automation

MDPI

Article

Rights-Based AI in Cyber–Physical Systems: A Governance Framework for Socio-Technical Resilience and Trust

Maral Niazi [1,*], Hossein Hassani [2] and Madison Lee [3]

[1] Balsillie School of International Affairs, University of Waterloo, Waterloo, ON N2L 6C2, Canada
[2] The International Institute for Applied Systems Analysis (IIASA), Schloßpl. 1, 2361 Laxenburg, Austria; hassani.stat@gmail.com
[3] Balsillie School of International Affairs, Wilfrid Laurier University, Waterloo, ON N2L 6C2, Canada; mlee@balsillieschool.ca
* Correspondence: mniazi@balsillieschool.ca

Abstract

AI-enabled cyber–physical systems (CPSs) are increasingly deployed in public governance contexts where they sense human populations, infer classifications or risks, and trigger interventions that can shape liberty, equality, and access to essential services. In these deployments, governance failures often arise not only from model error but from systems-level interactions across data generation, model updates, organizational practices, and downstream actuation. This paper introduces a Risk–Rights–Rules (3R) architecture that treats fundamental rights and legal rules as enforceable constraints on the sensing–inference–actuation loop, rather than as external ethical aspirations. Building on established risk-management baselines and safety engineering practice, we specify a testable assurance object, a structured 3R assurance case, that links rights claims to explicit assumptions, measurable evidence, and accountable control points across the lifecycle. The approach is designed to reduce "legitimacy drift" in stochastic decision pipelines by making uncertainty, demographic error, contestability, and procurement leverage auditable at the system level. The result is a governance blueprint for high-consequence public-sector AI deployments for governance failures, which is both technically robust and institutionally defensible.

Keywords: cyber–physical systems; socio-technical systems; public-sector AI governance; predictive AI; biometric identification; risk assessment; fundamental rights; assurance cases; Goal-Structuring Notation (GSN); STPA/STAMP; automation bias; procedural due process

Academic Editors: Quanyan Zhu and Abbas Yazdinejad

Received: 30 January 2026
Revised: 18 March 2026
Accepted: 25 March 2026
Published: 15 June 2026

1. Introduction

The traditional definition of a cyber–physical system (CPS) emphasizes the seamless integration of computational algorithms and physical components, characterized by a feedback loop in which embedded computers monitor and control physical processes, and where physical processes affect computations and vice versa [1,2]. Until recently, the CPS was largely confined to hard engineering domains—avionics, automotive control, and industrial robotics—where the primary objective was functional safety and the mitigation of kinetic hazards [2].

A consequential shift is now underway: CPS-like sensing–inference–actuation loops are increasingly deployed as governance infrastructures that manage people (e.g., policing, welfare administration, and border screening), not only machines. In these settings, the "plant" being regulated is partly a human population and its behavior un-

der institutional constraints, requiring socio-technical—not purely mechanistic—models of safety, accountability, and legitimacy [3,4]. In this emergent socio-technical CPS, the sensing–inference–actuation loop can operate as a cybernetic regulator of rights: measurements are transduced from the environment (e.g., biometric capture), processed through statistical learning systems to produce classifications or risk estimates, and translated into control actions that change a person's procedural or physical pathway (e.g., denial of entry, increased enforcement attention, or eligibility restriction) [5–9]. Because these systems intervene through institutions rather than only through mechanics, they embed normative commitments (e.g., what counts as "risk," what error costs are tolerated, and what contestation is available). The governance question therefore becomes systems-level: the relevant unit is not the model alone but the end-to-end pipeline and its operational controls [7,10,11].

A central concern in high-consequence deployments is the migration of decision authority from publicly contestable legal rules to opaque algorithmic configurations and institutional routines. Where coercive or high-friction interventions are shaped by automated inferences, the technical pipeline can function not merely as decision support but as a routinized form of governance—often without an equivalent routinization of procedural safeguards (notice, reasons-giving, and meaningful contestability). This is not a narrow "model bias" problem; it is an accountability and due-process problem created by how errors, uncertainty, and discretion propagate across socio-technical pipelines [6,9,11].

This transition creates an epistemic gap. Machine learning introduces both aleatoric uncertainty (irreducible randomness) and epistemic uncertainty (uncertainty arising from limited knowledge, data, or model misspecification) into decision environments that, legally and institutionally, depend on stable, reviewable reasons [12]. When uncertainty is not operationalized—measured, bounded, and linked to actuation constraints—institutions can produce decisions that are "confident in form but fragile in substance." That fragility undermines legal certainty because individuals cannot reliably understand, challenge, or correct the basis on which power is exercised [10,12].

Biometric identification provides a concrete example. When a face recognition pipeline exhibits demographically differentiated error rates, technical error margins can become a differential distribution of rights exposure, particularly when false matches trigger downstream interventions in policing or border workflows [13]. NIST's FRVT-testing program documents the demographic differentials in contemporary face recognition performance and provides a primary empirical basis for treating demographic error as a governance-relevant systems risk rather than a purely technical nuisance [13]. Even where the average performance appears "acceptable," legitimacy can fail when (i) error burdens fall disproportionately, (ii) sensing violates contextual constraints, or (iii) the affected individuals lack procedural pathways to contest the outcomes and obtain a correction [11].

To address these vulnerabilities, this paper proposes a 3R architecture, motivated by scholarship arguing that risk-based governance must be complemented by constitutional and rule-of-law constraints rather than treated as a purely technocratic optimization problem [14–16]. The paper's core claim is systems-theoretic: in high-consequence deployments, risk management and rights protection are interdependent because rights breaches constitute governance failures that undermine legitimacy and safe operation at the institutional level [14–16]. Rights constraints define non-negotiable boundaries for acceptable operation, while risk tools manage residual uncertainty within those boundaries [14–16].

The framework is grounded in authoritative baselines, including the NIST AI Risk Management Framework (AI RMF 1.0), which provides a lifecycle-oriented governance structure organized around GOVERN–MAP–MEASURE–MANAGE [17], and the EU Artificial Intelligence Act (Regulation (EU) 2024/1689), which establishes binding obligations

for high-risk systems and integrates fundamental-rights considerations through mechanisms such as the Fundamental Rights Impact Assessment (FRIA) [18,19]. However, these baselines can be difficult to operationalize into a single implementable artifact usable by engineers, procurement teams, auditors, and legal reviewers to test whether a deployment remains within acceptable bounds.

To that end, we introduce an implementable assurance object, the "3R Assurance Case," which translates rights and legal-rule claims into an auditable, testable argument structure. The 3R Assurance Case uses Goal-Structuring Notation (GSN) to link (i) top-level claims (e.g., non-discrimination, and due-process contestability) to (ii) explicit assumptions and boundary conditions, (iii) evidence artifacts (evaluation results, demographic error tests, logging/provenance records, red-team reports, and procurement audit rights), and (iv) named control points in the CPS loop where constraints must be enforced [20,21]. In parallel, we leverage systems safety reasoning (STAMP/STPA) to treat harmful outcomes as control failures and to derive "unsafe control actions" corresponding to rights-relevant hazards in socio-technical pipelines.

Figure 1 summarizes the study's four main contributions to governing high-consequence predictive and biometric AI in public workflows. It defines a scope-explicit governance target, reframes rights breaches as systems control failures at the actuation boundary, introduces the testable 3R Assurance Case to translate rights and legal rules into auditable claims and evidence, and specifies lifecycle control points where uncertainty, demographic error, and contestability must be measured and enforced.

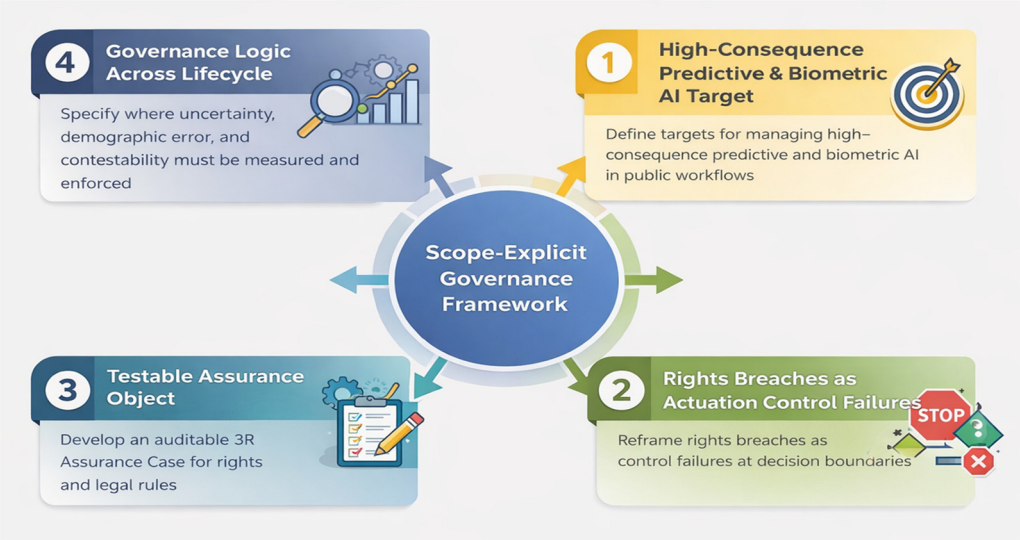

Figure 1. Scope, control, assurance, lifecycle: key contributions of this study.

The following list provides a more detailed summary of the paper's contributions, in addition to the four main contributions highlighted in Figure 1:

- This paper reframes the CPS as governance infrastructure: This paper extends the CPS sensing–inference–actuation model from engineering safety to socio-technical governance, where the "plant" includes people, institutions, and rights exposure—and where failures are legitimacy and due-process failures, not just technical errors.
- This paper defines a scope-explicit target for high-consequence AI: This paper positions predictive and biometric AI in public workflows as end-to-end pipelines (data → model → decision → actuation → feedback), making the system—not the algorithm—the unit of governance and accountability.

- This paper introduces the 3R architecture (Risk–Rights–Rules): This paper proposes a systems-theoretic framework where rights and legal rules operate as enforceable constraints on the CPS loop, while risk tools manage residual uncertainty within non-negotiable boundaries.
- This paper operationalizes governance via a "3R Assurance Case": This paper provides an implementable assurance object using Goal-Structuring Notation (GSN) to connect top-level rights/rule claims to assumptions, evidence artifacts, and specific control points in the loop.
- This paper adds a structured auditing and testing approach: This paper specifies the evidence package needed for auditability (e.g., demographic error testing, uncertainty bounds, logging/provenance, red-teaming, and procurement audit rights) and ties it to lifecycle checkpoints aligned with NIST AI RMF and the EU AI Act/FRIA.
- This paper derives control failures using systems safety methods: This paper applies STAMP/STPA to translate rights-relevant harms into unsafe control actions, enabling the systematic identification of "pinch points" where governance controls must intervene.
- This paper produces procurement and vendor-governance requirements: This paper converts the framework into actionable requirements for procurement, contracting, and ongoing oversight, supporting engineers, auditors, and legal reviewers with a testable definition of rights-based CPS governance.

The remainder of this paper proceeds as follows: Section 2 provides a short literature review in line with the Introduction; Section 3 establishes the technical and institutional substrate, analyzing CPSs as governance infrastructures and redefining resilience as sustained functioning under disruption without collapsing legitimacy constraints. Section 4 develops the 3R analytical framework, mapping the governance triad onto the CPS loop and specifying the evidentiary requirements for trust. Section 5 applies the framework to a portfolio of documented high-consequence deployments, using process tracing to identify governance pinch points. Section 6 derives operational requirements for procurement and vendor governance, establishing a testable definition of rights-based CPS governance.

2. Literature Review

2.1. Socio-Technical CPS and Systems-Level Governance

As CPS-like feedback loops migrate from physical engineering into public administration, the literature increasingly treats these deployments as socio-technical systems whose safety and legitimacy depend on the end-to-end pipeline, not solely on model performance. In such systems, sensing, inference, and actuation are embedded within institutional procedures, requiring governance frameworks that account for operational controls, accountability, and legitimacy across the full lifecycle [3,4,7,10,11].

2.2. Uncertainty, Accountability, and Legal Certainty

A core strand of scholarship focuses on how uncertainty—both aleatoric and epistemic—enters decision environments that require stable, reviewable reasons. When uncertainty is not explicitly measured and constrained at the actuation stage, organizations can produce decisions that appear definitive while being structurally fragile, undermining legal certainty and due process [10–12]. This work emphasizes that the accountability problem is not reducible to bias metrics: it arises from how uncertainty and error propagate through institutional routines and decision pathways [10–12].

2.3. Biometric Systems, Demographic Error, and Rights Exposure

Empirical work on biometric identification, particularly face recognition, demonstrates that model error is not evenly distributed across demographic groups, and that these differences can translate into unequal exposure to downstream interventions in policing and border workflows [13]. The NIST FRVT results provide a central empirical anchor for treating demographic error as a governance-relevant risk, motivating requirements for demographic testing, documentation, and oversight mechanisms that connect technical evaluation to real-world actuation [13].

2.4. Fairness Beyond Metrics: Feedback Loops and Institutional Context

A recurring critique in fair-ML and socio-technical research is the "abstraction trap": treating fairness as a property of a model or metric while ignoring the institutional context and feedback loops that shape real-world outcomes [7,22]. In the ML-enabled CPS, feedback effects are especially salient. Predictive policing and enforcement-allocation systems can produce endogenous evidence: increased surveillance yields more recorded incidents, feeding back into the models and reinforcing disproportionate intervention patterns in already over-surveilled communities [23,24]. This literature highlights that governance must address the full control loop and its dynamic impacts, not only static model metrics.

2.5. Human-in-the-Loop Oversight and the "Moral Crumple Zone"

Another body of research warns that "human-in-the-loop" safeguards can provide false assurance. Studies in human–automation interaction document automation bias and reduced verification under time pressure and complexity, leading to both commission and omission errors [25–27]. Related governance scholarship describes the "moral crumple zone," where formal responsibility collapses onto human operators who lack meaningful agency due to workflow constraints, tooling, and organizational incentives [28]. Together, these findings support the argument that accountability must be engineered into system design and organizational controls, rather than assumed through nominal human oversight.

2.6. Governance Baselines and Operationalization into Assurance Artifacts

Finally, emerging governance baselines—including the NIST AI RMF and the EU AI Act—provide lifecycle-oriented structures and binding obligations for high-risk systems, increasingly integrating fundamental-rights considerations such as impact assessment requirements [17–19]. A key challenge identified in practice-oriented scholarship is operationalizing these standards into implementable, auditable artifacts that align engineering, procurement, audit, and legal review. Assurance-case approaches (e.g., using GSN) offer a method for structuring claims, assumptions, evidence, and control points in a way that supports auditability and accountability across socio-technical pipelines [20,21].

3. Cyber–Physical Systems, Critical Infrastructure, and Rights-Risk Coupling

3.1. CPS as Governance Infrastructure: The Engineering of Public Law

In the context of public authority, a CPS is a coupled computational–physical control system that monitors and actuates physical processes through feedback loops [29,30]. When CPSs are deployed to exercise or support public power, they operate as a regulatory infrastructure. In these circumstances, governance objectives are pursued through cybernetic processes of (i) standard setting, (ii) continuous sensing/monitoring, and (iii) intervention/behavior modification, increasingly executed or mediated by algorithmic systems [31,32]. In that sense, the "control law" that structures conduct is no longer expressed only in legal syntax (rules, reasons, and justifications), but is partly operationalized as computational standards, decision thresholds, the fusion/aggregation of sensor and

database signals, and automated triggers, characteristic of code-driven and data-driven regulation [31–33]. Figure 2 summarizes how an AI-enabled CPS can become governance infrastructure and why rights–risk coupling matters for resilience.

