MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1022-1026

Artificial Intelligence in Criminal Justice Management: Ethical Issues, Challenges, and Implications for Fair Governance

Ms. Rimpy Bhardwaj
Assistant Professor, Department of Law Maharishi Markandeshwar (Deemed to be University) Mullana- Ambala
Vishant Saini
Assistant Professor, Department of Law, Maharishi Markandeshwar (Deemed to be University) Mullana - Ambala
Dr. Anil Vats
Assistant Professor, Department of Law, Maharishi Markandeshwar (Deemed to be University)
Dr. Rakhee Rani
Assistant Professor of English, Department of Law, Maharishi Markandeswar (Deemed to be University), Mullana - Ambala
S.P. Saini
Professor, Department of Law, Maharishi Markandeshwar (Deemed to be University) Mullana - Ambala
Ms. Suman Khanna
Assistant Professor, Department of Law Maharishi Maakadeshwar Deemed to be University, Mullana- Ambala

Abstract
The integration of artificial intelligence (AI) into criminal justice management represents one of the most consequential technological developments of the twenty-first century. From predictive policing and facial recognition systems to risk assessment algorithms and automated sentencing tools, AI applications are reshaping law enforcement, judicial processes, and correctional practices across the globe. While proponents argue that AI enhances efficiency, objectivity, and public safety, critics raise profound concerns about algorithmic bias, violations of privacy, lack of transparency, and threats to fundamental human rights and due process. This research article critically examines the ethical issues, governance challenges, and implications for fair governance arising from the deployment of AI in criminal justice systems. Drawing on secondary data from academic literature, government reports, policy documents, and international organization guidelines, the study employs qualitative, exploratory research design, utilizing thematic, content, and comparative analysis. Findings reveal systematic patterns of racial and socioeconomic bias embedded in AI tools, pervasive surveillance risks, accountability deficits, and inadequate regulatory frameworks. The article offers concrete policy recommendations for transparent, human-centered, and ethically governed AI in criminal justice.
Keywords: *artificial intelligence, criminal justice, algorithmic bias, predictive policing, ethical governance, facial recognition, COMPAS, surveillance, human rights, fair governance*

1. Introduction

Artificial intelligence has emerged as a transformative force across virtually every domain of public life, and criminal justice management is no exception. Over the past two decades, law enforcement agencies, courts, and correctional systems in numerous countries have increasingly adopted AI-driven technologies to augment or replace traditional human decision-making. These technologies encompass a broad spectrum: predictive analytics that forecast where crimes may occur and who may commit them, facial recognition systems deployed in public surveillance, algorithmic risk assessment tools used to guide bail, sentencing, and parole decisions, and automated monitoring platforms in correctional settings (Brayne, 2017; Ferguson, 2017).The appeal of AI in criminal justice is understandable. Proponents assert that machine learning models can process vast quantities of data faster and more consistently than human officers or judges, potentially reducing arbitrary or inconsistent decisions. Advanced surveillance technologies promise to improve public safety and accelerate criminal investigation. Risk assessment instruments claim to offer evidence-based, data-driven insights into recidivism probability, ostensibly reducing subjective bias that has historically plagued sentencing decisions (Dressel & Farid, 2018). In an era of constrained public budgets and rising public safety expectations, the efficiency arguments for AI adoption are compelling to policymakers and administrators alike.However, the rapid and often poorly regulated deployment of AI in criminal justice raises urgent and fundamental ethical questions. Research has consistently demonstrated that many AI tools reproduce and amplify existing social inequalities, disproportionately targeting racial minorities, the economically disadvantaged, and other marginalized groups (Angwin et al., 2016; Eubanks, 2018). The opacity of machine learning algorithms frequently described as "black boxes" makes it nearly impossible for defendants, attorneys, or judges to scrutinize or challenge automated decisions, threatening core principles of due process and the right to a fair trial (Pasquale, 2015). Mass surveillance enabled by AI erodes privacy rights and enables authoritarian social control. Meanwhile, questions of legal accountability when AI systems make consequential errors remain largely unresolved in most jurisdictions.These tensions highlight the critical importance of examining AI in criminal justice through the lens of ethics and governance. The stakes are extraordinarily high: decisions made by or with AI assistance can deprive individuals of liberty, affect livelihoods, and shape life trajectories. Inadequately governed AI systems can institutionalize discrimination at scale, undermine public trust in justice systems, and corrode the foundational principles of equality before the law. There is, therefore, an urgent need for systematic scholarly analysis of the ethical landscape of AI in criminal justice and for rigorous policy recommendations to guide its governance.This research article addresses that need. It critically examines how AI is currently deployed in criminal justice management, identifies and analyzes the principal ethical issues and governance challenges these deployments generate, and offers recommendations for ensuring that AI in criminal justice serves the values of fairness, accountability, transparency, and human dignity. The study is grounded in a qualitative, exploratory research design, drawing on a comprehensive review of academic literature, government reports, judicial decisions, policy papers, and reports from international human rights organizations.

