https://doi.org/10.37420/j.aiel.2026.001

The Triple Path of AI-Empowered Criminal Investigation Curriculum Construction

Shuchen Tang

School of Criminal Investigation, People's Public Security University of China

Corresponding author: tsc@ppsucedu.cn

Abstract

Artificial intelligence (AI) is changing criminal investigation from a predominantly experience-led practice into a data-intensive activity in which forecasts, similarity scores, rankings, and generated content increasingly shape investigative attention. Yet police and criminal-investigation curricula often treat AI as a collection of tools rather than as a sociotechnical system whose outputs require statistical interpretation, legal justification, and accountable human judgment. This conceptual paper develops a Triple-Path framework for curriculum construction. Path I, Technological Reconstruction, builds investigative AI literacy through probabilistic reasoning, model evaluation, data provenance, and explainability. Path II, Ethical and Legal Governance, integrates fairness, privacy, due process, contestability, evidentiary reliability, and algorithmic impact assessment. Path III, Practice-Oriented Human-AI Collaboration, uses simulations and laboratory work to calibrate trust, counter automation bias, and preserve meaningful human control. The framework converts broad calls for responsible AI into teachable competencies, learning activities, and assessment evidence. It also emphasizes that investigators should neither reject AI categorically nor defer to it uncritically; they should be able to explain what a system does, identify when it may fail, document how it influenced a decision, and justify the final investigative action under law and professional ethics. The proposed framework offers a practical foundation for modernizing criminal-investigation education while protecting fairness, accountability, and public trust.

Keywords

AI governance; criminal investigation curriculum; algorithmic bias; predictive policing; explainable AI; digital evidence; human-in-the-loop.

1 Introduction

The datafication of social life has altered both crime and its investigation. Mobile devices, platform records, network logs, location histories, biometric systems, and large-scale video collections can now become investigative leads or evidence. At the same time, law-enforcement organizations use statistical and machine-learning systems to prioritize locations, identify patterns, rank possible matches, and manage information. These developments do not simply add new instruments to an established profession. They change how investigators form hypotheses, allocate attention, evaluate uncertainty, and justify decisions (Brayne, 2017; Ferguson, 2017).

This transformation creates an educational problem. Conventional training is usually organized around substantive law, procedure, interviewing, physical evidence, and case experience. Those elements remain indispensable, but they do not by themselves prepare an investigator to ask how a model was trained, what population its error rate describes, whether a score is calibrated, which data are missing, how a vendor's system can be challenged, or when an automated recommendation should be overridden. General AI-literacy research similarly shows that competent use requires more than operational skill: learners need conceptual understand-

ing, critical evaluation, communication, and ethical awareness (Long & Magerko, 2020; Ng et al., 2021).

The resulting competence gap produces at least three risks. First, an epistemic risk arises when probabilistic outputs are treated as facts or when correlation is confused with causation. Second, a legitimacy risk arises when opaque, biased, or unlawfully deployed systems affect communities without adequate explanation and contestability. Third, an organizational risk arises when nominal human oversight becomes a ritual in which investigators routinely endorse machine outputs without independent assessment. Contemporary governance instruments increasingly require risk management, documentation, human oversight, and AI literacy, making these educational deficits operational as well as theoretical (European Parliament & Council, 2024; Tabassi, 2023).

This paper therefore proposes a Triple-Path framework for criminal-investigation curriculum construction: (1) Technological Reconstruction, moving from tool operation to logic and evidence comprehension; (2) Ethical and Legal Governance, moving from formal compliance to critical, rights-aware evaluation; and (3) Practice-Oriented Human-AI Collaboration, moving from dependence or rejection toward calibrated, documented teamwork. The paper is a conceptual synthesis rather than a systematic review. It draws on research in policing, algorithmic fairness, explainable AI, human factors, digital evidence, and professional education to translate responsible-AI principles into curriculum design.

2 Literature Review: The Algorithmic Turn in Policing

2.1 Predictive analytics and datafied policing

Data-driven policing ranges from descriptive crime mapping to place-based forecasting, network analysis, risk assessment, facial comparison, and decision-support systems. Some evaluations have reported operational benefits for narrowly specified forecasting tasks, including randomized field trials of place-based crime prediction (Mohler et al., 2015). Such findings, however, do not establish that every predictive system is accurate, fair, cost-effective, or suitable for every decision. Performance depends on the target, data-generating process, deployment context, and the intervention that follows a prediction.