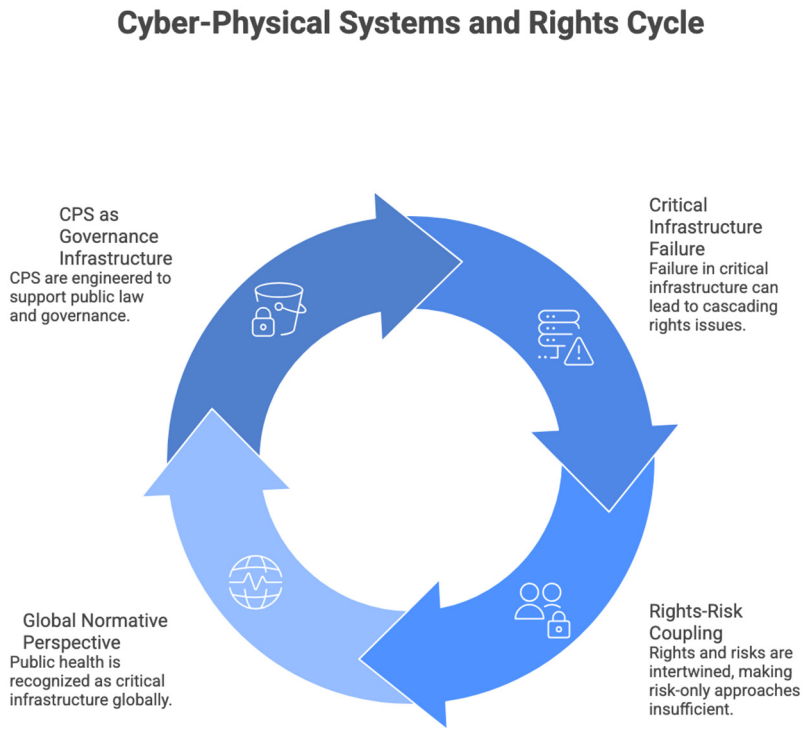

Figure 2. Cyber–physical systems and the rights cycle: conceptual schematic showing CPS as governance infrastructure, the coupling of critical-infrastructure failure modes with rights impacts, the rights–risk coupling that makes "risk-only" governance insufficient, and the global normative framing of public health as critical infrastructure.

In live biometric surveillance contexts, the operative control law includes not only a model and its threshold, but also where sensing is authorized and who is eligible to be sensed (watchlisting). When the siting and watchlisting criteria are under-specified, risk detection becomes the open-ended surveillance capacity, a rules failure at the sensing boundary, not an accuracy failure at the model boundary [31–33]. If rights are treated as external ethics, agencies can attempt to fix the model while leaving the sensing boundary unconstrained; if rights are treated as control constraints, legality and bounded discretion must be enforced before the loop is allowed to operate.

In traditional governance, officials apply qualitative standards, such as reasonable suspicion, through context-sensitive judgment and reasons that are contestable and subject to review. In a CPS-enabled workflow, discretion can be delegated to an inference layer that outputs probabilistic assessments or scores, which are then mapped to action through fixed or adaptive thresholds [34]. Where interventions are automatically triggered once a threshold is met, the decision process risks collapsing individualized assessment into category-based scoring, raising long-standing due-process concerns about transparency, contestability, and reason-giving in automated prediction systems used by public authorities [11]. This is the mechanism-level failure mode. The system's reasons become a score and a trigger rule, and, without enforceable transparency/contestability hooks, affected persons cannot meaningfully interrogate why authority was exercised over them.

3.2. Critical Infrastructure and the Cascading Failure of Rights

The governance stakes intensify when CPSs operate within critical infrastructure (CI), because CI is best understood as an interdependent system-of-systems in which failures propagate across coupled technical and organizational networks [35,36]. When CI functions are increasingly mediated by AI-enabled CPSs, the inference layer introduces additional, well-characterized failure modes, most notably concept drift/dataset shift (performance degradation as the data-generating process changes) and pipeline vulnerabilities/entanglement in production ML systems [37,38]. We refer to this condition as normative fragility: situations in which technical degradation in the inference-and-decision stack can result in service denials that implicate fundamental rights and legally protected entitlements. However, many AI-enabled CPS deployments depend on shared cloud, identity, and third-party data-integration layers before a denial event occurs [39,40].

For instance, when authentication becomes a gate to benefits, an "error" becomes a denial event. In biometric welfare systems, infrastructural fragility and authentication failures can lead to exclusion (people entitled by law are excluded in practice), including the denial of health-relevant benefits [41,42]. The analytic point is not that systems should be perfect; it is that CI systems must be governable under imperfection. In rights-relevant CI contexts, the absence of fallback procedures, review pathways, and evidence trails turns technical non-idealities into legally salient harms [41,42].

Interdependence is especially salient in contemporary deployments because many AI-enabled CPSs share a common digital stack: cloud-based compute and service layers that industrialize AI development and deployment pipelines, federated identity components that centralize access control, and third-party data ecosystems that supply or intermediate data used in automated decision systems [39,40]. In such architectures, disruptions in the inference layer, whether from drift or from a compromise of the training/validation pipeline, do not merely reduce "accuracy" or cause downtime; they can reshape access decisions in ways that deny essential services. Here, the control-theoretic and governance insights coincide. CI failures are rarely isolated. They propagate across coupled components [35,36], and ML introduces a non-stationary failure class (drift) that causes yesterday's assurance to decay unless continuously re-evidenced [37,38]. This is why risk-only governance is insufficient. Risk governance that treats the model as a static component ignores the systems reality that CI governance must manage drift, entanglement, and dependency coupling as ongoing operational conditions, not one-time deployment checks [37,38].

3.3. The Rights–Risk Coupling: Why "Risk-Only" Paradigms Fail

A core technical–normative claim of this framework is that risk reduction measures often have rights-externalities: in control and safety engineering, improved risk performance is frequently pursued through denser sensing, tighter feedback, and stronger control constraints; in socio-technical governance, those same moves can increase surveillance intensity and strengthen behavior-shaping interventions, thereby transforming "security risk" optimization into a rights risk unless bounded by legality, proportionality, and contestability [31,32]. Consider the following areas:

- Biometric Disparity as System Failure
    - A body of research on face recognition repeatedly finds that error rates and decision thresholds vary by demographic cohort, and that uniform thresholds can therefore impose unequal false-match burdens across groups [43,44]. Normatively, these differentials are best analyzed as distributional harms (disparate burdens of erroneous identification/verification) rather than mere implementation defects, because higher false match rates for a protected group predictably

translate into more frequent wrongful interventions, such as secondary screening, denial of access, and investigative escalation, for that group [7,13,45].

- When a uniform operational threshold is applied in a high-throughput setting, statistical error becomes a procedural burden, with more stops, more delays, and more escalations, borne unevenly by identifiable groups [13,45]. This is precisely why the 3R framework treats rights as control constraints. A rights-consistent loop must require (i) demographic disparity testing as a precondition for deployment, (ii) threshold governance tied to error distributions, and (iii) actuation constraints that prevent error-prone groups from absorbing the system's friction as a stable output [13,45].

- Actuation-Sensing Feedback Loops
  - In predictive policing and CPS actuation (e.g., reallocating patrols), these processes reshape the data-generating process (e.g., observed arrests, stops, and recorded incidents). The inference layer then treats these observations as updated evidence of "risk," generating a self-reinforcing loop that can concentrate policing in the same neighborhoods even when the true underlying rates do not justify it [46]. Technically, the system can appear "stable" because it continually confirms its own predictive target via endogenous data; normatively, it is unstable because it can amplify historically patterned disparities and convert prior over-policing into future "evidence" [24,47].
  - Studies show that predictive policing systems can reproduce enforcement concentration because the system learns from "discovered" incidents rather than underlying incidents [46]. Moreover, policing data is shaped by prior enforcement patterns and surveillance practices, which can embed civil-rights harms into the data pipeline itself [24,47]. This means risk is confounded with enforcement intensity. A risk-only paradigm fails because it treats the feedback signal as ground truth; a rights-aware control paradigm must constrain the loop so that actuation does not manufacture the evidence that justifies more actuation [24,46,47].

3.4. The Global Normative Perspective: Public Health as Critical Infrastructure

At the domestic level, public health-adjacent systems are explicitly treated as critical infrastructure in both the United States and Canada. In the U.S., healthcare and public health are formally designated as a DHS/CISA critical infrastructure sector [48]. In Canada, the federal government has classified "Health" as critical infrastructure, supporting the claim that health services are treated as an infrastructure whose disruption has cross-sector consequences [49].

At the global level, however, the normative definition of "critical infrastructure" is less settled; instead, global health security is more accurately conceptualized as a transboundary surveillance-and-response regime organized around early warning, notification, and coordinated response duties under the International Health Regulations (IHR), a legally binding instrument of international health governance implemented through the World Health Organization (WHO)-centered coordination and state-party core capacities [50–54].

The integration of AI-enabled CPSs into global health data collection and intervention, such as automated contact tracing, risk scoring, or AI-supported screening at points of entry, therefore intensifies assurance demands (security, integrity, and auditability) while simultaneously raising human rights demands (privacy, proportionality, and nondiscrimination) [55–59]. Historically, emergency rationales can expand administrative discretion and normalize exceptional measures, thereby threatening durable rights constraints [58]. Our framework accordingly treats rights-based constraints as more, not less,

necessary in crises: when interventions are executed through automated feedback loops, governance failures can become self-reinforcing and difficult to unwind, producing "run-away" intervention dynamics unless constrained by legally legible safeguards (traceability, contestability, and proportionality) [56–58].

4. Analytical Framework: The Risk–Rights–Rules (3R) Architecture

4.1. Systems-Theoretic Safety and the Formalization of Rights as Control Constraints

A recurring limitation of model-centric "accuracy-as-safety" AI governance is its reliance on component-level reliability, implicitly assuming that, if an AI model performs well, the overall system is safe [60]. However, in complex cyber–physical systems (CPSs), accidents—and, by extension, public-law harms—are frequently the result of unsafe interactions, incomplete feedback, and inadequate control authority across components rather than isolated component failures [8]. This point is the central safety-engineering reason that system safety methods focus on the control structure and interaction constraints rather than part performance alone. Figure 3 illustrates the Risk–Rights–Rules (3R) architecture as an iterative governance loop that connects system safety, enforceable legal rules, assurance evidence, and remedy.

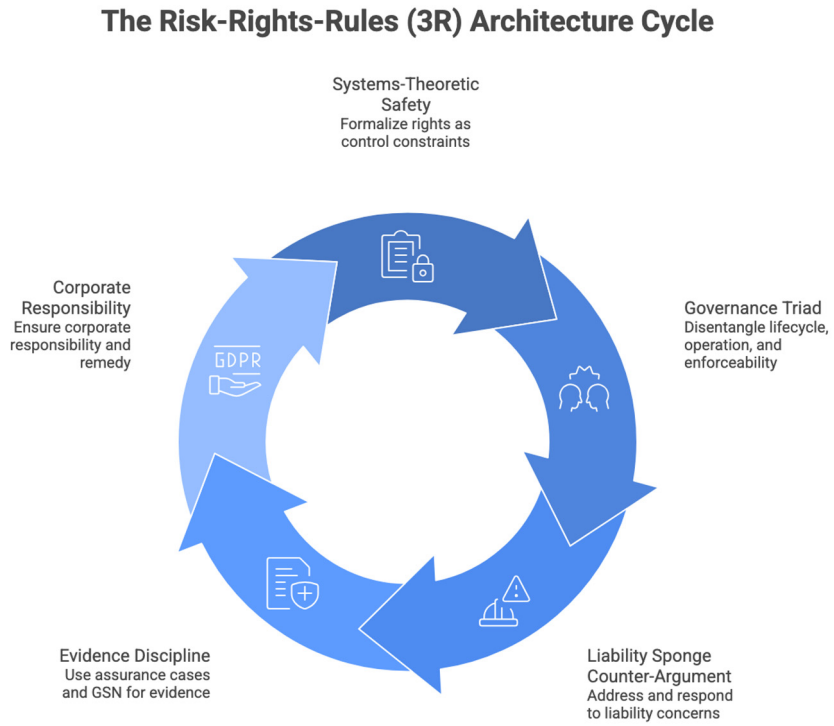

Figure 3. The Risk–Rights–Rules (3R) architecture cycle: conceptual schematic linking systems-theoretic safety (rights as control constraints), governance triad decomposition, the liability-sponge counter-argument, evidence discipline via assurance cases/GSN, and corporate responsibility and remedy as an iterative governance loop.

To address this, we adopt Systems-Theoretic Process Analysis (STPA), an engineering methodology that models safety as a hierarchical control problem and derives safety requirements as enforceable constraints on control actions [1,6,8]. The methodological payoff for governance is precision: STPA forces the analyst to name (i) the controller, (ii) the controlled process, (iii) the feedback signal, and (iv) the constraints under which control actions are permitted. This is exactly what is missing in "ethics-only" AI governance, which tends to state values without specifying where and how they bind the loop.

To make this move operational for governance-oriented CPSs, we formalize the socio-technical loop as a controlled dynamical system whose "safe operation" is defined not only by functional stability but also by the preservation of legally salient invariants.

In a border biometric workflow, R is violated when the system escalates to intrusive or coercive measures without meeting jurisdiction-specific public-law constraints (lawful basis, necessity, proportionality, and reviewability). R is therefore implemented as an actuation gate: actuation is permitted only if the preconditions are satisfied and recorded. This turns rights from a value statement into a control precondition and an evidence obligation.

In this framework, a rights violation is modelled as a rights-relevant Unsafe Control Action (UCA). In STPA, a UCA is a control action that becomes unsafe given the context, for example, when the action is provided when it should not be, not provided when it should be, or provided with unsafe timing/duration such that it can lead to a hazardous state [8,61]. For a rights-based CPS, we define the hazardous state as a breach of a legally salient invariant, such as privacy, due process safeguards, or equality/non-discrimination. Let Cr(x_t) denote the set of rights constraints that must hold in state x_t. A control action u_t is rights-unsafe if it is issued while the relevant constraints are not satisfied. Rights-consistent operation therefore requires that high-consequence actuation occurs only when applicable rights constraints are satisfied. For instance, in an automated border-control system, a rights-relevant UCA occurs if the inference engine issues an actuation command to escalate to an intrusive or coercive search without satisfying the control constraint of a lawful basis, necessity and proportionality, and effective contestability or human review. Accordingly, probable cause is not treated as a universal numeric threshold; instead, the constraint is specified in jurisdiction-appropriate public-law terms (legality, proportionality, and reviewability) and must be satisfied at the point of actuation.

A rights-relevant UCA occurs when a system escalates a person to secondary screening or police engagement on the basis of a match/score while (i) uncertainty is unbounded or degraded, (ii) the watchlist/scope rule is not legally constrained, or (iii) no contestability pathway exists at the point of intervention. Note what this does analytically: it locates failure at the control boundary (scope/actuation/feedback), not only at model performance.

4.2. The Governance Triad: Disentangling Lifecycle, Operation, and Enforceability

At the enforceability layer, rights-based AI governance is also anchored in emerging international public-law instruments, including the Council of Europe Framework Convention on Artificial Intelligence and related legal analysis [62]. To mitigate governance- and ethics-washing, i.e., treating voluntary corporate principles as reputational substitutes for enforceable public accountability, the 3R architecture separates responsibilities across lifecycle, operation, and enforceability, while identifying where documentation and accountability break down [63]. We refer to this as preserving an epistemic chain of custody: traceable provenance from data and design choices to decisions, actions, and outcomes.

The governance of AI (Lifecycle Stewardship) is a technical layer that concerns the stochastic integrity of the model and its dependencies. It requires standardized documentation artifacts, Model Cards and Datasheets for Datasets, and internal auditing processes that tie design decisions to deploy-time monitoring and remediation [64,65]. Without these artifacts, inference is practically non-auditable, increasing the likelihood that performance differentials, including demographic error-rate disparities documented in the face-recognition literature, propagate into actuation undetected [66].

Governance by AI (Algorithmic Authority) integrates the control law, whereby probabilistic scores function as coercive triggers through threshold-escalation rules. A critical gap in many "HITL solves accountability" narratives is that human oversight can become rubber-stamp oversight under time pressure, workload, and interface design

conditions that reliably produce automation bias and deference to automated recommendations [25,26,67,68]. In such settings, the human worker can be positioned as a "moral crumple zone," formally responsible yet operationally constrained, unless the system is designed for meaningful human control, real authority to pause/override, adequate time, and accountable procedures [25,28,69].

"Governance by" fails when the human is present but lacks (i) time to review, (ii) authority to pause/override, and (iii) accountable procedures that require verification rather than deference. Wagner shows how "quasi-automation" can turn humans into rubber stamps, thereby increasing liability without restoring control [69]. This is exactly the condition in which "moral crumple zones" emerge: humans absorb responsibility for system behavior they did not meaningfully control [28].

