AI systems in Corrections and their flaws

Adrian Luca
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
adrian.luca2904@stud.etti.upb.ro

Bogdan Ionescu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
bogdan.ionescu@upb.ro

***Abstract*—This paper provides a comparative analysis of artificial intelligence (AI) systems used in prison and correctional settings across the United States, the United Kingdom, EU and China. Focusing on three case studies: COMPAS (US), HART (UK), RISE AI (Finland), RISCANVI (Spain), and Smart Prisons (China), examines both successful and failed implementations of AI technologies such as risk assessment algorithms and surveillance systems. The analysis highlights key challenges including algorithmic bias, lack of transparency, and human rights concerns, while also identifying factors that contribute to responsible and ethical use. The findings underscore the importance of legal oversight, human-in-the-loop design, and the need for robust validation before deployment in correctional settings.**

***Index Terms*—AI Act, bias, data, transparency, corrections, predictive policing, human rights**

I. Introduction

Artificial Intelligence (AI) is increasingly used in prison and correctional systems to improve efficiency, assess inmate risk and support decision-making. Although these technologies offer potential benefits, such as reducing recidivism and improving rehabilitation planning—they also raise serious concerns about bias, transparency, and human rights.

This report compares AI systems deployed in correctional settings in the United States, the United Kingdom, China, and parts of the European Union. It focuses on tools such as COMPAS, HART, RISE AI, and AI-driven surveillance systems, evaluating their performance, ethical challenges, and regulatory environments. By analyzing successful and problematic implementations, the report identifies key factors that influence responsible AI use in corrections and highlights the importance of oversight, fairness, and legal safeguards.

II. Comparative Analysis of AI in Corrections

AI applications in corrections generally fall into two categories: risk assessment algorithms (used to predict recidivism or inmate behavior) and AI-driven surveillance/predictive tools (used to monitor or forecast security events). A comparison between regions reveals stark differences in adoption and oversight.

A. United States

Rapid Adoption, Persistent Bias, and Reactive Oversight

The U.S. has been a frontrunner in using AI risk assessment tools like COMPAS to predict recidivism and guide sentencing or parole. While widely adopted, COMPAS sparked controversy after a 2016 ProPublica report revealed it was nearly twice as likely to misclassify Black defendants as high-risk compared to white defendants. Further studies showed COMPAS(van Djick, 2022) was no more accurate than simple models or even untrained human judgment, raising doubts about its reliability. Despite these concerns, tools like PATTERN were introduced federally under the First Step Act to improve risk classification and support rehabilitation. PATTERN incorporates dynamic factors like program participation but has shown similar racial(Long, 2023) disparities—initially classifying only 7% of Black inmates as low risk versus 21% of white inmates. Even after adjustments, DOJ reviews found it still overestimated recidivism risk for minority groups. Alongside risk scoring, AI surveillance tools like Verus are now used in U.S. prisons to monitor inmate calls for threats or criminal activity. While credited with preventing incidents, these tools face criticism for misinterpretations, lack of transparency, and potential bias—especially in speech recognition for Black speakers.(Angwin, Larson, Mattu, & Kirchner, 2016) Overall, the U.S. has rapidly deployed AI in corrections with limited regulation, often addressing fairness concerns after implementation rather than proactively.(Dressel & Farid, 2018)

B. European Union

Rights-First, Transparent Approach

The EU has taken a cautious and transparency-oriented stance on AI in criminal justice. Unlike the U.S., where proprietary risk assessment tools like COMPAS dominate, most European countries rely on publicly developed, interpretable tools aimed at supporting—rather than replacing—human judgment(Fazel, 2022). Systems like OxRec(Johnson & Fazel, 2022) (validated in Sweden and the Netherlands) and RISCANVI (used in Catalonia) exemplify this approach, offering statistical or expert-based risk assessments with open methodologies and human oversight. In Finland, RISE AI assists staff in sentence planning, while remaining fully advisory. Pilots such as the UK's HART showed the potential for predictive risk models but also revealed risks of embedded socioeconomic bias. As a result, input features like postcode were removed, and the system was kept strictly as a decision-support tool.(UK, 2017) Failures like the Dutch SyRI fraud detection system (struck down for violating human rights) and

2nd Workshop on Artificial Intelligence for Multimedia, AI Multimedia Lab, CAMPUS Research Institute, POLITEHNICA Bucharest, Bucharest, Romania, 15 July, 2025

France's abandoned sentencing algorithm reflect Europe's low tolerance for opaque or biased AI. These cases have reinforced demand for transparency, fairness, and legal accountability. The EU Artificial Intelligence Act (2024)(Commission, 2024) codifies this ethos. It classifies AI used for risk prediction or sentencing as "high-risk," requiring: • Bias mitigation • Transparency and auditability • Human-in-the-loop control • Proof of added value over simpler models Coupled with GDPR, which limits profiling and sensitive data use(of Europe, 2024), these regulations ensure that AI in EU corrections remains accountable, transparent, and aligned with human rights. As a result, Europe has fewer full-scale prison AI systems, favoring small-scale, ethically guided pilots and regulatory oversight.

