REVIEW OF CRIME, PEACE AND SOCIETY
ISSN (O): 3079-6105, ISSN (P): 3079-6091
Vol. 3, No. 5, 2026, 353-369

Forecasting Crime, Forecasting Police: A Critical Review of Predictive Policing and Algorithmic Risk Assessment in American Criminal Justice

Kaukab Jamal Zuberi[1], Warda Maqsood[2], Areeba Azeem[3]

Abstract

Two decades have passed since the first place-based forecasting pilots, and we could not develop a fair system of algorithmic forecasting in American policing. The evidentiary record remains thin and contested. This is affecting every stage of the American justice system. The existing systems rely on historical data which consists of enforcement and not of offending. This is important when decisions are made to allocate resources to patrolling, pretrial releases, sentencing and awarding paroles to prisoners. This review considers the problems identified in present practices. These problems consist of using enforcement data instead of offending data, feedback loops in the forecasting models, disproportionate burden on minorities like Black, Latino and poor communities, and weak accountability. Feedback loops happen when the data generated by the biased datasets is automatically used by the systems to train themselves, which increases the bias. Accountability remains weak because the process is not transparent and the bias remains hidden in the core dataset used to initially train the model. This review also discusses literature on predictive risk assessment in courts and prisons, where the COMPAS controversy exposed competing definitions of fairness. Across both literatures the evidence of benefit is modest and uneven while the evidence of harm is credible. In the end this review closes with the National Institute of Justice's current emphasis on the evaluation of the use of AI in courts, policing and prisons. This review suggests more focus on the outcomes of statistical predictions instead of evaluating the accuracy of statistical predictions.

Article History
Received 15 june 2026
Revised 25 June 2026
Accepted 24 July 2026
Published 01 August 2026
Corresponding Author*
OPEN ACCESS

Keywords
predictive policing, risk assessment, algorithmic fairness, police data, feedback loops, accountability, artificial intelligence

1. Introduction

The National Institute of Justice convened its first symposium on the subject in 2009. It was considered a new idea that statistical machinery, which was used to predict earthquakes and consumer demand, might be used to predict crime.[4] For police forces with limited resources, the appeal was obvious. Crime concentrates in time and space. Using the technology to predict where the next burglary will take place would allow the necessary allocation of police officers. Within five years, hundreds of American police departments started using PredPol, HunchLab, and Palantir. In addition to these, several person-scoring instruments were used to score defendants for flight risk, prisoners for parole suitability, and probationers for supervision intensity.

The years that followed were not kind to the idea. In November 2019, Chicago decommissioned its person-based Strategic Subject List after the city's own Inspector General

1. Associate Dean Faculty of Social Sciences, Department of Criminology, Lahore Garrison University. dir.dfrsc@lgu.edu.pk
2. Lecturer Criminology, Lahore Garrison University.
3 Lecturer Criminology, Lahore Garrison University

4Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). Predictive policing: The role of crime forecasting in law enforcement operations (RR-233). RAND Corporation. https://www.rand.org/pubs/research_reports/RR233.html

questioned the program's efficacy and legal footing.[5] Among the earliest PredPol customers, The Los Angeles Police Department ended its program in April 2020 amid an internal audit that could not establish whether the software reduced crime.[6] PredPol's corporate successor, Geolitica, ceased independent operations at the end of 2023 after its acquisition by SoundThinking. The European Union, across the Atlantic, has gone further than any American jurisdiction has contemplated. Article 5(1)(d) of its AI Act, applicable since February 2, 2025, prohibits outright AI systems that use personality features or profile alone to forecast a person's likelihood of committing a crime.[7]

It would be a mistake, though, to read these developments as the end of the story. The underlying capabilities are still present. They are now part of integrated platforms, marketed as "data-informed community-focused policing," and integrated into department-owned gunshot detection and case management suites.[5] Using the technology is now easier and less expensive than it was when the controversies initially arose, and the federal government of the United States currently favors expansion over restraint. In the meantime, research on AI applications in law enforcement, courts, and prisons is being actively sought after by the National Institute of Justice, which explicitly calls for an analysis of both advantages and disadvantages.[6] The question facing the field is therefore not whether algorithmic forecasting will be used, but whether criminology can supply the evidence needed to govern it.

This review proceeds in three movements. It first describes what these systems do, distinguishing place-based forecasting, person-based forecasting, and decision-point risk assessment, because the three raise different empirical and normative questions that public debate tends to collapse. It then works through the central concerns: the quality of the data these systems learn from, the feedback dynamics they can set in motion, the distribution of their burdens, the opacity of their reasoning, and the diffusion of responsibility they enable. It closes with an assessment of governance responses and a research agenda. Throughout, it aims to separate what has been demonstrated from what is merely asserted, by critics as well as vendors.

