Sorting the Past, Not Predicting the Future: Probability, Uncertainty, and Algorithmic Risk in Criminal Justice

Mags Lesiak, Soumya Banerjee
University of Cambridge, United Kingdom
mkl35@cam.ac.uk,sb2333@cam.ac.uk

Abstract

Machine-learning risk assessment tools in criminal justice (such as COMPAS, HART, and the recent "Super Learner" for domestic homicide) are widely presented as predictive technologies. We argue that these systems do not generate individual-level forecasts in any epistemically coherent sense; rather, they perform epistemic simulation. The structural preconditions for probabilistic prediction—stable outcome spaces, stationarity, and repeatability, are violated in socially contingent open-world domains like human behaviour. As a result, numerical risk scores resemble probabilities but lack the warrant of probability theory, collapsing group-level classification into misleading claims about individual futures. This epistemic inflation suppresses genuine uncertainty, masking ignorance as knowledge and licensing high-stakes decisions under false confidence. We formalise this critique through the lens of Kolmogorov probability, Knightian uncertainty, and the reference class problem, and propose an epistemic honesty criterion: models should output intervals, sets, or abstentions when probability assumptions fail, rather than collapsing uncertainty into point scores. By shifting from simulation to epistemic intelligence, machine learning can better respect the limits of inference, communicate ambiguity transparently, and support decision-making without overstepping its epistemic boundaries.

Keywords: Epistemic Uncertainty; Epistemic Simulation; Criminal Justice; Probability Theory; Responsible AI

1 Introduction: The Appearance of Prediction

Across jurisdictions, machine-learning tools are being deployed to assess 'criminal risk': who might reoffend, who poses a threat, who should be detained or diverted. Presented as predictive technologies, these tools have come to influence decisions in courts, policing, and domestic violence response systems. But their promise of foresight masks a deeper epistemic confusion. Tools like COMPAS in the United States (Angwin et al. , 2016), HART in the United Kingdom (Oswald et al. , 2018), and newer classifiers such as the "super learner" for domestic homicide (Verrey et al. , 2023), do not—and cannot—generate meaningful predictions about individual futures. Their forecasts are simulations: artifacts of statistical form that resemble probability but do not satisfy its foundational conditions (Gigerenzer , 2002; Knight , 1921; Kolmogorov , 1933; Hajek , 2007).

The contemporary language of risk is not a neutral descriptor of uncertainty but a historically contingent construct shaped by capitalist rationality and biopolitical governance. Early articulations (emerging from maritime insurance and actuarial science) defined risk within closed, repeatable systems governed by calculable exposure and statistical regularity (Hacking , 1990). Twentieth-century

1st Workshop on Epistemic Intelligence in Machine Learning (EIML 2025) .

formalisation further aligned risk with variance, expected loss, and probabilistic rationality (Knight , 1921), establishing its utility as a tool for managing uncertainty through strategic abstraction.

Knight's foundational distinction between measurable risk and true uncertainty shows the limits of inference in singular or unstable contexts. Yet this boundary has been increasingly eroded. In criminal justice, risk has been redefined not through the axioms of probability theory (Kolmogorov , 1933) but through institutional demand. Tools claiming to measure an individual's "likelihood" of future harm transpose statistical semantics into settings that lack repeatability, stability, or defined outcome spaces (Breiman , 2001a; Gigerenzer , 2002). The result is a category error: risk is invoked where only uncertainty exists, producing simulation rather than prediction (Harcourt , 2007; Berk , 2021).

This article challenges the growing assumption that machine-learning risk assessments can produce personal predictions. It argues that the core methodology of these tools—classification on historical data evaluated through population-level metrics—is not consistent with interpreting individual risk scores as probabilities. The use of terms like "forecast," "likelihood," or "risk of reoffending" implies causal or probabilistic knowledge that these systems do not possess (Berk , 2021). This misrepresentation is not merely semantic; it constitutes what philosophers of probability identify as a category error—the conflation of reference-class frequencies with the probability of a singular future event (Hajek , 2007; Gillies , 2000). While group-level metrics (e.g., calibration or AUC) can be meaningful in aggregate, their translation into subject-level risk estimates is epistemically incoherent in contexts marked by open-world uncertainty and ill-defined outcome spaces (Knight , 1921; Gigerenzer , 2002). Prediction is not a universal capacity—it is a conditional claim on the future that requires structure, stability, and causal warrant. Where those conditions do not hold, the performance of prediction is not caution—it is error. If criminal justice systems are to retain public trust, they must respect the difference between sorting the past and knowing the future.