1.1 Research Objectives

- To examine the role and scope of AI in contemporary criminal justice management.
- To identify and analyze ethical issues associated with AI-based criminal justice systems.
- To assess challenges related to bias, transparency, accountability, and privacy.
- To suggest evidence-based policy recommendations for ethical AI governance in criminal justice.

1.2 Research Questions

- How is AI transforming criminal justice management globally?
- What ethical concerns arise from the use of AI in criminal justice systems?
- How do AI systems impact fairness, accountability, and human rights?
- What governance mechanisms can ensure ethical implementation of AI in criminal justice?

1.3 Scope and Limitations

This study focuses on AI applications in policing, judicial processes, and correctional systems, with primary reference to developments in the United States, United Kingdom, China, and India. The analysis is based exclusively on secondary data; no primary data collection or empirical experimentation was conducted. The study does not address AI in crime prevention education, forensic science laboratories, or cybercrime investigation in depth. Given the rapid pace of AI development, some specific technological details may evolve beyond the scope of the sources reviewed.

2. Literature Review

2.1 Concept of Artificial Intelligence. Artificial intelligence refers to the simulation of human cognitive processes by computer systems, including learning, reasoning, problem-solving, perception, and language understanding (Russell & Norvig, 2020). The term was first coined by John McCarthy in 1956 at the Dartmouth Conference, which is widely regarded as the founding moment of AI as an academic discipline. In the decades since, AI has evolved through several paradigms, from rule-based expert systems of the 1970s and 1980s, through the statistical learning approaches of the 1990s, to the current era dominated by machine learning and deep learning techniques driven by massive datasets and powerful computational resources (LeCun et al., 2015).
In the context of governance and law enforcement, AI applications typically rely on supervised machine learning, where algorithms learn from labeled historical data to make predictions about new cases; natural language processing for document analysis and information extraction; computer vision for facial and object recognition; and network analysis for identifying patterns of criminal association. These technologies are increasingly integrated into the operational systems of police departments, courts, and correctional institutions, often with limited public scrutiny or democratic oversight (Brayne, 2017).

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1022-1026

2.2 AI Applications in Criminal Justice Management

2.2.1 Predictive Policing: Predictive policing refers to the use of mathematical, predictive, and analytical techniques in law enforcement to forecast potential criminal activity and allocate police resources accordingly (Perry et al., 2013). Software systems such as PredPol (now Geolitica) and ShotSpotter analyze historical crime data, geographic information, and other variables to generate maps of high-risk areas, directing patrol officers to concentrate their presence in predicted hotspots. Research by Brayne (2017) documented the Los Angeles Police Department's use of Palantir's predictive policing software, noting that it drew on data from disparate sources including social media, license plate readers, and field interview cards, raising significant concerns about scope creep and discriminatory surveillance. Lum and Isaac (2016) critically demonstrated that predictive policing systems trained on historically biased crime data perpetuate and reinforce discriminatory policing patterns, creating feedback loops in which over-policed communities generate more recorded crime data, which in turn directs more police attention to those communities.

2.2.2 Facial Recognition Systems: Facial recognition technology (FRT) uses computer vision algorithms to identify or verify individuals by analyzing facial features captured in images or video footage. Law enforcement agencies in the United States, United Kingdom, China, and India have deployed FRT in public spaces, at border crossings, at mass events, and in connection with criminal investigations (Garvie et al., 2016). A landmark study by Buolamwini and Gebru (2018), known as the "Gender Shades" project, demonstrated that commercially available facial recognition systems exhibited substantially higher error rates for darker-skinned individuals, particularly women, compared to lighter-skinned individuals, revealing profound racial and gender bias in the technology. The National Institute of Standards and Technology (NIST, 2019) confirmed these findings in a large-scale evaluation of 189 facial recognition algorithms, finding that false positive rates were significantly higher for Black and Asian faces relative to White faces.

2.2.3 Risk Assessment Algorithms: Risk assessment instruments are algorithmic tools used in pretrial, sentencing, and parole decisions to estimate an individual's likelihood of reoffending or failing to appear for trial. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), developed by Equivant (formerly Northpointe), is among the most widely used such tools in the United States. An investigation by ProPublica journalists Angwin et al. (2016) analyzed COMPAS scores and actual recidivism outcomes for more than 7,000 defendants in Broward County, Florida, finding that Black defendants were nearly twice as likely as White defendants to be falsely flagged as high risk for future crime. Conversely, White defendants were more likely to be incorrectly assessed as low risk. The developers disputed these findings, triggering a significant academic debate about the appropriate metrics for evaluating algorithmic fairnessa debate that itself illuminated the fundamental value trade-offs embedded in any risk assessment system (Chouldechova, 2017; Kleinberg et al., 2016).