A note on terminology. It uses "predictive policing" to mean the application of statistical models to police-held data for the purpose of forecasting crime locations, times, victims, or perpetrators, and "risk assessment" to mean instruments that estimate an individual's tendency of being involved in any future event (rearrest, failure to appear, violence) to inform a custodial or

[5]City of Chicago Office of Inspector General. (2020). Advisory concerning the Chicago Police Department's predictive risk models. https://igchicago.org/wp-content/uploads/2020/01/OIG-Advisory-Concerning-CPDs-Predictive-Risk-Models-.pdf

[6]Los Angeles Police Commission, Office of the Inspector General. (2019). Review of selected Los Angeles Police Department data-driven policing strategies. https://www.oig.lacity.org/

[7]European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

supervisory decision. The boundary is blurry, and some of the most troubling applications, like Chicago's Strategic Subject List, sat squarely on it.

2. Three Technologies, Often Conflated

2.1 Place-based forecasting

The most widely deployed systems forecast when and where the crime is likely to happen, typically at the resolution of a 150-by-150-meter cell over a single shift. The philosophical lineage extends to environmental criminology: crime concentrates at micro-places, victimization that is repeated or almost repeated follows regular patterns, and the short-term risk to neighboring homes is quantifiably increased by yesterday's burglary. PredPol's core model was a self-exciting point process borrowed from seismology, treating crimes like earthquakes that produce aftershocks.[8] The promise was a more disciplined version of hot-spots policing, a strategy with a reasonably strong evaluation record of its own.

It matters that these systems forecast events recorded by police, not crime itself. That distinction, which I take up in Section 4, does most of the analytic work in critical literature.

2.2 Person-based forecasting

A second family of tools which ranks the individuals by their estimated risk of involvement in violence, as perpetrator, victim or both. Chicago's Strategic Subject List was the most prominent American example: an algorithm developed at the Illinois Institute of Technology scored individuals on factors such as earlier arrests, records of gang affiliation and co-arrest network position, and the highest-scoring people were referred to district commanders for preventive attention (Saunders et al., 2016). Kansas City, New Orleans, and Los Angeles (through its LASER program) ran variants. Person-based forecasting raises the sharpest civil liberties questions, because it attaches state suspicion to named individuals who may have done nothing actionable, and it is the application the EU has now prohibited.

2.3 Decision-point risk assessment

The third family is older than either, descending from the parole prediction tables of the 1920s. For perspective, that ancestry is important. Actuarial risk assessment was not created by the tech industry, and its fundamental discovery—that a straightforward statistical combination of a few variables predicts recidivism at least as well as clinical judgment—has been repeated for more than a century. Among the contemporary tools are COMPAS, the Public Safety Assessment, and the federal Post Conviction Risk Assessment. To create a risk classification utilized for pretrial release, sentence, or supervision decisions, they blend criminal history with demographic and occasionally attitudinal factors. Later generations add "criminogenic needs" that are meant to direct therapy in addition to categorization. The underlying data have not altered as much during the last ten years.

[8]Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., & Brantingham, P. J. (2015). Randomized controlled field trials of predictive policing. Journal of the American Statistical Association, 110(512), 1399–1411. https://doi.org/10.1080/01621459.2015.1077710

It is the scope, commercialization, and transfer of these ratings from parole boards, where they influenced discretionary release, to bail and sentencing, where they influence decisions about liberty earlier and with less stringent procedural protections. These tools also differ from patrol forecasting in that they operate inside formal legal proceedings, which entangles their outputs with due process in a way a heat map never faces. The leading American case remains the decision of Wisconsin Supreme Court in *State v. Loomis*,[9] which permitted sentencing courts to consult COMPAS despite its proprietary opacity so long as judges were warned of its limitations, and its reasoning has satisfied almost no one on either side of the debate.

The evidence for and against each family is different, which is why the three should not be conflated. The efficacy evidence is strongest for place-based forecasting, which also draws the weakest individual-rights objections. For person-based forecasting the situation is inverted, since the efficacy evidence there is weakest and the objections strongest. Decision-point risk assessment falls in the middle, backed by a long psychometric pedigree but carrying an unresolved fairness problem.

3. Does Any of It Work?

Occasionally, modestly and less frequently than the procurement documents suggested. That's the truthful response.
The baseline that these technologies were intended to improve upon must be the starting point for any fair evaluation, and it wasn't nothing. By the time the first predictive software was introduced, hot-spot policing had already established one of the best evaluation records in the field. Over an extended length of time, randomized and quasi-experimental investigations have demonstrated that focusing patrol at high-crime micro-places results in modest crime reductions, usually without merely moving crime around the corner. The bottom line of the business argument was that algorithms could choose those micro-places more quickly, with finer temporal resolution, and with less analyst work than the current approach. That is a true but purposefully limited assertion, and a large portion of the subsequent disappointment was predetermined. The algorithm's ceiling is the ceiling of hot-spot policing itself if it is a less expensive method of selecting boxes on a map. The important question is whether small advances in cluster selection endure contact with dispatch demands, officer discretion, and the typical entropy of patrol work, rather than whether the software can identify clusters, which is statistically uninteresting.

