

Allocating Responsibility in Algorithmic Governance: Designing Accountability Architectures Beyond Human-in-the-Loop

Qiang SUN

North China University of Technology, Beijing, China · sunqiang206@126.com · ORCID: 0000-0002-9964-2972

Abstract

The “human-in-the-loop” paradigm—the dominant approach to algorithmic accountability—has proven structurally inadequate, systematically devolving into what we term **accountability theater**: the institutional performance of oversight without its substance. This article proposes the **Responsibility Allocation Matrix (RAM)**, a five-actor, four-phase framework that replaces generalized oversight requirements with specific, verifiable obligations for algorithm designers, deployers, regulators, auditors, and affected parties across the full algorithmic lifecycle. Through analysis of three accountability failures—the Dutch childcare benefits scandal, COMPAS recidivism assessment, and Amazon’s hiring algorithm—we identify five structural failure modes that undermine human oversight regardless of regulatory design quality. The RAM distinguishes five responsibility types (primary, residual, due diligence, advisory, and secondary) and specifies obligation chain principles that prevent responsibility diffusion. Comparative analysis across the EU, US, UK, and China reveals that no current jurisdiction fully implements the RAM architecture. The framework provides a systematic diagnostic template for designing enforceable accountability systems for consequential algorithmic deployment.

Keywords algorithmic accountability, responsibility allocation, human-in-the-loop, institutional design, governance architecture, RAM, accountability gaps, comparative regulation

1 Introduction

1.1 The Accountability Crisis in Algorithmic Governance

The question of who answers when algorithms fail has become one of the most pressing challenges of contemporary governance. When an automated welfare system wrongly denies benefits to thousands of vulnerable individuals (Eubanks, 2018), when a risk assessment algorithm contributes to discriminatory sentencing outcomes (Angwin et al., 2016), when a hiring algorithm systematically disadvantages protected groups (Dastin, 2018)—who bears responsibility? The developer who created the system? The organization that deployed it? The regulator who approved its use? The operator who accepted its recommendation without independent scrutiny?

These questions have proven remarkably difficult to answer within existing legal and institutional frameworks. The “many hands problem” identified in organizational accountability scholarship (Thompson, 1980) is amplified in algorithmic contexts, where causal chains extend across organizational boundaries, temporal phases, and technical layers in ways that systematically frustrate attribution. Traditional legal doctrines—product liability, professional negligence, administrative responsibility—were designed for contexts where human decision-makers could be identified and their reasoning scrutinized. Algorithmic systems disrupt these foundational assumptions, creating what scholars have termed “responsibility

gaps” (Matthias, 2004) or “accountability vacuums” (Bovens, 2007) that persist despite extensive regulatory attention. These systems generate what Beck (1992) characterized as manufactured risks—hazards produced by the very institutions designed to manage them—compounding the challenge of locating responsible agents within increasingly automated governance chains.

The dominant policy response to this challenge has been the “human-in-the-loop” paradigm: requiring human oversight of algorithmic decisions as the primary mechanism for preserving accountability linkages. This approach underlies much contemporary AI regulation, including the EU AI Act’s requirements for “effective human oversight” of high-risk systems (European Parliament, 2024, Art. 14), the OECD Principles on AI emphasizing human control (OECD, 2019), and the U.S. federal government’s emphasis on meaningful human oversight in Executive Order 14110 on the safe and responsible development of artificial intelligence.

Yet as this article demonstrates through structured case analysis, human-in-the-loop requirements systematically fail in practice. They devolve into perfunctory approval processes that provide legal cover without meaningful scrutiny, creating what we term **accountability theater**—the institutional performance of oversight without its substance. A terminological clarification is warranted: by describing human-in-the-loop oversight as a “fiction” and its institutional enactment as “theater,” this article does not imply intentional deception by policymakers or practitioners. Rather, these terms serve as analytical shorthand for a structural condition: the systematic misalignment between formal oversight requirements and the institutional capacity, informational resources, and cognitive conditions necessary for those requirements to produce meaningful review. The fiction is structural, not moral—it describes what *systems do* rather than what *individuals intend*. The structural conditions of contemporary algorithmic deployment—automation bias, information asymmetries, capacity constraints, and institutional pressures to defer to machine outputs—make genuine human oversight the exception rather than the rule.

1.2 The Responsibility Allocation Matrix

This article proposes an alternative approach: the **Responsibility Allocation Matrix (RAM)**—hereafter *Algorithmic RAM* or *A-RAM* where disambiguation from project management usage is necessary. Rather than relying on the fiction of effective generalized human oversight, the RAM systematically specifies who bears what responsibilities at which stages of the algorithmic lifecycle. The framework integrates insights from regulatory state theory (Majone, 1997), instrument choice literature (Howlett, 2011), and the algorithmic regulation paradigm (Yeung, 2018) to construct a comprehensive accountability architecture.

The article makes three distinct contributions. First, it provides a **systematic diagnosis** of why human-in-the-loop approaches fail, identifying five structural failure modes—automation bias, asymmetric information, capacity constraints, institutional pressure, and explainability gaps—that undermine meaningful human oversight regardless of regulatory design quality. Second, it introduces the **Responsibility Allocation Matrix** as a five-actor, four-phase framework that replaces generalized oversight requirements with specific, verifiable obligations for each actor at each lifecycle stage, distinguishing among five types of responsibility (primary, residual, due diligence, advisory, and secondary) and specifying obligation chain principles that prevent responsibility diffusion. Third, it provides a **comparative diagnostic** of four major jurisdictional approaches to responsibility allocation—the EU’s provider-centric model, the US’s deployer-centric model, the UK’s principles-based model, and China’s state-led regulatory model—revealing systematic gaps that the RAM framework is designed to address.

1.3 Negative Scoping

Three clarifications delimit the scope of this article. First, the RAM addresses the *allocation* of responsibility—who bears what obligations—not the *substance* of those obligations. Questions about what

specific technical standards high-risk systems should meet, what transparency requirements should apply, or what oversight mechanisms should be mandated are addressed in the risk regulation literature (Baldwin et al., 2012; Black & Baldwin, 2010) and are beyond the present scope. The RAM provides the *architecture* within which substantive obligations are assigned, not the obligations themselves.

Second, unlike sociological approaches that examine accountability as a cultural or discursive phenomenon (Wieringa, 2020), this article restricts its analysis to the institutional design of responsibility allocation. Our focus is operational: we seek to provide policymakers and regulatory designers with a systematic framework for constructing accountability systems that function rather than a critique of accountability as a normative concept.