2.2.4 Automated Surveillance: AI-enabled surveillance systems have been deployed extensively in public spaces, combining video analytics, behavioral recognition, license plate scanning, social media monitoring, and communication interception. China's Social Credit System and urban Skynet project represent the most extensive implementation of AI surveillance in the world, integrating facial recognition, credit data, travel records, and behavioral metrics to assign citizens scores that affect access to transportation, education, and employment (Mozur, 2018). In the United Kingdom, the Information Commissioner's Office (ICO, 2019) raised concerns about the lawfulness of live facial recognition deployment by police forces, while civil liberties organizations have challenged surveillance programs in court. In India, the Automated Facial Recognition System (AFRS) proposed by the National Crime Records Bureau has attracted criticism from civil society groups over its implications for privacy and the rights of marginalized communities (Internet Freedom Foundation, 2020).

2.2.5 AI in Court and Sentencing Systems: Beyond risk assessment, AI is being explored for document analysis, legal research, bail determination support, and sentence recommendation. In Estonia, AI has been piloted for resolving small claims disputes. In the United States, some jurisdictions use algorithmic tools to guide decisions about pretrial detention, fundamentally affecting whether individuals await trial in custody or freedom (Stevenson & Doleac, 2019). The Wisconsin Supreme Court case of State v. Loomis (2016) addressed the use of COMPAS scores in sentencing, with the court ruling that while COMPAS could be considered, it could not be the determinative factora decision that nonetheless left significant ambiguity about the appropriate role of algorithmic tools in judicial decision-making.

2.3 Ethical Theories and AI in Criminal Justice

2.3.1 Utilitarianism: From a utilitarian perspective, the use of AI in criminal justice can be justified if it maximizes overall social welfare for example, by reducing crime rates, enhancing public safety, or allocating limited law enforcement resources more efficiently. However, utilitarian analysis must account for the harms imposed on individuals or groups who bear disproportionate costs of AI errors, particularly false positives in risk assessment or predictive policing that result in wrongful detention, surveillance, or stigmatization (Mill, 1863; Rawls, 1971). A pure utilitarian calculation that ignores distributional effects is insufficient when the burdens of AI errors fall systematically on already marginalized populations.

2.3.2 Deontological Ethics: Kantian deontological ethics holds that persons must be treated as ends in themselves, never merely as means. The use of AI risk scores to make incarceration decisions reduces individuals to data points and statistical probabilities, potentially violating their dignity and autonomy. Deontological reasoning also undergirds the right to a fair hearing before an impartial arbiter and the right to confront and challenge evidence rights fundamentally threatened when algorithmic scores are used without adequate transparency or opportunity for contestation (Dworkin, 1977; Kant, 1785/1998).

2.3.3 Justice and Fairness Theory: Rawls's (1971) theory of justice as fairness provides a powerful lens for evaluating AI in criminal justice. Under the difference principle, social and economic inequalities are only acceptable if they benefit the least advantaged members of society. AI systems that disproportionately burden racial minorities and economically disadvantaged individuals while providing net benefits primarily to advantaged groups cannot be justified within a Rawlsian framework. The veil of ignorance thought experiment suggests that rational actors designing justice systems without knowledge of their own social position would insist on strong protections against algorithmic discrimination.

2.3.4 Human Rights Perspective: International human rights law provides binding normative standards for evaluating AI in criminal justice. Relevant instruments include the Universal Declaration of Human Rights (UDHR, 1948), the International Covenant on Civil and Political Rights (ICCPR, 1966), the Convention against Torture, and regional instruments such as the European Convention on Human Rights. The UN Special Rapporteur on extreme poverty and human rights (Alston, 2019) warned that AI systems used in welfare and criminal justice contexts risk automating inequality and undermining human rights protections. The Council of Europe's European Commission for the Efficiency of Justice (CEPEJ, 2018) adopted Ethical Principles for AI in Justice, emphasizing that AI must respect fundamental rights, prohibit discrimination, and guarantee due process.

2.4 Research Gaps

Despite a growing body of scholarship on AI ethics in criminal justice, several significant gaps remain. First, most empirical studies are concentrated in the United States context, with limited comparative research on AI governance in Global South countries. Second, existing literature tends to address individual AI applications in isolation rather than examining the cumulative and systemic effects of multiple AI systems operating simultaneously within justice systems. Third, the voices of directly affected communitiesparticularly racial minorities and low-income populationsare underrepresented in academic and policy discourse on AI ethics. Fourth, there is limited longitudinal research examining how AI systems perform and evolve over time in real-world deployments, particularly regarding drift in model accuracy and changing social contexts (Richardson et al., 2019; Završnik, 2021).

3. Research Methodology

3.1 Research Design: This study adopts a qualitative, exploratory research design, which is appropriate for investigating complex social phenomena that are not yet fully understood and for which quantitative measurement would be insufficient to capture the normative, contextual, and interpretive dimensions of the subject matter (Creswell & Poth, 2018). The exploratory orientation reflects the relatively nascent state of empirical research on the governance and ethical implications of AI in criminal justice, and the need to map and theorize the terrain rather than test pre-specified hypotheses.