The strongest evidence for place-based forecasting comes from Mohler et al.,[10] who ran randomized controlled field trials in Los Angeles and Kent (UK) comparing algorithmic predictions against forecasts produced by experienced crime analysts. The algorithm identified roughly twice as much crime as the analysts at the same patrol budget, and patrol guided by it associated with measurable crime reductions. This is a real result, and critics who claim technology "doesn't work" at all overstate their case. But three qualifications matter. The comparison was algorithm versus analyst, not algorithm versus a well-run conventional hot-spots program. The authors included PredPol's founders, a conflict the paper disclosed but which counsels' independent replication. And

[9] *State v. Loomis, 881 N.W.2d 749 (Wis. 2016).*
[10] Mohler et al., 2015.

the outcome was short-term crime counts in treated cells, which says nothing about displacement, durability, or community effects.

Independent evaluations have been notably less encouraging. The RAND evaluation of Shreveport's experiments on predictive policing found no significant reduction in property crime in treatment districts.[11] Ratcliffe et al.[12] conducted the most rigorous independent test to date, a randomized experiment across twenty Philadelphia districts, and found that awareness of predicted grids by patrol officers produced no significant crime reduction; marked-car dosage produced at best a small, temporary effect on property crime that did not extend to violence. A systematic review by Meijer and Wessels[13] concluded that claimed benefits rested largely on weak designs and vendor-affiliated studies.

It is worse than the person-based record. When Saunders et al.[12] evaluated the initial iteration of Chicago's Strategic Subject List, they discovered that while those on the list were significantly more likely to be arrested for a shooting, they were neither more nor less likely to be shot than matched comparisons. This is likely because officers used the list as a roster of suspects when shootings took place. The program, in other words, did not deliver the promised prevention; it delivered additional enforcement exposure to the listed individuals. Most of the subsequent critical literature is anticipated by that one discovery.

Additionally, the accuracy claims for decision-point risk assessment should be examined. COMPAS receives 137 inputs, however Dressel and Farid[13] demonstrated that it was no more accurate at predicting recidivism than two-variable linear model or untrained internet workers who were provided with two information about a defendant. In this field, predictive performance appears to soon reach a low ceiling, and sophistication doesn't really help you get past it. A different concern is whether even a small amount of accuracy results in better choices. Stevenson examined Kentucky's[14] ten years of mandatory pretrial risk assessment in over a million cases and discovered that while the tool altered bail-setting procedures, it only slightly increased pretrial release as judges reverted to their previous practices. The impact of an algorithm is felt by the people who use it, and their compliance is selective. Based on bail statistics from New York City, Kleinberg et al.[15] contend that machine forecasts could enable significant concurrent drops in both imprisonment and crime. However, that outcome is a simulation of potential rather than an observation of practice, and the most challenging issues in the discipline are found in the space between the two.

---

[11]Hunt, P., Saunders, J., & Hollywood, J. S. (2014). Evaluation of the Shreveport predictive policing experiment (RR-531). RAND Corporation. https://www.rand.org/pubs/research_reports/RR531.html

[12]Ratcliffe, J. H., Taylor, R. B., Askey, A. P., Thomas, K., Grasso, J., Bethel, K. J., Fisher, R., & Koehnlein, J. (2021). The Philadelphia predictive policing experiment. Journal of Experimental Criminology, 17(1), 15–41. https://doi.org/10.1007/s11292-019-09400-2

[13]Meijer, A., & Wessels, M. (2019). Predictive policing: Review of benefits and drawbacks. International Journal of Public Administration, 42(12), 1031–1039. https://doi.org/10.1080/01900692.2019.1575664

After two decades, the record supports a plain summary. Place-based forecasting can concentrate patrol slightly more efficiently than human analysts, though independent tests put its crime effects anywhere from modest to undetectable. Person-based forecasting has no credible record of preventing violence. Decision-point risk assessment predicts about as well as simple checklists and changes decisions less than its sponsors hoped. None of this would justify investment on efficacy grounds alone. The technologies have persisted for institutional reasons, because they organize work, satisfy demands for "data-driven" management, and, as Brayne[14] documented inside the LAPD, because they serve organizational purposes well beyond crime reduction.

4. Dirty Data: The Input Problem

Every forecasting system in this domain learns from records generated by police activity, and police activity is not a neutral sensor. Crimes appear in the data when someone reports them or when an officer observes them; the latter channel, which dominates drug, weapon, and public-order offenses, reflects deployment choices as much as underlying behavior. Arrest data add a second filter, since arrest requires discretionary action by an officer. The result is that the dependent variable in most of these models is best described not as crime but as the intersection of crime and policing.

Lum and Isaac[15] made this concrete in a study that has become the canonical citation. Using public-health survey data to estimate the true geography of illicit drug use in Oakland, they showed that While drug usage was widespread throughout the city, drug arrests were primarily concentrated in low-income and predominately Black neighborhoods. When they ran PredPol's published algorithm on the arrest data, the software directed patrols overwhelmingly to those same neighborhoods. The model they observed predicted future policing rather than future crime. The phrase has been quoted so often that its precision gets lost: the claim is not that the algorithm was malicious or even miscoded, but that an accurate model of a biased measurement process reproduces the bias with statistical authority.