The central difficulty is that police data do not provide a neutral census of crime. Arrests, stops, field interviews, and discovered incidents are shaped by prior enforcement choices, reporting behavior, institutional priorities, and unequal exposure to surveillance. When those records are reused as training labels, a model may learn patterns of policing as though they were patterns of underlying offending. Empirical and theoretical work has shown how this process can produce self-reinforcing feedback loops in which increased patrol generates more recorded events, which then justify further patrol (Ensign et al., 2018; Lum & Isaac, 2016). Richardson et al. (2019) describe this as a 'dirty data' problem: civil-rights violations and distorted administrative records can become embedded in later predictions.

Accordingly, the educational objective should not be to teach that predictive analytics are either objective solutions or inherently illegitimate. Students need to separate technical validity from policy validity. A model may forecast a recorded event accurately and still be unsuitable because the target is normatively inappropriate, the intervention is disproportionate, the data are systematically incomplete, or the deployment creates unacceptable feedback effects. This distinction is foundational for responsible investigative practice.

2.2 The interdisciplinary education gap

Technical courses commonly explain algorithms without examining criminal procedure, evidentiary disclosure, or institutional power. Legal courses commonly address privacy and due process without teaching students how to interpret confusion matrices, calibration, dataset shift, or model documentation. Police training may demonstrate a product interface while leaving the product's assumptions and limitations implicit. These silos encourage a black-box mentality in which the investigator knows what button to press but cannot evaluate what the result means.

An effective curriculum must therefore integrate three forms of rationality. Instrumental rationality asks whether a system performs a defined task. Legal-ethical rationality asks whether the task, data, and intervention are lawful, proportionate, fair, and contestable. Practical rationality asks how a trained professional should combine a system's output with case context, contradictory evidence, and accountable judgment. The Triple-Path framework is designed to align these forms rather than teaching them as separate electives.

3 The Triple-Path Curriculum Framework

The framework treats AI competence as a professional capability that combines knowledge, judgment, and demonstrable practice. Each path has distinct learning outcomes, but the paths are intentionally interdependent. Technical literacy without governance can make harmful systems more efficient. Governance without technical literacy can become symbolic compliance. Practical skill without either can normalize unexamined reliance. Figure 1 presents the relationship among the three paths.