Governance for AI (Institutional Enforceability) comprises hard-law and institutional mechanisms that make "rules" binding design conditions, such as conformity assessment, post-market monitoring, and independent oversight. A scholarly analysis of the EU AI Act emphasizes the need to translate legal requirements into verifiable criteria and auditable evidence throughout the lifecycle, rather than merely aspirational compliance statements [70]. At the international level, the Council of Europe's Framework Convention on AI anchors enforceability beyond soft ethics by requiring lifecycle-consistent measures aligned with human rights, democracy, and the rule of law [71]. Scholarship assesses it as a rights-based, full-lifecycle instrument with both promise and governance trade-offs [72].

"Governance for" fails when oversight bodies cannot inspect evidence: logging, traceability, evaluation access, and update/change control are not optional; they are enforceability primitives. Mökander et al. make this explicit by mapping regulatory duties to auditable enforcement mechanisms (conformity assessment and post-market monitoring) rather than ethics statements [70]. Operationally, this implies audit readiness at the point of sensing and decision (logging, traceability, and review hooks), rather than post hoc investigations after harm [66,70].

4.3. The "Liability Sponge" Counter-Argument and Response

A common counter-argument is that stricter rights-based constraints (explainability, reviewability, and due-process-like safeguards at actuation) increase compliance costs and latency, potentially reducing efficiency in high-consequence sectors [73,74]. The 3R response is not to deny tradeoffs; it is to reframe them as control design choices with empirically predictable failure costs. If institutions save time by bypassing contestability and evidence logging, they externalize costs in the form of downstream harms: litigation, loss of legitimacy, program collapse, and chronic contestation that impairs operational viability.

Trust and legitimacy are not soft add-ons. Trust shapes the appropriate reliance under uncertainty [75], and legitimacy shapes cooperation and compliance with authority [76]. Rights-based AI principal mapping also supports treating procedural safeguards as governance requirements rather than voluntary ethics statements [77]. Moreover, responsibility gaps show that delegating control to opaque learning systems can create situations in which it is neither fair nor coherent to assign responsibility to operators for outcomes they could not foresee or prevent [78]. Empirically, this dynamic is amplified by oversight-by-policy (human-in-the-loop mandates) that legitimize deployment without enabling an effective review [67].

Therefore, rights constraints are stabilizing control constraints: they reduce the probability of high-impact harms, preserve legitimacy, and support the durable adoption of CPS governance infrastructures [79]. In practical terms, the cost of rights constraints should be compared to the cost of uncontrolled actuation: when the actuation boundary is not

gated by legality and evidence, efficiency gains can be purchased by systematically shifting burdens onto identifiable populations, a governance failure, not an optimization.

4.4. Evidence Discipline: Assurance Cases and Goal-Structuring Notation (GSN)

To bridge the gap between claims (e.g., "the CPS is rights-consistent") and evidence (e.g., audit logs, drift tests, disparity tests, and human-review traces), we use Goal-Structuring Notation (GSN), a graphical notation developed to improve the structure and assessability of safety arguments in safety-critical domains [20,21]. The Assurance Case Scholarship emphasizes that the distinctive value of assurance cases lies in the explicit argument linking claims to evidence and assumptions, and in the maintainability of that argument as systems evolve [80].

This relates to the reviewer-style question: "Where is the formal object engineers can implement/verify/test?" The formal object is the 3R Assurance Case, a GSN-structured, versioned artifact that binds rights claims to named control points and evidence obligations. It is implementable because each node corresponds to an auditable artifact (test report, log schema, review trace, drift monitor, disparity evaluation, and change-control record). In a 3R GSN structure, we find the following:

- Top-Level Goal (G1): The CPS deployment is rights-consistent.
- Context (C1): This refers to the applicable hard-law duties and rights safeguards [70].
- Strategy (S1): This includes decomposing rights into technical invariants and control constraints (sensing scope, inference auditability, and actuation gates) [8,61].
- Evidence (E1, E2, . . .): This includes STPA hazard/UCA logs, disparity tests from the peer-reviewed biometrics literature, drift monitoring results, and audit trails of human intervention/override [26,70].

For a biometric escalation workflow, the assurance case must include the following: (i) a scope-control instrument (where/when sensing is permitted, and who is eligible for watchlisting), (ii) threshold governance tied to the demographic error evaluation, (iii) an actuation-gate policy with logged preconditions, (iv) a human-review trace proving override is practical, and (v) a post-market monitoring plan that triggers pause-and-review under drift or anomaly. This turns "rights" into a testable assurance claim rather than a declaration.

This evidentiary discipline ensures that the assurance case is a living document: if the environment changes (e.g., dataset shift/concept drift), previously valid evidence may fail, and the safety/rights argument must be updated, potentially triggering a pause-and-review [80]. Rather than claiming "rights are proven," the framework operationalizes rights as testable, auditable claims that must remain defensible in the face of change.

4.5. Corporate Responsibility and Remedy in the CPS Ecosystem

Since many governance CPSs are procured from private vendors, business-and-human-rights scholarship treats the corporate responsibility to respect rights and provide access to remedy as operationally relevant across complex supply chains and public procurement [81]. We argue that the actuation gate should include a remedy-by-design requirement: the system must generate contestability-enabling information at the time of intervention, including (i) an actionable explanation and (ii) a clear pathway to challenge or seek review. Counterfactual explanations are particularly suited to this role because they can support contestation and recourse without requiring the disclosure of the model's full internal logic [82]. Where recourse is possible, formalized actionable recourse further clarifies the remedies required in practice [83]. The technical governance implication is precise: if a person is diverted, delayed, denied, or escalated, the system must produce a minimal

record of contestability sufficient to enable a review; otherwise, contestability exists only in theory.

Concretely, the actuation pathway should produce an auditable "adverse action packet": the decision, the minimally sufficient features/conditions that drove the escalation (in counterfactual form), and the review channel, including human review, appeal, or the oversight body [84]. This prevents the black box from becoming a black hole for accountability by embedding remedy and review hooks into the CPS control structure rather than relegating them to after-the-fact investigations [28,81].

5. Methodology: A Socio-Technical Systems (STSs) Approach to Governance Auditing

5.1. Research Design: Multi-Case Theory Building and Replication Logic

This study employs a qualitative, multi-case study design centered on analytic generalization. Unlike statistical generalization, which seeks to extrapolate findings from a sample to a population, analytic generalization aims to compare empirical results against a theoretical template, in this case, the Risk–Rights–Rules (3R) architecture. Following guidance on the case-study method, analytic generalization is implemented through pattern matching and replication across cases, rather than through population inference [85]. Figure 4 summarizes the socio-technical systems governance-auditing cycle used to connect case selection, mapping, evidence ranking, assurance argumentation, process tracing, and harm minimization.

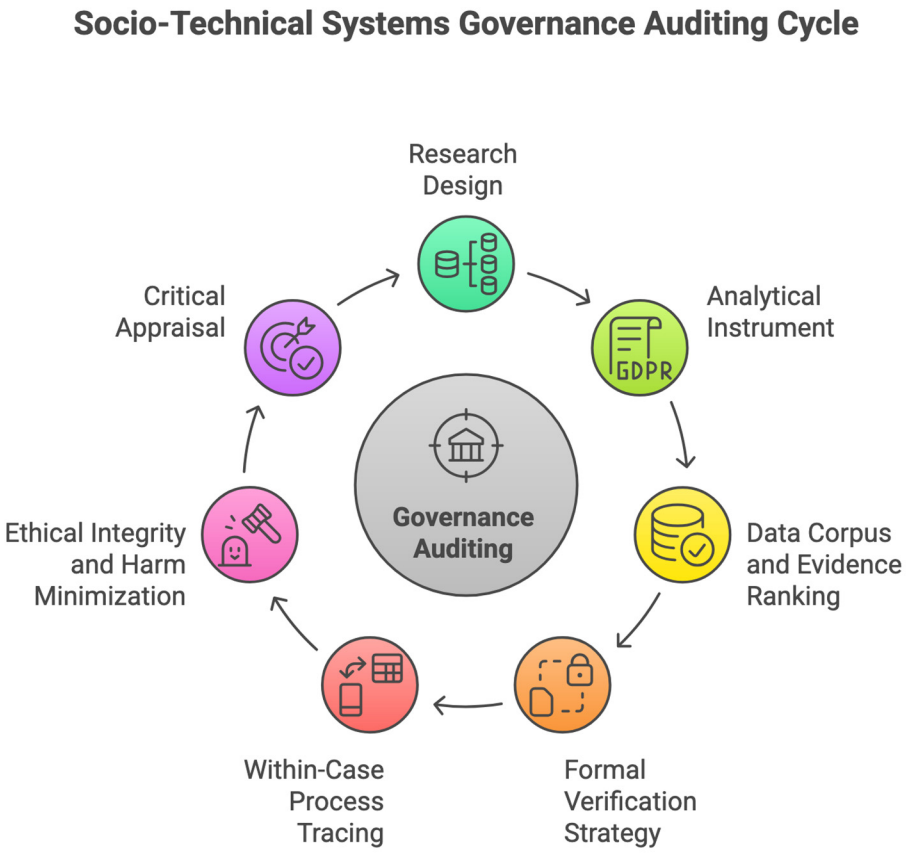

Figure 4. Socio-technical systems governance auditing cycle.

We treat each case as a theoretically informative "experiment" to refine and stress-test the 3R causal logic through case-based theory building. We employ Replication Logic, treating each case as an independent experiment. If two or more cases support the same theory (e.g., that the absence of an actuation gate leads to a rights violation),

literal replication is claimed. If cases produce contrasting results but for predictable reasons (e.g., a system with a strong "Governance for" layer avoids the harms seen in a system without one), theoretical replication is achieved [86].

The case inclusion criteria are specified ex ante to maximize comparability and avoid post hoc selection: (i) the system is deployed by (or in direct support of) public authority; (ii) an AI-enabled inference component materially influences a decision; (iii) the decision produces an administrative or kinetic actuation with plausible rights impacts; and (iv) sufficient documentary trace evidence exists to reconstruct the sensing–inference–actuation chain [85]. Because AI-enabled CPSs are context-dependent and socio-technical, the objective is not "universal applicability" in a statistical sense, but rather theoretical portability, which identifies recurring mechanisms and their boundary conditions across domains (policing, borders, and health) [87].

5.2. Analytical Instrument: The CPS-Governance Mapping Protocol

The core of this methodology is the CPS-Governance Mapping Protocol, a standardized analytical instrument designed to discretize and make auditable the complex interactions between software systems and legal-normative requirements. The protocol functions as the operational interface between engineering and jurisprudence: it translates abstract legal obligations into concrete system constraints, control points, and evidentiary artifacts that can be examined, tested, and contested. At a high level, the protocol enforces a traceability discipline across three domains, law, system design, and control behavior, so that governance claims can be evaluated against observable system properties rather than policy assertions. Formally, the protocol induces a mapping,

$$\mathcal{T}: \boldsymbol{L} \rightarrow \boldsymbol{C} \rightarrow \boldsymbol{U}$$

where $\boldsymbol{L}$ denotes legally binding obligations (e.g., rights, duties, and safeguards), $\boldsymbol{C}$ denotes implementable technical constraints and non-functional requirements, and $\boldsymbol{U}$ denotes prohibited or conditioned control actions. A governance failure is identified whenever a legally relevant obligation admits no corresponding technical constraint or control mechanism under this mapping. The methodological test is therefore binary at key points: either (i) an obligation has a traceable control, evidence, and enforcement hook, or (ii) it does not, and the absence is itself a governance finding.

Technical State-Space Mapping. The first step in the protocol is a control-oriented decomposition of the CPS into functional blocks: Sensing (measurement and data capture), Inference (stochastic estimation or classification), and Actuation (kinetic or administrative output). This decomposition follows the established practice in feedback-systems analysis and control theory [88]. The objective is not merely an architectural description, but the identification of where authority is exercised within the loop. For example, in an automated screening or biometric verification system, the control law is instantiated as an operational decision rule, such as a similarity-score threshold or a risk-score cutoff, that governs when inference outputs are translated into interventions. The protocol explicitly records where this rule is calibrated, how it is monitored over time, and under what conditions it can be overridden or suspended. This enables a later assessment of whether probabilistic estimates are functioning as advisory signals or as de facto commands.

Normative Constraint Identification. The second step translates legal and rights-based requirements into operationalizable compliance constraints and non-functional requirements (NFRs). Rather than relying on ad hoc ethical mapping, the protocol draws on established methods from legal-compliance requirements engineering to derive traceable obligations from statutes, regulations, case law, and binding policy instruments [89–91]. Concretely, legal principles are decomposed into implementable technical controls. For

instance, the principle of purpose limitation is operationalized through enforceable data-flow and access constraints, such as purpose-bound access control, the separation of processing contexts, logging requirements, and retention limits, rather than the narrower, often misleading claim that it is satisfied by simple data siloing. Each derived constraint is explicitly linked back to its legal source, preserving bidirectional traceability.

Hazard Identification via STPA. The third step applies Systems-Theoretic Process Analysis (STPA) to identify rights-relevant hazards and Unsafe Control Actions (UCAs). STPA models safety as a hierarchical control problem and treats hazards as violations of control constraints arising from inadequate or ineffective control, rather than as isolated component failures [8,92]. Within the CPS-Governance Mapping Protocol, STPA is used to ask governance-critical questions, such as whether a control action was issued without legal authorization. Was a required safeguard, such as human review, notice, or contestability, not provided? Was actuation triggered under conditions of excessive uncertainty or degraded evidence? By treating rights-relevant failures as system hazards rather than software bugs, the protocol captures interaction failures and governance gaps that conventional model audits and performance evaluations systematically overlook.

Protocol Outputs and Replicability. The outputs of the CPS-Governance Mapping Protocol are standardized to support analytical rigor and cross-case replicability. Each application of the protocol produces the following: (i) a justified system-boundary statement; (ii) a sensing inference actuation control-structure diagram; (iii) a legal-obligation-to-technical-constraint traceability matrix; (iv) a UCA and hazard register documenting rights-relevant failure modes; and (v) a constraint catalogue specifying the actuation gates and control conditions required for rights-consistent operation. Together, these artifacts provide an audit-ready evidentiary substrate for evaluating whether an AI-enabled CPS can legitimately exercise authority under conditions of uncertainty, coupling, and drift.

5.3. Data Corpus and Evidence Ranking (Auditability Constraint)

A critical challenge in AI governance research is information asymmetry, in which private vendors and agencies restrict access to system logic, sometimes invoking trade secrecy and opacity. This is not a methodological inconvenience but a substantive governance variable: secrecy and opacity shape accountability, contestability, and audit feasibility [84,93,94]. To ensure evidence-based integrity, this methodology employs a scholarly, defensible evidence hierarchy and a triangulation strategy grounded in document analysis and case-study best practices [95]:

- Tier 1. Primary governance records (used as case evidence, not as "scholarly citations")
  We prioritize judicial decisions, statutory instruments, official audit records, and regulator findings as primary data and analyze them using systematic document analysis procedures [95]. Where the case involves live facial recognition governance, we supplement primary legal materials with peer-reviewed scholarly analyses of the Bridges litigation and UK LFR governance [96,97].
- Tier 2. Peer-reviewed technical and socio-legal benchmarks
  We use peer-reviewed studies as benchmarks for performance and harm to avoid reliance on vendor claims. For face recognition disparities, we rely on independent, peer-reviewed evaluations that demonstrate demographic differences and threshold effects [44,98–100]. For predictive policing feedback loops and endogenous data generation, we rely on peer-reviewed work demonstrating runaway reinforcement mechanisms and their civil rights implications [24,47,101].
- Tier 3. Non-peer-reviewed corroboration (strictly scoped)
  We do not treat journalism/NGO reports as evidentiary anchors. Where used, they serve only to identify timelines, surface contested claims, and locate missing artifacts,

each of which is then treated as a governance risk indicator requiring corroboration in Tier 1–2 materials. These sources could also include investigative reports used solely to establish timelines and identify evidence gaps, which we then analyze as Governance Risks [95,102].