Richardson, Schultz, and Crawford[16] extended the point from measurement error to something closer to evidentiary contamination. Examining thirteen jurisdictions that deployed predictive systems while under federal investigation or consent decree for unconstitutional policing, they documented cases, Chicago, New Orleans, and Maricopa County among them, where data generated during periods of adjudicated civil rights violations flowed directly into predictive systems. Records produced by unlawful stops, falsified reports, or discriminatory arrests carry no marker in a database. They occupy the same fields as everything else. And no technical debiasing procedure can

[14]Brayne, S. (2017). Big data surveillance: The case of policing. American Sociological Review, 82(5), 977–1008. https://doi.org/10.1177/0003122417725865

[15]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

[16]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, 15–55. https://www.nyulawreview.org/online-features/dirty-data-bad-predictions-how-civil-rights-violations-impact-police-data-predictive-policing-systems-and-justice/

recover ground truth from records of conduct a court has found unconstitutional, because the needed counterfactual, what enforcement would have recorded under lawful policing, was never observed.

Two complications deserve more attention than they usually receive. First, the severity of the input problem varies by crime type. Homicide and motor vehicle theft are reported and recorded at high rates regardless of police deployment; drug and nuisance offenses are almost entirely deployment driven. A forecasting system trained on burglary reports inherits far less enforcement bias than one trained on drug arrests, which is why blanket claims that "the data are biased" are as imprecise as blanket claims that they are not. Brantingham, Valasik, and Mohler,[17] Reexamining arrest data from the Los Angeles trial revealed no discernible difference in the racial makeup of arrests between algorithm-directed and analyst-directed patrols. This finding supports the theory that the algorithm did not introduce bias beyond what was present in the baseline practice for the violent and property crimes that the system targeted. The finding is narrow, vendor-affiliated authorship again counsels caution, and "no worse than the status quo" is a low bar, but it is part of the record and should be cited alongside Lum and Isaac rather than suppressed.

Second, the input problem afflicts person-based tools through criminal history. Skeem and Lowenkamp,[18] studying the federal PCRA on a sample of nearly 35,000 offenders, found little evidence of test bias in the psychometric sense: a given score meant roughly the same recidivism probability for white and black offenders. But higher score on average were achieved by black offenders, and two-thirds of that gap ran through criminal history, a variable that encodes prior enforcement patterns and that sentencing systems already count. The instrument can be unbiased as a measurement device and still function as a conveyor belt moving historical enforcement disparities into future decisions. That distinction, between test bias and disparate impact, structures the entire fairness debate, and conflating the two has produced a great deal of crosstalk.

5. Feedback Loops: When Predictions Manufacture Their Own Confirmation

The static input problem becomes dynamic once a system's output influences the data it is retrained on. Send officers to a forecasted cell and they will observe more of whatever is visible there; record those observations; retrain; forecast the same cell with greater confidence. Ensign, Friedler, Neville, Scheidegger, and Venkatasubramanian[19] formalized this intuition using Pólya urn models and proved that under realistic conditions regardless of the true crime rate, a predictive system fed its own discovered incidents converges on repeatedly dispatching police to the same neighborhoods with the uncertainty of the estimates shrinking even as their accuracy does not improve. They also showed the loop is corrigible in principle: down-weighting incidents discovered

[17]Brantingham, P. J., Valasik, M., & Mohler, G. O. (2018). Does predictive policing lead to biased arrests? Results from a randomized controlled trial. Statistics and Public Policy, 5(1), 1–6. https://doi.org/10.1080/2330443X.2018.1438940

[18]Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. Criminology, 54(4), 680–712. https://doi.org/10.1111/1745-9125.12123

[19]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. https://proceedings.mlr.press/v81/ensign18a.html

through algorithm-directed patrol, so that the system learns mainly from reported crime, prevents the runaway dynamic.

Three observations about this literature. First, the mathematical result is about discovered incidents, officer-initiated observations, which again means the loop is most vicious for drug and order offenses and weakest for well-reported crimes. Second, the empirical footprint of feedback loops in deployed systems is harder to establish than the theory, because departments rarely log which recorded incidents trace to algorithm-directed patrols; the absence of evidence here reflects an absence of instrumentation, not reassurance. Third, the loop has a human-capital analogue that ethnographers have captured better than statisticians. Brayne,[20] embedded with the LAPD, described how point-based systems like LASER instructed officers to accumulate contact cards on flagged individuals, each contact raising the person's score, each raised score inviting further contact. The individuals at the center of this cycle experienced it as harassment without accusation. The LAPD Inspector General's 2019 audit found that LASER's "chronic offender" designations were applied inconsistently, that some flagged individuals had minimal or no violent history, and the program was terminated shortly thereafter.

Feedback loops also operate at the decision-point level, though they are less discussed there. Pretrial detention itself increases the probability of conviction and future arrest; a risk score that contributes to detention thus participates in producing the outcome it predicted, and the next generation of training data will count that outcome as a successful prediction. The methodological literature has barely begun to grapple with this form of performativity, and I flag it as among the most important open problems in the field.

6. Distribution of Burdens: Over-Policing and the Communities That Absorb It

The technical concerns above have consequences that are anything but technical. The costs fall on identifiable neighborhoods, which are primarily Black, Latino, and impoverished, when forecasting algorithms focus police attention on areas where enforcement has already reached saturation. Stops, citations, arrests for minor infractions, and the accumulating criminal records that feed all downstream instruments are examples of these costs.