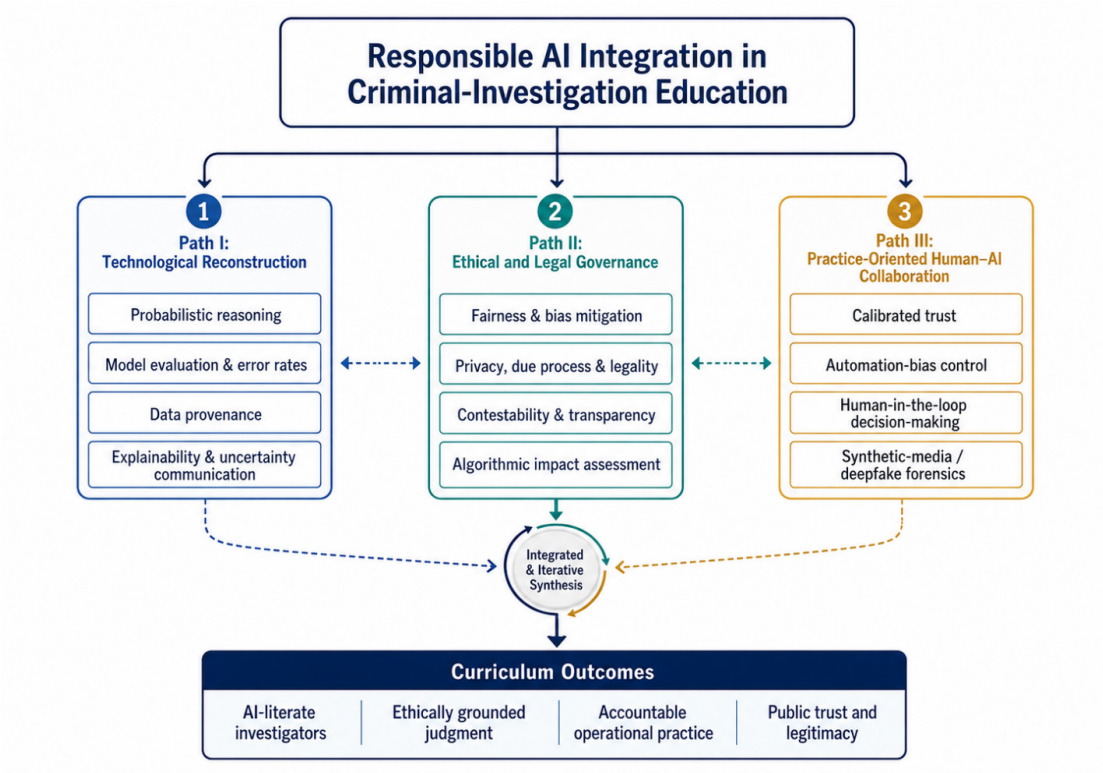

Figure 1.TheTriple-Pathframeworkfor responsible AI integration in criminal-investigation education.

Table 1. Curriculum domains, learning activities, and assessment evidence

Path	Core competencies	Illustrative learning activities	Assessment evidence
I. Technological Reconstruction	Probabilistic reasoning; model metrics; data provenance; explainability; uncertainty communication	Audit a hotspot model; compare subgroup error rates; inspect model cards; reconstruct a data pipeline	Technical audit memo; metric interpretation test; oral explanation of limitations
II. Ethical and Legal Governance	Fairness trade-offs; privacy; proportionality; due process; evidentiary reliability; impact assessment	COMPAS debate; biometric-use impact assessment; disclosure and challenge exercise; governance hearing	Reasoned legal-ethical opinion; algorithmic impact assessment; stakeholder briefing
III. Practice-Oriented Collaboration	Trust calibration; automation-bias control; independent verification; documentation; synthetic-media forensics	Adversarial simulations; blind-first analysis; red-team/blue-team exercise; deepfake authentication workflow	Scenario performance; decision log; after-action review; portfolio of validated findings

3.1 Path I: Technological Reconstruction

The first path replaces product-specific training with durable investigative AI literacy. Students do not need to become machine-learning engineers, but they must understand enough of the data and model lifecycle to identify assumptions, request documentation, interpret outputs, and recognize conditions under which performance claims do not transfer to a case.

3.1.1 Probabilistic reasoning and evidential interpretation

AI systems usually produce scores, rankings, classifications, or probability estimates rather than legal conclusions. Investigative reasoning has always involved uncertainty; the pedagogical challenge is to prevent numerical precision from being mistaken for certainty. A 95% similarity score, for example, is not automatically a 95% probability that a person is the source of an image. Its meaning depends on the system's definition, the comparison population, threshold, image quality, and relevant base rate.

Students should learn conditional probability, sensitivity, specificity, false-positive and false-negative rates, positive predictive value, calibration, confidence intervals, and threshold trade-offs. Bayesian reasoning can be introduced as a disciplined method for updating a hypothesis in light of evidence:

$$P(H \mid E) = [P(E \mid H) \times P(H)] / P(E)$$

The purpose is not to calculate a numerical probability of guilt. Rather, students should learn to distinguish the likelihood of evidence under competing hypotheses from the ultimate legal judgment, and to identify how biased priors, selective data, or an inappropriate reference population can distort an AI-assisted inference. Exercises should explicitly address the prosecutor's fallacy, base-rate neglect, and the difference between investigative leads and admissible proof.

3.1.2 Model performance, demographic differentials, and data provenance

Performance must be disaggregated by task and population. NIST testing has demonstrated that face-recognition error differentials can vary substantially across algorithms, demographic groups, image sources, and operational settings; therefore, a single claim that facial recognition is either 'accurate' or 'biased' is analytically inadequate (Grother et al., 2019). Independent audits have also found intersectional disparities in commercial facial-analysis systems, especially for darker-skinned women (Buolamwini & Gebru, 2018). These studies should be taught together: one illustrates large-scale vendor variation, while the other demonstrates why evaluation must include groups that aggregate accuracy can conceal.