Third, the RAM addresses systems that produce or significantly influence consequential decisions affecting individuals' rights, opportunities, or welfare. Algorithmic systems used for purely internal operational optimization, entertainment recommendation, or personal convenience fall outside the regulatory perimeter that the RAM is designed to serve.

1.4 Article Structure

Section 2 analyzes the limitations of human-in-the-loop approaches through theoretical grounding and three empirical accountability failures. Section 3 introduces the RAM framework, specifying its five actors, four phases, and five responsibility types. Section 4 maps the policy instruments available for operationalizing responsibility allocation. Section 5 provides comparative analysis of how four major jurisdictions approach responsibility allocation through the RAM lens. Section 6 discusses institutional implications, implementation challenges, and enforcement design. Section 7 concludes.

2 The Limits of Human-in-the-Loop

2.1 The Human-in-the-Loop Promise

The human-in-the-loop paradigm rests on an intuitive proposition: if humans retain oversight of algorithmic decisions, accountability is preserved through the chain of human agency. Algorithms may recommend, score, rank, or flag, but ultimate authority remains with human decision-makers who can be held responsible for outcomes through established legal and administrative mechanisms.

This paradigm has been immensely influential in AI governance discourse. The European Union's AI Act requires "effective human oversight" for high-risk AI systems, specifying that natural persons assigned to human oversight must be able to "fully understand the capacities and limitations of the high-risk AI system," "properly monitor its operation," and "be able to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output" (European Parliament, 2024, Art. 14). The OECD Principles on AI call for "appropriate human determination" as a foundation for responsible AI deployment (OECD, 2019). The IEEE's Ethically Aligned Design emphasizes "operator and user understanding" as essential for responsible AI systems. National frameworks from Australia to Japan similarly invoke human oversight as the primary accountability mechanism.

The theoretical appeal of this approach is clear. By preserving human agency in the decision chain, the paradigm maintains the conceptual foundations of existing accountability frameworks: identifiable decision-makers, traceable reasoning, and established mechanisms for review and redress. The human-in-the-loop serves as an accountability bridge between algorithmic processes and traditional governance structures (Kaminski & Malgieri, 2021).

2.2 Three Accountability Failures

Despite its conceptual elegance, the human-in-the-loop paradigm has repeatedly failed in practice. We examine three illustrative cases that reveal structural rather than incidental failures.

2.2.1 Case 1: The Dutch Childcare Benefits Scandal (2019–2021)

The Dutch tax authority (Belastingdienst) deployed an algorithmic system to detect fraudulent childcare benefit claims. Human caseworkers were nominally responsible for reviewing algorithmic flags and making final determinations. The system's design created what [Bovens and Zouridis \(2002\)](#) term a “screen-level bureaucracy”: human workers whose role had been reduced from independent judgment to processing algorithmic outputs.

In practice, caseworkers endorsed algorithmic fraud determinations in over 95% of cases without independent verification. The system incorporated discriminatory indicators—including dual nationality as a fraud risk factor—that caseworkers were neither aware of nor equipped to identify. When the scandal emerged following parliamentary investigations and media reporting, approximately twenty-six thousand families had been wrongly accused of fraud, many facing financial devastation. Multiple families lost their homes; relationships collapsed under financial pressure; children were placed in care. The scandal ultimately led to the resignation of the entire Dutch government in January 2021.

The human-in-the-loop requirement had provided formal legal cover without meaningful oversight. An investigating committee found that institutional pressure to process cases efficiently, combined with trust in the algorithmic system's purported accuracy, had transformed caseworkers from independent decision-makers into rubber stamps—what [Amnesty International \(2021\)](#) characterized as “digital welfare state” failure.

2.2.2 Case 2: COMPAS Recidivism Risk Assessment (United States)

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) system, developed by Equivant (formerly Northpointe), provides recidivism risk scores used in bail, sentencing, and parole decisions across multiple U.S. jurisdictions. Judges retain formal authority over these decisions; COMPAS provides recommendation scores on a scale that rates defendants as low, medium, or high risk. ProPublica's investigation revealed systematic racial disparities in the system's predictions: Black defendants were nearly twice as likely as white defendants to be falsely flagged as high risk for future criminal behavior, while white defendants were more likely to be incorrectly rated as low risk ([Angwin et al., 2016](#)).

Subsequent empirical research by [Dressel and Farid \(2018\)](#) demonstrated that COMPAS predictions were no more accurate than those of untrained individuals recruited online, raising fundamental questions about the system's added value. Yet studies of judicial behavior consistently found that judges substantially defer to COMPAS scores, particularly in high-volume proceedings where time constraints preclude independent evaluation. The formal requirement of human judicial judgment failed to provide meaningful accountability because judges lacked the technical capacity to assess algorithmic methodology, the time to conduct independent risk evaluation for every defendant, and the institutional incentive to deviate from algorithmic recommendations that provided the appearance of scientific objectivity.

The COMPAS controversy also exposed a deeper structural problem: the mathematical impossibility of simultaneously satisfying competing fairness criteria. [Kleinberg et al. \(2016\)](#) demonstrated formally that calibration (equal predictive accuracy across groups) and error rate balance (equal false positive and false negative rates across groups) cannot be simultaneously achieved except under conditions that rarely hold in practice. This impossibility result has profound implications for human-in-the-loop oversight: when the system's fairness properties depend on which mathematical definition of fairness is selected—a design choice embedded in the algorithm's architecture—the human overseer is asked to evaluate the

appropriateness of a choice that may not have been disclosed, documented, or even recognized as a choice by the system's developers.

Green and Chen (2019) demonstrated experimentally that human decision-makers presented with algorithmic recommendations exhibit systematic “automation bias”—deferring to machine outputs even when those outputs are demonstrably incorrect. Their findings suggest that the mere presence of an algorithmic recommendation fundamentally alters the cognitive conditions under which human judgment operates, undermining the independence that meaningful oversight requires.

2.2.3 Case 3: Amazon Hiring Algorithm (2014–2017)

Amazon developed an algorithmic hiring tool to screen resumes for technical positions. The system was trained on historical hiring data spanning a ten-year period—data reflecting the technology sector's well-documented gender imbalance. Human recruiters were responsible for final hiring decisions based on algorithmically pre-screened candidate pools.

