machine Learning Architecture for Securing IoT Networks with Digital Governance in the AI Era of Cyber-Resilient Smart Cities. In *2025 International Conference on AI-Driven STEM Education and Learning Technologies (AISTEMEDU)* (pp. 1-7). IEEE.

[23] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92.

[24] Movva, C. K., & Mupparapu, H. K. (2026). MVVM architectural patterns for enterprise Android kiosk applications: State management and lifecycle handling. *International Journal of Emerging Trends in Computer Science and Information Technology*, 7(2), 320–330.

[25] Potharaju, A. (2021). Designing dynamic questionnaire engines for risk-based account review workflows. *International Journal of Applied Data Science & Modern Computing*, 4(1), 1–37.

[26] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). ACM.

[27] Yuvaraj, N., & Kumar, M. S. (2025). Agentic AI for zero-touch customer experience: A governance-constrained framework for autonomous service systems. *American International Journal of Computer Science and Technology*, 7(2), 100–111.

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

[29] Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Proceedings of Machine Learning Research, Vol. 81, pp. 149–159). PMLR.

[30] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). ACM.

[31] Aluri, G. N. C., & Manohar, V. (2025). Generative AI integration patterns for enterprise workflow automation: A practitioner framework. *International Journal of Emerging Research in Engineering and Technology*, 6(3), 165–172.

[32] Harcourt, B. E. (2007). *Against prediction: Profiling, policing, and punishing in an actuarial age*. University of Chicago Press.

[33] Kumar, M. S., Yuvaraj, N., & Manoj, M. S. (2026). Intelligent orchestration of master data and privacy compliance via agentic AI for customer lifecycle optimization. In *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 1715–1723). IEEE.

[34] Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, May 23). How we analyzed the COMPAS recidivism algorithm. *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[35] Potharaju, B., & Ravipudi, J. (2025). Automated data validation and resilient infrastructure design for enterprise regulatory reporting platforms. *International Journal of Computer Science and Engineering Innovations*, 1(1), 46–64.

[36] Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS risk scales: Demonstrating accuracy equity and predictive parity*. Northpointe Inc.

[37] Rahman, M. M., Computer, M. M. R., Das, N., Uddin, M. S., Shammah, R. S., & Rahman, M. M. (2025, May). An Investigation of Face Detection and Feature Extraction Techniques in Deepfake Detection Using a Multi-Model Machine Learning Approach. In *2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS)* (pp. 1-8). IEEE.

[38] Manohar, V., & Mupparapu, H. K. (2025). Scalable data pipeline architecture for real-time supply chain analytics using PySpark and Snowflake. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 6(3), 203–212.

[39] Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to "Machine bias". *Federal Probation, 80*(2), 38–46.

[40] Mallempati, A., Jaladi, D. S., & Jegajothi, B. (2026). HDLF-DGC: A hybrid deep learning framework for intelligent cybersecurity and data governance in distributed web systems. In *2026 Third International Conference on Innovations in Cybersecurity and Data Science (ICICDS)* (pp. 16–21). IEEE.

[41] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Proceedings of Machine Learning Research, Vol. 81, pp. 77–91). PMLR.

[42] Jaladi, D. S., Mallempati, A., & Jegajothi, B. (2026). A hybrid federated privacy-aware deep learning framework for secure data governance in cloud systems. In *2026 4th International Conference on Sustainable Computing and Smart Systems (ICSCSS)* (pp. 1629–1636). IEEE.

[43] Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 429–435). ACM.

[44] Kumar, M. S., & Yuvaraj, N. (2024). Predictive customer experience orchestration using governed data pipelines and intelligent service signals. *International Journal of Emerging Trends in Computer Science and Information Technology*, 5(1), 206–215.

[45] Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

[46] Mupparapu, H. K., & Movva, C. K. (2020). Serverless migration patterns for compliance-critical financial applications on Microsoft Azure. *International Journal of Emerging Research in Engineering and Technology*, 1(2), 85–96.

[47] Rahman, M. S., Bhowmik, P. K., Hossain, B., Tannier, N. R., Amjad, M. H. H., Chouksey, A., & Hossain, M. (2023). Enhancing fraud detection systems in the USA: A machine learning approach to identifying anomalous transactions. *Journal of Economics, Finance and Accounting Studies*, *5*(5), 145-160.

[48] Potharaju, B. (2026). Architecting national-scale regulatory reporting platforms: A high-availability infrastructure framework for compliance-critical systems. *Journal of International Crisis & Risk Communication*.

[49] Shahiduzzaman, M., Sourav, M. S. A., Kabtia, M., Shammah, R. S., Islam, M. A., Chakraborty, P., ... & Adnan, M. (2026). Attention-guided quantum-inspired extreme learning machine with grey wolf optimization for rare-attack detection in IIoT networks. *Discover Artificial Intelligence*.

[50] Movva, C. K., & Aluri, G. N. C. (2025). Enterprise mobile payment gateway architecture on IoT handheld devices: Security and reliability design. *International Journal of AI, BigData, Computational and Management Studies*, 6(1), 232–241.

[51] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

[52] Rahman, M. M., Computer, M. M. R., Das, N., Uddin, M. S., Shammah, R. S., & Rahman, M. M. (2025, May). An Investigation of Face Detection and Feature Extraction Techniques in Deepfake Detection Using a Multi-Model Machine Learning Approach. In *2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS)* (pp. 1-8). IEEE.

[53] Green, B., & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 90–99). ACM.

[54] Shammah, R. S., Sultana, S., Faruq, O., Kowser, S., Uddin, A., & Haque, S. A. (2026, July). Data-Driven Decision Support Framework for Engineering and IT Project Management Using Integrated Business Analytics and Management Information Systems. In *2026 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)* (pp. 1-6). IEEE.

[55] Stevenson, M. (2018). Assessing risk assessment in action. *Minnesota Law Review, 103*(1), 303–384.

[56] Potharaju, A. (2026). Policy-driven authorization architectures for modern financial regulatory platforms. *Journal of International Crisis and Risk Communication Research*, 9(4), 36–44.

[57] Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Proceedings of Machine Learning Research, Vol. 81, pp. 134–148). PMLR.

[58] Aluri, G. N. C., & Manohar, V. (2026). Model context protocol servers as orchestration layers for generative AI agents in enterprise software delivery. *International Journal of AI, BigData, Computational and Management Studies*, 7(2), 337–345.