5.4. The Formal Verification Strategy: Assurance Cases and GSN

This study does not characterize Goal-Structuring Notation (GSN) as "formal verification"; rather, it is structured assurance argumentation that links claims to evidence and explicit assumptions. GSN is a graphical modelling language used in safety-critical industries (e.g., nuclear and aerospace) to provide a clear, auditable argument that a system meets its safety goals [20,21,80]. The purpose of using GSN is to move from assurance-by-assertion ("We promise our AI is fair") to assurance-by-evidence (an explicit argument showing which evidence supports which rights claims, under which operating assumptions). In our case analysis, we construct a counterfactual assurance structure for each system: "What specific evidence would have been required to support the claim that the system was rights-consistent under the stated operating conditions?" This makes the governance deficit explicit as an evidentiary delta between what was claimed and what was independently auditable. The resulting templates are designed to be usable as audit blueprints under risk-based regulatory regimes by specifying evidence obligations, monitoring triggers, and review hooks [66,84].

5.5. Within-Case Process Tracing (CPT)

To understand how a governance design leads to a specific form of harm, we use process tracing to reconstruct the causal chain within the CPS loop and to test the mechanism evidence in each case. Rather than a "minute-by-minute" narration, the unit of analysis is the sequenced mechanism: the decision points, thresholds, handoffs, and override conditions linking sensing $\rightarrow$ inference $\rightarrow$ actuation [103]. We identify "pinch points" where discretion is constrained by the interface design, workload, or deference to decision support. This is grounded in the automation bias literature, which documents the systematic commission and omission errors when decision aids are relied upon, particularly under time pressure and multitasking [25–27]. The outputs are actuation-level recommendations specifying where enforceable actuation gates (pause-and-review, override authority, escalation protocols, or mandatory second-person checks) should be inserted to restore meaningful control and contestability. In terms of coding and reliability discipline, we use a pre-specified codebook for (i) system functions, (ii) legal/rights constraints, (iii) UCAs/hazards, and (iv) evidence types; we maintain analytic memos and an audit trail; and we double-code a subset of materials with adjudication rules to reduce interpretive drift [104].

5.6. Ethical Integrity and Harm Minimization

This methodology operates under the "Security-by-Design" and "Rights-by-Design" ethos. Given the sensitive nature of surveillance and the exercise of coercive power, we apply Ethical Scoping to our reporting. We do not expose operational vulnerabilities that could be exploited by malicious actors to circumvent security controls. Instead, we concentrate on institutional accountability structures. This aligns with the UN Guiding Principles on Business and Human Rights (UNGP), ensuring that our research supports the "Protect, Respect, and Remedy" framework by providing the technical–legal roadmap for institutional oversight [105].

5.7. Critical Appraisal of the Methodology: Strengths, Weaknesses, and Counter-Arguments

The first methodological strength is technical–normative commensurability: STPA enables rights-relevant harms to be represented as constraint violations within a control

structure, thereby allowing the ex ante identification of hazards and UCAs rather than relying solely on post hoc litigation. A second strength is evidentiary discipline: GSN forces explicit claim–evidence–assumption links, making unverifiable assertions and missing artifacts visible [20,21,80]. The third strength is auditability as an analytic variable: by treating opacity and missing artifacts as governance findings rather than inconveniences, the method avoids "confidence by absence of evidence" and makes audit feasibility part of the causal story [93–95].

We identify three limitations and mitigation connections: (i) Information asymmetry can prevent the diagnosis of internal causes even when governance gaps are detectable [93,94]. We mitigate this by using triangulation and by explicitly documenting evidence deficits in the results, rather than filling gaps with conjecture [95]. (ii) Boundary sensitivity is a limitation, as findings depend on disciplined system-boundary justification in socio-technical systems engineering [87]. We mitigate it by requiring an explicit boundary statement and by stress-testing alternative boundary choices as a sensitivity check (e.g., does the mechanism still hold if the boundary shifts from "model" to "workflow and vendor update pipeline"?) [87,103]. (iii) Non-random case selection can bias conclusions. We mitigate this by using replication logic, explicit inclusion criteria, and transparent negative-case reasoning ("what would falsify the mechanism in this domain?") [85,86,103].

6. Case Portfolio: Auditing Coercive-Power CPS Through the 3R Lens

6.1. Case 1. Live Facial Recognition (LFR) in the UK—The Sensing Scope Failure

Live Facial Recognition (LFR) policing in the United Kingdom provides a primary evidentiary record of how discretionary migration occurs within a CPS loop. In R (Bridges) v Chief Constable of South Wales Police, the Court of Appeal examined the lawfulness of a pilot program that used biometric sensors to scan public spaces for individuals on a watchlist of persons of interest [10,106]. Crucially, the Court's reasoning makes the governance mechanism legible. The legality problem was not that the algorithm was inaccurate, but that key public-law constraints were not sufficiently determinate at the sensing boundary; i.e., the conditions under which people could be scanned, and the criteria by which individuals could be placed on watchlists were not adequately bounded to meet rule-of-law demands of foreseeability and constraint on discretion [10,106].

- Technical Audit of the Loop
  The sensing layer captured real-time biometric data from CCTV feeds in public spaces; the inference layer generated probabilistic match outputs; and the actuation layer triggered officer attention/engagement based on match alerts [107]. The 3R-relevant point is that "control" was exercised at two upstream points that typical model audits ignore: (i) siting (where/when the cameras and matching were activated) and (ii) watchlisting (who was eligible to be flagged). These are not peripheral parameters; they are the governance-relevant control-law of the system [10,97,108].
- The Governance Pinch Point
  The core deficiency was not simply the model performance but a rules/constraints failure at the sensing and watchlisting boundary: peer-reviewed legal analyses of Bridges emphasize that the deployment lacked sufficiently clear and bounded criteria governing where LFR could be used and who could be placed on the watchlist, creating an excessive discretion problem incompatible with the rule-of-law requirements of legality, foreseeability, and non-arbitrariness. In 3R terms, this is a Sensing-Scope (Rules) failure: absent hard constraints on sensing scope and watchlist construction, "risk detection" defaults into the open-ended biometric surveillance capacity [10,97,108]. The system performed a high-impact control action (population-level biometric sens-

ing) without a sufficiently specified constraint set governing when that action is permitted [8,10].

- Socio-Technical Consequences
  The system's institutional viability depended on enforceable constraints at the sensing boundary; absent those constraints, the deployment was found to be unlawful and could not be stabilized through accuracy improvements alone [10]. This provides a sharp inference for your framework: when the legality of the sensing scope is underdetermined, "better accuracy" is non-responsive to the legal failure mode because the harm mechanism is discretion, not error. This is why the minimum-viable control is not only a performance dossier but a scope-control instrument (siting and watchlisting rules) that is auditable and enforceable [10,70]. This demonstrates that public trust and legitimacy are central to police LFR acceptability and governance, independently of technical claims [107].

6.2. Case 2. Border Facial Biometrics (US)—The Actuation Gate Failure

The U.S. deployment of facial biometrics at ports of entry illustrates how border automation and biometric verification systems institutionalize high-throughput identity decisions within asymmetric power settings, where choice and contestability are structurally constrained. Rather than treating "opt-out" as a binary, we treat the key governance variable as whether meaningful contestability and human review are feasible at the point of actuation, given the operational tempo [109,110]. GAO assessments document recurring privacy-notice and opt-out information issues in operational environments; i.e., the governance substrate at the moment of actuation is often thinner than the "policy" suggests [111,112].

- Technical Audit of the Loop
  The system performs a facial comparison for identity verification using gallery-based matching and/or verification workflows (e.g., comparing a live capture to pre-staged image galleries associated with travel documents/manifests), then acts on the outcome (clearance vs. diversion to secondary screening) [111,112]. The "clearance vs. diversion" routing is the rights-relevant actuation gate: it operationalizes a probabilistic inference as a constraint on movement and time, and it is where lawful-basis, proportionality, and reviewability must bind in practice [111,112].
- The Governance Pinch Point
  The principal risk is actuation-gate erosion: under throughput and workload pressure, "algorithm-in-the-loop" designs can induce automation bias, leading operators to defer to probabilistic outputs as de facto commands rather than as uncertain signals, unless systems are designed for meaningful human control (time, authority, override, and accountable procedure) [25,113]. Accordingly, the 3R audit treats the actuation gate as a control constraint that must specify (and log) the following: (i) when secondary screening is legally warranted; (ii) what evidence must be recorded to enable an after-the-fact review (including confidence/uncertainty and any escalation rationale); and (iii) what review/override path is available under the real tempo (who can pause; what triggers a mandatory second-person review; and what the recourse channel is) [84,111–113]. This reframes "HITL" from a slogan into an auditable control property: oversight must be demonstrable in logs and workflow traces, not assumed [67,84].
- Socio-Technological Consequences
  Even where systems are operationally effective, biometric error differentials documented in the peer-reviewed biometrics literature imply distributional procedural burdens: higher false non-match or false match rates for some demographic groups can translate into disproportionate delays, repeated screening, or escalations [13].

We term this accumulation of unequal procedural burden "rights debt": a measurable governance externality in which the system's throughput optimization is partly achieved by offloading friction onto identifiable groups over time. A rights-consistent actuation gate must include disparity-sensitive monitoring and a corrective control loop (threshold governance plus review triggers plus remedial pathways); otherwise, inequality becomes an operational invariant of the system [13,70].

6.3. Case 3. Chicago Strategic Subject List (SSL)—The Feedback Loop Failure

The Chicago Police Department's Strategic Subject List (SSL), also known as the heat list, illustrates a distinct failure mode: endogenous feedback in person-based predictive policing, where deployment practices shape the data that later justify further deployment [23,46]. The system aimed to estimate an individual's likelihood of involvement in gun violence as a victim or offender [114]. The SSL documentation is valuable as Tier-1 evidence because it specifies the system's operational aim (shooting-involvement risk) and links the scoring pipeline to a defined data period used to generate risk scores [114].

- Technical Audit of the Loop
  The inference engine leveraged administrative/policing data and social-network features; the actuation layer included targeted interventions (e.g., custom notifications and other forms of intensified attention) [114]. This is a structurally feedback-prone control loop. Actuation (increased attention/contact) changes what is discovered and recorded, which can then be re-ingested as evidence of future risk [23,46].
- The Governance Pinch Point
  Scholarly analyses of the SSL emphasize limited transparency, weak evidence of effectiveness, and accountability deficits in the generation and use of scores [115]. Critically, the system architecture is vulnerable to endogenous bias: increased patrol attention or contact can increase the number of detected incidents and recorded police interactions, which the inference layer may then treat as confirmation of elevated risk, producing a self-reinforcing allocation dynamic [23,46,115]. This is not just bias; it is a control failure in which the feedback signal is contaminated by the controller's own interventions.
- Socio-Technological Consequences
  This dynamic is consistent with formal and empirical demonstrations of runaway feedback loops in predictive policing, in which the model repeatedly allocates attention to the same communities when trained on "discovered" policing data [23,46]. In 3R terms, this is a Rules-and-Rights failure at the feedback boundary. The system lacks constraints that distinguish "risk" from "policing intensity," thereby undermining the trust and legitimacy required for effective public safety governance [24,47]. The minimum viable control is therefore not only transparency, it is a feedback-discrimination protocol: (i) the explicit separation of enforcement-intensity features from risk targets, (ii) monitoring that detects endogeneity, and (iii) governance rules that prevent self-generated data from being treated as independent evidence [23,24,46].

6.4. Comparative Analysis: Synthesis of Socio-Technical Failure Mechanisms

Table 1 synthesizes the cross-case mechanisms: the dominant failure modes arise at distinct control points, scope (LFR), actuation (border), and feedback (SSL), each producing a different pathway from inference to rights injury and institutional instability.

Table 1. Cross-case synthesis of socio-technical failure mechanisms (3R lens).

Feature	Case 1. UK LFR	Case 2. U.S. Border	Case 3: Chicago SSL
Primary failure mode	Sensing-scope (rules) failure	Actuations-gate (control) erosion	Feedback-loop endogeneity
Governance deficit	Unbounded watchlisting/siting discretion	Deference under throughput; weak contestability at actuation	Endogenous data generation; weak effectiveness evidence
Normative injury	Privacy/legality; foreseeability	Procedural fairness/equity (distributional burdens)	Liberty/fairness via intensified attention
Resilience outcome	Legally unstable (judicially invalidated)	Operationally sustained with distributive frictions ("rights debt")	Institutionally unstable (legitimacy/effectiveness contested)

- Scope–Speed–Feedback Trade-Offs (Mechanism Clarification)
  Across cases, the recurring mechanism is not bad AI, but missing constraints at distinct control points: (i) scope constraints on sensing and watchlisting (Case 1), (ii) procedural gates that preserve contestability under throughput (Case 2), and (iii) feedback controls that prevent endogenous data from being misread as objective risk (Case 3) [96,109]. The comparative implication is actionable: each failure mode maps to a different minimum viable control class, scope-control instruments, actuation-gate controls, and feedback-discrimination controls, each with distinct evidence obligations in the assurance case [70,80].
- Socio-Technological Consequences: The Chilling Effect and Trust Degradation
  Where a coercive-power CPS becomes opaque and difficult to contest, the empirical literature on the surveillance-induced chilling effects documents the deterrence of lawful information seeking and democratic discourse [116,117]. The 3R connection is not abstract. The chilling effects are a population-level feedback mechanism—reduced engagement, reduced reporting, and reduced civic participation—that degrades institutional legitimacy and weakens the quality of the very data streams governance systems rely upon. This is why trust degradation is a systems risk, not an external social concern [75,76,118,119].
- Cross-Case Inference
  Across the portfolio, the recurring failures are control-structure failures, missing constraints, weak feedback discrimination, and actuation without procedural gates, rather than isolated model defects. These minimum-viable controls should be translated into procurement-ready and oversight-ready assurance artifacts: (i) scope-control instruments (siting and watchlisting rules); (ii) threshold/performance dossiers (including disparity tests and drift monitoring); (iii) logging and intervention audit schemas; (iv) feedback-monitoring protocols distinguishing risk from enforcement intensity; and (v) enforceable governance-for mechanisms (audit access, sanctions, and remedy/contestability) [25,68,70,80].

7. Operationalizing Rights as Resilience Constraints: From Normative Claims to Engineering Invariants

The transition from a theoretical framework to an operational system represents the core implementation gap in AI governance. To make the 3R framework technically actionable, Rights are operationalized as resilience constraints: explicit, testable boundaries on what the system may do across sensing–inference–actuation transitions [31,120]. In resilience engineering, constraints are the mechanism by which systems prevent hazardous

trajectories under surprise, drift, and coupling. They are the conditions that keep the system from entering brittle regimes [120]. This highlights coupled socio-technical systems. When the speed and coupling increase, small interpretive errors can cascade unless explicit circuit breakers and supervisory controls are in place [121]. By framing rights in this manner, rights protection becomes a precondition for a sustained legitimate operation rather than an optional trade-off against performance [79,120].

In the UK LFR example, the critical constraint is the sensing scope (where and when biometric sensing is permitted, and who is watchlist-eligible). Without it, the system is legally unstable regardless of accuracy [10]. In U.S. border biometrics, the critical constraint is the actuation gate (what justifies diversion/escalation, what is logged, and what review path exists under the tempo) [111,112]. In predictive policing systems, the critical constraint is feedback discrimination (preventing the system from treating enforcement intensity as evidence of risk) [23,24,46].

7.1. The Engineering of Rights: Requirements Decomposition

Operationalization begins with requirements engineering (RE), specifically the decomposition of high-level governance goals into operational requirements and verifiable assertions [122]. Within the governance of the AI layer, rights are treated as non-functional system requirements that constrain admissible deployment behavior rather than aspirational design principles. In this section, we focus on the right to non-discrimination as a paradigmatic case, because it admits concrete technical operationalization while remaining normatively central to public-authority CPSs. Concretely, non-discrimination is operationalized as a deploy-time admissibility condition for inference-driven actuation. Let $g \in G$ denote protected groups, and let FP_g and FN_g denote group-conditional false-positive and false-negative rates, respectively. Following established fairness formulations, we monitor the disparities across groups using group-conditional error metrics drawn from the equalized odds and equal opportunity families [123]. For a given deployment context, we define a disparity measure Δ as follows:

$$\Delta = max\{g,g' \in G\} \; |FPg - FP\{g'\}| \; or \; |FNg - FN\{g'\}|.$$

This monitoring is not treated as a one-time certification exercise but as a continuous internal auditing obligation. Continuous auditing pipelines are implemented during training and post-deployment operation, generating auditable artifacts such as time-stamped performance reports, drift indicators, and subgroup evaluations rather than static fairness snapshots [66]. These artifacts constitute part of the evidentiary basis for continued high-consequence actuation. If the monitored disparity Δ exceeds a predefined tolerance threshold τ, the system must trigger a safety-and-rights circuit-breaker:

$\Delta > \tau \Rightarrow$ high-consequence actuation is paused or throttled pending review and remediation. This mechanism ensures that emerging or amplified inequities are treated as control failures rather than latent technical imperfections. The system is permitted to continue inference and monitoring, but it is not permitted to translate inference into coercive or high-friction action when the admissibility condition is violated.