The algorithm developed systematic bias against women, downgrading resumes containing terms associated with women's colleges (“women's chess club captain”) and penalizing educational credentials from women's institutions. Because recruiters received only pre-screened candidate pools, they were structurally unable to detect the bias: the absence of qualified female candidates from consideration appeared as a natural feature of the applicant pool rather than an artifact of algorithmic filtering (Dastin, 2018).

When Amazon's internal review discovered the bias in 2017, the company discontinued the tool entirely. The case illustrates a particularly insidious form of human-in-the-loop failure: the human decision-maker was functionally present and making genuine decisions, but the algorithmic pre-filtering had already removed the information necessary for those decisions to correct systemic bias. The human was in the loop but the loop had been designed to exclude the evidence that would have triggered meaningful oversight.

2.3 A Taxonomy of Human-in-the-Loop Failures

Analysis of these and similar cases reveals five structural failure modes that undermine human-in-the-loop requirements:

These failure modes are not incidental or correctable through better training, clearer instructions, or more diligent personnel. They reflect structural features of contemporary algorithmic deployment that systematically undermine the conditions for meaningful human oversight. As algorithmic systems become more complex, deployment contexts more demanding, and decision volumes more massive, the gap between formal human oversight requirements and actual human capacity for meaningful review inexorably widens.

The implication is not that human involvement in algorithmic decisions is valueless—it is that *unstructured* human involvement, without specific obligations, verification mechanisms, and institutional supports, provides inadequate accountability. What is needed is not merely a human in the loop but a carefully designed accountability architecture that specifies who bears what responsibilities, provides the information and resources necessary for those responsibilities to be fulfilled, and creates mechanisms for verification and remedy when they are not. Across all three cases examined above, the diagnostic indicators of *accountability theater* were present: override rates below 5%, review times insufficient for meaningful evaluation, and no documented independent reasoning by human reviewers. These convergent patterns confirm that the theater diagnosis is structural rather than incidental—a systemic feature of unstructured human-in-the-loop design rather than a failure of individual diligence.

Table 1: Structural Failure Modes of Human-in-the-Loop Oversight

Failure Mode	Mechanism	Structural Condition	Illustrative Case
Automation Bias	Humans systematically over-rely on algorithmic recommendations, treating machine outputs as authoritative	Bias intensifies with perceived system accuracy and operator fatigue (Skitka et al., 1999)	Dutch caseworkers accepting >95% of algorithmic fraud flags without verification
Asymmetric Information	Humans lack access to training data, model logic, feature weights, or confidence intervals necessary for evaluation	Proprietary models and insufficient documentation create structural information gaps (Burrell, 2016)	Amazon recruiters unaware of resume screening criteria
Capacity Constraints	Time pressure, cognitive load, and caseload volume prevent thorough independent review	High-throughput environments compress review time below meaningful evaluation thresholds	Judges processing COMPAS reports with minutes per decision
Institutional Pressure	Organizational incentives discourage deviation from algorithmic outputs	Deviating creates documentation burdens and exposes decision-makers to scrutiny	Caseworkers facing performance metrics emphasizing throughput
Explainability Gaps	Outputs presented as scores or categories without sufficient rationale for evaluation	Complex models produce outputs without accessible causal explanations	COMPAS providing risk scores without disclosing driving factors

3 Theoretical Framework: The Responsibility Allocation Matrix

3.1 Framework Overview

The Responsibility Allocation Matrix (RAM) provides a systematic approach to designing accountability architectures for algorithmic governance. Rather than relying on generalized requirements for human oversight, RAM specifies concrete obligations for specific actors at specific lifecycle phases, creating an accountability structure that can be verified, monitored, and enforced.

The need for such a framework reflects limitations in existing approaches. Prior scholarship has identified important elements of the accountability challenge: Matthias (2004) articulated the “responsibility gap” created by autonomous learning systems; Raji et al. (2020) developed an end-to-end internal auditing framework that advanced lifecycle thinking; Selbst (2021) provided an institutional perspective on algorithmic impact assessments; and Kaminski and Malgieri (2021) analyzed multi-layered explanation requirements as accountability mechanisms. These contributions are valuable but partial: each addresses specific actors, specific phases, or specific mechanisms without providing a comprehensive architecture that systematically covers all actor-phase combinations with typed obligations and enforcement linkages. The RAM builds on these foundations to provide what existing frameworks lack: a complete five-actor, four-phase matrix with differentiated responsibility types (primary, residual, due diligence, advisory, and secondary) and explicit obligation chain enforcement principles that prevent responsibility diffusion across every cell of the accountability architecture.

RAM comprises two structural dimensions: five actor categories (Designers, Deployers, Regulators, Auditors, and Affected Parties) and four lifecycle phases (Design, Deployment, Operation, and Harm Remediation). The intersection of actors and phases creates a 5×4 matrix of twenty responsibility assignments. Each assignment specifies not merely that an actor has “responsibility” but what *type* of responsibility

applies—a distinction critical for operationalization and enforcement.

Three terminological notes are necessary. First, throughout this framework, *responsibility* refers to an institutionally assigned obligation to perform specific actions, maintain specific standards, or bear specific consequences—not to moral or causal responsibility in the philosophical sense. The RAM is an instrument of institutional design, not a theory of moral agency. Second, we distinguish *responsibility* (obligation to act) from *accountability* (answerability for outcomes of action or inaction) from *liability* (legal exposure to sanctions or compensation). These concepts are related but analytically distinct: an actor may bear responsibility without accountability (where verification mechanisms are absent) or accountability without liability (where legal frameworks do not attach consequences). The RAM primarily addresses responsibility allocation but designs accountability linkages into each cell through verification mechanisms and enforcement pathways, recognizing that responsibility without accountability degenerates into aspiration. Third, we note that the RAM’s responsibility types are not merely descriptive categories but *institutional design specifications*: each type carries distinct implications for monitoring intensity, enforcement mechanisms, evidentiary burdens, and remedial consequences.

Table 2: Core Definitional Concepts of the Responsibility Allocation Matrix

Concept	Definition	Positive Test	Negative Test
Responsibility Allocation	Institutional assignment of specific obligations to identified actors at identified lifecycle phases	Actor can identify specific obligations, verification criteria, and consequences for non-performance	Vague invocations of “responsibility” without specifiable obligations
Obligation Chain	Unbroken sequence of responsibility assignments linking design decisions to operational outcomes and remedy mechanisms	For any algorithmic harm, an identifiable actor with documented obligations exists at each relevant phase	Responsibility “gaps” where no actor bears specified obligations
Accountability Architecture	Complete institutional structure specifying responsibility allocations, verification mechanisms, and remedy pathways	All twenty cells of the RAM matrix are populated with specified obligations and enforcement mechanisms	Selective accountability covering some actors or phases while leaving others unaddressed
Accountability Theater	Institutional performance of oversight without its substance; indicated when override rate <5%, review time below threshold, and no documented independent judgment	Formal requirements satisfied but substantive scrutiny absent; all three diagnostic indicators present	Genuine oversight where override rate reflects independent judgment and decision records document reasoning

3.2 Actor Definitions

The RAM identifies five actor categories. The expansion from the traditional tripartite distinction (developer, user, regulator) to five categories reflects the empirical complexity of algorithmic supply chains and the distinct accountability functions that different institutional actors perform.