C. Other Regions

Control vs. Caution in AI Corrections

China presents one of the most expansive uses of AI in corrections, emphasizing state control and constant surveillance. Under the country's "Smart Prison" initiative, high-security facilities are equipped with AI-enabled cameras, facial recognition, and emotion detection systems that monitor inmates 24/7. These tools automatically flag behaviors such as pacing or agitation, and in some prisons, emotion recognition is used to detect potential self-harm or violence before it occurs. Companies like Taigusys claim to monitor over 60,000 cameras across 300 facilities, using AI to analyze biometric and facial cues. However, these systems raise serious ethical concerns, particularly around mental privacy, dignity, and ethnic profiling. Reports confirm that some Chinese systems tag individuals by ethnicity – features that would be illegal in the EU. Until recently, China lacked comprehensive data protection laws, and public transparency remains minimal, with national security interests prioritized over individual rights. By contrast, countries like Australia have adopted a cautious, health-oriented approach. Traditional tools like the LSI-R and Violence Risk Scale remain dominant and are administered by trained professionals. Recent initiatives focus on AI for well-being, such as a 2024 project by the University of Wollongong developing contactless radar+AI to detect distress and prevent suicide in at-risk inmates(Wollongong, 2024b). This system is guided by ethical oversight and multidisciplinary collaboration, designed for care rather than control.(Wollongong, 2024a) Australia lacks a dedicated national AI law for criminal justice, but general privacy and human rights statutes apply. Its cautious, small-scale implementation echoes European principles—favoring explainability, fairness, and human-in-the-loop design. Importantly, no major scandals involving prison AI have emerged in Australia, in part because adoption is limited, transparent, and grounded in pilot research.

III. Case studies

A. US Case study - COMPASS Risk Assessment

COMPAS, developed by Northpointe (now Equivant), is a widely used risk assessment tool in U.S. courts that predicts recidivism using over 130 factors. It gained national attention in 2016 when a ProPublica investigation revealed significant racial bias: Black defendants were more likely to be incorrectly labeled as high-risk, while white defendants who reoffended were often rated low-risk. Although the tool showed similar overall accuracy across racial groups (65The proprietary nature of COMPAS prevents public scrutiny of its algorithm, raising concerns about transparency and due process. In Loomis v. Wisconsin, the court allowed COMPAS but warned judges not to rely solely on its output. The backlash prompted some jurisdictions to limit its use, and California voters rejected a proposal to replace cash bail with a statewide algorithm due to similar bias concerns. COMPAS remains in use but is now seen as a cautionary example of how opaque and biased AI can undermine fairness in criminal justice.

B. Ethical and Transparent AI in European Corrections

Europe has taken a cautious and rights-based approach to AI in criminal justice, focusing on transparency, fairness, and human oversight. Several systems illustrate how algorithmic tools can support decision-making without replacing human judgment. In the UK, the Harm Assessment Risk Tool (HART) was piloted by Durham Police and Cambridge researchers to predict reoffending risk. Using historical police data, it classified suspects as low, moderate, or high risk to help divert suitable individuals into rehabilitation rather than prosecution. Early concerns arose when postcode data—a proxy for socioeconomic status—was found to strongly influence predictions. To address this, developers removed or downweighted postcode factors, reducing potential bias. Importantly, HART served as a decision-support tool, not an autonomous actor. The project ended as a limited pilot but is viewed as a model of responsible experimentation, showing how ethical red flags can be addressed proactively. In Finland, the RISE AI system aids prison staff in designing personalized sentence plans. It recommends interventions based on structured risk data but remains non-binding and fully interpretable. This aligns with Finland's broader emphasis on rehabilitation and dignity in corrections. OxRec, developed by Oxford researchers, is another transparent tool designed to predict violent reoffending. Validated in Sweden and the Netherlands, it is open-source, statistically robust, and built with clear ethical safeguards—offering a strong contrast to proprietary U.S. tools like COMPAS. Meanwhile, RISCANVI, used in Catalonia (Spain) since 2011, helps assess risk of in-prison violence or self-harm. While not yet AI-driven, it reflects a structured, expert-guided approach that's now under review for possible AI integration. Crucially, it operates with clear oversight and accountability. Together, these European cases show how AI can be responsibly used to support rehabilitation, risk management, and fairness. By prioritizing transparency, human control, and legal compliance—as reinforced by the EU Artificial Intelligence Act—Europe has avoided many of the pitfalls seen in less-regulated environments.