This design constitutes a control-theoretic translation of the idea of "rights as constraints". Rather than optimizing solely for predictive performance, the CPS enforces a hard boundary between inference and actuation when its own internal evidence indicates that actuation would be systematically illegitimate. In this sense, rights protection functions as a dynamic stability constraint: it prevents the system from entering a regime in which a biased or degraded performance is operationally amplified through actuation [66,124–128].

7.2. Actuation Gates as Dynamic Stability Mechanisms

Under the governance of an AI layer, the actuation gate is the primary mechanism for socio-technical resilience. High-velocity, tightly coupled socio-technical systems are structurally prone to cascading failure when small interpretive errors meet tight coupling and time pressure, a core insight of the Normal Accident Theory [121]. In such settings, probabilistic inference outputs can be prematurely converted into coercive or high-friction actions, amplifying the consequences of uncertainty. The operationalization of the Risk–Rights–Rules (3R) framework, therefore, requires context-aware actuation gates that explicitly mediate between stochastic inference and real-world intervention.

Rather than implementing a binary "allow/deny" logic, a rights-aware actuation gate evaluates uncertainty and shift indicators such as predictive uncertainty, concept drift alarms, and out-of-distribution (OOD) detection signals and selects an actuation mode accordingly [125–130]. Formally, the gate functions as a supervisory controller that constrains high-consequence actions when epistemic conditions are degraded or when rights-relevant constraints cannot be verified at runtime. This design reframes actuation as a graded-control problem rather than a deterministic threshold-crossing problem. When uncertainty exceeds a predefined tolerance or when drift or OOD conditions are detected, the 3R constraint mandates a circuit breaker effect. Specifically, the system must (i) generate a contestability-enabling explanation at the point of decision and (ii) open a mandatory human override window with genuine authority to pause, override, or escalate the decision. This requirement directly addresses automation bias by preventing probabilistic outputs from functioning as de facto commands under time pressure. Counterfactual explanations are particularly well-suited to this role because they articulate the minimal changes required for an alternative outcome without exposing the internal logic or parameters of the underlying model [82,129].

Algorithmically, this mechanism can be expressed as a Rights-Aware Actuation Gate (RAAG), which mediates between inference and intervention by enforcing rights constraints at runtime (Algorithm 1):

Algorithm 1: Rights-Aware Actuation Gate (RAAG)

```
Input: inference output ŷ_t, uncertainty estimate σt, drift/OOD flag dt, rights
constraints Cr
 Output: actuation decision u_t
If dt = TRUE or σt > σmax:
        block high-consequence actuation
        generate counterfactual explanation
        trigger mandatory human review with override authority
        log rights-check failure
 Else if violates Cr(xt):
        block actuation
        generate contestability packet
        log constraint violation
 Else:
        allow bounded actuation
        log justification and evidence
```

This operationalizes safe interruptibility in governance practice: the system must remain interruptible by design whenever the environment, evidence base, or legality conditions are uncertain. By embedding interruptibility, explanation, and human authority directly into the actuation pathway, the gate functions as a dynamic stability mechanism

that preserves rights-consistent operation under uncertainty and prevents the CPS from entering runaway or illegitimate control regimes.

7.3. The "Governance for AI" Layer: Enabling Innovation Through Enforceable Rules

A common counterargument is that rigorous operationalization stifles innovation. The stronger, evidence-based claim is conditional: innovation is more likely to be sustained when compliance boundary conditions are legible and when standards reduce transaction costs and uncertainty across complex ecosystems [131]. In other words, the innovation vs. regulation frame is incomplete because well-designed standards can lower adoption friction by clarifying what must be built, logged, and audited.

What formal procedures or standards exist? In practice, enforceability is anchored by (i) risk-management and governance baselines (e.g., lifecycle governance in NIST AI RMF) [17], and (ii) management-system and risk-management standards that define auditable organizational controls, most notably ISO/IEC 42001 (AI management systems) and ISO/IEC 23894 (AI risk management guidance) [132,133]. These standards matter because they specify organizational control expectations (policy, monitoring, incident handling, and improvement cycles) that can be bound to technical evidence artifacts (logs, evaluations, and change control).

Who provides oversight, and at what stages? Where binding law exists, oversight typically attaches to pre-deployment conformity or risk processes and post-deployment monitoring duties. The EU AI Act highlights this lifecycle logic. High-risk systems require risk management, documentation, and monitoring. The point is enforceability via auditable evidence rather than ethics statements [18,70].

How does governance differ on Big Tech infrastructure vs. independent systems? Cloud-based AI often creates an algorithmic supply chain. Multiple actors (cloud provider, model provider, integrator, and deploying agency) contribute to the system, which complicates accountability unless contracts specify audit rights, evaluation access, update control, and evidence retention. The accountability literature shows that governance must explicitly address supply-chain roles and access to evidence, or responsibility diffuses across multiple actors [134]. Procurement must contract for logging, monitoring access, change control, and exit rights; otherwise, the assurance case cannot be maintained [66,134].

If no comprehensive regulator exists, how is the gap addressed? In fragmented governance environments, the practical substitute is evidence discipline and enforceable procurement: the system is engineered so that oversight can be executed through audit artifacts (assurance cases, and traceability records) even when institutional authority is distributed [66,80,134]. An audit-ready architecture is the governance mechanism that remains viable under institutional fragmentation.

7.4. Socio-Technological Consequences of Failed Operationalization

If we fail to operationalize rights as constraints, the potential consequences are not merely legal fines but Socio-Technical Collapse. We identify three Critical Failure States that must be avoided:

- The Legitimacy Crisis and Adversarial Adaptation
  When systems are opaque, unreviewable, or self-reinforcing, the affected populations and organizations rationally adapt by withholding cooperation, disputing outputs, or gaming signals, reducing data quality and degrading system performance. In 3R terms, the loss of legitimacy becomes a negative feedback shock that undermines the evidence base on which the system relies [75,76].
- The "Liability Sponge" Backlash and Institutional Deskilling
  When humans are held responsible even though they are operationally unable to

control automated processes, responsibility collapses onto frontline operators [28]. The Post Office Horizon IT Inquiry documents how the presumption of system reliability, coupled with weak contestability and governance failures, produced catastrophic accountability outcomes and prolonged institutional denial [135]. If the evidence cannot be audited and challenged, the system becomes a black box tribunal, and legitimacy fails.

- Kinetic Escalation or High-Friction Escalation
  When due-process safeguards are absent at actuation, errors translate into coercive outcomes (detention, denial, and intensified scrutiny), which then trigger predictable second-order effects—injunctions, litigation, or program termination—forcing systems offline. This is exactly what the UK LFR case demonstrates: scope constraints were missing, so the system became legally unstable and could not be "patched" by accuracy claims [10].

7.5. Standards, Oversight, and Enforcement Gaps

Formal "ethics-by-design" practice in industry is increasingly implemented through auditable governance procedures rather than aspirational principles: organizations adopt management-system and risk-management standards (e.g., ISO/IEC 42001 for an AI management system and ISO/IEC 23894 for AI risk management) that institutionalize policy, roles, monitoring, incident handling, and continuous improvement, often paired with lifecycle risk frameworks such as the NIST AI RMF and evidence artifacts (e.g., documentation, evaluation reports, monitoring logs, and change-control records) to demonstrate that the system remains within approved operating bounds [17,38,66,132,133]; legally, oversight attaches at defined stages: ex ante impact assessment and design constraints (e.g., General Data Protection Regulation (GDPR) requires a Data Protection Impact Assessment before high-risk processing, including the large-scale monitoring of public areas), pre-market and pre-deployment obligations for high-risk systems (technical documentation, risk controls, and, where applicable, conformity assessment), and post-deployment monitoring, incident reporting, and corrective action [18,70,136,137].

Enforcement bodies are therefore stage-specific (e.g., data protection authorities under the GDPR; and AI Act enforcement through national competent/market-surveillance structures, with EU-level coordination mechanisms), and the practical governance question is whether an agency can produce audit-ready evidence at the moment of actuation and throughout the process of change over time [18,70,136,138]. Governance also differs materially when systems are built on Big Tech/cloud infrastructure versus when they are built independently. Cloud-based AI typically creates an algorithmic supply chain (cloud provider, model provider, integrator, and deploying institution) in which accountability depends on contractual and technical control points (audit rights, evaluation access, logging/telemetry, update/change-control, data-processing terms, and exit/portability), whereas independent development collapses some dependencies but does not remove legal exposure: providers and deployers remain subject to applicable cross-border regimes when placing systems on regulated markets or processing personal data (e.g., the AI Act's scope reaches providers placing systems in the EU regardless of location; GDPR obligations attach to the processing itself) [18,134,136].

Where no single comprehensive regulator exists, the gap is typically bridged (imperfectly) by a patchwork of sector regulators, procurement-based governance, standards certification, and assurance-case evidence discipline (so that oversight can be executed through inspectable artifacts even under a fragmented authority). Establishing a new authority, in practice, requires either legislation that grants investigative/audit powers and sanctions, while coordinating with product-safety and data-protection enforcement, or

multilateral institutionalization (treaty- or convention-based duties plus cross-border supervision mechanisms) so that rights-as-constraints become enforceable design conditions rather than voluntary claims [62,70,71,80,105].

7.6. Realizing the 3R Framework: The "Audit-Ready" Architecture

The final step in operationalization is the shift to an audit-ready architecture, meaning that the evidence required for assurance arguments is generated continuously by the system rather than reconstructed after harm [80.66]. The minimum audit-ready event record specifies what each high-friction actuation must generate. Each actuation event is linked to a traceability artifact containing: (i) the model and version identifier, (ii) the relevant monitoring status (e.g., drift and shift flags), (iii) the rule and constraint that authorized the intervention, and (iv) the human action record (accept and override, plus justification where required). This is the minimal technical substrate for accountability under distributed responsibility and ML system entanglement [38,134]. The theoretical payoff is precise: innovation is compatible with high-performing, complex models, provided that the control structure enforces rights-relevant constraints and keeps the system interruptible, auditable, and contestable under drift and uncertainty. In this sense, governance becomes part of the control law of the digital state, an engineered set of constraints and feedback mechanisms that preserve legality and legitimacy under automation.

8. Discussion: Designing Legitimacy into AI-Enabled CPS

The empirical record from the UK Bridges litigation, U.S. border biometrics deployments, and Chicago's SSL supports a single systemic diagnosis: a legitimacy gap—i.e., operational capability (continuous sensing, probabilistic inference, and high-tempo actuation) expands faster than the institutional mechanisms required for legality, reviewability, and contestability [10,23,46,111,112,139–142]. The cases show that legitimacy fails at different control points, scope (LFR), actuation (border routing), and feedback (SSL). Coercive authority is exercised through a loop whose justificatory practices are thinner than the consequences it produces. In the public sector, this is a rule-of-law problem because automated systems can displace or attenuate reason-giving, procedural regularity, and accountable review, the practices that make coercive decisions publicly justifiable [11,134,143]. The 3R contribution is therefore to treat legitimacy as a design requirement—to engineer legality, contestability, and evidence-traceability into the sensing–inference–actuation loop as enforceable constraints and audit-ready hooks, not as post hoc policy rhetoric [70,80,144].

8.1. Bridging the Stochastic–Deterministic Gap: Probabilistic Authority

The core technical governance tension is that AI-enabled CPSs produce probabilistic inferences, while public authority requires publicly justifiable grounds for coercive or high-friction action, grounds that can be explained, contested, and reviewed [11,143,145]. Designing legitimacy, therefore, requires calibrating what we call probabilistic authority. Probabilistic scores must be treated as uncertain evidence inputs that trigger bounded procedures, not as self-executing legal conclusions [141,143]. Case 2 makes the mechanism concrete: in border biometrics, a similarity score is a likelihood estimate, not a fact. Legitimacy improves when low-confidence matches trigger verification protocols (additional checks, mandatory review, and contestability packet) rather than default escalation [111–113]. This is a control necessity under uncertainty because high-tempo actuation is where automation bias converts probabilistic outputs into de facto commands [25,113,144]. When the system treats correlation as authority (from score to escalation), it risks statistical injustice: adverse outcomes imposed because a person resembles a data pattern rather than because legally relevant conditions have been established through reasons and reviewable procedures. As

a critique of actuarial governance in criminal-justice contexts, prediction can reshape the epistemic basis of punishment and policing, even when it appears efficient [145].

8.2. Countering the "Efficiency Trap": Legitimacy as a Stabilizing Feedback

A common narrative of deployment frames rights constraints as an efficiency tax. The cross-case evidence supports the opposite pattern: short-run throughput optimization without constraints produces long-run instability through contestation, legal invalidation, and trust collapse [10,139,141]. Case 1 demonstrates the strongest form of this inversion. LFR could not be stabilized by improvements in accuracy because the failure was due to a rules deficit at the sensing boundary. The Court invalidated the deployment on the grounds of inadequate constraints on discretion and foreseeability [10]. Case 3 demonstrates the operational analogue. Predictive policing systems can generate apparently self-confirming "success" by reallocating enforcement and thereby manufacturing the data that later justifies further enforcement; this can concentrate intervention even if underlying rates do not justify it, undermining legitimacy and institutional viability [23,24,46,47].

From a CPS perspective, legitimacy behaves like a stabilizing feedback variable: constraints on the sensing scope, actuation gates, and feedback discrimination prevent the system from entering a runaway regime of biased amplification. Because the governed population is part of the plant, the loss of legitimacy can feed back into reduced cooperation, degradation of data quality, and adversarial adaptation, thereby directly harming system performance [75,76,139].

8.3. Cross-Analysis of Failure Modes: The Function Creep and Discretion Leak

The comparative analysis identifies two recurring governance failure mechanisms that should be addressed at design time:

- Function Creep (Sensing Layer)
  Function creep, the expansion of a system beyond its originally justified purpose without transparent authorization, is a mature conceptual and legal phenomenon [140]. In Case 1, the failure mechanism is function-creep-ready by construction. If siting and watchlisting rules are under-specified, the same sensing infrastructure can be extended across contexts (new locations, new watchlists, and broader purposes) without the procedural burdens that normally constrain public power. The design implication is that creep can be made auditably visible and procedurally contestable through scope-control instruments: purpose-bound authorization, watchlist governance logs, and immutable records of when/where sensing was activated [10,66,140].
- Discretion Leak (Inference Layer)
  Discretion leak occurs when policy discretion is silently displaced into model updates, threshold adjustments, feature changes, or vendor-controlled tuning without public-law safeguards governing changes in the decision criteria [38,134,144]. Modern deployments often depend on multi-actor service models where upstream providers can update components that materially affect outcomes, while downstream public agencies have limited visibility (accountability horizon) [134]. Legitimacy therefore requires treating material model/threshold updates as governance events, documentation, reasons, evidence refresh, and oversight triggers, because a threshold shift can function as a de facto policy change even if no statute changed [70,80,144].

8.4. Designing Contestability into the Actuation Pathway

The most consequential shift is from explainability-as-insight to contestability-by-design. For affected persons, a heatmap or feature-importance plot is rarely a remedy. Contestability requires actionable grounds for challenge, review, and correction [11,146]. The actuation gate can be an insertion point where probabilistic inference becomes lived consequence. Accordingly, legitimacy is strengthened when the system generates a contestation artifact ("adverse action packet") at the moment of intervention: (i) the decision and its legal basis category, (ii) the minimal contestable conditions (often counterfactual), (iii) the model/version + monitoring status, and (iv) the review channel and deadlines [66,82,84]. Counterfactual explanations are useful because they support recourse without requiring full model disclosure, but they are not sufficient unless paired with a real review pathway and evidentiary access. The right to contest under GDPR Article 22 is best understood as a due-process-like remedy that implies concrete transparency and procedural mechanisms, not merely a narrative explanation [146].

