

HUMAN, ALGORITHM, JUSTICE.

THE LIMITS OF AI IN
RECIDIVISM FORECASTING



The Promise of Big Data

Uncovering hidden behavioral patterns through vast datasets.

Predicting recidivism risk to allocate correctional resources.

Standardizing decisions to reduce subjective human prejudice.

Fundamental Human Rights

Human Dignity and Equality
(Constitutional & ECHR Arts. 6 & 8).

Right to a Fair Trial and
Procedural Justice.

Informational Autonomy and the
right to human intervention.

Can technologies that automate, evaluate, and forecast remain subordinate to the idea of human dignity?

The European Architecture of Algorithmic Protection

Algorithmic Transparency

The imperative for a meaningful explanation of the system's role in a decision.

CEPEJ 2018 Ethical Charter

Requires respect for fundamental rights, non-discrimination, transparency, impartiality, and systems operating strictly under user control.

GDPR & Directive (EU) 2016/680

Regulates data in criminal justice. Article 11 explicitly bans purely automated decisions carrying adverse legal effects without guaranteed human intervention.

Constitutional Law & ECHR

Human Dignity, Right to a Fair Trial

The EU AI Act Draws a Hard Boundary

Prohibited Practices (Art. 5.1.d)

The Ban: From February 2025, AI systems assessing criminal risk solely on personality traits or profiling are strictly prohibited.

Exception: Systems supporting human assessment based on objective, verifiable facts tied to criminal activity.

High-Risk Systems (Annex III)

The Designation: Recidivism scoring and systems assisting judicial interpretation of facts are classified as High-Risk.

Strict Requirements: Mandates rigorous data quality control, technical documentation, human oversight, transparency, and fundamental rights impact assessments by public authorities.

The Psychological Trap of Automation Bias



The Deskilling Effect

The illusion of algorithmic objectivity leads practitioners to doubt their own professional experience and intuition.

Ignoring Protective Factors

Algorithms excel at weighing historical risk factors (crimes committed) but severely undercalculate dynamic protective factors like new support systems or therapy breakthroughs.

The Fiction of Oversight

Formal human in the loop policies fail if the practitioner becomes psychologically anchored to the machine's output, transforming judges into mere algorithmic rubber stamps.

The Automation Spectrum and the Human in the Loop

Full Automation (Replacement)

AI → Final
Decision

Removes the human entirely.
— **Incompatible** with ECHR and EU AI Act (Art. 14) requirement for human oversight in liberty deprivation.

Suggestive Semi-Automation (Anchoring)

AI Proposes → Human Approves

High vulnerability to **cognitive anchoring**; the human merely rubber-stamps the algorithmic output.

Confrontational Semi-Automation (Evaluation)

Human Drafts → AI Evaluates → Human Finalizes

The algorithm acts as a **secondary auditor**, checking for missed variables or inconsistencies only after the human has formed an independent judicial opinion.

The Methodological Minefield of Actuarial Justice



Overfitting

The model memorizes training data noise instead of general patterns, failing when exposed to new real-world cases.



Class Imbalance

The model favors the majority class, producing high overall accuracy but poor minority recognition.



Concept Drift

The relationship between inputs and recidivism changes over time due to shifts in penal policy or crime structure.



Data Bias

Algorithms amplify historical prejudices (e.g., disproportionate policing) under a veneer of objectivity.



Target Variable Proxy

We cannot measure true re-offending. Systems actually predict the risk of re-arrest, reflecting police activity rather than human behavior.



Metric Incompatibility

Perfect fairness across disparate groups is statistically impossible. Proceed to the mathematical proof.

The Mathematical Impossibility of Perfect Fairness

Equal Calibration (The Developer's View)

A risk score of 7 means the exact same probability of recidivism, regardless of the defendant's demographic group.



Equal Error Rates (The Critic's View)

The system should produce the exact same rate of False Positives and False Negatives across all demographic groups.