3.2 Data Collection: The study relies exclusively on secondary data sources, which include: peer-reviewed academic articles from journals such as Science, Nature, the Harvard Law Review, the Journal of Criminal Law and Criminology, and AI & Society; government reports and official documents from agencies including the U.S. Department of Justice, the U.K. Home Office, and the European Union Agency for Fundamental Rights; policy papers and reports from civil society organizations including the American Civil Liberties Union, Amnesty International, and the AI Now Institute; international organization documents from the United Nations, the Council of Europe, and the International Association of Chiefs of Police; and significant judicial decisions addressing AI tools in criminal justice proceedings. Sources were identified through systematic database searches in Google Scholar, JSTOR, LexisNexis, and the SSRN repository, using search terms including "predictive policing," "algorithmic bias," "facial recognition law enforcement," "risk assessment criminal justice," and "AI ethics governance."

3.3 Data Analysis: Three complementary analytical approaches were employed. Thematic analysis was used to identify recurring themes, patterns, and tensions across the literature regarding ethical concerns, governance challenges, and proposed solutions (Braun & Clarke, 2006). Content analysis was applied to policy

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1022-1026

documents and official reports to systematically examine the principles, commitments, and regulatory approaches articulated by governmental and intergovernmental bodies. Comparative analysis was used to examine commonalities and differences in AI governance frameworks, ethical challenges, and policy responses across different national contexts, enabling identification of transferable lessons and persistent cross-jurisdictional challenges.

4. Ethical Issues in AI-Based Criminal Justice

4.1 Algorithmic Bias and Discrimination

4.1.1 Racial Bias: The most extensively documented ethical issue in AI-based criminal justice is racial bias. Angwin et al.'s (2016) ProPublica investigation revealed that COMPAS assigned higher risk scores to Black defendants than White defendants with comparable criminal histories and demographic profiles, and that Black defendants falsely labeled as high risk at nearly twice the rate of White defendants. Dressel and Farid (2018) demonstrated that COMPAS's predictive accuracy was no better than that of untrained human volunteers asked to predict recidivism based on brief case descriptions, undermining claims of algorithmic superiority. Facial recognition studies by Buolamwini and Gebru (2018) and NIST (2019) established that error rates are substantially higher for Black and Asian individuals, creating elevated risks of false identification and wrongful investigation. These biases are not incidental technical errors; they reflect the historical data on which AI systems are trained. When algorithms learn from records of arrests, prosecutions, and convictions that are themselves products of racially discriminatory policing and judicial practices, they internalize and reproduce those patterns at scale. Richardson et al. (2019) coined the term "dirty data" to describe how problematic police practicesincluding unconstitutional stops and unlawful arrests disproportionately targeting minority communitiescontaminate the training datasets for predictive policing systems, creating self-reinforcing cycles of discriminatory surveillance.

4.1.2 Gender and Socioeconomic Bias: Bias in criminal justice AI is not limited to race. Socioeconomic factors are often proxies for protected characteristics in risk assessment algorithms: variables such as employment history, educational attainment, neighborhood of residence, and family criminal history are correlated with poverty and structural disadvantage. Eubanks (2018) documented how algorithmic systems used in social welfare and criminal justice contexts systematically disadvantage poor communities, creating what she terms a "digital poorhouse." Gender bias has also been documented in facial recognition systems (Buolamwini & Gebru, 2018), and feminist scholars have raised concerns about the potential use of AI surveillance to monitor and control women in domestic violence and reproductive rights contexts (Citron, 2014).

4.2 Privacy and Surveillance Concerns: The deployment of AI-powered surveillance in criminal justice contexts poses profound threats to privacy rights recognized under international and domestic law. Mass facial recognition in public spaces effectively eliminates the expectation of anonymity that has historically existed in public life, creating persistent records of individual movements, associations, and activities without the knowledge or consent of those surveilled (Garvie et al., 2016). The aggregation of data from multiple surveillance modalitiesfacial recognition, license plate readers, social media monitoring, mobile device tracking, and financial transaction recordsenables law enforcement to construct extraordinarily detailed profiles of individuals, far exceeding what any single data source would permit (Brayne, 2017). The chilling effects of pervasive surveillance on constitutionally protected activitiespolitical association, religious practice, journalistic inquiry, and lawful protestrepresent a distinct category of harm beyond individual privacy invasion. Research has demonstrated that awareness of surveillance reduces participation in civic activities and the exercise of expressive rights (Penney, 2016). In the United States, the American Civil Liberties Union (ACLU, 2020) documented numerous instances in which law enforcement agencies used social media monitoring and geofence warrants to identify and surveil participants in political protests, including racial justice demonstrations.

4.3 Lack of Transparency and Explainability: Many AI systems used in criminal justice function as "black boxes," producing outputsrisk scores, suspect matches, resource allocation recommendationswithout generating understandable explanations of the factors and reasoning that produced those outputs (Pasquale, 2015). This opacity creates acute problems in justice contexts. Defendants and their attorneys cannot effectively challenge AI-generated evidence or recommendations without access to detailed information about how the algorithm functions, what data it uses, how variables are weighted, and how errors are identified and corrected. Courts are ill-equipped to evaluate the reliability and validity of AI tools without adequate disclosure of model architecture, training data, validation methodology, and performance metrics (Stevenson & Doleac, 2019).