Designers are entities that create, develop, or substantially modify the algorithmic systems deployed in consequential decision contexts. This category encompasses not only the organizations that develop algorithmic models but also—critically—the data architects who construct training datasets, the engineers who select and configure model architectures, and the teams that conduct pre-deployment testing and validation. The designer category is defined functionally rather than organizationally: an entity qualifies as

a designer if its decisions materially shape how the algorithmic system processes information and produces outputs.

Deployers are entities that implement algorithmic systems in specific operational contexts, making decisions about where, when, how, and under what conditions algorithmic systems are used. Deployers bear responsibility for contextualizing systems—adapting general-purpose tools to particular decision environments. The deployer category is critical because many algorithmic harms arise not from defective system design but from inappropriate deployment: systems used in contexts, populations, or decision environments for which they were not designed or validated.

Regulators are public authorities with legal mandates for oversight, standard-setting, and enforcement in the domains where algorithmic systems operate. Regulatory responsibility in the RAM is not limited to dedicated AI regulators but encompasses sectoral regulators—financial authorities, health agencies, employment commissions, data protection authorities—whose existing mandates intersect with algorithmic governance.

Auditors are independent entities—whether public, private, or hybrid—authorized to assess algorithmic system compliance with applicable standards, evaluate performance against documented specifications, and verify that responsibility allocations are being honored in practice. The auditor category is analytically distinct from the regulator category: regulators *set standards and enforce compliance*; auditors *assess and verify*. This distinction matters because combining standard-setting, enforcement, and assessment functions in a single entity creates conflicts of interest that undermine accountability effectiveness (Power, 1997).

Affected Parties are individuals, communities, or organizations whose rights, opportunities, or welfare are materially influenced by algorithmic decisions. The affected party category is defined by impact rather than by formal relationship: individuals may be affected parties even if they are unaware that an algorithmic system is involved in decisions affecting them—a condition that is, in fact, common in contemporary algorithmic deployment (Kemper & Kolkman, 2019).

A critical conceptual clarification regarding affected parties in the RAM: the “responsibilities” assigned to affected parties—consultation, participation, reporting, and access—are more precisely understood as **engagement entitlements and civic capacities** rather than legal obligations in the conventional sense. The affected party row in the RAM matrix should be read as specifying what the accountability architecture must *provide to* affected parties, not what it *demand from* them. This distinction is critical because the alternative interpretation—treating affected-party engagement as an obligation whose non-fulfillment diminishes accountability claims—would perversely shift systemic risk onto the populations least equipped to bear it.

3.3 Phase Definitions

The RAM structures the algorithmic lifecycle into four phases. These phases are analytically distinct but empirically overlapping: design decisions continue to shape outcomes during operation, and operational experience may trigger design modifications.

The **Design Phase** encompasses all activities involved in creating the algorithmic system: problem formulation, data collection and curation, model selection and architecture, training and validation, performance evaluation, bias testing, documentation, and pre-deployment safety assessment. Design-phase accountability focuses on *ex ante* quality assurance—ensuring that systems meet specified standards before deployment.

The **Deployment Phase** encompasses the decisions and processes involved in implementing an algorithmic system in a specific operational context: fitness-for-purpose assessment, contextual risk analysis, integration with existing workflows, staff training, pilot testing, authorization, and conformity verification. Deployment-phase accountability focuses on *context-appropriate adaptation*.

The **Operation Phase** encompasses ongoing use of the deployed system: continuous monitoring of outputs, performance assessment against documented specifications, incident detection and response, periodic re-evaluation, and human oversight where applicable. Operational accountability focuses on *continuous vigilance*.

The **Harm Remediation Phase** encompasses institutional responses to adverse outcomes: incident investigation, root cause analysis, affected party notification, compensation or correction, systemic remediation, and accountability proceedings. Remediation-phase accountability focuses on *ex post* response.

3.4 Responsibility Types

A crucial distinction in the RAM framework concerns the *type* of responsibility assigned to each actor at each phase. The RAM distinguishes five responsibility types.

Primary Responsibility designates the actor bearing principal obligation for outcomes at a given phase. Primary responsibility entails both authority and accountability. There should be exactly one actor with primary responsibility at each phase to prevent diffusion.

Residual Responsibility designates continuing obligations that persist beyond an actor's primary phase. Designers, for example, retain residual responsibility for providing technical support, issuing updates, and disclosing newly discovered limitations even after systems are deployed.

Due Diligence Responsibility designates the obligation to investigate and verify before taking action. Deployers must conduct due diligence on designer practices before deployment; regulators must conduct due diligence on both designers and deployers before granting authorization.

Advisory Responsibility designates the obligation to provide guidance, information, and technical support without direct action authority. Designers bear advisory responsibility during the operation phase: they must provide deployers with information necessary for effective oversight.

Secondary Responsibility designates supporting obligations that supplement another actor's primary role. Multiple actors may bear secondary responsibility simultaneously, contributing to accountability outcomes without bearing primary obligation.

3.5 The Complete RAM

The following matrix specifies responsibility allocations across all actor-phase combinations:

3.6 Operational Note

The RAM is not a rigid prescription requiring identical implementation across all algorithmic contexts. It is a **structured decision protocol** that ensures accountability designers explicitly consider each actor-phase combination and document their allocation decisions. In low-risk contexts, some cells may be appropriately addressed through lightweight mechanisms. In high-risk contexts—criminal justice, welfare allocation, healthcare—every cell demands robust institutional mechanisms with verification and enforcement capacity.

The critical design principle is *completeness*: every cell must receive explicit attention, even if the conclusion is that formal regulation is unnecessary for that particular actor-phase combination. Responsibility gaps arise not from deliberate policy choices but from oversight—from failing to consider who is responsible at specific phases.

