Assessing AI in Action

A Case Study on the Use of AI-Powered Classification Tools in State Corrections

September 2026

Task Force and Case Studies Context

The Council on Criminal Justice Task Force on Artificial Intelligence is a national, nonpartisan initiative to develop standards and evidence-based recommendations to guide the safe, ethical, and effective use of AI in the criminal justice system. In March 2026, the Task Force released a User Decision Framework to provide guidance for responsible AI procurement and deployment in the criminal justice field.

This case study is one of three that extends the work of that framework by demonstrating its guidance in action across different AI tools and hypothetical use cases, with a focus on substantive decision-making considerations.

These case studies follow the logic and structure of the framework and draw on information from practitioners, technology vendors, and publicly available research and product documentation. The assessments also reflect the insights and expertise of Task Force members, whose perspectives helped identify key questions, risks, opportunities, and implementation considerations. Together, these sources informed the case studies' deliberations and conclusions.

More Case Studies

Executive Summary

Corrections classification decisions shape the day-to-day conditions of confinement in state and federal prisons and can affect safety, access to services, release preparation, and institutional management. These decisions are highly consequential, influencing where

people are housed, their custody level, which programs they can access, how they move across the system, when they are dynamically reclassified, and other dimensions of prison life. Incorporating an AI tool into the process can help corrections stakeholders improve accuracy and consistency, reduce over-classification, identify safety concerns more reliably, support better programming decisions, and make classification decisions more auditable. But the use of AI technology in correctional settings also raises serious concerns about liberty.

This case study examines AI-powered inmate classification systems used to support custody-level determination, housing assignment, programming eligibility, and reclassification. The hypothetical use case considered here is a state Department of Corrections managing approximately 15,000 to 25,000 incarcerated people across multiple facilities. The agency is considering whether to use AI to modernize classification processes, with placement officers serving as the primary end-users responsible for housing and custody-level decisions. A distinct consideration for users in this scenario is how AI-powered classification compares to the limitations of existing processes.

Applying the five decision phases outlined in the User Decision Framework to this case study, the key takeaways are:

Phase 1: Foundation and Readiness

The goal is to reduce administrative burden and improve accuracy at scale more effectively than non-AI alternatives would allow. Organizational readiness largely depends on the department's ability to navigate variables like disparate-impact testing and human review.

Phase 2: Classification

The context and use case do not trigger limitations from the prohibited-use screening as long as disparate-impact testing and substantive contest preconditions can be met. The use case presents both substantial risk and substantial opportunity, which triggers "**Careful Implementation**."

Phase 3: Procurement

Resources and processes for thoughtful implementation, such as integration with existing information and records management systems, should be prepared in advance. Enhanced contractual protections should be used to address the substantial risk of the use case and tools, including additional terms related to validation.

Phase 4: Implementation

The careful implementation label resulting from the use case's substantial risk and opportunity classifications requires enhanced implementation, including incorporation of relevant equal protection and Title VI rights protections.

Phase 5: Ongoing Management and Reassessment

Scheduled reassessment should take place annually, in keeping with the Task Force recommendation for systems and use cases that are classified as substantial risk. Reassessment beyond the annual schedule could be triggered by developments such as product updates and expanded use cases.

Some questions raised by this case study can only be resolved by considering the specific product and jurisdiction in question. This is especially true for questions concerning transparency, validation, disparate impact, data portability, and procedural fairness.

Glossary

Automation Bias: The tendency to over-rely on AI-generated outputs without sufficient critical evaluation.

CJIS (Criminal Justice Information Services): The largest division of the FBI, which serves as a centralized hub for law enforcement and national security data.

Due Process: A concept in the U.S. Constitution, largely rooted in the Fifth and Fourteenth Amendments, that guarantees that no one may be deprived of life, liberty, or property without proper legal procedures.

Equal Protection: A clause in the Fourteenth Amendment of the U.S. Constitution that requires state governments to apply their laws and protections to people equally, unless there is a legitimate and legal reason to do otherwise.

ML: Machine Learning. A specific subset of AI that identifies patterns from algorithms and data and uses those patterns to make predictions rather than relying on explicit scenario coding.