The mechanism's "biased algorithm" is the least intriguing variation, and accuracy is important. If recorded occurrences cluster in a public housing complex, a place-based system that directs patrol there is technically error-free. Asymmetric visibility is the source of injury. Depending on how frequently cops are present to witness it, the same behavior such as teenage drug usage, an open container, or an outstanding traffic warrant—carries drastically different repercussions. The following round of intensification is justified by the fact that ordinary activity in forecasted regions is converted into recorded violations at a higher rate than the same behavior elsewhere. This is what Ferguson[21] refers to as the accumulation of "small data indignities into big data injustice," and it explains why community protests continue even when departments can honestly claim that there is

[20]Brayne, 2017. Brayne, 2021.

[21]Ferguson, A. G. (2017). The rise of big data policing: Surveillance, race, and the future of law enforcement. NYU Press.

no race variable in the model. Because location, criminal history, and network position incorporate race, it is not necessary for it to be a part of the model for it to be a characteristic of the outcome.

At its peak the Chicago SSL contained hundreds of thousands of entries, and Upturn's analysis of released data found that more than half of Black men in their twenties in Chicago appeared on it. Inclusion was not accompanied by notice, explanation, or any mechanism for removal, and as Saunders et al.[22] showed, its operational effect was elevated arrest exposure rather than protection. When a state instrument touches a majority of a demographic group in a city while remaining invisible to those it touches, the appropriate analytic frame is no longer prediction policy but surveillance governance.

There is also a system-level cost that the patrol-allocation literature rarely prices in: the effect of concentrated, algorithmically justified enforcement on the willingly cooperation of residents with police at all. Desmond, Papachristos, and Kirk[23] showed that a single extremely publicized incident of police violence against a Black man who was unarmed in Milwaukee was followed by a sharp, sustained drop in 911 calls from Black neighborhoods, an estimated tens of thousands of lost calls over the following year. The mechanism, legal cynicism deepening into withdrawal, bears directly on predictive systems for two reasons. Normatively, it identifies harm that never appears in treated-cell crime counts. Methodologically, it threatens the data pipeline itself: reported crime is the input these models depend on most and trust most, and reporting is endogenous to how policing is experienced. On its own terms, a forecasting system that weakens the community's desire to notify the police is deteriorating its best sensor. As far as I'm aware, there isn't a deployed system that tracks reporting propensity in areas with high forecasts, and there isn't an assessment that has considered it an outcome. That disparity reveals a lot about whose expenses the field has taken into account.

I want to resist, however, the occasional implication in critical literature that the alternative to algorithmic policing is unproblematic policing. The discrepancies that these systems replicate arise from human judgment; the pin maps and unofficial suspect lists that came before the software encoded the same patterns with even less reviewability and uniformity. Formalization creates an auditable artifact and allows for measured correction, but it also launders discretion through an idiom of objectivity, scales it, and hardens it against challenge. This is a serious question raised by scholars such as Selbst.[24]

The empirical answer likely varies by context, and we have remarkably few studies designed to provide it.

[22]Saunders et al., 2016.

[23]Desmond, M., Papachristos, A. V., & Kirk, D. S. (2016). Police violence and citizen crime reporting in the Black community. American Sociological Review, 81(5), 857–876. https://doi.org/10.1177/0003122416663494

[24]Selbst, A. D. (2017). Disparate impact in big data policing. Georgia Law Review, 52(1), 109–195. https://doi.org/10.2139/ssrn.2819182

7. Identifying Suspects: Facial Recognition and the Anatomy of a Wrongful Arrest

Forecasting where crime will occur and identifying who committed a crime are different tasks, but they are converging in practice, and the identification side of the ledger has produced the clearest documented harms to named individuals. Police use of facial recognition deserves treatment in any review of algorithmic policing because it exhibits, in compressed form, every failure mode this article has discussed: demographically uneven error, opaque process, and accountability that arrives, if at all, only through litigation.

The technical predicate is well established. The evaluation of 189 face recognition algorithms by National Institute of Standards and Technology found demographic differentials in the large majority of them, with false positive rates for African American and Asian faces exceeding those for white faces by factors of 10 to 100 in one-to-one matching, depending on the algorithm, and elevated false positive rates for Black women in the one-to-many searches that police actually run against mugshot and license databases.[25] A false positive in a one-to-many search is not an abstraction; it is a name and a face delivered to an investigator with the imprimatur of a match.

What happens next depends on investigative discipline, and the documented record shows how thin that discipline can be. In January 2020, After a facial recognition search matched a blurry surveillance footage of a shoplifter to his expired driver's license photo; Robert Williams was arrested by Detroit police in front of his familyand then he was held for roughly thirty hours before the error unraveled.[26] His was the first publicly reported wrongful arrest traceable to a false face recognition match, and the known subsequent cases share a stark pattern: nearly all of the publicly documented victims have been Black. Every agency policy describes technology as generating "investigative leads" requiring independent corroboration; in the documented failures, the lead effectively became the case, with corroboration reduced to rituals, a photo lineup seeded by the algorithm's own output, that recycle rather than test the machine's hypothesis. Williams's federal lawsuit ended in a June 2024 settlement that imposed what the ACLU describes as the strongest departmental constraints on the technology in the country, including a bar on arrests based on facial recognition results alone.[27] That the constraint had to be created by settlement, four years after the arrest, is itself the accountability finding.