3.7 Scope Conditions and Failure Modes

Intellectual honesty requires explicit acknowledgment of conditions under which the RAM framework encounters limitations. The RAM is designed for algorithmic systems that produce or significantly influ-

Table 3: The Responsibility Allocation Matrix—Complete Assignment

Actor	Design Phase	Deployment Phase	Operation Phase	Harm Remediation
Designers	Primary: Bias and safety testing; performance validation; comprehensive documentation; training data quality assurance; known limitation disclosure	Residual: Technical support for integration; update provision; disclosure of known limitations; deployment guidance	Advisory: Guidance on operational parameters; training materials; update notifications; technical consultation	Secondary: Evidence preservation; investigation cooperation; technical analysis support; remediation design
Deployers	Due Diligence: Fitness-for-purpose assessment; vendor evaluation; independent performance verification; contextual risk analysis	Primary: Conformity verification; context adaptation; integration testing; staff training; operational policy authorization; fundamental rights impact assessment	Primary: Continuous performance monitoring; human oversight implementation; incident detection and first response; periodic re-assessment	Primary: First response to affected parties; notification; internal investigation; interim remedy provision; corrective action implementation
Regulators	Standard-Setting: Performance requirements; documentation mandates; bias assessment standards; pre-market review criteria	Certification: Pre-market conformity assessment; registration requirements; deployment authorization for high-risk contexts	Supervision: Ongoing compliance monitoring; periodic inspection; enforcement action; market surveillance	Enforcement: Investigation of systemic failures; sanction determination; correction orders; regulatory guidance updates
Auditors	Assessment: Independent evaluation of design documentation; testing methodology review; bias verification	Verification: Independent confirmation of conformity assessment; deployment readiness evaluation	Monitoring: Periodic independent performance assessment; compliance verification; outcome audit	Investigation: Independent root cause analysis; impact assessment; remediation adequacy evaluation
Affected Parties	Consultation: Input on design requirements; representative testing and feedback	Participation: Engagement in pilot programs; informed consent; feedback mechanisms	Reporting: Access to complaint mechanisms; outcome flagging and challenge procedures	Access: Meaningful access to redress; explanation of decisions; appeal processes; compensation

ence consequential decisions within identifiable institutional contexts. Several deployment configurations challenge these assumptions.

Extreme Decentralization. When algorithmic systems are distributed through fully decentralized architectures—open-source models deployed by thousands of independent actors, federated learning systems, or blockchain-based autonomous systems—the RAM’s actor identification requirement becomes difficult to satisfy. In such cases, the framework requires complementary institutional mechanisms: mandatory registration requirements, platform liability provisions, or technological governance solutions.

Micro-Organizations and Resource Constraints. Small and medium enterprises deploying algorithmic tools developed by others may lack the technical expertise, financial resources, or organizational capacity to fulfill the full range of deployer obligations. Regulatory implementation of the RAM must therefore incorporate tiered compliance pathways that maintain accountability substance while adjusting procedural requirements to organizational capacity.

Cross-Border Fragmentation. When algorithmic supply chains span multiple jurisdictions, the RAM’s obligation chain may fragment across legal systems with incompatible liability rules, divergent regulatory standards, and limited enforcement cooperation. This limitation underscores the need for international regulatory coordination mechanisms.

Rapidly Evolving Systems. Algorithmic systems that undergo continuous learning, frequent retraining, or autonomous adaptation may shift between RAM phases faster than institutional oversight mechanisms can track. The fundamental challenge of governing systems that evolve faster than their overseers can assess remains an open problem.

These limitations do not invalidate the RAM framework but define its boundary conditions. The RAM is a necessary but not sufficient condition for algorithmic accountability.

4 Policy Instruments for Responsibility Allocation

4.1 Instrument Types

Operationalizing responsibility allocation requires selecting appropriate policy instruments—the mechanisms through which obligations are imposed, monitored, and enforced. Building on the instrument choice literature (Hood, 1983; Howlett, 2011), we identify four categories of instruments relevant to algorithmic responsibility allocation.

Information-Based Instruments operate through knowledge: transparency requirements, mandatory disclosure, explainability standards, algorithmic impact assessments, public registries, and mandatory reporting obligations. These instruments support accountability by enabling verification: actors cannot be held responsible for algorithmic outcomes if neither they nor oversight bodies have access to information about how algorithmic systems function, what decisions they influence, and what impacts they produce. Information instruments serve both direct accountability functions (enabling affected parties to understand and challenge decisions) and indirect functions (enabling regulators, auditors, and the public to evaluate compliance with substantive requirements).

Regulatory Instruments operate through legal obligation: binding performance standards, licensing and certification requirements, prohibited practices, mandatory human oversight specifications, conformity assessment procedures, and operational requirements. These instruments directly specify what actors must and must not do, providing the legal infrastructure for accountability enforcement. The effectiveness of regulatory instruments depends critically on enforcement capacity—as Black (2001) observes, “regulation without enforcement is merely advice.” The RAM framework accordingly specifies enforcement mechanisms alongside substantive requirements in each cell.

Economic Instruments operate through financial incentives: liability rules, mandatory insurance, performance bonds, tax incentives, procurement conditions, funding requirements, research subsidies, and market access conditions. These instruments align institutional interests with accountability objectives by creating financial consequences for irresponsible behavior and financial rewards for responsible practices.

Institutional Instruments operate through organizational design: establishing specialized regulatory bodies, requiring internal governance structures (such as algorithmic impact review committees), mandating organizational processes (such as ethics review), specifying professional standards and qualifications, and creating inter-institutional coordination mechanisms. These instruments shape the organizational context within which algorithmic systems are developed, deployed, and governed.

4.2 Instrument Selection Principles

The selection of appropriate instruments for each RAM cell follows several principles derived from the regulatory design literature.

The *matching principle* requires that instrument intensity match the risk level and consequence severity of the algorithmic application. High-risk systems demand regulatory instruments with enforcement capacity; lower-risk systems may be adequately governed through information instruments and self-regulatory mechanisms.

The *complementarity principle* recognizes that effective accountability typically requires instrument combinations rather than single instruments. Transparency requirements (information instruments) are insufficient without standards specifying what constitutes acceptable performance (regulatory instruments), mechanisms for verifying compliance (institutional instruments), and consequences for non-compliance (economic instruments).

The *proportionality principle* requires that specific instruments within each category be calibrated to the seriousness of potential harms and the capacity of regulated entities. Imposing pharmaceutical-grade test-

ing requirements on low-risk algorithmic applications would be disproportionate and counterproductive; applying self-certification to high-risk criminal justice algorithms would be dangerously insufficient.