Static Actuarial Tool: A tool that uses pre-programmed algorithms and guidelines to

automate specific tasks, requiring manual reprogramming and new data to make adjustments.

Title VI: A provision of the Civil Rights Act of 1964 that prohibits discrimination based on race, color, or national origin in programs or services that receive federal funding.

Workflow Walkthrough

Phase 1: Foundation and Readiness

Define the Problem

The Framework at a Glance: *What is the agency? What problem does it face? Why is this product being considered? What alternatives to AI exist? What would success look like?*

The Agency and the Problem

The agency considered in this case study is a state Department of Corrections responsible for 15,000 to 25,000 incarcerated people across multiple facilities. The primary end-user is the placement officer, who makes or supports decisions about housing, custody level, programming eligibility, and reclassification. These decisions shape the day-to-day conditions of confinement and can affect safety, access to services, and release preparation for incarcerated people, as well as institutional management.

The Opportunity

The department is considering AI-powered classification systems because classification is operationally consequential and difficult to manage consistently at scale. Classification decisions influence where people are housed, what custody level they receive, which programs they can access, how they move across the system, and when they are dynamically reclassified. If a tool improved accuracy and consistency, it could reduce over-classification, identify safety concerns more reliably, support better programming decisions,

and make classification decisions more auditable.

Defining Success

The status-quo comparison frame is critical in this context. Existing classification systems already use seemingly arbitrary or under-explained scoring rules, such as one-year versus five-year look-back windows with no published rationale. Many systems still rely on paper-based forms and produce little routinely disaggregated outcome tracking. The right comparison is not AI versus ideal human judgment, but rather AI versus the agency's existing classification process.

Realistic non-AI alternatives should be evaluated. These include purely clinical judgment, paper-based objective classification, simple point-based systems, and rule-based decision trees. Success would mean more accurate placements, including fewer false positives that lead to over-classification and fewer false negatives that lead to safety failures. It would also mean reduced disparities, auditable decisions, and stronger legal defensibility.

Organizational Readiness

The Framework at a Glance: *What capacities does this product compel beyond generic technological readiness?*

Legal Infrastructure

AI-powered classification systems require capacities that go beyond ordinary software procurement. The agency should be able to manage proprietary-algorithm contracts in a setting where liberty interests are directly affected. Trade-secret defenses can block individual review and constrain agency oversight. Legal counsel should be prepared to address that possibility before the agency enters into a contract, not after litigation or appeals arise.

Outcome Tracking

The agency also would need the capacity to run disaggregated outcome tracking. Many state departments of corrections do not routinely produce classification outcome data by race, gender, age, custody level, override pattern, appeal outcome, or misconduct outcome. Without those data, the agency cannot determine whether the AI tool improves accuracy, reduces disparities, or simply reproduces existing inequities in a more technical form.

Community Engagement

Because of the liberty risk, the agency should also be prepared to support a Community Advisory Committee with representation from directly-impacted people. Many departments have not built this infrastructure for classification specifically. The committee should include formerly incarcerated people, families, civil rights organizations, and others with relevant lived experience and expertise.

Data Infrastructure

Data-portability readiness is another distinct consideration. Legacy contracts can cement vendor lock-in, making it difficult for an agency to change systems, audit historical decisions, or retain access to its own data in usable formats. The agency's capacity to negotiate against lock-in and preserve data ownership, portability, and audit rights is a binding constraint.

Phase 2: Classification

Assessing Specific Products

The product category includes both established actuarial instruments and newer AI or machine learning-enabled platforms. These systems differ in purpose, architecture, validation history, data requirements, and deployment profile. Some tools are proprietary actuarial instruments that combine official records, interviews, and self-report questionnaires to generate risk scores. Others are management systems that may include configurable classification workflows, housing matching, or business-intelligence features. Emerging platforms may incorporate computer vision, pattern analysis, or predictive modules that require separate scrutiny.