Trade secrecy claims by private vendors compound transparency problems. Companies marketing AI tools to law enforcement and courts frequently resist disclosure of proprietary algorithms, claiming competitive confidentiality. In Loomis v. Wisconsin (2016), the defendant was denied access to the source code of the COMPAS algorithm, with the court accepting the vendor's trade secret claim, notwithstanding the tool's role in a consequential sentencing decision. The European Union's General Data Protection Regulation (GDPR) provides individuals with a right to explanation in automated decision-making (Article 22), but implementation in law enforcement contexts has been inconsistent and contested (Wachter et al., 2017).

4.4 Accountability Challenges: The diffusion of decision-making authority across human and algorithmic actors in AI-assisted criminal justice creates significant accountability gaps. When an algorithmic risk score influences a judge's sentencing decision, and that decision later proves unjust, it may be unclear whether responsibility lies with the algorithm developer, the agency that deployed the tool, the official who relied on the score, or the legislative framework that permitted its use (Citron, 2014). This diffusion of responsibilitysometimes described as a "responsibility gap"can effectively shield individual decision-makers from accountability by allowing them to attribute consequential choices to the algorithm.

Legal frameworks for accountability in AI-assisted decisions remain underdeveloped in most jurisdictions. Most existing law was not designed to address automated decision-making systems, and there is significant uncertainty about whether algorithmic tools constitute "evidence" subject to disclosure requirements, whether their use can constitute unconstitutional delegation of judicial authority, and what evidentiary standards should govern their admission (Stevenson & Doleac, 2019). The absence of mandatory audit requirements, independent oversight mechanisms, and meaningful redress procedures for individuals harmed by AI errors represents a critical governance deficit.

4.5 Human Rights and Due Process Concerns: The use of AI in criminal justice raises fundamental questions about due process, equality before the law, and the right to a fair trial. Due process protections in adversarial legal systems are premised on the right of defendants to know and challenge the evidence against them, to confront witnesses, and to have decisions made by impartial human adjudicators. AI-generated evidence and recommendations that cannot be meaningfully examined, contested, or explained potentially violate these foundational guarantees (Pasquale, 2015). The International Covenant on Civil and Political Rights (Article 14) and the Universal Declaration of Human Rights (Article 10) guarantee the right to a fair and public hearing by an independent and impartial tribunalrights that are threatened when opaque algorithmic tools substantially influence judicial outcomes.

The principle of equality before the law is undermined when AI systems that function as gatekeepers to liberty and punishment are demonstrably more accurate for some demographic groups than others. Individuals from groups with higher algorithmic error rates effectively receive a lower quality of justicetheir liberty interests adjudicated by tools that are less reliable for people who look like them. This represents a systemic form of discrimination that violates both domestic equal protection guarantees and international non-discrimination norms (Alston, 2019; Human Rights Watch, 2019).

5. Challenges in Implementation: Beyond the ethical concerns identified above, the practical implementation of AI in criminal justice faces a range of technical, institutional, and social challenges that complicate efforts to govern these technologies responsibly.

5.1 Technical and Data Quality Challenges: AI systems in criminal justice are only as reliable as the data on which they are trained. Historical crime data are systematically distorted by patterns of over-policing in minority communities, underreporting of crimes in certain contexts, and inconsistent data collection practices across agencies and jurisdictions (Richardson et al., 2019). Training AI systems on such data produces models that embed and amplify historical injustices. Additionally, AI models trained in one context may perform poorly when deployed in different settings with different population demographics, crime patterns, and policing practicesa phenomenon known as "distribution shift" or "model drift" (Mehrabi et al., 2021). Regular validation and monitoring of model performance across demographic groups are necessary but often absent in practice.

5.2 Cybersecurity Risks: Criminal justice AI systems present attractive targets for malicious actors. Adversarial attacks subtle manipulations of input data designed to cause AI systems to produce incorrect outputs could potentially allow sophisticated actors to manipulate facial recognition systems, evade predictive policing surveillance, or corrupt risk assessment scores (Goodfellow et al., 2015). The centralization of sensitive personal data in law enforcement AI systems also creates significant risks of data breaches, with potentially severe consequences for the individuals whose information is exposed.

5.3 Legal and Regulatory Gaps: Most existing legal frameworks were not designed to govern AI systems, and regulatory catch-up has been slow and uneven. The United States lacks comprehensive federal AI governance legislation, relying instead on a patchwork of sector-specific regulations, state laws, and local ordinances. Several U.S. cities, including San Francisco, Oakland, and Boston, have banned municipal use of facial recognition technology, while others continue

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1022-1026

to deploy it without restriction (Garvie et al., 2016). The European Union's AI Act, adopted in 2024, represents the world's most comprehensive AI governance framework, classifying law enforcement AI applications as high-risk and imposing stringent requirements for transparency, human oversight, and conformity assessment. However, implementation and enforcement across diverse national contexts remain significant challenges.