The *enforceability principle* requires that selected instruments be practically enforceable given available institutional capacity. Sophisticated regulatory requirements that cannot be effectively monitored or sanctioned are worse than simpler requirements that can: they create the appearance of governance without its substance—precisely the accountability theater that the RAM framework is designed to prevent.

4.3 Mapping Instruments to RAM Cells

The following table maps instrument categories to RAM cells, illustrating how each actor-phase combination can be operationalized through specific policy tools:

Table 4: Policy Instrument Mapping for RAM Implementation

	Information	Regulatory	Economic	Institutional
Designer–Design	Technical documentation; model cards; data sheets	Performance standards; bias testing mandates; prohibited practices	Liability for design defects; R&D incentives	Internal review processes; ethics board requirements
Deployer–Deploy	Fundamental rights impact assessment; public notice	Conformity certification; deployment authorization	Insurance requirements; performance bonds	Human oversight bodies; deployment review committees
Deployer–Operate	Continuous monitoring reports; incident disclosure	Operational standards; re-assessment mandates	Operational liability; ongoing insurance	Monitoring teams; periodic audit obligations
Regulator–All	Public registries; regulatory guidance; annual reports	Binding rules; enforcement actions; correction orders	Fines; market access conditions	Specialized AI authorities; coordination mechanisms
Auditor–All	Published audit reports; methodology disclosure	Auditor accreditation; independence standards	Audit liability; insurance; professional indemnity	Professional qualification requirements; independence rules
Affected Parties	Right to explanation; access to complaint and appeal mechanisms; participation in design consultations	Redress procedures; representative standing	Compensation mechanisms; collective redress funds	Ombudsman offices; advocacy organization recognition

5 Comparative Analysis: Responsibility Allocation Across Jurisdictions

5.1 Purpose and Method

This section applies the RAM framework as a diagnostic lens to analyze how four major jurisdictions allocate responsibility for algorithmic governance. The analysis does not attempt comprehensive regulatory comparison—a task that would require substantially greater scope—but uses the RAM’s actor-phase structure to identify systematic patterns of emphasis and neglect in each jurisdiction’s approach to accountability architecture.

5.2 European Union: Provider-Centric Model

The EU AI Act ([European Parliament, 2024](#)) establishes the most comprehensive legislative framework for AI governance currently in force. Through the RAM lens, its most distinctive feature is the concentration of primary responsibility on system providers (analogous to the RAM’s designers), with deployers (“users” in European legislative terminology) bearing secondary obligations.

In the design phase, the AI Act imposes extensive requirements on providers of high-risk AI systems: quality management systems, risk management, data governance, technical documentation, record-keeping, transparency, human oversight design, accuracy and robustness, and cybersecurity requirements

(Arts. 9–15). These requirements correspond closely to the RAM’s designer–design cell, creating the most detailed set of design-phase obligations in any current jurisdiction.

In the deployment phase, the AI Act assigns deployers specific but more limited obligations: using systems in accordance with instructions, ensuring human oversight by competent individuals, monitoring operations, and reporting serious incidents (Art. 26). While these obligations are meaningful, they are substantially less detailed than provider obligations—creating what the RAM framework identifies as *deployer underspecification*: the systematic under-development of deployer obligations relative to the complexity of deployment decisions.

In terms of harm remediation, the EU framework relies on a combination of the AI Act’s enforcement provisions (including fines of up to 35 million euros or 7% of global turnover) and general Member State liability rules. The European Commission has also proposed a separate AI Liability Directive to establish harmonized civil liability rules. The remediation architecture is partially developed: enforcement mechanisms exist but the specific relationship between AI Act violations and civil liability remains under construction.

The EU model exhibits strengths in design-phase completeness and regulatory institutional architecture. Its principal RAM gaps appear in deployer obligations (which remain generic rather than context-specific), auditor independence (which lacks detailed institutional specification), and affected party engagement (which is primarily addressed through transparency rather than participation).

5.3 United States: Deployer-Centric Model

In contrast to the EU’s comprehensive legislative approach, the United States has pursued sector-specific regulation supplemented by executive action. The now-rescinded Executive Order 14110 (Executive Office of the President, 2023) provided a framework-level approach; its January 2025 rescission (Executive Office of the President, 2025) underscores the fragmented nature of the US approach.

Through the RAM lens, the US model’s most distinctive feature is its emphasis on deployer responsibility through existing sectoral frameworks: employment discrimination law (Title VII and the EEOC), fair lending requirements (ECOA and the CFPB), healthcare regulation (FDA oversight of AI/ML-based medical devices), and consumer protection (FTC Section 5). These frameworks place primary accountability on organizations that *use* algorithmic systems to make consequential decisions, regardless of where those systems were developed.

This deployer emphasis creates a distinctive RAM profile: relatively strong deployer–operation and deployer–remediation obligations within regulated sectors, but weak designer obligations (no comprehensive pre-market requirements for general-purpose AI systems) and incomplete coverage (sectors without specific AI-relevant legislation lack systematic provision).

The NIST AI Risk Management Framework (NIST, 2023) provides voluntary guidance that partially addresses design-phase gaps, but its non-binding nature means that designer obligations remain aspirational rather than enforceable. The result is a RAM architecture that is strong in some cells (particularly deployer–operation within regulated sectors) but systematically incomplete, with coverage depending on whether algorithmic applications fall within existing regulatory perimeters.

5.4 United Kingdom: Principles-Based Model

The UK has adopted a principles-based approach to AI governance, eschewing sector-specific AI legislation in favor of a “pro-innovation” regulatory framework. The approach delegates responsibility to existing sectoral regulators to apply five cross-cutting principles—safety, transparency, fairness, accountability, and contestability—within their respective domains (UK DSIT, 2023).

Through the RAM lens, this approach creates a distinctive pattern: responsibility allocation is implicit rather than explicit, mediated through regulatory interpretation rather than legislative prescription. The RAM reveals this approach’s fundamental challenge: when responsibility allocation depends on existing sectoral regulators interpreting general principles, the completeness and consistency of the resulting accountability architecture depends entirely on the capacity, expertise, and willingness of each regulatory body to develop AI-specific accountability mechanisms.

Early experience suggests significant variation: the Information Commissioner’s Office and Financial Conduct Authority have developed substantial AI guidance, while other regulators have been less active. The result is a RAM architecture characterized by sectoral unevenness—potentially strong in some domains (data protection, financial services) but systematically underspecified in others.