2 What These Tools Actually Do

Machine-learning tools used in criminal justice settings are not predictive instruments. They operate as historical classifiers trained on past data, outputting calibrated scores or categorical bins, and evaluated through population-level performance metrics. Understanding what these tools actually do requires a step-by-step breakdown of the standard machine learning pipeline and a precise interpretation of the metrics that emerge from it.

The foundational step in risk modelling is to train a classifier on labelled historical data. In a typical supervised learning setup, the model is exposed to a dataset comprising input features (e.g., prior convictions, age, postcode) and outcome labels (e.g., reoffended: yes/no) from past cases. The objective is to learn a function that approximates the mapping between these inputs and outputs. In random forest classifiers, for example—used in tools like HART—this involves aggregating decisions from multiple decision trees trained on bootstrapped samples of the data, each split on different features to minimise overfitting (Breiman , 2001a). What is "learned" is not a causal theory of offending, but a statistical association between feature patterns and label distributions.

The labels used in training are not inherent properties of individuals but post hoc administrative outcomes: whether a person was rearrested, charged, or convicted. These outcomes reflect not only individual behaviour but also structural factors such as policing patterns, reporting rates, and institutional discretion (Lum and Isaac , 2016; Berk , 2012). The model thus encodes regularities from historical institutional behavior—not a neutral record of crime, and certainly not individual propensity.

2.1 Model Outputs and Performance Metrics: The Illusion of Predictive Power

Once trained, machine-learning classifiers output raw scores that rank examples by their estimated likelihood of class membership. These scores, however, are not inherently calibrated probabilities. As Zadrozny and Elkan (2001) show, standard classifiers like decision trees or naive Bayes often produce biased or high-variance estimates—particularly in sparse feature regions—leading to overconfident outputs near 0 or 1 that do not reflect true class probabilities. Calibration techniques such as Platt scaling, isotonic regression, or histogram binning can transform raw scores into better-aligned posterior estimates (Platt , 1999), but calibration alone does not confer epistemic validity.

Even well-calibrated models can fail to generalise if the training data is biased or the target context unstable.

Performance is typically evaluated using population-level metrics such as AUC-ROC, precision, recall, and F1 score. These do not assess whether a prediction about a specific individual is correct, but rather how well the model sorts known cases in aggregate. AUC, for instance, measures relative ranking ability—not causal accuracy or personal probability (Hanley and McNeil , 1982). Precision and recall quantify the proportion of correct positive predictions and the coverage of actual positives, respectively, but can be misleading under class imbalance or distributional drift (Powers , 2011). These metrics are retrospective descriptors, not prospective forecasts. As such, they offer no epistemic warrant for treating a model's output as a credible estimate of individual likelihood—especially in open, socially contingent domains like criminal justice.

3 Impossibility of Individual Prediction under Open-World Conditions

In probability theory, a prediction requires the existence of a well-defined probability space $(\Omega, \mathcal{F}, P)$ where Ω is the set of all possible outcomes, $\mathcal{F}$ is a collection of measurable events, and P is a probability measure satisfying Kolmogorov's axioms. A model produces a valid prediction of an event if the following assumptions hold (Kolmogorov , 1933):

- **Stable outcome space**: the set of possible outcomes Ω must be fixed and well-defined.
- **Stationarity**: the conditional distribution $P(Y \mid X)$ must be invariant across time.
- **Repeatability**: the event Y_i must be subject to repeated trials (so that probability can be empirically verified).

In socially contingent domains such as criminal justice, conditions (1)–(3) fail. Therefore, $P(Y_i = 1 \mid X_i)$ is not mathematically well-defined as an epistemic probability. Any numerical "prediction" is a simulation of probability, not a valid measure.

4 The Reference Class Problem

Assigning an individual-level probability of crime runs directly into the reference class problem (Gillies , 2000; Hajek , 2007). Any given person simultaneously belongs to many groups—for example: age 20–30 (reoffense rate 20%), prior convictions (40%), and residents of a high-policing district (60%). Which of these group rates should define that person's "true" probability of reoffending? Unless there is a principled way to privilege one reference class over all others—a condition almost never met in social domains—the assignment of a single probability to an individual becomes arbitrary. From a machine learning perspective, this is not a case of ordinary statistical noise (aleatoric uncertainty), but of deeper epistemic ignorance: the absence of a unique, stable distribution that can be used to ground personal forecasts (Gillies , 2000; Hajek , 2007). For criminology, the implication is that so-called "risk scores" do not reflect intrinsic likelihoods of individual behavior, but rather choices about which population groupings we use—and those choices are always contestable.