Students should be able to inspect a confusion matrix, compare subgroup performance, ask whether a test set resembles operational data, and evaluate the consequences of a threshold. They should also trace data provenance: who collected the data, for what purpose, under which legal authority, with what labeling process, and with what known gaps. Auditing itself can create privacy and representation risks, so the curriculum should also examine ethical limits on collecting sensitive benchmark data (Raji et al., 2020).

3.1.3 Explainability, documentation, and uncertainty communication

Explainability is not a single technical property. Investigators may need a global description of a system, a local explanation of a particular output, documentation of training and validation, or an account of how a result affected a human decision. LIME and SHAP can help illustrate local feature contributions (Lundberg & Lee, 2017; Ribeiro et al., 2016), but post-hoc explanations do not prove that a model is fair, causal, robust, or legally justified. They can also be unstable or misunderstood.

The learning outcome should therefore be 'explanation for a purpose.' Students should identify the audience—investigator, supervisor, court, affected person, auditor, or public—and provide information appropriate to that audience. They should be able to state the system's intended use, data source, validation setting, known failure modes, threshold, uncertainty, and role in the final decision. Model cards, data sheets, audit reports, and decision logs should be treated as operational evidence, not administrative paperwork.

3.1.4 Learning design for technological reconstruction

The technical path should use low-code laboratories and case-based analysis. A student might compare two hotspot models trained on reported incidents and police-discovered incidents; evaluate how patrol allocation changes the next training cycle; or test a facial-comparison system on images degraded by lighting, compression, and camera angle. Assessment should reward correct interpretation and justified skepticism rather than programming sophistication. A technically competent graduate should be able to explain what a system can support, what it cannot establish, and what additional evidence is required.

3.2 Path II: Ethical and Legal Governance

The second path embeds value rationality into technical practice. In high-stakes policing, fairness, privacy, legality, and accountability cannot be added after deployment. They influence the choice of target, training data, performance metric, threshold, intervention, disclosure, retention period, and appeal mechanism. Governance education should therefore be organized around decisions throughout the system lifecycle.

3.2.1 Fairness as a sociotechnical and normative problem

Bias can enter through problem formulation, sampling, measurement, labels, model design, deployment, and human response. It is also possible for a technically well-calibrated system to produce an unjust outcome because the underlying policy is discriminatory or the burden of error is unequally distributed. Students should learn to distinguish data bias, measurement bias, model error, disparate impact, and institutional inequity. They should also recognize the limits of abstract metrics when removed from the social process that produced the data (Selbst et al., 2019).

Formal fairness criteria frequently conflict when base rates differ across groups. Kleinberg et al. (2017) and Chouldechova (2017) show that calibration and equalized error rates generally cannot all be satisfied simultaneously except under restrictive conditions. This is not a reason to abandon fairness; it is a reason to make the trade-off explicit. Curriculum exercises should require students to identify who bears false positives and false negatives, connect metric choices to legal and policy objectives, and defend a threshold rather than invoking 'fair AI' as a slogan.

3.2.2 The COMPAS case and the limits of metric-only debate

The controversy surrounding the COMPAS recidivism instrument is an effective teaching case. ProPublica reported racial differences in false-positive and false-negative classifications (Angwin et al., 2016), while subsequent scholarship explained how competing fairness definitions can yield different judgments about the same score (Chouldechova, 2017; Corbett-Davies et al., 2017). Students should reproduce a simplified analysis, then move beyond the arithmetic to ask whether the target is appropriate, whether the score should influence detention or sentence severity, what disclosure is required, and how a person can contest the result.

3.2.3 Explanation, contestability, and procedural justice

Claims that the GDPR creates a broad, unqualified right to explanation require greater precision. The existence and scope of such a right under the GDPR have been extensively debated; Articles 15 and 22 and Recital 71 provide relevant safeguards, but their application is limited and contested (Wachter et al., 2017). The EU AI Act now includes a more specific right to an explanation for certain individual decisions based on high-risk AI systems, while also imposing documentation, human-oversight, and risk-management obligations in defined contexts (European Parliament & Council, 2024).

For investigators, the broader educational principle is contestability. A person affected by an AI-assisted decision should have a meaningful route to identify the system's role, challenge inaccurate data, test the reliability of an output, and obtain review by an accountable human. Source-code disclosure is not always

necessary or sufficient. Depending on the dispute, meaningful challenge may require validation studies, error rates, data provenance, threshold settings, version history, or access to an expert. Scholarship on accountable algorithms and proprietary criminal-justice technologies illustrates the tension between trade secrecy and due process (Kroll et al., 2017; Wexler, 2018).