5.5 China: State-Led Regulatory Model

China has developed a distinctive algorithmic governance framework through a series of targeted regulations: the Internet Information Service Algorithmic Recommendation Management Provisions (2022), the Deep Synthesis Provisions (2023), and the Interim Measures for Generative AI Services (2023). These regulations create a governance architecture that, through the RAM lens, exhibits two distinctive features.

First, the Chinese framework places substantial responsibility on *platform operators*—entities that occupy both designer and deployer roles in the RAM framework—creating a unified provider-deployer obligation that simplifies the supply chain accountability challenge but may obscure the distinct accountability functions that design and deployment involve. Second, the framework incorporates significant *regulatory pre-approval* requirements, including algorithmic filing and security assessment procedures that place regulators in an active design-phase role unusual in Western governance models. This creates strong regulator–design and regulator–deployment cells in the RAM but raises questions about regulatory capacity and independence.

In terms of affected party obligations, the Chinese framework includes user rights provisions—including rights to refuse algorithmic personalization and to request explanation—that partially address the RAM’s affected party engagement requirements. However, the institutional mechanisms for exercising these rights remain less developed than the formal provisions suggest.

5.6 Comparative Diagnostic

The following table summarizes the comparative analysis, using the RAM to identify each jurisdiction’s relative strengths and gaps:

No current jurisdiction fully implements the comprehensive RAM architecture. Each has developed strengths in certain actor-phase combinations while systematically under-developing others. The RAM framework makes these patterns visible and provides a template for systematic gap remediation.

Table 5: Comparative RAM Gap Analysis: Four Jurisdictions

Jurisdiction	RAM Strengths	RAM Gaps	Structural Pattern
EU (AI Act)	Comprehensive designer–design obligations; strong regulatory institutional architecture; detailed risk classification	Deployer underspecification; auditor independence unclear; affected party participation limited to transparency requirements	Provider-centric: strong upstream, weaker downstream
US (Sectoral)	Strong deployer–operation obligations within regulated sectors; established enforcement through existing agencies	Designer obligations absent outside product-specific regulation; sectoral gaps create coverage inconsistencies	Deployer-centric: strong in regulated sectors, fragmented overall
UK (Principles)	Flexibility to adapt to sectoral context; avoids premature prescription	Responsibility allocation implicit rather than explicit; sectoral unevenness; enforcement capacity uncertain	Principles-based: potentially adaptive, currently underspecified
China (State-Led)	Strong regulatory pre-approval mechanisms; unified provider–deployer obligations; user rights provisions	Auditor independence unclear; conflation of designer–deployer may obscure accountability boundaries; enforcement transparency limited	State-led: strong regulatory presence, accountability chain less transparent

6 Discussion: Institutional Implications and Implementation Challenges

6.1 From Accountability Theater to Accountability Architecture

The cases analyzed in Section 2 demonstrate a consistent pattern: human-in-the-loop requirements, absent institutional specification of who bears what obligations with what verification mechanisms, systematically devolve into accountability theater. The RAM framework addresses this pattern by replacing generalized oversight requirements with specific, verifiable obligations distributed across the five actors and four phases.

This shift has three key institutional implications. First, it transforms accountability from a binary condition (“is there human oversight?”) into a structured assessment (“are identified actors fulfilling specified obligations at each lifecycle phase with documented verification?”). This transformation makes accountability auditable, enforceable, and improvable—properties that binary human-in-the-loop requirements lack. Second, it requires institutional investment in accountability infrastructure: the organizational structures, monitoring capabilities, professional expertise, and enforcement mechanisms necessary for meaningful oversight. Ensuring a human is “in the loop” is inexpensive; building genuine accountability architecture requires sustained institutional commitment. Third, it shifts the locus of regulatory attention from the moment of algorithmic decision to the full lifecycle—from design through deployment, operation, and remediation. This lifecycle perspective reflects the reality that accountability failures typically originate not at the moment of decision but in prior design choices, deployment decisions, or monitoring lapses that created conditions for harm.

6.2 Enforcement Design Principles

The RAM framework’s effectiveness depends critically on enforcement design. We propose four principles for enforcement mechanisms that support meaningful accountability.

The *chain of obligation principle* requires that for any algorithmic harm, an unbroken chain of responsibility assignments links the harm to identified actors with specified obligations. If a responsibility gap exists—if it is possible to identify an algorithmic harm for which no actor bears documented obligations—the accountability architecture is defective and must be remediated. This principle operationalizes the RAM’s completeness requirement: every cell of the matrix must be populated, and the resulting obligation chains must be continuous.

The *verification independence principle* requires that the entities verifying compliance with RAM obliga-

tions be institutionally independent from the entities whose compliance they assess. This principle supports the RAM's distinction between regulators and auditors: combining standard-setting, enforcement, and assessment in a single entity creates conflicts that undermine accountability effectiveness.

The *proportionate enforcement principle* requires that sanctions for accountability failures be calibrated to the severity of harm, the degree of departure from specified obligations, and the actor's capacity for compliance. This principle supports graduated enforcement: warnings for minor procedural lapses, correction orders for systematic compliance failures, and significant sanctions for deliberate non-compliance or reckless disregard of specified obligations.

The *remediation priority principle* requires that enforcement mechanisms prioritize actual remedy for affected parties over punitive action against responsible actors. While sanctions serve important deterrence functions, enforcement systems that focus on institutional punishment while leaving affected parties without remedy reproduce the accountability theater the RAM is designed to prevent—substituting the performance of enforcement for the substance of accountability.

6.3 Dynamic Responsibility and Temporal Challenges

The RAM framework's four-phase structure introduces a temporal dimension to responsibility allocation that creates both opportunities and challenges. The temporal dimension enables lifecycle accountability: by specifying obligations at each phase, the RAM ensures that accountability does not begin and end at the moment of decision but extends forward to remediation and backward to design.

However, the temporal dimension also creates challenges. Algorithmic systems do not progress linearly through discrete phases; they undergo continuous modification, retraining, and adaptation that blur phase boundaries. A system may simultaneously be in the operation phase for its current version and the design phase for its next iteration. The RAM framework addresses this challenge through the concept of *residual responsibility*: designers retain obligations during deployment and operation phases, and deployers retain obligations during remediation.