Years earlier, Garvie, Bedoya, and Frankle[28] had estimated that the faces of roughly half of American adults already sat in law enforcement-searchable networks, almost entirely without

[25]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

[26]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

[27]American Civil Liberties Union. (2024). Williams v. City of Detroit: Face recognition false arrest. https://www.aclu.org/cases/williams-v-city-of-detroit-face-recognition-false-arrest

[28]Garvie, C., Bedoya, A. M., & Frankle, J. (2016). The perpetual line-up: Unregulated police face recognition in America. Georgetown Law Center on Privacy & Technology. https://www.perpetuallineup.org/

legislative authorization, audit requirements, or accuracy standards. The wrongful arrests that followed were, in that light, less a malfunction than a forecast fulfilled. State legislatures have begun imposing guardrails, and the pattern of regulation arriving only after demonstrable injury, rather than before deployment, is the governance lesson that generalizes to every technology reviewed here.

8. The Fairness Impasse: What COMPAS Actually Taught Us

The modern debate was detonated by ProPublica's 2016 investigation of COMPAS scores in Broward County, Florida. Angwin, Larson, Mattu, and Kirchner[29] found that while white reoffenders were more likely to have been classified as low risk, Black defendants were almost twice as likely as white defendants to have been classified as high risk among those who did not commit new crimes. Northpointe, the vendor, replied that the instrument was calibrated, meaning that at any given score, Black and white defendants reoffended at similar rates. Both claims were essentially correct, and the apparent paradox produced the most consequential theoretical result in the field. Chouldechova[30] and Kleinberg, Mullainathan, and Raghavan[31] independently proved that When the base rate of the anticipated result varies across two groups, no instrument can satisfy calibration and equal error rates across groups at the same time. Therefore, the contradiction that ProPublica revealed was a mathematical feature of prediction under unequal base rates rather than a flaw in a single product.

The impossibility results pushed the debate onto healthier ground. The question stopped being whether the algorithm is biased and became which fairness property we should sacrifice, and who gets to decide. That is not a statistical question. Equalizing false positive rates protects non-reoffenders in the higher-base-rate group from erroneous detention. Preserving calibration protects the meaning of the score for decision-makers. The choice between them allocates errors across racial groups and cannot be delegated to engineers. Berk, Heidari, Jabbari, Kearns, and Roth[32] catalogue the trade-offs and conclude that criminal justice applications force explicit value choices that current procurement and deployment practice leaves implicit, unexamined, and effectively made by default.

Two further lessons from the COMPAS episode are underappreciated. First, base rates themselves are measured with the dirty data of Section 4, rearrest standing in for reoffending, so

---

[29]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

[30]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

[31]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–Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[32]Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research, 50(1), 3–44. https://doi.org/10.1177/0049124118782533

even the framing "different base rates" partially encodes enforcement disparity. Second, as Dressel and Farid[33] showed, the entire controversy concerned an instrument whose accuracy a two-item checklist could match. We conducted a national argument about the fairness of a black box that contained, functionally, age and prior record. That fact should discipline both the promotional and the dystopian registers of this debate.

9. Explainability and Accountability: Who Answers for an Algorithm's Influence?

The opacity objection comes in two distinct forms. The first is technical, since some model classes do not yield human-followable reasons for individual outputs. The second is proprietary, when vendors claim trade confidentiality over models that are rather easy to understand. The second form has been important in criminal justice. Loomis focused on Northpointe's refusal to reveal how COMPAS weighed its inputs, and the court's chosen remedy warning label attached to presentence reports—has come under fire for being ceremonial rather than protective. Rudin[34] makes a strong case that the concept of "explaining black boxes" as a whole makes too many concessions. The accuracy-transparency trade-off that vendors describe is basically fictional in this area because interpretable models perform essentially as well as opaque ones for high-stakes choices on this type of tabular data. Nothing in the COMPAS performance record indicates that complexity was purchasing accuracy, and the criminological evidence backs her up.

Because explainability is a prerequisite for contestation, which is where American practice is most deficient, explainability is important. A defendant scored high-risk in most jurisdictions cannot learn the inputs, cannot correct errors in them, and cannot cross-examine a developer. Consider the minimum that contestation would require. A defendant would need notice that a score was used, access to the inputs, a mechanism for correcting factual errors, and a forum in which the instrument's relevance to this defendant can be argued. The error-correction piece matters most, since criminal history records are famously error-prone and an error in the input is the most tractable kind of algorithmic injustice. Not a single American jurisdiction supplies all four as a matter of course, and most supply none. A resident of a chronically forecasted block has even less than a defendant does. Patrol allocation has never been subject to individualized due process, so no procedural vehicle exists at all, and the algorithm inherits that vacuum. The disparity is seen with other forensic technology. Because they enter via the courtroom door, breathalyzers, DNA techniques, and radar guns are subject to admissibility issues, calibration records, and defense examination. Outcomes are shaped upstream of any adversarial process by risk scores and patrol projections. Because of this, procurement—the choice to purchase and implement—has emerged as the de facto site of governance. Agencies that engage in procurement are not required to consult the people who will bear the consequences.