C. AI in Surveillance vs. Care – China and Australia

In China, AI has been integrated into prisons as a tool for total surveillance and behavioral control. "Smart Prison"

systems, like those at Yancheng Prison, use networks of CCTV cameras, facial recognition, and behavioral algorithms to monitor inmates around the clock. The AI flags "abnormal" activity—such as loitering, restlessness, or small gatherings—for immediate human review. Officials claim this makes escapes impossible and deters misconduct. Beyond tracking movement, many Chinese facilities now deploy emotion-recognition AI, developed by firms like Taigusys, which claims to detect agitation or despair through micro-expressions and biometric cues. The stated goal is early intervention—to prevent violence or suicide—but in practice, this form of psychological surveillance raises severe ethical concerns. Prisoners effectively lose all privacy, including emotional autonomy, and false positives could result in punitive or unnecessary actions. Moreover, systems that classify inmates by race or ethnicity (e.g., labeling Uyghurs) suggest embedded state-led profiling, drawing sharp criticism from human rights groups. China's legal safeguards remain minimal, with AI deployed without public oversight, governed largely by security imperatives rather than civil rights. By contrast, Australia takes a care-based and cautious approach. While it still uses traditional risk assessments (like LSI-R), recent developments show interest in AI for inmate health and suicide prevention. A 2024 pilot led by the University of Wollongong is testing contactless radar with AI to monitor inmates' vital signs and detect distress in real time. Framed explicitly as a health intervention, this system is being co-developed with psychologists, nurses, and ethicists, and is subject to privacy safeguards. Australia lacks a national AI law for justice, but its prison AI use remains limited, research-based, and transparent, aligned with European principles of human oversight, explainability, and care over control.

IV. Conclusions

The global use of AI in correctional systems reveals a stark divide in values, implementation, and oversight. In the United States and China, AI has often prioritized control and efficiency—frequently at the expense of fairness, transparency, or human rights. Tools used and smart prison surveillance systems demonstrate the risks of opaque algorithms and unchecked surveillance. In contrast, the European Union and countries like Australia have pursued a more cautious, ethically grounded path. Their emphasis on human oversight, transparency, and legal safeguards illustrates how AI can be aligned with rehabilitation and rights, also taking into the consideration the EU AI Act, coming into force in 2024. Ultimately, these case studies show that AI in corrections is not just a technical matter—it reflects deeper societal choices. Where accountability and human dignity are prioritized, AI can support justice. Where unchecked, it can reinforce bias and control. Future adoption must be guided not only by capability, but by values.

References

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

Commission, E. (2024). *Artificial intelligence act - european union legislation.* (https://artificialintelligenceact.eu)

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

Fazel, S. A. . F. T., S. (2022, Jun). Towards a more evidence-based risk assessment for people in the criminal justice system: the case of oxrec in the netherlands. *Eur J Crim Policy Res*. (https://doi.org/10.1007/s10610-022-09520-y)

Johnson, T., & Fazel, S. (2022). Risk prediction for violence in criminal offenders: A prospective validation study of oxrec in sweden and the netherlands. *European Journal on Criminal Policy and Research*. (10.1007/s10610-022-09520-y)

Long, B. (2023, Oct). A pattern of bias in pattern. *The Criminal Law Practitioner*. (Available at: https://www.crimlawpractitioner.org/post/a-pattern-of-bias-in-pattern)

of Europe, C. (2024). *Recommendation cm/rec(2024)1 of the committee of ministers to member states on the use of artificial intelligence in the criminal justice system.* (https://rm.coe.int/recommendation-ai-criminal-justice/1680abdc0d)

UK, W. (2017). *Police are using software to predict crime. is it a smart idea?* (https://www.wired.co.uk/article/police-using-algorithms-to-predict-crime)

van Djick, G. (2022, jun). Predicting recidivism risk meets ai act. *Springer Nature*, *28*(7), 407-423. doi: 10.1007/s10610-022-09516-8

Wollongong, U. (2024a). *Ai-powered radar system could help prevent inmate suicide in prisons.* (https://www.uow.edu.au/media/2024/ai-radar-prison-inmate-health.html)

Wollongong, U. (2024b, Feb). Uow researchers to develop ai technology for suicide prevention in prisons. *University of Wollongong, Australia*. (uow.edu.auuow.edu.au)