5.4 Public Trust and Legitimacy: The legitimacy of criminal justice institutions depends in substantial part on public trust and perceived fairness. Survey research consistently shows that many communities particularly those most directly affected by criminal justice AI, including racial minorities and low-income communities are sceptical of AI tools and concerned about their fairness and accuracy (Tyler, 2006). When communities perceive AI systems as instruments of discriminatory surveillance and control rather than public safety tools, resistance and non-cooperation with law enforcement may increase, potentially undermining the effectiveness of AI-assisted policing. Building legitimacy requires meaningful community engagement, transparent communication about AI tools and their limitations, and demonstrated commitment to accountability when systems fail.

6. Case Studies

6.1 COMPAS Algorithm in the United States: The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm, developed by Equitant (formerly Northpointe Inc.), is used in pretrial, sentencing, and parole contexts across numerous U.S. states. Following ProPublica's 2016 investigation documenting its racial disparities, COMPAS became the central exhibit in national debates about algorithmic fairness in criminal justice. The Wisconsin Supreme Court's decision in State v. Loomis (2016) affirmed that judges could consult COMPAS scores in sentencing while denying defendants access to the algorithm's proprietary source code, illustrating the tension between transparency rights and trade secrecy protections. Subsequently, scholars including Chouldechova (2017) and Kleinberg et al. (2016) demonstrated mathematically that no algorithm can simultaneously satisfy multiple competing fairness criteria a finding with profound implications for the design and governance of risk assessment tools.

6.2 Facial Recognition in China: China's deployment of facial recognition technology represents the most extensive state surveillance apparatus in the world. The Skynet and Sharp Eyes systems combine hundreds of millions of cameras with AI-powered facial recognition to monitor urban populations in real time, integrated with data from mobile devices, financial transactions, and public records (Mozur, 2018). The Social Credit System uses AI-analyzed behavioral data to assign citizens scores that affect access to transportation, financial services, education, and employment. While Chinese authorities present these systems as tools for public safety and social governance, international human rights bodies and scholars have documented their use to monitor and suppress religious and ethnic minorities, particularly Uyghur Muslims in Xinjiang, in what UN experts have described as potentially constituting crimes against humanity (United Nations Human Rights, 2022).

6.3 Facial Recognition and AI Surveillance in the United Kingdom and India

The United Kingdom's Metropolitan Police Service has conducted live facial recognition trials in London since 2018, generating significant controversy and legal challenge. A landmark ruling by the Court of Appeal in R (Bridges) v. Chief Constable of South Wales Police (2020) found that the police's use of automated facial recognition violated human rights law and data protection requirements, lacking sufficiently clear legal authorization and independent oversight. The court's ruling established important precedents for human rights compliance in AI surveillance deployment. In India, the proposed Automated Facial Recognition System (AFRS) of the National Crime Records Bureau has been challenged by civil society organizations on grounds of privacy, discrimination, and the absence of legal safeguards, highlighting governance gaps in developing country contexts (Internet Freedom Foundation, 2020; Parsheera, 2019).

7. Findings and Discussion

The thematic analysis of the literature and case studies reveals several cross-cutting patterns that characterize the ethical landscape of AI in criminal justice and have significant implications for governance.First, algorithmic bias is not an anomaly or a technical malfunction is a predictable and systematic outcome of deploying machine learning systems trained on historically discriminatory data in contexts where historical inequities persist. The racial disparities documented in COMPAS risk scores, facial recognition error rates, and predictive policing deployments reflect and reinforce structural racism in criminal justice systems (Angwin et al., 2016; Buolamwini & Gebru, 2018; Richardson et al., 2019). Governance responses that treat bias as an engineering problem amenable to purely technical solutions are therefore likely to prove insufficient; meaningful remediation requires confronting the structural social conditions that generate biased training data.

Second, the current governance deficit in AI-assisted criminal justice is severe and multi-dimensional. It encompasses insufficient transparency in algorithmic decision-making, inadequate legal frameworks for accountability, weak or absent independent oversight mechanisms, and limited meaningful participation by affected communities in governance decisions. This deficit is particularly acute in jurisdictions where AI tools are developed and deployed by private vendors under trade secrecy protections, insulating them from democratic accountability and independent scrutiny (Pasquale, 2015; Wachter et al., 2017).

Third, the cross-national comparison reveals significant variation in governance approaches, with the European Union's AI Act representing the most comprehensive regulatory framework and serving as a potential model for other jurisdictions. However, even the EU framework faces challenges of consistent implementation and enforcement, particularly in law enforcement contexts where national security exemptions may dilute protection for fundamental rights. The absence of comparable international standards creates the risk of "governance arbitrage," in which AI developers and deployers migrate to less regulated jurisdictions or exploit gaps in cross-border enforcement. Fourth, the existing literature confirms that affected communities particularly racial minorities, economically marginalized populations, and immigrants bear disproportionate burdens from AI errors and harms while having the least influence over AI governance decisions. Democratizing AI governance to include these communities in design, oversight, and accountability mechanisms is both an ethical imperative and a practical prerequisite for building the public legitimacy that effective governance requires (Tyler, 2006; Eubanks, 2018).