The dispersion of responsibility along the sociotechnical chain is the underlying accountability issue. The officer points to the model when an algorithmically flagged stop goes awry, the department points to the vendor, the vendor points to the data, and the data points back to

[33]Dressel & Farid, 2018.

[34]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

decades of practice that no one in the office is now permitted. Each actor's contribution is real and partial; the institutional arrangements ensure that no actor's contribution is sufficient for blame. Brayne[35] adds an ethnographic irony: officers themselves resist these systems partly because the data trails track them, and managers deploy "the algorithm said so" as cover in both directions, to justify aggressive deployment and to deflect criticism of it. Accountability structures designed for human discretion do not map onto distributed human-machine decisions, and almost no American jurisdiction has redesigned them. The few exceptions, algorithmic accountability ordinances in cities like New York, and the surveillance technology oversight ordinances that several California municipalities used to ban predictive policing outright beginning with Santa Cruz in 2020, remain young, uneven, and largely unevaluated.

The European response is instructive as contrast. The AI Act classifies law-enforcement risk assessment systems as high-risk, imposing documentation, human oversight, and audit obligations, and flatly prohibits individual crime prediction based solely on profiling.[36]

Whether prohibition or regulated use better serves the affected communities is an empirical question Europe and America are now running as a natural experiment, and criminologists should treat it as one.

10. A Research Agenda: What NIJ Should Buy

The National Institute of Justice's current solicitation for evaluation and research of artificial intelligence across policing, courts, and corrections asks applicants to examine operational suitability, impacts, and downstream risks.[37] Having reviewed the record, I would put that investment toward five gaps, in rough order of how badly the field needs them filled.

The first gap: evaluate decisions rather than predictions. The literature is saturated with area-under-the-curve statistics and starved of studies that measure what changes when a forecast reaches a human. Stevenson's[38] Kentucky study and the Philadelphia experiment[39] are nearly alone in examining implementation at scale, and in both the outcome was determined by human response, judicial drift in one case, officer disengagement in the other. Every funded evaluation should treat the human-algorithm dyad, not the model, as the unit of analysis.

The second gap is instrumentation of the feedback loop. The runaway dynamics proven by Ensign et al.[40] cannot be detected in deployed systems unless departments tag which recorded incidents arise from algorithm-directed activity. Requiring that instrumentation as a condition of funding would convert a theoretical literature into an empirical one, at modest cost.

---

[35]Brayne, 2021.

[36]European Union, 2024.

[37]National Institute of Justice, 2025.

[38]Stevenson, M. T. (2018). Assessing risk assessment in action. Minnesota Law Review, 103, 303–384. https://minnesotalawreview.org/article/assessing-risk-assessment-in-action/

[39]Ratcliffe et al., 2021.

[40]Ensign et al., 2018.

The third gap is community-side outcomes. Twenty years in, we have almost no rigorous estimates of what sustained algorithmic targeting does to stop rates, low-level arrest rates, perceived legitimacy, or willingness to call police in forecasted areas. Crime counts in treated cells are the easiest outcome to measure and very nearly the least informative.

Fourth, test interpretable baselines. Rudin's[41] challenge has a simple evaluative implication: every procurement of a proprietary instrument should be benchmarked against a transparent checklist of the kind Dressel and Farid[42] constructed. If the black box cannot beat the checklist, and so far it generally cannot, jurisdictions are paying in opacity for nothing.

Fifth, study governance experiments. The EU prohibition, municipal bans, consent-decree data exclusions, and the post-2023 absorption of forecasting into integrated platforms each create comparison conditions. The discipline that built an evidence base for hot-spots policing can build one for algorithm governance; it has simply not been asked to.

A final word on scope. The technologies reviewed here are the ones with a research record, but the solicitation that frames this agenda is broader, and deliberately so.[43] Agencies are already experimenting with generative systems that draft incident reports, summarize body-camera footage, and triage investigative leads, applications that will reshape the data pipeline feeding every predictive tool discussed above before most evaluations of those tools are complete. The lesson of the past fifteen years should be applied prospectively once: the time to build evaluation, audit logging, and community consultation into these systems is at adoption, not after the first scandal. Criminology spent a decade documenting harm that better-designed studies, demanded earlier, might have prevented or at least surfaced sooner. The discipline will not get a better invitation than the present one to avoid repeating that sequence.

11. Conclusion

The predictive turn in criminal justice rested on a syllogism. Crime is patterned, patterns can be modeled, and therefore modeling will reduce crime. Every premise is accurate. Because the model interacts with the world through data that documents enforcement rather than offenses, through institutions that selectively act on projections, and through communities that take in attention the forecasts direct, the conclusion still does not follow. The vendor's assertion that algorithms eliminate human prejudice is not supported by any of the records examined here, nor is it supported by the most compelling argument that they only automate it. These methods clearly distribute accountability for the outcomes while concentrating, formalizing, and lending statistical authority to whatever practices produced their training data.