3.2.4 Privacy, legality, and evidentiary reliability

AI does not suspend ordinary rules governing search, seizure, authorization, disclosure, reliability, and chain of custody. Instead, it complicates them. Aggregating location, association, and biometric data can reveal more than any isolated record. In United States doctrine, Carpenter v. United States (2018) and scholarship on the mosaic theory illustrate why prolonged or aggregated digital surveillance may create privacy interests that are not captured by examining each data point separately (Kerr, 2012). Comparable analysis in other jurisdictions should begin with local constitutional, statutory, and data-protection law.

The exclusionary consequences of unlawful data collection vary by jurisdiction and cannot be inferred merely from a vendor's violation of platform terms. Students should instead work through a disciplined sequence: identify the source and legal authority for collection; determine whether processing and sharing were authorized; preserve provenance and integrity; assess whether an algorithm materially transformed the evidence; disclose limitations; and evaluate admissibility under the governing rules. Where proprietary systems generate testimonial-like outputs, courts may also need evidence about validation and operation rather than accepting a machine result as self-authenticating (Roth, 2017).

3.2.5 Algorithmic impact assessment

An algorithmic impact assessment (AIA) provides a practical bridge between ethics and administration. Before deployment, students should examine the proposed purpose, affected populations, data sources, performance evidence, alternative methods, human-oversight design, foreseeable harms, procurement terms, security, retention, complaint mechanisms, and conditions for suspension or withdrawal. Research on AIAs emphasizes that impacts are co-constructed through organizational practice; completing a checklist is not equivalent to accountability (Metcalf et al., 2021).

A capstone assignment could require teams to assess a proposed facial-identification, gunshot-detection, or investigative triage system. Teams would present separate technical, legal, community-impact, and operational analyses before producing a reasoned recommendation: deploy, modify, pilot under constraints, or reject. The assessment should specify measurable safeguards and a post-deployment review schedule.

3.3 Path III: Practice-Oriented Human-AI Collaboration

The third path turns knowledge into behavior under uncertainty and time pressure. Human oversight is meaningful only when a person has the competence, information, authority, time, and organizational support to disagree with a system. Merely placing an officer 'in the loop' can legitimize automation without reducing risk.

3.3.1 Automation bias and calibrated reliance

Automation bias includes commission errors, in which a user follows an incorrect automated recommendation, and omission errors, in which a user fails to act because the system did not issue an alert. Foundational studies show that users can over-rely on decision aids even when contradictory information is available (Skitka et al., 1999). Later human-factors research explains how attention, workload, system reliability, and task design influence complacency and bias (Parasuraman & Manzey, 2010).

Training should not simply warn students that algorithms make mistakes. It should expose them to plausible, consequential errors and require independent reasoning. Useful techniques include blind-first analysis, in which students record an initial assessment before seeing the AI output; forced consideration of an alternative hypothesis; verification checklists tied to the system's known failure modes; and structured prompts

asking what evidence would falsify the recommendation. Experimental work indicates that cognitive forcing functions can reduce overreliance, although they may add friction and should be designed carefully (Buçinca et al., 2021).

3.3.2 The investigator as a human-AI team leader

The 'centaur' metaphor is useful only when it denotes deliberate task allocation rather than generic human-machine cooperation. AI may be effective at searching large datasets, ranking similarities, detecting anomalies, and maintaining consistency. Human investigators remain responsible for defining the question, evaluating lawful authority, interpreting context, interviewing people, resolving contradictions, considering proportionality, and accepting accountability for action. Hybrid-intelligence research similarly emphasizes combining complementary strengths rather than assuming that either humans or machines are universally superior (Dellermann et al., 2019).

Students should practice three roles: operator, reviewer, and decision owner. As operators, they must use the system within validated conditions. As reviewers, they must test outputs against independent evidence and document disagreement. As decision owners, they must explain how much weight the output received and why the final action was justified. Human-AI interface guidelines support this approach by emphasizing clear communication of capabilities, feedback, correction, and recovery from error (Amershi et al., 2019).