8. Recommendations

Based on the analysis of ethical issues, governance challenges, and comparative case studies, the following recommendations are advanced for policymakers, criminal justice practitioners, AI developers, and civil society.

8.1 Establish Transparent AI Frameworks: Governments should enact legislation requiring full public disclosure of AI tools used in criminal justice, including detailed information about model architecture, training data, validation methodology, performance metrics disaggregated by demographic group, and known limitations. Vendors should be prohibited from asserting trade secrecy protections to resist disclosure of AI tools used in consequential justice decisions. Model documentation standards such as the Algorithmic Impact Statements proposed by Reisman et al. (2018) should be mandated for all AI tools deployed in law enforcement and judicial contexts.

8.2 Mandate Human-Centered Governance: AI tools in criminal justice should be designed and deployed to support, not replace, human judgment and accountability. Legislation should prohibit fully automated decisions about detention, sentencing, or supervision conditions, and should require that human decision-makers be capable of independently reviewing and overriding algorithmic recommendations. The EU AI Act's requirement for "meaningful human oversight" provides a useful model (European Parliament, 2024). Training programs for law enforcement officers, prosecutors, defenders, and judges should include education about AI capabilities, limitations, and the risks of automation bias.

8.3 Implement Mandatory Ethical Auditing: Independent, mandatory pre-deployment and ongoing auditing of AI tools in criminal justice should be required by law. Audits should assess bias across demographic groups, accuracy and validity in deployment contexts, compliance with applicable legal and ethical standards, and alignment with stated policy objectives. Auditors should be independent of both the AI developer and the deploying agency, should have full access to model documentation and training data, and audit results should be published in accessible form. Algorithmic Impact Assessments analogous to Environmental Impact Assessments should be required prior to any new AI deployment in criminal justice (Reisman et al., 2018).

8.4 Strengthen Data Protection Frameworks: Comprehensive data protection legislation should govern the collection, retention, use, and sharing of personal data in criminal justice AI systems. Such legislation should establish strict purpose limitations, data minimization requirements, rights of access and correction for individuals, meaningful consent requirements, and strong independent enforcement. Law enforcement agencies should be subject to the same data protection rules as other public and private sector entities, with any exemptions narrowly defined, clearly justified, and subject to independent oversight. The EU's Law Enforcement Directive provides one model for this framework (European Parliament, 2016).

8.5 Build Capacity and Professional Training: Effective governance of AI in criminal justice requires substantial investment in capacity-building across the justice ecosystem. Law enforcement agencies need training in the appropriate and responsible use of AI tools, the interpretation of algorithmic outputs, recognition of AI limitations and error modes, and the exercise of independent professional judgment when AI recommendations are inconsistent with other evidence. Legal professionals' judges, prosecutors, and defense attorneys need training in the technical concepts required to effectively evaluate AI evidence and arguments. Judicial training programs and bar associations should develop specialized curricula on AI in criminal justice.

8.6 Develop International Ethical Standards: The global deployment of AI in criminal justice requires international coordination to establish minimum ethical and governance standards, prevent regulatory arbitrage, and address cross-border data flows and investigative cooperation involving AI tools. The United Nations

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1022-1026

should lead a multilateral process to develop binding international standards for AI in law enforcement, drawing on the expertise of member states, civil society, AI researchers, and affected communities. Regional human rights bodies including the Council of Europe, the African Commission on Human and Peoples' Rights, and the Inter-American Commission on Human Rights should develop specific guidance on AI compatibility with regional human rights obligations. International police cooperation agreements, such as those governing Interpol, should be updated to incorporate AI governance requirements.

9. Conclusion

This research article has undertaken a systematic examination of the ethical issues, governance challenges, and implications for fair governance arising from the deployment of artificial intelligence in criminal justice management. The analysis confirms that AI technologies are rapidly transforming policing, judicial decision-making, and correctional practices in ways that hold both transformative potential and profound risks for justice, equality, and human rights.

The evidence reviewed consistently demonstrates that the benefits of AI in criminal justice are not equitably distributed, while the harms fall disproportionately on communities already subject to systemic disadvantage and discrimination. Racial bias in predictive policing, facial recognition, and risk assessment algorithms is not an isolated technical flaw but a reflection of historical injustices embedded in the training data and institutional practices on which AI systems depend. The opacity of algorithmic decision-making threatens due process guarantees; mass surveillance erodes privacy rights and chills political freedom; and accountability deficits allow consequential errors to occur without meaningful redress for those harmed.

At the same time, the analysis reveals significant variation in governance responses across jurisdictions, with some, particularly the European Union taking meaningful steps toward comprehensive AI regulation while others leave large governance gaps. This variation underscores both the feasibility and the urgency of stronger governance action. The recommendations advanced in this article for transparent AI frameworks, human-centered governance, mandatory ethical auditing, strengthened data protection, professional capacity-building, and international standards reflect a vision of AI in criminal justice that places human dignity, equality, and accountability at its center. Realizing this vision will require sustained commitment from policymakers, practitioners, developers, civil society, and the communities most directly affected by criminal justice AI. The stakes measured in individual freedoms, systemic justice, and the integrity of public institutions could not be higher. Future research should prioritize longitudinal studies of AI system performance in real-world deployments, participatory research methodologies that center the voices of affected communities, comparative analyses of governance frameworks across Global South countries, and empirical evaluation of the effectiveness of proposed governance interventions. The development of AI in criminal justice is proceeding rapidly; the development of the scholarship and governance frameworks needed to guide it must proceed with equal urgency.