5 Individual Unpredictability: The Problem of a Single Trial

The Galton board demonstrates the limits of probability. When hundreds of balls are dropped through the pins, they form a neat bell-shaped curve at the bottom. The overall distribution is highly predictable. We can say with confidence that most balls will cluster around the centre and fewer will land at the edges. The Galton board does not predict where any one ball will land—it only tells us the most likely distribution after many trials. For an individual ball, its final position depends on random perturbations at each peg. The problem is that, mathematically, there is no way to determine the exact path of a single ball in advance (Galton , 1889; Feller , 1971).

To illustrate this, consider a single ball's trajectory as a stochastic process. This is a random walk: each step depends on a sequence of independent binary choices. The final position of the ball is unknown until the process unfolds. While we can compute its expected value and variance, these provide no useful information about an individual case—only the broader statistical tendencies. The

same holds for human behaviour. At the population level, we can observe patterns—for example, poverty, trauma, or peer networks may correlate with crime. But for an individual person, their choices are contingent, context-dependent, and cannot be assigned a stable "probability."

Tool	Claimed Epistemic Role	What It Actually Produces	Why It Could Not Work (Structural Limitation)
COMPAS (US courts)	An individual's probability of reoffending, guiding bail, parole, sentencing.	A proprietary score mapping past administrative data into bins that correlate with group-level recidivism.	Probability requires a stable outcome space and repeatability. Individual futures are non-repeatable, socially contingent, and label-dependent. The premise of a well-defined "probability of reoffending" is incoherent.
HART (Durham Constabulary, UK)	Forecasts risk category (low/medium/high) to inform proactive policing.	A random forest trained on police-recorded events; ranks cases by historical resemblance.	Policing practices shape the data: outcomes are endogenous, not independent. The reference class problem makes any individual "risk" assignment underdetermined; prediction collapses into classification of the past.
Super Learner for Domestic Homicide (UK, Verrey et al. 2023)	"Excellent" prediction of domestic homicide, with strong AUC/recall.	An ensemble classifier outputting calibrated-like scores on finite samples.	Even with high AUC, outputs are retrospective sorting, not prospective forecasts. In an open-world domain like homicide, distributions drift and outcome spaces are unstable: no coherent probability measure exists for singular futures.

6 Why This Is Not Individual Prediction

Prediction, in an epistemological sense, refers to the act of generating knowledge claims about future states of the world that are both falsifiable and informationally non-trivial (Popper , 2005; Shmueli , 2010). It presupposes an intelligible mapping between known antecedents and not-yet-realised consequences, mediated by a model whose structure encodes assumptions about regularity, causality, or statistical association (Frigg and Hartmann , 2020). What can be predicted, therefore, is not simply a function of model sophistication, but of the ontological properties of the target phenomenon. In domains governed by physical law or ergodic statistical processes—such as planetary motion, radioactive decay, or aggregate mortality—predictive power is attainable because the systems in question are sufficiently closed, stationary, and insensitive to perturbation (Cartwright , 1983). Techniques such as regression, time-series forecasting, and Bayesian updating can meaningfully generate predictions in these settings, because the underlying distributions are stable and the sample spaces are well-defined (Kolmogorov , 1933).

Thus, predictive power is not a general feature of models, but a relational property that emerges only when the structure of the model aligns with the causal architecture of the system it targets (Cartwright , 1983; Frigg and Hartmann , 2020). A tool designed to forecast insurance claims may have predictive power in that domain but fail entirely when applied to crime or recidivism, not because the algorithm is flawed, but because the latter phenomena are non-ergodic, structurally contingent, and embedded in feedback-rich environments (Perdomo et al. , 2020; Lum and Isaac , 2016). In this light, the use of predictive language in criminal justice tools such as COMPAS or Homicide Super Learner does not indicate the presence of predictive power, but the semantic appropriation of scientific authority, whereby statistical classifiers are misrepresented as generators of epistemically valid future knowledge (Amoore , 2011; Harcourt , 2007). The confusion arises not merely from technical misuse, but from a deeper philosophical misunderstanding: to predict is not merely to extrapolate—it is to know under what conditions extrapolation is meaningful.