8.5. Critical Reflection: The Future of "Rule-Enforced" Innovation

Designing legitimacy into CPSs is not an anti-innovation posture; it is the condition for durable deployment under scrutiny. Assurance cases and reviewability-oriented design increase the probability that systems remain deployable by making normative stability auditable rather than asserted [70,80]. If the legitimacy gap is not bridged, we risk a predictable socio-technical schism: technically capable CPSs that are socially rejected or legally invalidated [10,139]. The Post Office/Horizon scandal provides a cautionary primary analogue: institutional collapse followed from the presumption of computer reliability combined with weak contestability, poor audit discipline, and prolonged governance failure [135,147], exactly the conditions your 3R blueprint is designed to prevent (audit-ready artifacts; constraints; and enforceable review hooks). This risk is amplified in algorithmic supply chains where accountability can dissipate across vendors, integrators, and updates unless audit rights, evaluation access, and change control are contractually and technically enforced [70,134]. Therefore, legitimacy is not a soft ethical add-on; it is a core control objective for the next generation of cyber–physical governance infrastructure.

8.6. Risk Types, Sources, and Governance Implications in Socio-Technical CPS

Table 2 presents a formal risk taxonomy for high-consequence predictive and biometric AI used in public workflows, structured around the end-to-end cyber–physical governance loop (sensing → inference → actuation → feedback) and aligned with a 3R (Risk–Rights–Rules) perspective. It enumerates key risk types—from sensing scope and data quality failures to uncertainty mismanagement, demographic performance differentials, feedback reinforcement, and lifecycle/supply-chain accountability gaps—and links each to its primary source, the mechanisms by which it arises, and the concrete governance implications that must be enforceable and auditable in practice (e.g., actuation gates, demographic testing, drift monitoring, contestability pathways, and traceability logs). By pairing each risk with typical empirical anchors such as benchmark testing and documented case evidence, the taxonomy is intended to support implementation: it helps practitioners translate abstract governance requirements into specific controls, evidence artifacts, and inspection points that can be used by engineers, procurement teams, auditors, and legal reviewers.

Table 2. Risk taxonomy for high-consequence predictive and biometric AI in public workflows (3R lens).

Risk ID	Risk Type	Primary Source	Typical Sources/Mechanisms	Governance Implications	Empirical Grounding
R1	Sensing scope and authorization failure	Sensing boundary	Overbroad siting, watchlisting, or data capture without bounded purpose/criteria; open-ended surveillance capacity	Scope-control instruments (purpose limitation, siting and watchlisting rules), activation logs, authorization lineage; audit rights over scope changes	Bridges/LFR as scope-control failure (legality not patchable by accuracy)
R2	Data quality and representativeness risk	Sensing + lifecycle	Skewed/dirty administrative data; non-representative samples; measurement bias	Data provenance, sampling justification, dataset documentation, minimum quality gates; evidence of population validity	Documented in governance literature on data quality and administrative data limits
R3	Privacy/unlawful processing risk	Sensing + rules	Large-scale monitoring, sensitive data processing, weak DPIA/FRIA alignment	DPIA/FRIA (as applicable), lawful-basis mapping, retention and minimization controls; auditable compliance artifacts	EU AI Act lifecycle governance plus FRIA; DPIA practice in data protection regimes
R4	Uncertainty mismanagement (probabilistic authority)	Inference to actuation	Aleatoric/epistemic uncertainty treated as fact; scores become self-executing decisions	Uncertainty measurement; actuation gates linked to uncertainty thresholds (pause/review/verification); adverse action packet includes confidence and basis	Framed in the manuscript as a core cause of legal certainty erosion in stochastic loops
R5	Demographic performance differentials (rights exposure)	Inference to actuation	Differential false negative/false positive rates by group; threshold effects; unequal procedural burden (rights debt)	Mandatory demographic evaluation; threshold governance tied to disparity tests; remedial controls (review triggers, alternative verification, monitoring)	NIST FRVT as empirical anchor for demographic differentials in face recognition performance
R6	Model validity/specification mismatch	Inference	Model trained for one context used in another; proxy targets for legally relevant standards	Task/construct validity review; intended-use constraints; periodic re-validation and documentation	Observed across high-consequence ML deployments when constructs and legal standards diverge

Table 2. *Cont.*

Risk ID	Risk Type	Primary Source	Typical Sources/Mechanisms	Governance Implications	Empirical Grounding
R7	Drift/dataset shift/out-of-distribution degradation	Lifecycle (operations)	Performance decay as environment changes; previously valid evidence becomes stale	Post-deployment monitoring, drift triggers, evidence-refresh rules, controlled rollback; assurance case updated as a living artifact	Recognized in operational ML as a recurring failure mode; addressed via monitoring and change control
R8	Actuation-gate erosion (automation bias)	Actuation + human factors	Operators defer to alerts/scores under tempo; human-in-the-loop becomes rubber stamp	Meaningful human control evidenced in workflow traces: time-to-review, override authority, second checks, justification logging	Human–automation interaction research on automation bias; operational border/security workflows
R9	Due-process/contestability failure (black box tribunal)	Actuation + governance	No notice, no reasons, weak appeal/correction pathways; opaque evidence	Contestability-by-design: adverse action packets, reasons-giving, appeal channels, correction SLAs; auditable recourse records	Common governance failure in high-friction public decisions shaped by automated inferences
R10	Feedback endogeneity/runaway reinforcement	Feedback loop	Interventions change what gets observed; discovered data treated as objective truth; self-reinforcing patrol/attention	Feedback-discrimination protocol: separate enforcement intensity from risk; monitor endogeneity; constrain retraining inputs	Predictive policing literature; Chicago SSL used as an archetype of feedback-loop failure
R11	Auditability and traceability failure (evidence gaps)	Lifecycle + oversight	Missing logs/versioning; cannot reconstruct decisions; evidence not inspectable	Audit-ready event record: model/version ID, monitoring status, authorizing rule/constraint, human action record; retention and access	Audit practice and governance baselines emphasize traceability; discussed in the manuscript
R12	Update/change-control discretion leak	Supply chain + lifecycle	Vendor/integrator changes thresholds/models/features as de facto policy changes	Change-control governance: materiality thresholds, approval workflows, evidence refresh, external audit rights for updates	Observed in multi-actor service models; highlighted as discretion migration into configuration

Table 2. *Cont.*

Risk ID	Risk Type	Primary Source	Typical Sources/Mechanisms	Governance Implications	Empirical Grounding
R13	Accountability diffusion (multi-actor supply chain)	Supply chain + rules	Cloud/model provider/integrator/deployer split responsibilities; no one holds evidence	Contract for evaluation access, audit rights, telemetry/logging, exit/portability; named responsibility matrix	Supply-chain governance literature; procurement leverage emphasized in the manuscript
R14	Governance-washing (non-enforceable controls)	Rules and institutions	Principles without verifiable criteria; ethics statements substitute for enforceable evidence	Bind claims to evidence via 3R Assurance Case (GSN); enforceability primitives (conformity assessment, post-market monitoring)	NIST AI RMF and EU AI Act provide baselines; assurance-case approach operationalizes them
R15	Legitimacy collapse/chilling effects as systems risk	Population-level feedback	Opacity and unfair burdens reduce cooperation; adversarial adaptation; trust erosion degrades system inputs	Treat legitimacy as a design requirement: transparency, contestability, bounded scope, auditability; monitor trust proxies	Socio-technical governance literature; framed as destabilizing feedback in the manuscript

8.7. Evidence Map (Claim → Evidence Type → Strength → Transferability)

Table 3 classifies the main claim families in the manuscript by (i) claim role (conceptual/methodological vs. deployment–empirical), (ii) evidence tier (as defined in the paper's Data Corpus and Evidence Ranking), (iii) strength of support, (iv) transferability, and (v) key limitations and boundary conditions.

Table 3. Evidence map linking claim families to evidence type, support strength, transferability, and limitations.

Claim Family	Claim (Short)	Claim Role	Evidence Tier(s)	Strength	Transferability	Key Limitations/Boundary Conditions	Primary Anchors in Paper
Systems-theoretic framing	Rights breaches should be treated as control failures at the actuation boundary (system-level, not model-only).	Conceptual + methodological	Tier 2 + Tier 1 (case mechanism tests)	Moderate → Strong (case-dependent)	High (design principle); medium (implementation details)	Strength increases when actuation pathways and safeguards are documented; weaker when internal decision logic is inaccessible.	STPA/STAMP + CPS-Governance Mapping Protocol; replication logic across cases.
Uncertainty and legal certainty	Uncertainty must be operationalized and linked to actuation gates to avoid 'confident form, fragile substance'.	Conceptual + methodological	Tier 2 + Tier 3 + Tier 1 (where available)	Moderate	High (general); medium (threshold selection)	Operational thresholds depend on task, harm severity, and workflow tempo; requires monitoring and override evidence.	Uncertainty literature + actuation-gate requirements; audit-ready event record.

Table 3. *Cont.*

Claim Family	Claim (Short)	Claim Role	Evidence Tier(s)	Strength	Transferability	Key Limitations/Boundary Conditions	Primary Anchors in Paper
Biometrics disparity	Demographic error differentials in face recognition can create differential rights exposure in policing/border workflows.	Empirical (benchmark) + governance implication	Tier 2 (benchmarks) + Tier 1 (governance records)	Strong (existence of differentials); Moderate (downstream magnitude)	Medium	Magnitude depends on population, sensing conditions, thresholds, and downstream intervention design; requires context-specific testing.	NIST FRVT + peer-reviewed disparity studies; case linkage to escalation workflows.
Predictive policing feedback	Actuation reshapes the data-generating process, producing runaway feedback and endogenous evidence in enforcement allocation.	Empirical (mechanism)	Tier 2 + Tier 1 (where documented)	Strong (mechanism); Moderate (local magnitude)	Medium	Severity depends on coupling between enforcement intensity and recorded incidents; retraining practices and feature design.	Runaway feedback loop studies; 'dirty data' and civil-rights implications literature.
HITL and accountability	Human-in-the-loop can be a governance fig leaf under automation bias; accountability can collapse onto operators.	Empirical (human factors) + governance implication	Tier 2 + within-case tracing	Strong (bias tendency); Moderate (operational rates)	High (risk class); medium (severity)	Depends on workload, UI design, incentives, and time pressure; must be evidenced in workflow traces.	Automation bias literature; 'moral crumple zone'; process tracing of actuation pinch points.
Auditability as governance variable	If evidence cannot be audited (logs/versioning/provenance missing), governance fails; opacity is a governance finding.	Methodological	Tier 1 + Tier 3	Moderate → Strong	High	Trade secrecy and access constraints can block diagnosis of internal causes; mitigated by evidentiary deltas and counterfactual assurance.	Evidence hierarchy section; assurance-by-evidence; audit-ready architecture requirements.
3R Assurance Case operationalization	A 3R Assurance Case (GSN) translates rights and legal rules into auditable claims, assumptions, and evidence across lifecycle control points.	Methodological (implementation artifact)	Tier 3 + Tier 2 (assurance practice)	Moderate	High	Effectiveness depends on enforceable procurement/audit rights and keeping the artifact live under drift and updates.	NIST AI RMF + EU AI Act lifecycle logic; GSN assurance-case method; post-market monitoring hooks.

9. Conclusions

9.1. Synthesis of the 3R Framework and the Triadic Governance Model

The central contribution of this paper is to translate legitimacy from a normative aspiration into an implementable control objective for high-consequence AI-enabled CPSs. Specifically, we formalize the Risk–Rights–Rules (3R) architecture as a stabilizing control law as a set of enforceable constraints and evidence obligations that bind the sensing–inference–actuation loop under uncertainty, drift, and institutional coupling [8,31,32,70,80]. We argue that the recurring governance failures documented across

domains are not reducible to model error. They are system-design failures resulting from missing constraints at specific control points (scope, actuation, and feedback) and from the absence of audit-ready evidence hooks that make coercive authority reviewable [10,23,46,66,80,134]. Integrating 3R with the triadic model, "Governance of, by, and for AI", yields a layered defence that is operational rather than rhetorical:

- Governance of AI (Risk) secures lifecycle integrity through documentation, evaluation, monitoring, and internal auditing, making performance and drift auditable over time [38,64–66].
- Governance by AI (Rights) constrains probabilistic authority at the actuation boundary by requiring legality, contestability, and uncertainty-aware gates; i.e., inference may proceed, but high-friction action is admissible only when constraints and evidence conditions are satisfied [11,25,82,84].
- Governance for AI (Rules) makes constraints enforceable by binding them to oversight mechanisms, auditability primitives, evaluation access, logging, change control, and remedies, so compliance is testable rather than asserted [18,70,80,134].

The resulting blueprint bridges the gap between stochastic and deterministic approaches. It specifies how probabilistic inference can be used by public authorities without displacing the rule-of-law practices (reasonableness, reviewability, and contestability) of legitimate coercive power [11,31,32,143,144].

9.2. Lessons from the Case Portfolio: The Cost of Normative Fragility

The three-case portfolio demonstrates normative fragility as an empirically observable property: when rights-based constraints are not engineered into the loop, the system becomes unstable, legally, institutionally, or both. In the UK, Bridges shows that a sensing scope without enforceable rules (bounded siting and bounded watchlisting) is judicially invalidated; accuracy improvements cannot repair a scope-control deficit because the failure mechanism is discretionary authority, not prediction error [10,106–108]. In U.S. border biometrics deployments, the dominant vulnerability is actuation-gate erosion under tempo. Routing decisions (clearance vs. secondary screening) can become de facto commands if systems are not designed to resist automation bias and to generate reviewable records at the moment of intervention [25,111–113]. This is where rights debt accumulates: distributional procedural burdens (delays, escalations, and repeated screening) borne disproportionately by identifiable groups when demographic performance differentials meet high-throughput actuation without compensating constraints [13,111,112].

Chicago's SSL critically illustrates the endogeneity of the feedback loop. Actuation reshapes the data-generating process, producing self-reinforcing allocation dynamics unless the system explicitly distinguishes risk from "enforcement intensity" [23,24,46,47,114,115]. Across cases, the consistent inference is that risk-only optimization can turn the technology into a threat multiplier because unconstrained loops convert uncertainty and error into coercive outcomes while thinning out review and justification. The downstream consequence is not only individual error, it is predictable trust degradation and chilling effects that weaken institutional legitimacy and cooperation, inputs that are functionally necessary when the population is part of the plant [75,76,116–119].

9.3. High-Impact Call to Action: Designing for Contestability and Remedy

This paper's action plan is deliberately implementable: it specifies minimum-viable controls that agencies can procure, engineers can implement, and auditors can test. The three pillars are as follows:

- Mandatory control-point constraints (scope, actuation, and feedback). Every high-consequence AI-enabled CPS must implement enforceable constraints at the spe-

cific control points where authority is exercised: (i) scope-control instruments (siting + watchlisting rules) for sensing; (ii) rights-aware actuation gates that are uncertainty-sensitive and legally bounded; and (iii) feedback-discrimination controls that prevent endogenous data from being misread as objective risk [10,23,46,70,80]. Different failure modes require different controls, and "accuracy upgrades" cannot substitute for missing constraints.

- Contestability-by-design through auditable adverse action packets. Whenever the system produces a high-friction intervention, it must emit a contestability artifact at the moment of actuation: (i) the decision and the authorizing rule/constraint, (ii) a minimally sufficient, action-guiding explanation (often counterfactual), and (iii) the review/appeal channel. Counterfactual explanations can support recourse without full model disclosure, but only when paired with real review pathways and logged evidence [11,82,84].
- Audit-ready architectures and enforceable procurement in algorithmic supply chains. Public-sector procurement must contract for audit rights, evaluation access, change control (versioning/rollback/release gates), and continuous logging; otherwise, the assurance case cannot be maintained in the face of drift, vendor updates, and system entanglement [38,66,70,80,134]. This is the engineering meaning of rule-enforced innovation. Innovation remains deployable when normative stability is auditable rather than presumed [70,79,80,120].