An agency should consider key product-specific questions, including what the tool does, what architecture it uses, what data it requires, how it has been validated, and how it would be deployed in the agency's facilities. It should also distinguish between classification functions and broader platform capabilities, especially where vendors market surveillance, predictive-violence, contraband-detection, or digital-twin modules alongside classification or housing tools. Across AI tools, several considerations should be treated as central procurement and implementation issues rather than minor technical details:

- **Classification use cases**: Which classification use cases are explicitly supported, discouraged, or restricted?
- **Accuracy and performance**: What are the calibration metrics, predictive-validity metrics, and disaggregated accuracy by race, gender, and age? Do these metrics include sensitivity, specificity, positive predictive value, and negative predictive value? How are errors identified and communicated to users? What recourse exists for incarcerated people who wish to challenge their classification?
- **Data governance**: How are incarcerated people's data handled, retained, protected, and shared? Are customer data used to retrain or improve vendor models? Who owns the data if the agency switches vendors?
- **Evaluations**: What external benchmarks and internal evaluations exist, including disparate-impact and fairness analyses across demographic groups?

Prohibited Use Screening

The Framework at a Glance: *The prohibited-use screening asks whether the AI system poses an unacceptable risk to fundamental rights and should be prohibited in the criminal justice context. If the answer is yes to any screening question, the system is presumptively prohibited unless mitigation is documented.*

This screen is deliberately categorical, so each answer records only whether the use crosses a prohibited-use red line, not how risky it is overall. "No" means the use does not trigger that category; "no at screening" means it does not trigger the category but raises a concern that is not waived and instead carries forward into the graded risk-and-opportunity analysis later in this phase.

In this case, the screen for AI-powered classification systems in state corrections settings

does not produce a direct prohibited-use finding if disparate-impact testing and substantive contest can be negotiated and executed. The agency should proceed to complexity, sector context, and risk-opportunity classification if the preconditions are met. If not, the system is prohibited under Appendix B until mitigated.

Screening Question	Answer	Reasoning
Q1: Does the system make autonomous decisions about liberty (e.g., detention, sentencing) without the possibility of substantial human review?	NO	Most current implementations have a placement-officer in the loop, but vendors increasingly market "automation." Implementation must preserve documented human review and override.
Q2: Does the system eliminate or impair a person's right to contest a pending decision, or appeal a decision that's already been made affecting their rights?	YES with mitigation required	*Loomis*[1] allows use with warnings, but proprietary algorithms can block meaningful contest. Mitigation is possible only if individuals can access substantive information about how the score was used.
Q3: Does the system circumvent or undermine established legal or constitutional protections (e.g., due process, equal protection)?	NO if *Loomis*-compliant	YES if the agency cannot articulate what process is due above the constitutional floor
Q4: Does the system perform individualized tracking and surveillance of or otherwise have a chilling effect on a group engaging in lawful, constitutionally protected activities (e.g., First Amendment-related activities)?	NO for classification per se	Flag if the broader vendor platform includes surveillance modules (e.g., the NUCLEUS tool's computer-vision and predictive-violence components warrant separate scrutiny)

Screening Question	Answer	Reasoning
Q5: Does the system target or select people based on protected characteristics and create unjustified discriminatory effects because of race, gender, religion, national origin, disability, or another legally prohibited ground?	YES with mitigation required	ProPublica/COMPAS[2] and DOJ/PATTERN[3] have documented disparate impact. Agencies should proceed only if the vendor agrees to Title VI–grounded disparate-impact testing as a contract term and the agency commits to disaggregated outcome monitoring.
Q6: Does the system systematically undermine human dignity or value (e.g., by publicly shaming or humiliating people or stripping them of all agency)?	NO	These tools do not introduce novel categories or restrictions to the corrections environment.

System Complexity and Interpretability

The Framework at a Glance: *How does this system score across Appendix C's four dimensions of decision transparency, predictability, validation difficulty, and adaptability?*

AI-powered classification tools should be treated as higher-complexity systems under the framework's four dimensions of decision transparency, predictability, validation difficulty, and adaptability. The tools' higher complexity triggers enhanced oversight under Appendix C, requiring stronger validation, case-specific rationales, monitoring, and auditability before deployment.

Dimension	Assessment	Reasoning
Decision transparency	Varies	Static actuarial tools (e.g., LSI-R) more interpretable; COMPAS tool is proprietary; NUCLEUS tool is opaque. Agencies should demand case-specific decision rationales as a contract requirement.