References

Alston, P. (2019). Report of the Special Rapporteur on extreme poverty and human rights (A/74/48037/Add.1). United Nations Human Rights Council. https://undocs.org/A/74/493
American Civil Liberties Union. (2020). The surveillance of Black Lives Matter: Documenting law enforcement monitoring of racial justice protests. ACLU. https://www.aclu.org/report/surveillance-black-lives-matter
Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa
Brayne, S. (2017). Big data surveillance: The case of policing. American Sociological Review, 82(5), 977–1008. https://doi.org/10.1177/0003122417725865
Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81, 77–91.
Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163. https://doi.org/10.1089/big.2016.0047
Citron, D. K. (2014). Hate crimes in cyberspace. Harvard University Press.
Council of Europe, European Commission for the Efficiency of Justice (CEPEJ). (2018). European Ethical Charter on the use of artificial intelligence in judicial systems and their environment. Council of Europe Publishing.
Creswell, J. W., & Poth, C. N. (2018). Qualitative inquiry and research design: Choosing among five approaches (4th ed.). SAGE Publications.
Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580
Dworkin, R. (1977). Taking rights seriously. Harvard University Press.
Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.
European Parliament and The European Council, R. (2016). Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016, on the Protection of Natural Persons with Regard to the Processing of Personal. *Off. J. Eur. Union*, *119*, 89-131.
European Parliament. (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.
Ferguson, A. G. (2017). The rise of big data policing: Surveillance, race, and the future of law enforcement. New York University Press.
Garvie, C., Bedoya, A., & Frankle, J. (2016). The perpetual line-up: Unregulated police face recognition in America. Georgetown Law, Center on Privacy and Technology.
Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. Proceedings of the International Conference on Learning Representations (ICLR 2015). https://arxiv.org/abs/1412.6572
Human Rights Watch. (2019). How China's government is using AI to profile a minority. Human Rights Watch. https://www.hrw.org/report/2019/05/01/china-algorithms-repression
Information Commissioner's Office (ICO). (2019). ICO opinion: The use of live facial recognition technology by law enforcement in public places. UK Information Commissioner's Office.
Internet Freedom Foundation. (2020). Concerns about NCRB's Automated Facial Recognition System (AFRS): A review of the Request for Proposal. Internet Freedom Foundation Policy Brief.
Kant, I. (1998). Groundwork of the metaphysics of morals (M. Gregor, Trans.). Cambridge University Press. (Original work published 1785)
Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). https://arxiv.org/abs/1609.05807
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539
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
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607
Mill, J. S. (1863). Utilitarianism. Parker, Son, and Bourn.
Mozur, P. (2018, July 8). Inside China's dystopian dreams: A.I., shame and lots of cameras. The New York Times.
National Institute of Standards and Technology. (2019). Face recognition vendor test (FRVT) Part 3: Demographic effects (NISTIR 8280). U.S. Department of Commerce. https://doi.org/10.6028/NIST.IR.8280
Parsheera, S. (2019). A gendered perspective on data protection and privacy in India (NIPFP Working Paper No. 252). National Institute of Public Finance and Policy. https://doi.org/10.2139/ssrn.3452459
Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.
Penney, J. W. (2016). Chilling effects: Online surveillance and Wikipedia use. Berkeley Technology Law Journal, 31(1), 117–182.
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. RAND Corporation.
R (Bridges) v. Chief Constable of South Wales Police [2020] EWCA Civ 1058. Court of Appeal (England and Wales).
Rawls, J. (1971). A theory of justice. Harvard University Press.
Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. AI Now Institute.
Richardson, R., Schultz, J., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. New York University Law Review Online, 94, 192–233.
Russell, S. J., & Norvig, P. (2020). Artificial intelligence: A modern approach (4th ed.). Pearson Education.
State v. Loomis, 881 N.W.2d 749 (Wis. 2016).
Stevenson, M. T., & Doleac, J. L. (2019). Algorithmic risk assessment in the hands of humans (IZA Discussion Paper No. 12853). Institute of Labor Economics.
Tyler, T. R. (2006). Why people obey the law. Princeton University Press.
United Nations Human Rights. (2022). OHCHR assessment of human rights concerns in the Xinjiang Uyghur Autonomous Region, People's Republic of China. Office of the United Nations High Commissioner for Human Rights.
Universal Declaration of Human Rights. (1948). United Nations General Assembly Resolution 217A (III).
Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.
Završnik, A. (2021). Criminal justice, artificial intelligence systems, and human rights. ERA Forum, 20(4), 567–583. https://doi.org/10.1007/s12027-020-00602-0