9.4. Barriers to Real-World Implementation of the 3R Framework

While the 3R (Risk–Rights–Rules) architecture and the 3R Assurance Case offer a practical path to making rights constraints enforceable across the sensing–inference–actuation loop, several barriers may slow or distort real-world adoption. First, institutional resistance is likely where automated decision pipelines have become embedded in routine operations: introducing actuation gates, contestability hooks, and evidence discipline can be perceived as reducing speed, increasing scrutiny, or shifting power away from operational discretion—especially in high-tempo domains such as border processing, policing, and eligibility determination. Second, compliance and implementation costs are non-trivial: maintaining audit-ready logging, continuous monitoring for drift and demographic error, and version-controlled assurance artifacts requires sustained investment in tooling, governance workflows, and cross-functional coordination (engineering, legal, procurement, and operations). Third, auditing capacity constraints remain a bottleneck: even well-designed assurance cases depend on competent reviewers and institutions with the time, authority, and technical literacy to evaluate the evidence, interpret uncertainty bounds, and challenge "assurance-by-assertion." Finally, vendor ecosystem constraints can undermine enforceability: opaque models, trade-secret claims, limited access to evaluation data, and restricted change-control visibility weaken the evidence chain needed to keep the assurance case valid over time—particularly when systems are procured as part of a multi-actor supply chain (cloud provider, model vendor, integrator, and deploying agency). Addressing these barriers will require not only technical templates, but also procurement leverage (audit rights, evidence access, and change-control obligations), workforce development for independent review, and governance incentives that treat auditability and contestability as operational requirements rather than optional add-ons.

9.5. Future Research Directions: Filling the Gap

While this study establishes a minimum-viable blueprint for rights-based CPS governance, several gaps remain. First, proportionality and necessity require operational research agendas: how to formalize these public-law tests into implementable actuation constraints

and evidence thresholds in high-velocity loops without collapsing them into simplistic score cutoffs [31,144]. Second, comparative cross-jurisdictional audits are needed to test how the 3R blueprint interacts with different constitutional baselines and enforcement architectures (e.g., high-risk obligations, post-market monitoring, and audit access conditions), including Global South contexts where capacity constraints and data-extractive dependencies can reshape what "enforceability" means in practice [18,70,134]. Third, human factors research must move beyond "human-in-the-loop" slogans to empirically test which contestability artifacts and override procedures actually reduce automation bias under tempo [25,67,113]. Finally, the "liability sponge" problem requires deeper empirical study: how accountability collapses onto frontline actors when auditability is weak, and how assurance cases and contestability-by-design change responsibility allocation and institutional learning over time [28,78]. Answering these questions will determine whether the next-generation CPS governance infrastructure is not only technically advanced but institutionally durable and normatively legitimate under real-world stressors and change [80,120,139,141].

Author Contributions: Conceptualization, M.N., H.H. and M.L.; methodology, M.N., H.H. and M.L.; resources, M.N., H.H. and M.L.; writing—original draft preparation, M.N., H.H. and M.L.; writing—review and editing, M.N., H.H. and M.L.; visualization, M.N., H.H. and M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new datasets were generated in this study. The analysis relies on publicly available legal, regulatory, audit, technical, and scholarly materials cited in the reference list.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. National Academies of Sciences, Engineering, and Medicine. *A 21st Century Cyber-Physical Systems Education*; The National Academies Press: Washington, DC, USA, 2016. [CrossRef]
2. Lee, E.A. The Past, Present and Future of Cyber-Physical Systems: A Focus on Models. *Sensors* **2015**, *15*, 4837–4869. [CrossRef] [PubMed]
3. Eubanks, V. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*; St. Martin's Press: New York, NY, USA, 2018.
4. Narayanan, A.; Kapoor, S. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*; Princeton University Press: Princeton, NJ, USA, 2024.
5. Dressel, J.; Farid, H. The Accuracy, Fairness, and Limits of Predicting Recidivism. *Sci. Adv.* **2018**, *4*, eaao5580. [CrossRef]
6. Crawford, K.; Schultz, J. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston Coll. Law Rev.* **2014**, *55*, 93–128. Available online: https://bclawreview.bc.edu/articles/620?utm_source=chatgpt.com (accessed on 15 December 2025).
7. Selbst, A.D.; Boyd, D.; Friedler, S.A.; Venkatasubramanian, S.; Vertesi, J. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT* '19), Atlanta, GA, USA, 29–31 January 2019*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 59–68. [CrossRef]
8. Leveson, N.G. *Engineering a Safer World: Systems Thinking Applied to Safety*; MIT Press: Cambridge, MA, USA, 2012.
9. Citron, D.K.; Pasquale, F. The Scored Society: Due Process for Automated Predictions. *Wash. Law Rev.* **2014**, *89*, 1–33. Available online: https://digitalcommons.law.uw.edu/wlr/vol89/iss1/2/ (accessed on 11 December 2025).
10. R (On the Application of Edward Bridges) v the Chief Constable of South Wales Police and Others, [2020] EWCA Civ 1058, Case No. C1/2019/2670 (Court of Appeal (Civil Division). 11 August 2020. Available online: https://caselaw.nationalarchives.gov.uk/ewca/civ/2020/1058 (accessed on 11 December 2025).
11. Calo, R.; Citron, D.K. The Automated Administrative State: A Crisis of Legitimacy. *Emory Law J.* **2021**, *70*, 797.
12. Hüllermeier, E.; Waegeman, W. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Mach. Learn.* **2021**, *110*, 457–506. [CrossRef]
13. Grother, P.; Ngan, M.; Hanaoka, K. *Face Recognition Vendor Test (FRVT), Part 3: Demographic Effects*; NISTIR 8280; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2019. [CrossRef]
14. Smuha, N.A. Beyond the Individual: Governing AI's Societal Harm. *Internet Policy Rev.* **2021**, *10*, 1–32. [CrossRef]

15. De Gregorio, G.; Dunn, P. The European risk-based approaches: Connecting constitutional dots in the digital age. *Common Mark. Law Rev.* **2022**, *59*, 473–500. [CrossRef]
16. Ebers, M. Truly Risk-based Regulation of Artificial Intelligence: How to Implement the EU's AI Act. *Eur. J. Risk Regul.* **2025**, *16*, 684–703. [CrossRef]
17. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*; NIST AI 100-1; NIST: Gaithersburg, MD, USA, 2023. [CrossRef]
18. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828. *Off. J. Eur. Union* **2024**, 2024/1689. Available online: https://data.europa.eu/eli/reg/2024/1689/oj (accessed on 10 December 2025).
19. Mantelero, A. The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, legal obligations and key elements for a model template. *Comput. Law Secur. Rev.* **2024**, *54*, 106020. [CrossRef]
20. Kelly, T.P.; Weaver, R.A. The Goal Structuring Notation–A Safety Argument Notation. In Proceedings of the Workshop on Assurance Cases: Best Practices, Possible Obstacles and Future Opportunities (DSN 2004), Florence, Italy, 28 June–1 July 2004.
21. Hawkins, R.; Kelly, T. *A Software Safety Argument Pattern Catalogue*; Technical Report YCS-2013-482; Department of Computer Science, University of York: York, UK, 2013. Available online: https://www-users.york.ac.uk/ rdh2/papers/patternCatalogue.pdf (accessed on 12 December 2025).
22. Green, B.; Hu, L. The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. In Proceedings of the Machine Learning: The Debates Workshop, 35th International Conference on Machine Learning (ICML 2018), Stockholm, Sweden, 14 July 2018. Available online: https://api.semanticscholar.org/CorpusID:49544563 (accessed on 10 December 2025).
23. Ensign, D.; Friedler, S.A.; Neville, S.; Scheidegger, C.; Venkatasubramanian, S. Runaway Feedback Loops in Predictive Policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT* 2018), PMLR: 2018, Proceedings of Machine Learning Research, New York, NY, USA, 23–24 February 2018*; Friedler, S.A., Wilson, C., Eds.; Volume 81, pp. 160–171. Available online: https://proceedings.mlr.press/v81/ensign18a.html (accessed on 16 December 2025).
24. Richardson, R.; Schultz, J.; Crawford, K. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *NYU Law Rev.* **2019**, *94*, 192–233.
25. Parasuraman, R.; Manzey, D.H. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Hum. Factors* **2010**, *52*, 381–410. [CrossRef]
26. Skitka, L.J.; Mosier, K.L.; Burdick, M. Does Automation Bias Decision-Making? *Int. J. Hum.-Comput. Stud.* **1999**, *51*, 991–1006. [CrossRef]
27. Lyell, D.; Coiera, E. Automation Bias and Verification Complexity: A Systematic Review. *J. Am. Med. Inf. Assoc.* **2017**, *24*, 423–431. [CrossRef]
28. Elish, M.C. Moral Crumple Zones: Cautionary Tales in Human–Robot Interaction. *Engag. Sci. Technol. Soc.* **2019**, *5*, 40–60. [CrossRef]
29. Baheti, R.; Gill, H. Cyber-physical Systems. In *The Impact of Control Technology*; Samad, T., Annaswamy, A.M., Eds.; IEEE Control Systems Society: Piscataway, NJ, USA, 2011; pp. 161–166.
30. Lee, E.A. Cyber Physical Systems: Design Challenges. In *Proceedings of the 11th IEEE International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing (ISORC 2008), Orlando, FL, USA, 5–7 May 2008*; IEEE Computer Society: Los Alamitos, CA, USA, 2008; pp. 363–369. [CrossRef]
31. Hildebrandt, M. Law as Computation in the Era of Artificial Legal Intelligence: Speaking Law to the Power of Statistics. *Univ. Tor. Law J.* **2018**, *68*, 12–35. [CrossRef]
32. Yeung, K. Algorithmic Regulation: A Critical Interrogation. *Regul. Gov.* **2018**, *12*, 505–523. [CrossRef]
33. Bayamlıoğlu, E.; Leenes, R. The "Rule of Law" Implications of Data-Driven Decision-Making: A Techno-Regulatory Perspective. *Law Innov. Technol.* **2018**, *10*, 295–313. [CrossRef]
34. Kehl, D.; Guo, P.; Kessler, S. *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing*; Responsive Communities Initiative, Berkman Klein Center for Internet & Society, Harvard Law School: Cambridge, MA, USA, 2017.
35. Ouyang, M. Review on Modeling and Simulation of Interdependent Critical Infrastructure Systems. *Reliab. Eng. Syst. Saf.* **2014**, *121*, 43–60. [CrossRef]
36. Rinaldi, S.M.; Peerenboom, J.P.; Kelly, T.K. Identifying, Understanding, and Analyzing Critical Infrastructure Interdependencies. *IEEE Control Syst. Mag.* **2001**, *21*, 11–25. [CrossRef]
37. Gama, J.; Žliobaitė, I.; Bifet, A.; Pechenizkiy, M.; Bouchachia, A. A Survey on Concept Drift Adaptation. *ACM Comput. Surv.* **2014**, *46*, 44. [CrossRef]

38. Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.-F.; Dennison, D. Hidden Technical Debt in Machine Learning Systems. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; pp. 2503–2511.
39. Van der Vlist, F.; Helmond, A.; Ferrari, F. Big AI: Cloud Infrastructure Dependence and the Industrialisation of Artificial Intelligence. *Big Data Soc.* **2024**, *11*, 1–16. [CrossRef]
40. De Laat, P.B. Big Data and Algorithmic Decision-Making: Can Transparency Restore Accountability? *ACM SIGCAS Comput. Soc.* **2017**, *47*, 39–53. [CrossRef]
41. Carswell, G.; De Neve, G. Transparency, Exclusion and Mediation: How Digital and Biometric Technologies Are Transforming Social Protection in Tamil Nadu, India. *Oxf. Dev. Stud.* **2022**, *50*, 126–141. [CrossRef]
42. Dixon, P. A Failure to "Do No Harm"—India's Aadhaar Biometric ID Program and Its Inability to Protect Privacy in Relation to Measures in Europe and the U.S. *Health Technol.* **2017**, *7*, 539–567. [CrossRef]
43. Cavazos, J.G.; Phillips, P.J.; Castillo, C.D.; O'Toole, A.J. Accuracy Comparison across Face Recognition Algorithms: Where Are We on Measuring Race Bias? *IEEE Trans. Biom. Behav. Identity Sci.* **2021**, *3*, 101–111. [CrossRef]
44. O'Toole, A.J.; Phillips, P.J.; An, X.; Dunlop, J. Demographic Effects on Estimates of Automatic Face Recognition Performance. *Image Vis. Comput.* **2012**, *30*, 169–176. [CrossRef]
45. Barocas, S.; Selbst, A.D. Big Data's Disparate Impact. *Calif. Law Rev.* **2016**, *104*, 671–732. [CrossRef]
46. Lum, K.; Isaac, W. To Predict and Serve? *Significance* **2016**, *13*, 14–19. [CrossRef]
47. Brayne, S. Big Data Surveillance: The Case of Policing. *Am. Sociol. Rev.* **2017**, *82*, 977–1008. [CrossRef]
48. Cybersecurity and Infrastructure Security Agency (CISA). Healthcare and Public Health Sector. Available online: https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/critical-infrastructure-sectors/healthcare-and-public-health-sector (accessed on 10 December 2025).
49. Canada's Critical Infrastructure (CI). Public Safety Canada. 2025. Available online: https://www.publicsafety.gc.ca/cnt/ntnl-scrt/crtcl-nfrstrctr/cci-iec-en.aspx (accessed on 12 December 2025).
50. Gostin, L.O.; Katz, R. The International Health Regulations: The Governing Framework for Global Health Security. *Milbank Q.* **2016**, *94*, 264–313. [CrossRef]
51. Katz, R.; Fischer, J.E. The Revised International Health Regulations: A Framework for Global Pandemic Response. *Glob. Health Gov.* **2010**, *3*, 1–18. Available online: https://blogs.shu.edu/ghg/files/2011/11/Katz-and-Fischer_The-Revised-International-Health-Regulations_Spring-2010.pdf (accessed on 10 December 2025).
52. Katz, R. Use of Revised International Health Regulations during Influenza A (H1N1) Epidemic, 2009. *Emerg. Infect. Dis.* **2009**, *15*, 1165–1170. [CrossRef]
53. World Health Organization (WHO). International Health Regulations (2005): As amended in 2014, 2022 and 2024. 2025. Available online: https://apps.who.int/gb/bd/pdf_files/IHR_2014-2022-2024-en.pdf (accessed on 10 December 2025).
54. World Health Organization (WHO). IHR Core Capacities. Available online: https://www.emro.who.int/international-health-regulations/about/ihr-core-capacities.html (accessed on 11 December 2025).
55. Martinez-Martin, N.; Wieten, S.; Magnus, D.; Cho, M.K. Digital Contact Tracing, Privacy, and Public Health. *Hastings Cent. Rep.* **2020**, *50*, 43–46. [CrossRef]
56. Klenk, M.; Duijf, H. Ethics of Digital Contact Tracing and COVID-19: Who Is (Not) Free to Go? *Ethics Inf. Technol.* **2021**, *23*, 69–77. [CrossRef]
57. Shachar, C.; Gerke, S.; Adashi, E.Y. AI Surveillance during Pandemics: Ethical Implementation Imperatives. *Hastings Cent. Rep.* **2020**, *50*, 18–21. [CrossRef]
58. Orzechowski, M.; Schochow, M.; Steger, F. Balancing Public Health and Civil Liberties in Times of Pandemic. *J. Public Health Policy* **2021**, *42*, 145–153. [CrossRef]
59. Rinaldi, A.; Teo, S.A. The Use of Artificial Intelligence Technologies in Border and Migration Control and the Subtle Erosion of Human Rights. *Int. Comp. Law Q.* **2025**, *74*, 61–89. [CrossRef]
60. Delfos, J.; Zuiderwijk, A.M.G.; van Cranenburgh, S.; Chorus, C.G.; Dobbe, R.I.J. Integral System Safety for Machine Learning in the Public Sector: An Empirical Account. *Gov. Inf. Q.* **2024**, *41*, 101963. [CrossRef]
61. Patriarca, R.; Chatzimichailidou, M.; Karanikas, N.; Di Gravio, G. The Past and Present of System—Theoretic Accident Model And Processes (STAMP) and Its Associated Techniques: A Scoping Review. *Saf. Sci.* **2022**, *146*, 105566. [CrossRef]
62. Rotenberg, M. Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (Council Eur.). *Int. Leg. Mater.* **2025**, *64*, 859–902. [CrossRef]
63. Bietti, E. From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 210–219. [CrossRef]

64. Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 220–229. [CrossRef]
65. Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J.W.; Wallach, H.; Daumé, H., III; Crawford, K. Datasheets for Datasets. *Commun. ACM* **2021**, *64*, 86–92. [CrossRef]
66. Raji, I.D.; Smart, A.; White, R.N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; Barnes, P. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 33–44. [CrossRef]
67. Green, B. The Flaws of Policies Requiring Human Oversight of Government Algorithms. *Comput. Law Secur. Rev.* **2022**, *45*, 105681. [CrossRef]
68. Green, B.; Chen, Y. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* **2019**, *3*, 50. [CrossRef]
69. Wagner, B. Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. *Policy Internet* **2019**, *11*, 104–122. [CrossRef]
70. Mökander, J.; Axente, M.; Casolari, F.; Floridi, L. Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *arXiv* **2021**, arXiv:2111.05071. [CrossRef]
71. Council of Europe. *Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*; CETS No. 225; Council of Europe: Strasbourg, France, 2024. Available online: https://rm.coe.int/1680afae3c (accessed on 12 December 2025).
72. Van Kolfschooten, H.; Shachar, C. The Council of Europe's AI Convention (2023–2024): Promises and pitfalls for health protection. *Health Policy* **2023**, *138*, 104935. [CrossRef]
73. Chan, K.J.D.; Papyshev, G.; Yarime, M. Balancing the tradeoff between regulation and innovation for artificial intelligence: An analysis of top-down command and control and bottom-up self-regulatory approaches. *Technol. Soc.* **2024**, *79*, 102747. [CrossRef]
74. Gikay, A.A. Risks, innovation, and adaptability in the UK's incrementalism versus the European Union's comprehensive artificial intelligence regulation. *Int. J. Law Inf. Technol.* **2024**, *32*, eaae013. [CrossRef]
75. Lee, J.D.; See, K.A. Trust in automation: Designing for appropriate reliance. *Hum. Factors* **2004**, *46*, 50–80. [CrossRef]
76. Tyler, T.R. *Why People Obey the Law*; Princeton University Press: Princeton, NJ, USA, 2006.
77. Fjeld, J.; Achten, N.; Hilligoss, H.; Nagy, A.; Srikumar, M. *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*; Berkman Klein Center for Internet & Society, Harvard University: Cambridge, MA, USA, 2020. Available online: https://cyber.harvard.edu/publication/2020/principled-ai (accessed on 12 December 2025).
78. Matthias, A. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics Inf. Technol.* **2004**, *6*, 175–183. [CrossRef]
79. Floridi, L. Establishing the rules for building trustworthy AI. In *Ethics, Governance, and Policies in Artificial Intelligence*; Floridi, L., Ed.; Springer: Cham, Switzerland, 2021; pp. 41–45. [CrossRef]
80. Rushby, J. *Understanding and Evaluating Assurance Cases*; Contractor Report NASA/CR–2015-218802; NASA Langley Research Center: Hampton, VA, USA, 2015.
81. Hoekstra, J.; Diker-Vanberg, A. Can AI Tools Enhance Access to Remedies as Envisaged under the UN Guiding Principles on Business and Human Rights: Yay or Nay? A Critical Assessment. *Int. Rev. Law Comput. Technol.* **2025**, *39*, 1–22. [CrossRef]
82. Wachter, S.; Mittelstadt, B.; Russell, C. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harv. J. Law Technol.* **2018**, *31*, 841–887. [CrossRef]
83. Ustun, B.; Spangher, A.; Liu, Y. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19), Atlanta, GA, USA, 29–31 January 2019*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 10–19. [CrossRef]
84. Kroll, J.A.; Huey, J.; Barocas, S.; Felten, E.W.; Reidenberg, J.R.; Robinson, D.G.; Yu, H. Accountable Algorithms. *Univ. Pa. Law Rev.* **2017**, *165*, 633–705. Available online: https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/ (accessed on 14 December 2025).
85. Yin, R.K. *Case Study Research and Applications: Design and Methods*, 6th ed.; SAGE Publications, Inc.: Thousand Oaks, CA, USA, 2018. Available online: https://uk.sagepub.com/en-gb/eur/case-study-research-and-applications/book250150 (accessed on 13 December 2025).
86. Eisenhardt, K.M. Building Theories from Case Study Research. *Acad. Manag. Rev.* **1989**, *14*, 532–550. [CrossRef]
87. Baxter, G.; Sommerville, I. Socio-technical Systems: From Design Methods to Systems Engineering. *Interact. Comput.* **2011**, *23*, 4–17. [CrossRef]
88. Åström, K.J.; Murray, R.M. *Feedback Systems: An Introduction for Scientists and Engineers*; Princeton University Press: Princeton, NJ, USA, 2008. Available online: https://authors.library.caltech.edu/records/yzs24-xsx88 (accessed on 13 December 2025).

89. Breaux, T.D.; Antón, A.I. Analyzing Regulatory Rules for Privacy and Security Requirements. *IEEE Trans. Softw. Eng.* **2008**, *34*, 5–20. [CrossRef]
90. Otto, P.N.; Antón, A.I. Addressing Legal Requirements in Requirements Engineering. In *Proceedings of the 15th IEEE International Requirements Engineering Conference (RE 2007), New Delhi, India, 15–19 October 2007*; IEEE: Piscataway, NJ, USA, 2007; pp. 5–14. [CrossRef]
91. Ghanavati, S.; Amyot, D.; Rifaut, A. Legal Goal-Oriented Requirement Language (Legal GRL) for Modeling Regulations. In *Proceedings of the 6th International Workshop on Modeling in Software Engineering (MiSE 2014), Hyderabad, India, 31 May–7 June 2014;* Association for Computing Machinery: New York, NY, USA, 2014; pp. 1–6. [CrossRef]
92. Chaal, M.; Valdez Banda, O.A.; Glomsrud, J.A.; Basnet, S.; Hirdaris, S.; Kujala, P. A Framework to Model the STPA Hierarchical Control Structure of an Autonomous Ship. *Saf. Sci.* **2020**, *132*, 104939. [CrossRef]
93. Pasquale, F. *The Black Box Society: The Secret Algorithms That Control Money and Information*; Harvard University Press: Cambridge, MA, USA, 2015.
94. Burrell, J. How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms. *Big Data Soc.* **2016**, *3*, 1–12. [CrossRef]
95. Bowen, G.A. Document Analysis as a Qualitative Research Method. *Qual. Res. J.* **2009**, *9*, 27–40. [CrossRef]
96. Purshouse, J.; Campbell, L. Automated Facial Recognition and Policing: A Bridge Too Far? *Leg. Stud.* **2022**, *42*, 209–227. [CrossRef]
97. Urquhart, L.; Miranda, D. Policing Faces: The Present and Future of Intelligent Facial Surveillance. *Inf. Commun. Technol. Law* **2022**, *31*, 194–219. [CrossRef]
98. Cook, C.M.; Howard, J.J.; Sirotin, Y.B.; Tipton, J.L.; Vemury, A.R. Demographic Effects in Facial Recognition and Their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems. *IEEE Trans. Biom. Behav. Identity Sci.* **2019**, *1*, 32–41. [CrossRef]
99. Buolamwini, J.; Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT* 2018), New York, NY, USA, 23–24 February 2018*; Friedler, S.A., Wilson, C., Eds.; PMLR, 2018; Volume 81, pp. 77–91. Available online: https://proceedings.mlr.press/v81/buolamwini18a.html (accessed on 6 December 2025).
100. Klare, B.F.; Burge, M.J.; Klontz, J.C.; Vorder Bruegge, R.W.; Jain, A.K. Face Recognition Performance: Role of Demographic Information. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 1789–1801. [CrossRef]
101. Hung, T.-W.; Yen, C.-P. Predictive Policing and Algorithmic Fairness. *Synthese* **2023**, *201*, 1–29. [CrossRef]
102. Brauneis, R.; Goodman, E.P. Algorithmic Transparency for the Smart City. *Yale J. Law Technol.* **2018**, *20*, 103–176. [CrossRef]
103. Beach, D.; Pedersen, R.B. *Process-Tracing Methods: Foundations and Guidelines*, 2nd ed.; University of Michigan Press: Ann Arbor, MI, USA, 2019.
104. Miles, M.B.; Huberman, A.M.; Saldaña, J. *Qualitative Data Analysis: A Methods Sourcebook*, 3rd ed.; SAGE: Thousand Oaks, CA, USA, 2014.
105. Office of the United Nations High Commissioner for Human Rights (OHCHR). *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*; United Nations: New York, NY, USA, 2011.
106. Kotsoglou, K.N.; Oswald, M. The Long Arm of the Algorithm? Automated Facial Recognition as Evidence and Trigger for Police Intervention. *Forensic Sci. Int. Synerg.* **2020**, *2*, 86–89. [CrossRef]
107. Bradford, B.; Yesberg, J.A.; Jackson, J.; Dawson, P. Live Facial Recognition: Trust and Legitimacy as Predictors of Public Support for Police Use of New Technology. *Br. J. Criminol.* **2020**, *60*, 1502–1522. [CrossRef]
108. Gikay, A.A. Regulating Use by Law Enforcement Authorities of Live Facial Recognition Technology in Public Spaces: An Incremental Approach. *Camb. Law J.* **2023**, *82*, 414–449. [CrossRef]
109. Lisle, D.; Bourne, M. The Many Lives of Border Automation: Turbulence, Coordination and Care. *Soc. Stud. Sci.* **2019**, *49*, 682–706. [CrossRef]
110. Khan, N.; Efthymiou, M. The Use of Biometric Technology at Airports: The Case of Customs and Border Protection (CBP). *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100049. [CrossRef]
111. U.S. Government Accountability Office. *Facial Recognition: CBP and TSA Are Taking Steps to Implement Programs, but CBP Should Address Privacy and System Performance Issues*; GAO-20-568; GAO: Washington, DC, USA, 2020.
112. U.S. Government Accountability Office. *Facial Recognition Technology: CBP Traveler Identity Verification and Efforts to Address Privacy Issues*; GAO-22-106154; GAO: Washington, DC, USA, 2022.
113. Cummings, M.L. Automation Bias in Intelligent Time Critical Decision Support Systems. In Proceedings of the AIAA 1st Intelligent Systems Technical Conference, Chicago, IL, USA, 20–22 September 2004. [CrossRef]
114. Strategic Subject List—Historical (SSL) Dataset Documentation. Available online: https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List-Historical/4aki-r3np/about_data (accessed on 27 January 2026).
115. Ferguson, A.G. *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law* Enforcement; NYU Press: New York, NY, USA, 2017.
116. Penney, J.W. Chilling Effects: Online Surveillance and Wikipedia Use. *Berkeley Technol. Law J.* **2016**, *31*, 117–182.

117. Stoycheff, E. Under Surveillance: Examining Facebook's Spiral of Silence Effects in the Wake of NSA Internet Monitoring. *J. Mass Commun. Q.* **2016**, *93*, 296–311. [CrossRef]
118. Murray, D.; Fussey, P.; Hove, K.; Wakabi, W.; Kimumwe, P.; Saki, O.; Stevens, A. The Chilling Effects of Surveillance and Human Rights: Insights from Qualitative Research in Uganda and Zimbabwe. *J. Hum. Rights Pract.* **2024**, *16*, 397–412. [CrossRef]
119. Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*; Public Affairs: New York, NY, USA, 2019.
120. Woods, D.D. Four Concepts for Resilience and the Implications for the Future of Resilience Engineering. *Reliab. Eng. Syst. Saf.* **2015**, *141*, 5–9. [CrossRef]
121. Perrow, C. *Normal Accidents: Living with High-Risk Technologies*; Basic Books: New York, NY, USA, 1984.
122. Van Lamsweerde, A. *Requirements Engineering: From System Goals to UML Models to* Software Specifications; Wiley: Chichester, UK, 2009.
123. Hardt, M.; Price, E.; Srebro, N. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems, Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016*; Curran Associates Inc.: Red Hook, NY, USA, 2016.
124. Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.V.; Lakshminarayanan, B.; Snoek, J. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. In *Advances in Neural Information Processing Systems, Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019*; Curran Associates Inc.: Red Hook, NY, USA, 2019.
125. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the 33rd International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016.
126. Hendrycks, D.; Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv* **2017**, arXiv:1610.02136.
127. Hendrycks, D.; Dietterich, T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *arXiv* **2019**, arXiv:1903.12261. [CrossRef]
128. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete Problems in AI Safety. *arXiv* **2016**, arXiv:1606.06565. [CrossRef]
129. Orseau, L.; Armstrong, S. Safely Interruptible Agents. In *Proceedings of UAI, Jersey City, NJ, USA, 25–29 June 2016*; AUAI Press: Arlington, VA, USA, 2016.
130. Marusich, L.R.; Bakdash, J.Z.; Zhou, Y.; Kantarcioglu, M. Using AI Uncertainty Quantification to Improve Human Decision-Making. In *Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR 2024, Vienna, Austria, 21–27 July 2024*; JMLR: Norfolk, MA, USA, 2024; pp. 34949–34960.
131. Blind, K.; Münch, F. The Interplay between Innovation, Standards and Regulation in a Globalising Economy. *J. Clean. Prod.* **2024**, *445*, 141202. [CrossRef]
132. *ISO/IEC 42001:2023*; Information Technology—Artificial Intelligence—Management System. International Organization for Standardization: Geneva, Switzerland, 2023.
133. *ISO/IEC 23894:2023*; Information Technology—Artificial Intelligence—Guidance on Risk Management. International Organization for Standardization: Geneva, Switzerland, 2023.
134. Cobbe, J.; Veale, M.; Singh, J. Understanding Accountability in Algorithmic Supply Chains. In *Proceedings of FAccT*; Association for Computing Machinery: New York, NY, USA, 2023.
135. Post Office Horizon IT Inquiry. *Volume 1 of the Post Office Horizon IT Inquiry Final Report*; UK Inquiry Report: London, UK, 2025.
136. European Union. Regulation (EU) 2016/679 (General Data Protection Regulation). *Off. J. Eur. Union* **2016**, *L 119*, 1–88.
137. European Commission. When Is a Data Protection Impact Assessment (DPIA) Required? Available online: https://commission.europa.eu/law/law-topic/data-protection/rules-business-and-organisations/obligations/when-data-protection-impact-assessment-dpia-required_en (accessed on 6 January 2026).
138. European Commission. AI Act. Available online: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (accessed on 6 January 2026).
139. König, P.D.; Wenzelburger, G. The Legitimacy Gap of Algorithmic Decision-Making in the Public Sector: Why It Arises and How to Address It. *Technol. Soc.* **2021**, *67*, 101688. [CrossRef]
140. Koops, B.-J. The Concept of Function Creep. *Law Innov. Technol.* **2021**, *13*, 29–56. [CrossRef]
141. Grimmelikhuijsen, S.; Meijer, A. Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response. *Perspect. Public Manag. Gov.* **2022**, *5*, 232–252. [CrossRef]
142. Ruschemeier, H.; Hondrich, L. Automation Bias in Public Administration—An Interdisciplinary Perspective from Law and Psychology. *Gov. Inf. Q.* **2024**, *41*, 101953.
143. Binns, R. Algorithmic Accountability and Public Reason. *Philos. Technol.* **2018**, *31*, 543–556. [CrossRef]
144. Strandburg, K.J. Rulemaking and Inscrutable Automated Decision Tools. *Columbia Law Rev.* **2019**, *119*, 1851–1885.

145. Harcourt, B.E. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*; University of Chicago Press: Chicago, IL, USA, 2007.
146. Bayamlıoğlu, E. The Right to Contest Automated Decisions under the General Data Protection Regulation: Beyond the So-Called "Right to Explanation". *Regul. Gov.* **2022**, *16*, 1058–1078. [CrossRef]
147. Christie, J. The Post Office Horizon IT Scandal and the Presumption of the Dependability of Computer Evidence. *Digit. Evid. Electron. Signat. Law Rev.* **2020**, *17*, 49–70. [CrossRef]