Dimension	Assessment	Reasoning
Predictability	Partially Predictable	The same inputs typically produce the same outputs for rule-based components; ML components may drift with model updates.
Validation difficulty	Harder to Validate	Inputs include unstructured interview content; outputs interact with dynamic reclassification; ground truth ("would this person have been violent under different supervision?") partially unobservable
Adaptability	Varies	Static instruments are fixed; ML platforms may continue to learn.

Sector Context

The Framework at a Glance: *What is the current-practice baseline, and how does AI change risk and opportunity at the margin?*

The relevant legal framework is both federal and state law, not only federal constitutional law.

Fourteenth Amendment

Equal protection under the Fourteenth Amendment requires a showing of discriminatory intent in disparate-treatment claims, which is extremely difficult to prove for facially neutral algorithms. For that reason, federal constitutional doctrine alone is an inadequate accountability frame for classification systems that may produce disparate outcomes.

Title VI

Title VI matters because statutory disparate impact is effects-based and does not require proof of intent. In principle, algorithm-driven discrimination is actionable under this frame. In practice, however, disparate-impact claims remain difficult because outcome testing is not routine, disaggregated outcome data are often unavailable to the user agency, and proprietary vendor models make it difficult to identify less discriminatory alternatives.

State Law

State law may go further than the federal baseline in some jurisdictions. Emerging algorithmic decision-making rules in New York City, Illinois, and elsewhere show that state and local governments may impose additional transparency, testing, or accountability requirements.[4] Idaho's 2019 transparency law for pretrial risk assessment is a model the working group should consider.[5]

Due Process

Due process is also central, but the constitutional minimum is not the right benchmark. *Loomis*-style warnings and prohibitions on sole-basis use may satisfy a low procedural floor, but classification decisions should aim higher. Operationally, going above the constitutional floor could mean access to the basis of the score, disaggregated accuracy data, an opportunity to challenge inputs, an appeal path with documented review of patterns, and periodic public reporting of outcomes by demographic group. *Wilkinson v. Austin*[6] and *Sandin v. Conner*[7] set a low procedural bar, but more robust protections can be prioritized.

Risk, Opportunity, and Classification

The Framework at a Glance: *With sector context and use case in mind, assess the proper classification level and implicated next steps.*

All things considered, the classification for this use case is substantial risk and substantial opportunity.

Substantial Risk

Classification decisions have direct liberty impacts across custody level, housing, and programming eligibility. Documented racial disparities in existing tools create a serious fairness concern. The "impossibility of fairness" theory[8] also means that design choices have ethical content and should require deliberate stakeholder input, supporting the need for a

Community Advisory Committee. These risks exist with non-AI classification processes as well. But incorporating an AI-assisted classification system produces the added possibility of long-term over-reliance on the tools, cementing poor or problematic baselines.

Substantial Opportunity

Evidence suggests that structured, validated classification tools may improve consistency and support more targeted placement, supervision, and programming decisions, which can lower recidivism and improve institutional behavior and other important correctional outcomes.[9] AI-assisted classification tools may further support these outcomes through enhanced reliability and accuracy. The proper comparison frame is AI versus existing arbitrariness or bias, not AI versus idealistically principled human judgment.

Under the user decision framework, substantial risk combined with substantial opportunity triggers the Careful Implementation path, which means the agency should proceed with all Level 1 and Level 2 (enhanced) implementation requirements. These include formal legal analysis, a Community Advisory Committee with directly affected representation, independent validation, real-time monitoring, decision logs, and override audit trails.

The classification could shift toward Generally Avoid if vendors refuse to disclose disaggregated accuracy data on trade-secret grounds, if independent validation shows meaningful disparity levels without a credible mitigation path, if the agency cannot negotiate data-portability and audit-rights terms, or if the agency cannot stand up a Community Advisory Committee with genuine authority and representation of directly affected populations.