7 Algorithmic Contribution

We propose that classifiers deployed in criminal justice (and other open-world domains) should satisfy an *epistemic honesty* criterion: they should only output probability-like numbers when the assumptions of probability theory—such as stable outcome spaces and repeatable conditions—are satisfied. When these assumptions fail, as they often do in social and legal contexts, systems should instead communicate structured forms of uncertainty: for example, giving ranges of plausible values, sets of possibilities, or explicitly abstaining from prediction. This is not merely aspirational: existing ML techniques already allow this. Conformal prediction can provide guaranteed ranges that reflect distribution-free uncertainty (Vovk et al. , 2005); credal classification offers interval-valued probabilities rooted in imprecise probability theory (De Cooman and Zaffalon , 2004); and selective prediction methods allow a model to decline making a forecast when its confidence is unreliable (Chow , 1970; Geifman and El-Yaniv , 2017). Likewise, evaluation must go beyond conventional metrics such as AUC or precision, which measure accuracy in aggregate but say nothing about epistemic reliability at the individual level. Instead, we should assess how often uncertainty ranges are valid (coverage), how efficiently they balance informativeness with reliability (Angelopoulos and Bates , 2021), and how abstentions reduce harmful errors in practice (Cortes , 2016). Embedding these practices would allow ML systems in criminal justice to acknowledge their epistemic limits, improving both technical robustness and institutional trust.

For example, Verrey *et al.* (2023) report a "super learner" with AUC = 0.71, recall = 77.6%, and precision = 18.6%. Epistemically, this means that if 100 people are flagged "high risk," only 18 commit homicide while 82 do not. Such scores simulate probability but inflate confidence. By contrast, interval-based methods (e.g. conformal prediction) could mark 66 cases with informative ranges and abstain on 34, explicitly signalling uncertainty. Reporting should thus emphasise false positive burden, interval coverage, and abstentions (not only AUC) so models communicate what they do not know.

References

Amoore, L. (2011). Data Derivatives. Theory, Culture & Society.

Amoore, L. (2013). The politics of possibility: Risk and security beyond probability. Duke University Press.

Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against Blacks. ProPublica.

Berk, R. (2008). Forecasting methods in crime and justice. Annual Review of Law and Social Science.

Berk, R., Sherman, L., Barnes, G., Kurtz, E., & Ahlman, L. (2009). Forecasting murder within a population of probationers and parolees: A high stakes application of statistical learning. *Journal of the Royal Statistical Society: Series A*, 172(1), 191–211.

Berk, R. (2012). *Criminal justice forecasts of risk: A machine learning approach*. Springer.

Berk, R., Heidari, H., & Jabbari, S. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1), 3–44.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*.

Cartwright, N. (1983). How the Laws of Physics Lie.

Chow, C. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1), 41–46.

Cortes, C., DeSalvo, G., & Mohri, M. (2016). Learning with rejection. In *International Conference on Algorithmic Learning Theory* (pp. 67-82). Springer.

De Cooman, G., & Zaffalon, M. (2004). Updating beliefs with incomplete observations. *Artificial Intelligence*, 159(1-2), 75–125.

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

Ericson, R. V., & Haggerty, K. D. (1997). *Policing the risk society*. Oxford University Press. DOI:10.1093/oso/9780198265535.001.0001.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.

Feller, W. (1971). *An introduction to probability theory and its applications* (Vol. 963). New York: Wiley.

Frigg, R., & Hartmann, S. (2020). Models in science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020 Edition). Retrieved from https://plato.stanford.edu/archives/fall2020/entries/models-science/

Galton, F. (1889). *Natural inheritance* (Vol. 42). Macmillan.

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 30.

Gigerenzer, G. (2002). *Reckoning with risk: Learning to live with uncertainty*. Penguin.

Gillies, D. (2000). *Philosophical Theories of Probability*. Routledge.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning*.

Hacking, I. (1990). *The taming of chance* (No. 17). Cambridge University Press.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29–36.

Harcourt, B. E. (2007/2019). *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. University of Chicago Press.

Harcourt, B. E. (2015). Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter*, 27(4), 237–243.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Knight, F. H. (1921). *Risk, uncertainty and profit*. Houghton Mifflin.

Kolmogorov, A. N. (1933/1956). *Foundations of the Theory of Probability*. (2nd English ed., 2018 reprint). Dover.

Lum, K., & Isaac, W. (2016). To predict and serve?. *Significance*, 13(5), 14–19.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. ICML 2005.

Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: Lessons from the Durham HART model and 'experimental' proportionality. *Information, Communication & Society*, 21(8), 1284–1301.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., & Hardt, M. (2020). Performative prediction. In *International Conference on Machine Learning* (pp. 7599-7609). PMLR.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*.

Popper, K. (2005). *The logic of scientific discovery*. Routledge. (reprint)

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59-68).

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310.

Stevenson, M. T. (2018). Assessing risk assessment in action. *Minnesota Law Review*, 103, 303.

Verrey, J., Ariel, B., Harinam, V., & Dillon, L. (2023). Using machine learning to forecast domestic homicide via police data and super learning. *Scientific Reports*, 13(1), 22932.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Boston, MA: Springer US.

Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *ICML* (Vol. 1).

Hájek, A. (2007). The reference class problem is your problem too. *Synthese*, 156(3), 563–585.