3.3.3 Synthetic media and generative-AI forensics

Generative AI creates both new evidence sources and new opportunities for deception. Synthetic video, audio, images, and text may be used for impersonation, fraud, fabricated admissions, false alibis, or evidence contamination. Detection remains an adversarial problem: artifacts that identify one generation method may disappear in the next. Reviews of deepfake creation and detection therefore stress generalization limits and the need for multiple forensic signals (Mirsky & Lee, 2021; Tolosana et al., 2020; Verdoliva, 2020).

A synthetic-media module should combine technical detection with evidence authentication. Students should inspect metadata and compression history, compare independent recordings, verify source and transmission path, examine temporal and acoustic consistency, use content-provenance information where available, and document tool versions and confidence. The goal is not to teach a list of visual 'tells' but a repeatable workflow that remains useful as generation methods evolve.

3.3.4 Laboratories, simulations, and after-action review

Practice should culminate in realistic scenarios that combine technical, legal, and ethical pressures. One exercise might provide a high-confidence facial candidate together with an exculpatory timestamp. Another might present a hotspot forecast generated from enforcement-biased data. A third might include a partially synthetic video embedded in an otherwise authentic evidence package. Students must decide what additional steps are lawful and necessary, record the AI system's role, and defend their decision before a mock supervisor, court, or community-review panel.

Every simulation should end with an after-action review that separates outcome from process. A lucky decision based on poor reasoning should not receive the same evaluation as a well-documented decision that appropriately handled uncertainty. Rubrics should assess issue spotting, use of independent verification, legal authority, interpretation of metrics, communication of uncertainty, documentation, and willingness to override the system when justified.

4 Curriculum Implementation and Evaluation

4.1 Spiral design and prerequisite structure

The Triple-Path framework should be implemented as a spiral curriculum rather than as a single AI elective. Introductory courses can establish data literacy and basic probability. Intermediate modules can integrate AI into evidence, procedure, and ethics. Advanced courses can use complex simulations and impact assessments. Repeated exposure allows students to revisit the same concepts—uncertainty, bias, provenance, oversight—at increasing levels of professional responsibility.

A feasible sequence is: (1) foundations of data and AI in investigation; (2) model evaluation and investigative statistics; (3) law, ethics, and governance of algorithmic systems; (4) human-AI decision making and communication; and (5) an integrated laboratory or capstone. Adult-learning research in police education supports problem-centered instruction that connects theory to professional experience rather than relying exclusively on lectures (Birzer, 2003).

4.2 Faculty, infrastructure, and procurement literacy

Interdisciplinary teaching teams are essential. Computer-science faculty can explain data and models; legal scholars can address authority, procedure, and evidence; ethics and social-science faculty can analyze institutional effects; and experienced investigators can connect abstractions to operational practice. The program should also involve defense lawyers, judges, auditors, community representatives, and vendors in carefully structured sessions so students encounter competing perspectives.

Laboratories do not require access to sensitive operational systems. Public or synthetic datasets, open-source models, redacted procurement documents, and simulated interfaces can teach most core competencies. Procurement literacy should be included because many practical safeguards depend on contracts: access to validation evidence, audit rights, incident reporting, version-change notice, data retention, subcontractors, security testing, and termination or data-return provisions.

4.3 Competency-based assessment

Evaluation should measure what graduates can explain and do, not merely what terminology they can recall. A program-level portfolio might include a model audit, an AIA, a legal-ethical opinion, a synthetic-media authentication report, and a simulation decision log. Oral defense is valuable because investigators must communicate uncertainty to supervisors, prosecutors, courts, and the public.

Institutional evaluation should also examine transfer to practice. Relevant indicators include the quality of AI-related documentation, frequency and justification of overrides, error reporting, disclosure compliance, subgroup performance monitoring, complaint resolution, and whether graduates can identify out-of-scope uses. These indicators should not create incentives to maximize or minimize overrides; the objective is appropriate reliance supported by evidence.

5 Discussion

The framework makes three contributions. First, it reframes AI literacy as investigative epistemology. The central question is not whether students can operate software but whether they can understand how a machine-generated output changes the evidentiary state of a case. Second, it treats law and ethics as design constraints that shape technical and operational choices. Third, it replaces vague human-in-the-loop language with assessable capabilities: independent analysis, calibrated reliance, authority to override, and traceable responsibility.