Phase 3: Procurement

Budget and Resources

The Framework at a Glance: *What costs exist across procurement, deployment, and ongoing management?*

Procurement should, at minimum, account for the full cost of acquisition, integration, training, ongoing monitoring, technical fixes, and community engagement. Although an agency's ability to proactively prepare for procurement depends on vendor transparency and readily available information, key considerations for a correctional system's use of classification AI tools include:

- The agency should expect costs to extend beyond the initial software purchase since the systems may be deployed across large correctional populations and multiple facilities.
- Integration costs are likely to be substantial. The system may need to connect with jail management systems, electronic health records, programming records, geographic information systems for housing matching, and vendor-specific data-mapping processes. These integrations matter because classification decisions often depend on information drawn from multiple systems, and errors or gaps in data mapping can affect placement recommendations.
- Training and change management should focus on placement officers and supervisors. Placement officers need training on score interpretation, override documentation, and automation-bias resistance. Supervisors need training on reviewing override patterns, especially by demographic group, and on detecting patterns that suggest overreliance on the tool or inconsistent human review.
- Ongoing monitoring should include real-time disparate-impact tracking, periodic validation, and annual reassessment. Community engagement is recommended since the system carries substantial risk. The agency should budget for a Community Advisory Committee, including compensation for committee members.

Contract Negotiation and Essential Terms

The Framework at a Glance: *Which Appendix F categories apply, what does each look like for this specific product, and what vendor pushback should the agency expect?*

The applicable contract terms should be drawn from Appendix F's Level 1 and Level 2 (enhanced) requirements.

Appendix F Category	Product-specific term that matters most	Notes
L1 System Requirements	Define classification decisions explicitly as supported vs. excluded; constrain auto-accept configurations	Vendors may prefer broad scope
L1 Performance Standards	Disaggregated accuracy data by race, age, gender as a contract requirement	Vendors may resist disclosure; trade secret arguments
L1 Evidence and Validation	Right to independent audit and validation as a renewal condition	Frame as validation, not trade-secret audit
L1 Transparency and Explainability	Algorithm-transparency commensurate with IP protection; case-specific decision rationales required for higher-complexity systems	Vendor pushback likely on algorithm disclosure
L1 Data Governance and Privacy	Agency retains data ownership and portability—counter to long-term lock-in; vendor must certify NOT training on agency data without written consent	Vendors may prefer proprietary formats and long lock-in
L1 Bias Testing and Fairness	Disparate-impact testing obligation grounded in Title VI; disclosure of any known bias issues and mitigation strategies	Vendors may resist standardized testing
L1 Human Oversight and Override	Documented justification for both following AND overriding the system's recommendation; periodic supervisor review by demographic group	Vendor pushback unlikely
L1 Training and Support	Vendor training on score interpretation, automation-bias resistance, override documentation	Vendor pushback unlikely
L1 System Changes and Updates	Mandatory notification for model changes; right to refuse updates that change risk profile; re-validation required after updates	Vendors may prefer unilateral update authority
L1 Liability and Indemnification	Vendor accepts liability for system errors; indemnifies agency for civil rights violations caused by the system	Vendor pushback likely

Appendix F Category	Product-specific term that matters most	Notes
L1 Termination	Termination for convenience, performance failure, or rights concerns—with data-portability provisions intact	Vendor pushback likely to mirror standard termination power pushback; lock-in concerns
L2 Enhanced Validation	Independent validation tied to renewal; pilot success criteria pre-set	Vendor pushback likely
L2 Enhanced Rights Protections	Formal legal analysis covering Equal Protection, Title VI, *Loomis*-line, *Wilkinson/Sandin*; documented procedure for individuals to challenge classification	Vendor pushback likely to mirror standard burden arguments
L2 Community Engagement Support	Vendor participation in community meetings and explanations	Vendor pushback likely to mirror standard burden arguments
L2 Enhanced Auditability	Algorithm transparency with appropriate IP protections; audit rights for independent experts	Vendor pushback likely

Phase 4: Implementation

The Framework at a Glance: *How will the system function in your environment? How will you ensure it performs as intended?*

Level 1 Requirements For All Systems

These baseline recommendations apply to every AI system deployed in criminal justice settings.

Human-Centered Design and Training

Implementation should begin with how the product presents its outputs. The classification system should not only present a score; it should also present recommendations alongside the relevant information informing each recommendation so that placement officers can understand the key factors driving the score. This is especially important for higher-complexity systems, where case-specific decision rationales are needed to support substantive review.

Automation-bias risk is significant. Placement officers may defer to an algorithmic score even when their independent judgment differs. To counter this risk, override tracking should require documented justification both when an officer follows the system's recommendation and when an officer overrides it. Regular supervisor review should examine decision patterns and justification quality, including by demographic group.