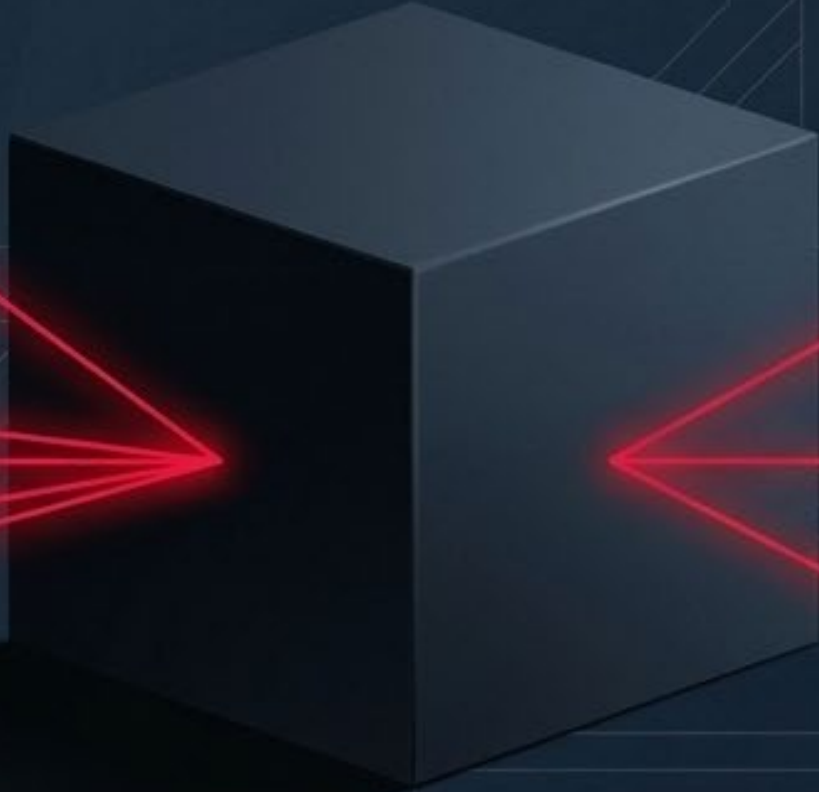
The Proof (Chouldechova / Kleinberg, 2017)

If base recidivism rates differ between two groups, it is mathematically impossible to achieve both Equal Calibration and Equal Error Rates simultaneously. Choosing the definition of a fair algorithm is a normative, legislative decision, not a software engineering task.

Global Implementations: A Diagnostic Comparison

System & Origin	Core Philosophy	Data Transparency	Mechanism
COMPAS (USA)	Actuarial Prediction & Punishment	Black Box (Commercial trade secret)	137-question proprietary profile; heavily relies on socio-economic proxy variables.
U-Case / KODA (Norway)	Case Management & Normalcy	Glass Box (Public institution integration)	Centralized repository linking prison, probation, and medical data to support daily rehabilitation plans.
PSORR-PL (Poland)	Risk-Needs-Responsivity (4th Gen)	Transparent & Multi-source	Integrates psychological, penitentiary, and probation data to monitor dynamic risk changes over time.

The Danger of the Black Box: The COMPAS Warning



Unequal Error Distribution

ProPublica analysis revealed minority defendants who did not re-offend were nearly twice as likely to be falsely classified as high-risk compared to white defendants.

Limited Efficacy

Research showed COMPAS accuracy (~65%) is comparable to random internet users or a simple 2-variable model of age and prior convictions.