All of this does not support giving up on prediction. For a century, risk has been evaluated in criminal justice, and no one considers unassisted human judgment to be the gold standard. These are consequential interventions into the distribution of state coercion, and they should be held to the evidentiary and accountability standards that apply to any such intervention. This is the stance that

[41]Rudin, 2019.
[42]Dressel & Farid, 2018.
[43]National Institute of Justice, 2025.

the evidence supports. Deployment ought to be postponed until after independent assessment. There should be notice and contestability associated with scoring. Systems shouldn't be retrained using data generated by their own deployment, and jurisdictions shouldn't purchase opacity that transparent models can match. Technically speaking, none of these criteria are difficult.

The role of the research community in providing evidence that makes evasion visible is more important now than it was when the first heat maps were created because all of them are institutionally demanding.

References

American Civil Liberties Union. (2024). Williams v. City of Detroit: Face recognition false arrest. https://www.aclu.org/cases/williams-v-city-of-detroit-face-recognition-false-arrest

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

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research, 50(1), 3–44. https://doi.org/10.1177/0049124118782533

Braga, A. A., Papachristos, A. V., & Hureau, D. M. (2014). The effects of hot spots policing on crime: An updated systematic review and meta-analysis. Justice Quarterly, 31(4), 633–663. https://doi.org/10.1080/07418825.2012.673632

Brantingham, P. J., Valasik, M., & Mohler, G. O. (2018). Does predictive policing lead to biased arrests? Results from a randomized controlled trial. Statistics and Public Policy, 5(1), 1–6. https://doi.org/10.1080/2330443X.2018.1438940

Brayne, S. (2017). Big data surveillance: The case of policing. American Sociological Review, 82(5), 977–1008. https://doi.org/10.1177/0003122417725865

Brayne, S. (2021). Predict and surveil: Data, discretion, and the future of policing. Oxford University Press.

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

City of Chicago Office of Inspector General. (2020). Advisory concerning the Chicago Police Department's predictive risk models. https://igchicago.org/wp-content/uploads/2020/01/OIG-Advisory-Concerning-CPDs-Predictive-Risk-Models-.pdf

Desmond, M., Papachristos, A. V., & Kirk, D. S. (2016). Police violence and citizen crime reporting in the Black community. American Sociological Review, 81(5), 857–876. https://doi.org/10.1177/0003122416663494

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. https://proceedings.mlr.press/v81/ensign18a.html

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

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

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

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

Hunt, P., Saunders, J., & Hollywood, J. S. (2014). Evaluation of the Shreveport predictive policing experiment (RR-531). RAND Corporation. https://www.rand.org/pubs/research_reports/RR531.html

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. Quarterly Journal of Economics, 133(1), 237–293. https://doi.org/10.1093/qje/qjx032

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–Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Los Angeles Police Commission, Office of the Inspector General. (2019). Review of selected Los Angeles Police Department data-driven policing strategies. https://www.oig.lacity.org/

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

Meijer, A., & Wessels, M. (2019). Predictive policing: Review of benefits and drawbacks. International Journal of Public Administration, 42(12), 1031–1039. https://doi.org/10.1080/01900692.2019.1575664

Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., & Brantingham, P. J. (2015). Randomized controlled field trials of predictive policing. Journal of the American Statistical Association, 110(512), 1399–1411. https://doi.org/10.1080/01621459.2015.1077710

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. Annual Review of Clinical Psychology, 12, 489–513. https://doi.org/10.1146/annurev-clinpsy-021815-092945

National Institute of Justice. (2025). NIJ FY25 research and evaluation of artificial intelligence for criminal justice purposes (Funding opportunity O-NIJ-2025-172615). U.S. Department of Justice, Office of Justice Programs. https://nij.ojp.gov/funding/opportunities/o-nij-2025-172615

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). Predictive policing: The role of crime forecasting in law enforcement operations (RR-233). RAND Corporation. https://www.rand.org/pubs/research_reports/RR233.html

Ratcliffe, J. H., Taylor, R. B., Askey, A. P., Thomas, K., Grasso, J., Bethel, K. J., Fisher, R., & Koehnlein, J. (2021). The Philadelphia predictive policing experiment. Journal of Experimental Criminology, 17(1), 15–41. https://doi.org/10.1007/s11292-019-09400-2

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, 15–55. https://www.nyulawreview.org/online-features/dirty-data-bad-predictions-how-civil-rights-violations-impact-police-data-predictive-policing-systems-and-justice/

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

Saunders, J., Hunt, P., & Hollywood, J. S. (2016). Predictions put into practice: A quasi-experimental evaluation of Chicago's predictive policing pilot. Journal of Experimental Criminology, 12(3), 347–371. https://doi.org/10.1007/s11292-016-9272-0

Selbst, A. D. (2017). Disparate impact in big data policing. Georgia Law Review, 52(1), 109–195. https://doi.org/10.2139/ssrn.2819182

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. Criminology, 54(4), 680–712. https://doi.org/10.1111/1745-9125.12123

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

Stevenson, M. T. (2018). Assessing risk assessment in action. Minnesota Law Review, 103, 303–384. https://minnesotalawreview.org/article/assessing-risk-assessment-in-action/