Training should cover score interpretation, override-documentation expectations, and automation-bias resistance. Placement officers should understand what the system is designed to do, what it is not designed to do, how errors may arise, and how to document their own judgment. Supervisors should be trained to identify patterns of overreliance, inconsistent overrides, or demographic disparities in classification outcomes.

Transparency, Accountability, and Privacy

The agency should publish information about what the system does, how it is used, the demographic distribution of classifications, override rates, and the appeals process. Public notification is important because classification decisions shape conditions of confinement, and the people affected by the system need to know how it operates and how to contest its use.

The agency should designate a named Department of Corrections official accountable for classification outcomes and disparate-impact monitoring. Accountability should not rest solely with individual placement officers if system design, vendor updates, scoring logic, or data-quality issues affect recommendations. A named official should be responsible for ensuring that the system's use remains legally defensible, auditable, and consistent with agency policy.

Incarcerated people should have meaningful access to the basis of their score, the ability to challenge inputs, and an appeal path with documented review. As a concrete implementation approach responsive to the COMPAS proprietary-algorithm finding, the contract should

require that, on individual contests, the agency can produce the specific factors and weights that drove the score for that person. This is the case-specific decision-rationale recommendation under Appendix C.

Privacy and data security should include CJIS compliance, minimum-necessary data collection, retention with automated deletion, and no cross-system repurposing without a new assessment. Classification systems often draw on sensitive correctional, health, programming, and behavioral data. The agency should ensure that data are used only for the assessed purpose and not repurposed for surveillance, discipline, or other uses without additional review.

Pilot Program

The pilot should define scope, duration, success criteria, comparison methodology, and a formal decision-making process. Because the system is classified as substantial risk, Appendix G recommends a minimum six-month pilot. In certain contexts, such as needing adequate time to measure institutional misconduct and reclassification effects, a longer 12-month pilot is recommended.

Key considerations should include whether to pilot at one facility or system-wide, how to compare the tool with the current classification approach, and whether the sample size is large enough to detect demographic disparities. Success criteria should include classification accuracy compared with baseline, misconduct rates by classification level, override rates by demographic group, appeal outcomes, and demographic parity in custody-level placements.

The comparison methodology should ideally involve parallel running of baseline and AI classification processes on the same intakes or matched samples across facilities. The framework recommends a formal go/no-go decision after the pilot. If the pilot reveals unacceptable disparities, poor accuracy, inadequate appeal procedures, weak documentation, or inability to audit recommendations, the agency should not proceed to broader deployment without mitigation.

Level 2 Enhanced Requirements For Substantial Risk Systems

Because the system is classified as substantial risk, the agency must apply additional rights protections, community engagement, evidence and validation requirements, auditing, and

technical safeguards. Important additional safeguards include:

- Rights protection should begin with formal legal analyses covering Equal Protection, Title VI, *Loomis*-line doctrine, *Wilkinson*, *Sandin*, and applicable state algorithmic-decisioning laws. The agency should also conduct a human rights impact assessment with a mitigation plan. Individual contest procedures should provide meaningful access to the score's basis, the ability to challenge inputs, and an appeal with documented review of patterns.
- Community engagement should include a Community Advisory Committee with directly affected representation, including formerly incarcerated people, family members, and civil rights organizations. The committee should have genuine authority to recommend changes. Committee members should be compensated, and the agency should provide multiple feedback channels with particular effort to include people directly affected by the criminal justice system. The agency should also publish regular reports on classification outcomes by demographic group.
- Evidence and validation requirements should include independent validation by experts not affiliated with the vendor or agency. The agency should also conduct a formal alternatives analysis comparing AI-powered classification with paper-based objective classification, simpler point-based systems, and rule-based decision trees. That analysis should explain why AI is preferable despite its risks.
- Auditing and enhanced oversight should include extended training and continuing-education requirements, whistleblower protections for staff who flag bias or pressure to accept recommendations, real-time monitoring of disaggregated override rates by demographic group, score-distribution drift, and data-quality flags in input fields, and decision logs. Agencies should make decision logs and complete audit trails accessible for legal review and appeals.
- Additional technical safeguards should include adaptive system monitoring with anomaly alerts, the ability to freeze deployment if concerns arise, and regular revalidation as models are migrated.