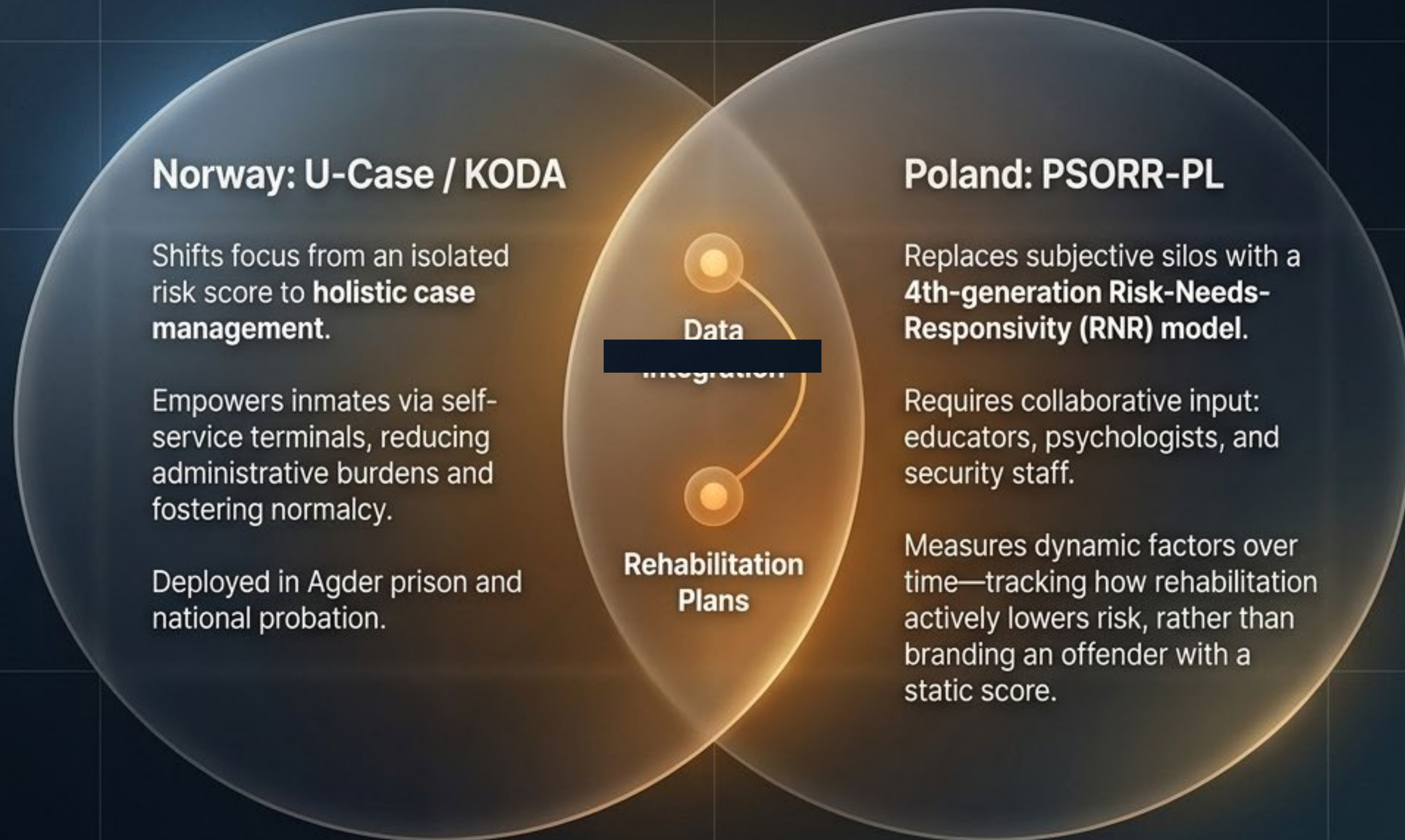
Proxy Variables

Relies on employment, housing, and peer groups—effectively punishing socio-economic status.

The Legal Compromise: *State v. Loomis* (Wis. 2016)

The court allowed COMPAS for sentencing if accompanied by a **warning label**. However, warning labels **do not fix the fundamental information asymmetry** between the defendant and a proprietary algorithm.

The Transparent Approach: Rehabilitation over Prediction



The Ideal Ecosystem: Continuous Data Integration



Eliminate Silos

Stop re-collecting the same data at every stage.

Dynamic Updating

Risk assessments must evolve in real-time as the offender completes therapy or gains employment.

Strict Access Controls

Data must be minimized and strictly tethered to the specific role of the user (judge vs. therapist vs. probation officer).

A Roadmap for Ethical Algorithmic Justice

Defendant Appeal Mechanisms

Guarantee technical transparency. The defense must have the right and the data necessary to meaningfully challenge algorithmic outputs.

Defendant Treat Lookers

Train judges, prosecutors, and probation officers not just on software mechanics, but on the cognitive dangers of automation bias.

Practitioner Training

Train judges, prosecutors, and probation officers not just on software mechanics, but on the correct deployment.

Mandatory Bias Audits

Enforce independent, national-population validation and demographic error-rate reporting before any deployment.

Statutory Frameworks

Align domestic criminal procedure with the EU AI Act and Directive 2016/680. Legislate the "assist, do not replace" mandate.

**Algorithms predict how groups behave.
Humans decide what is just for the individual.**

Data and AI are powerful tools for the desk, but they must never become a substitute for the gavel. We must build systems that rigorously assist, but fundamentally do not replace, the human conscience in the administration of justice.