The framework also has limits. Legal requirements differ across jurisdictions, and some technologies described here may be prohibited, restricted, or unavailable in particular systems. Empirical evidence on the long-term effects of AI-specific police education remains limited. The curriculum should therefore be evaluated through pilots and revised in response to learner performance, technological change, judicial decisions, community impact, and documented operational failures. Responsible education is itself an iterative governance process.

Finally, the framework should not be interpreted as an endorsement of every law-enforcement AI system. Critical competence includes the ability to conclude that a proposed use lacks a legitimate purpose, sufficient validation, lawful data, effective safeguards, or acceptable social value. The most responsible decision may be constrained deployment, continued research, or non-use.

6 Conclusion

AI-enabled investigation requires a new educational settlement. Investigators must understand statistical outputs without confusing them with legal conclusions; evaluate data and model limitations without assuming that technology is neutral; apply privacy, fairness, due-process, and evidentiary principles throughout the system lifecycle; and work with AI without surrendering professional responsibility.

The Triple-Path framework provides a coherent route to these goals. Technological Reconstruction develops the capacity to interpret, audit, and communicate algorithmic evidence. Ethical and Legal Governance turns rights and public values into operational requirements. Practice-Oriented Human-AI Collaboration builds habits of verification, contestability, documentation, and calibrated trust. Together, the paths prepare investigators not merely to use AI, but to govern its role in the pursuit of justice.

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Paper 3, pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bahner, J. E., Hüper, A.-D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. International Journal of Human-Computer Studies, 66(9), 688–699. https://doi.org/10.1016/j.ijhcs.2008.06.001

Birzer, M. L. (2003). The theory of andragogy applied to police training. Policing: An International Journal of Police Strategies & Management, 26(1), 29–42. https://doi.org/10.1108/13639510310460288

Brayne, S. (2017). Big data surveillance: The case of policing. American Sociological Review, 82(5), 977–1008. https://doi.org/10.1177/0003122417725865

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), Article 188, 1–21. https://doi.org/10.1145/3449287

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 77–91.

Carpenter v. United States, 585 U.S. 296 (2018).

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

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 797–806). Association for Computing Machinery. https://doi.org/10.1145/3097983.3098095

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. Business & Information Systems Engineering, 61(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (pp. 160–171). Proceedings of Machine Learning Research, 81.

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down

harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689.

Ferguson, A. G. (2017). The rise of big data policing: Surveillance, race, and the future of law enforcement. New York University Press.

Grother, P., Ngan, M., & Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT), Part 3: Demographic effects (NISTIR 8280). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8280

Kerr, O. S. (2012). The mosaic theory of the Fourth Amendment. Michigan Law Review, 111(3), 311–354.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou (Ed.), 8th Innovations in Theoretical Computer Science Conference (ITCS 2017) (Article 43, pp. 43:1–43:23). Schloss Dagstuhl–Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165(3), 633–705.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Paper 598, pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

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

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765–4774).

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021). Algorithmic impact assessments and accountability: The co-construction of impacts. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 735–746). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445935

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. ACM Computing Surveys, 54(1), Article 7, 1–41. https://doi.org/10.1145/3425780

Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., & Brantingham, P. J. (2015). Randomized controlled field trials of predictive policing. Journal of the American Statistical Association, 110(512), 1399–1411. https://doi.org/10.1080/01621459.2015.1077710

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, Article 100041. https://doi.org/10.1016/j.caeai.2021.100041

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410. https://doi.org/10.1177/0018720810376055

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 145–151). Association for Computing Machinery. https://doi.org/10.1145/3375627.3375820

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Richardson, R., Schultz, J. M., & 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, 15–55.

Roth, A. (2017). Machine testimony. Yale Law Journal, 126(7), 1972–2053.

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, Account-

ability, and Transparency (pp. 59–68). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human-Computer Interaction, 36(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? International Journal of Human-Computer Studies, 51(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). DeepFakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131–148. https://doi.org/10.1016/j.inffus.2020.06.014

Verdoliva, L. (2020). Media forensics and deepfakes: An overview. IEEE Journal of Selected Topics in Signal Processing, 14(5), 910–932. https://doi.org/10.1109/JSTSP.2020.3002101

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. International Data Privacy Law, 7(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wexler, R. (2018). Life, liberty, and trade secrets: Intellectual property in the criminal justice system. Stanford Law Review, 70(5), 1343–1429.