Phase 5: Ongoing Management and Reasses

The Framework at a Glance: *How should ongoing monitoring and periodic reassessment be leveraged to ensure AI tools continue to function as intended?*

Ongoing management should define the scheduled reassessment cadence, triggered reassessment events, and product-specific indicators that would require review. Because the AI system in this case is substantial risk, scheduled reassessment should occur annually. The agency should monitor indicators that reveal accuracy, fairness, usability, and rights impact, including but not limited to:

- Override rates by demographic group
- Score-distribution drift
- Data-quality flags in input fields
- Override-justification quality
- Misconduct rates by classification level
- Reclassification appeals and outcomes
- Demographic disparities in custody-level placements
- Outcomes for incarcerated people classified at the margin

Triggered reassessment should occur after underlying model migration if key product or context changes occur, such as:

- The vendor updates the model, retrains it, or adds features
- The agency expands use to new populations or use cases, such as tightening the reclassification cadence
- Disparities widen, override rates spike, or misconduct rates diverge from prediction
- Title VI complaints, litigation, or Community Advisory Committee escalation occurs
- New state algorithmic-decisioning laws or caselaw change the legal environment

Acknowledgments

This case study is from the Council on Criminal Justice Task Force on Artificial Intelligence, a national, nonpartisan initiative developing standards and evidence-based recommendations to guide the safe, ethical, and effective use of AI in the criminal justice system. The case study is a product of its members, who graciously shared their time and expertise.

Jesse Rothman produced this report, with support from Cameryn Farrow, Andrew Page, Jonathan Wroblewski, and others from the Council on Criminal Justice team.

James Anderson and RAND serve as research partners to the Task Force.

Support for the Task Force on Artificial Intelligence comes from the Georgia Power Foundation, Heising-Simons Foundation, The Just Trust, Microsoft, and The Tow Foundation, as well as the John D. and Catherine T. MacArthur Foundation and other CCJ general operating contributors.

Suggested Citation

Council on Criminal Justice. (2026). *Assessing AI in action: A case study on the use of AI-powered classification tools in state corrections*. https://counciloncj.org/assessing-ai-in-action-a-case-study-on-the-use-of-ai-powered-classification-tools-in-state-corrections/

Endnotes

[1] *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016). https://law.justia.com/cases/wisconsin/supreme-court/2016/2015ap000157-cr.html; Joh, E. E. (2019). *Algorithms and sentencing: What does due process require?* Brookings Institution. https://www.brookings.edu/articles/algorithms-and-sentencing-what-does-due-process-require/

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

[3] National Institute of Justice. (2024). *2023 review and revalidation of the First Step Act risk assessment tool.* U.S. Department of Justice, Office of Justice Programs. https://nij.ojp.gov/library/publications/2023-review-and-revalidation-first-step-act-risk-assessment-tool

[4] Anderson, H., Reem, N., & Susas, J. (2025). Automated Decision Making Emerges As An

Early Target of State AI Regulation. White and Case. https://www.whitecase.com/insight-alert/automated-decision-making-emerges-early-target-state-ai-regulation

[5] Idaho Legislature. (2019). House Bill No. 118, *65th Legislature, 1st Regular Session.* https://legislature.idaho.gov/wp-content/uploads/sessioninfo/2019/legislation/H0118E2.pdf

[6] *Wilkinson v. Austin,* 545 U.S. 209 (2005). https://www.law.cornell.edu/supct/pdf/04-495P.ZO

[7] *Sandin v. Conner,* 515 U.S. 472 (1995). https://tile.loc.gov/storage-services/service/ll/usrep/usrep515/usrep515472/usrep515472.pdf

[8] Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). *Fairness in criminal justice risk assessments: The state of the art* (Working Paper No. 2017-1.0). University of Pennsylvania Department of Criminology. https://crim.sas.upenn.edu/sites/default/files/2017-1.0-Berk_FairnessCrimJustRisk.pdf

[9] Duwe, G. (2017). *The use and impact of correctional programming for inmates on pre- and post-release outcomes* (NCJ 250476). National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. https://www.ojp.gov/pdffiles1/nij/250476.pdf