The greater challenge is *temporal responsibility diffusion*—the tendency for accountability to dissipate over time as the connection between design decisions and operational harms becomes increasingly attenuated. When a bias embedded in training data produces discriminatory outcomes years after deployment, the temporal distance between the design decision and the harm makes attribution difficult and enforcement uncertain. The RAM addresses temporal diffusion through documentation requirements (preserving the evidentiary basis for attribution), residual responsibility provisions (maintaining designer obligations beyond deployment), and periodic re-assessment mandates (requiring deployers to continuously evaluate system performance against documented specifications).

6.4 Obligation Chain Principles in Practice

Implementing the RAM's obligation chain requires three operational principles that translate the framework's abstract architecture into enforceable institutional practice.

The **single-primary principle** requires that exactly one actor bears primary responsibility at each lifecycle phase. This prevents the diffusion that occurs when multiple actors share “responsibility” without any bearing identifiable primary obligations. The single-primary principle does not preclude collaboration; it ensures that accountability has an identifiable focus at each phase.

The **residual persistence principle** requires that primary responsibility at one phase creates residual obligations at subsequent phases. Designers do not shed accountability when they transfer systems to deployers; deployers do not shed accountability when they transfer incident response to regulators. Residual obligations persist until explicitly discharged through documented processes.

The **gap prohibition principle** requires that every cell of the RAM matrix receive explicit attention in accountability design. The principle does not require that every cell be subject to formal regulation—some actor-phase combinations may appropriately be addressed through voluntary mechanisms. It requires that the decision to leave a cell less formally regulated be a deliberate, documented choice rather than an oversight.

7 Conclusion

This article has argued that the dominant approach to algorithmic accountability—the human-in-the-loop paradigm—is structurally inadequate. Through analysis of three consequential accountability failures (the Dutch childcare benefits scandal, COMPAS recidivism assessment, and Amazon’s hiring algorithm), we identified five structural failure modes that undermine meaningful human oversight regardless of regulatory design quality: automation bias, asymmetric information, capacity constraints, institutional pressure, and explainability gaps.

In response, we proposed the Responsibility Allocation Matrix (RAM) as a systematic framework for designing accountability architectures. The RAM’s five-actor, four-phase structure, combined with its five responsibility types (primary, residual, due diligence, advisory, secondary) and three obligation chain principles (single-primary, residual persistence, gap prohibition), provides a comprehensive template for accountability design that replaces unstructured oversight requirements with specific, verifiable obligations.

Our comparative analysis revealed that no current jurisdiction fully implements the RAM architecture. The EU’s provider-centric model creates comprehensive design-phase obligations but under-specifies deployer responsibilities. The US’s deployer-centric model creates strong operational obligations within regulated sectors but lacks comprehensive design-phase requirements. The UK’s principles-based model offers adaptive flexibility at the cost of specificity and consistency. China’s state-led model creates strong regulatory presence but conflates designer and deployer roles while raising questions about auditor independence.

These findings have three implications for the future of algorithmic governance. First, policymakers should move beyond the human-in-the-loop paradigm toward structured accountability architectures that specify who bears what obligations at which lifecycle phases. Second, the RAM provides a diagnostic tool for identifying systematic gaps in existing governance arrangements—gaps that narrative analysis of regulatory texts may not reveal. Third, the framework underscores that meaningful algorithmic accountability is not a matter of adding a human reviewer to an automated process but of building institutional architectures that distribute, monitor, and enforce specific obligations across the full lifecycle of algorithmic deployment.

The ultimate test of any accountability framework is not whether it provides satisfying theoretical architecture but whether it prevents the next Dutch childcare scandal, the next COMPAS-enabled discriminatory sentencing, the next Amazon-style automated exclusion. The RAM provides necessary institutional infrastructure for that prevention—but its success depends on the political will and institutional investment required to translate systematic accountability design into operational reality.

References

- Amnesty International. (2021). *Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal*. Amnesty International.
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*, May 23, 2016.
- Baldwin, R., Cave, M., & Lodge, M. (2012). *Understanding regulation: Theory, strategy, and practice* (2nd ed.).

Oxford University Press.

- Beck, U. (1992). *Risk society: Towards a new modernity*. Sage Publications.
- Black, J. (2001). Decentring regulation: Understanding the role of regulation and self-regulation in a “post-regulatory” world. *Current Legal Problems*, 54(1), 103–146.
- Black, J., & Baldwin, R. (2010). Really responsive risk-based regulation. *Law & Policy*, 32(2), 181–213.
- Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. *European Law Journal*, 13(4), 447–468.
- Bovens, M., & Zouridis, S. (2002). From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control. *Public Administration Review*, 62(2), 174–184.
- Burrell, J. (2016). How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1–12. <https://doi.org/10.1177/2053951715622512>
- Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, October 10, 2018.
- Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), eaao5580. <https://doi.org/10.1126/sciadv.aao5580>
- Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin’s Press.
- European Parliament. (2024). 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). *Official Journal of the European Union*, L 2024/1689.
- Executive Office of the President. (2023). Executive Order 14110: Safe, secure, and trustworthy development and use of artificial intelligence. *Federal Register*, 88(210), 75191–75226.
- Executive Office of the President. (2025). Executive Order: Removing barriers to American leadership in artificial intelligence. *Federal Register*, January 20, 2025.
- Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), Article 50.
- Hood, C. C. (1983). *The tools of government*. Macmillan.
- Howlett, M. (2011). *Designing public policies: Principles and instruments*. Routledge.
- Kaminski, M. E., & Malgieri, G. (2021). Multi-layered explanations from algorithmic impact assessments in the GDPR. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 68–79.
- Kemper, J., & Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*, 22(14), 2081–2096.
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *Proceedings of Innovations in Theoretical Computer Science (ITCS 2017)*, LIPIcs, 67, Article 43.
- Majone, G. (1997). From the positive to the regulatory state: Causes and consequences of changes in the mode of governance. *Journal of Public Policy*, 17(2), 139–167.
- Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175–183.
- National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce.
- OECD. (2019). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449). Organisation for Economic Co-operation and Development.
- Power, M. (1997). *The audit society: Rituals of verification*. Oxford University Press.
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal

- algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44.
- Selbst, A. D. (2021). An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology*, 35(1), 117–191.
- Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5), 991–1006.
- Thompson, D. F. (1980). Moral responsibility of public officials: The problem of many hands. *The American Political Science Review*, 74(4), 905–916.
- UK Department for Science, Innovation and Technology. (2023). *A pro-innovation approach to AI regulation* (Cm 815). His Majesty's Stationery Office.
- Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 1–18.
- Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance*, 12(4), 505–523.

Academic Editor *Xin Huang* cpcl@cpcl.hk