

Scaling, Lock-In, and Proxy Compliance: A Political Economy of Responsible AI

Florian A. D. Burnat¹, Brittany I. Davidson¹

¹University of Bath, Bath, UK
fadb20@bath.ac.uk, bid23@bath.ac.uk

Abstract

AI accountability at scale is an institutional problem: who can observe, verify, and change deployed systems. We develop a sequential political-economy model in which an AI vendor chooses auditability and substantive mitigation, a deployer monitors after adoption while facing switching costs, and enforcement depends on verifiable evidence. Anticipating the deployer’s monitoring response, the vendor may stop at an observable procurement floor while mitigating below the social first best, producing a *proxy-compliance equilibrium*. We characterize the unique interior equilibrium and the corner in which harm is fully mitigated. Independent audit rights raise enforcement exposure directly; portability restores deployer leverage; incident reporting adds a regulator-visible evidence channel; and outcome-linked liability creates incentives that do not depend on vendor-controlled detection. The results explain why documentation and standardized evaluations can coexist with persistent post-deployment harms, and generate testable implications for monitoring, mitigation, and the gap between formal compliance and operational outcomes.

1 Introduction

Motivating puzzle. Responsible AI checklists, model cards, conformity documentation, and standardized evaluations have become common; however, recurring harms and accountability gaps remain. Four observations sharpen this puzzle. *First*, the EU AI Act places extensive documentation and risk-management obligations on providers, but many relevant inputs—training data, evaluation protocols, red-team logs, and access conditions—remain under provider control (Veale and Zuiderveen Borgesius 2021; Schuett 2023). *Second*, independent reproduction of vendor-reported safety and fairness claims is often infeasible under black-box or contractually restricted access (Mitchell et al. 2019; Mokander et al. 2023; Casper et al. 2024). *Third*, audits can be scoped around negotiated data, metrics, and publication rights rather than independently imposed tests (Raji and Buolamwini 2019; Costanza-Chock, Raji, and Buolamwini 2022). *Fourth*, once a deployer has integrated a model into workflows, fine-tuning pipelines, prompt libraries, and downstream tooling, switching providers can be costly. Evidence from frontier-AI firms’ voluntary commitments is consistent with the resulting gap

between visible reporting and underlying performance (Wang et al. 2025). The puzzle is therefore not merely why organizations fail to implement an otherwise complete ethics framework, but why visible compliance and limited mitigation may coexist as rational choices.

Thesis. We argue that proxy compliance can arise when enforcement requires verifiable evidence, the vendor controls part of the evidence-generating infrastructure, and post-integration switching costs weaken the deployer’s recourse. Lock-in reduces monitoring; weaker monitoring lowers verifiable detection; and lower enforcement exposure reduces the vendor’s return to substantive mitigation. Lock-in is an amplifier rather than a necessary primitive; constrained independent detection can sustain the mechanism even when switching costs are moderate.

Contributions. First, we develop a vendor–deployer model in which auditability, mitigation, and monitoring are endogenous, conditional on institutional parameters for procurement, switching costs, and enforcement. Second, we solve the sequential game while allowing the vendor to anticipate how mitigation changes the deployer’s monitoring response. This yields a unique interior equilibrium, a transparent full-mitigation corner, and policy comparative statics. Third, we connect the formal levers to current governance institutions and specify empirical measures and settings in which the mechanism can be tested. The contribution is a tractable institutional explanation of the gap between observable responsibility signals and harm reduction, not a claim that any single governance instrument is sufficient. In this paper, scale is the institutional setting rather than an endogenous variable; the formal analysis isolates how auditability, switching costs, and evidence-dependent enforcement shape mitigation once AI systems are embedded in organizational workflows.

Roadmap. Section 2 positions the argument. Section 3 introduces the model. Section 4 characterizes the equilibrium. Section 5 studies policy levers and welfare. Section 6 develops empirical implications, and Section 7 discusses the scope and limitations.

2 Related Work

This argument relates to responsible AI governance, algorithmic auditing, the economics of switching costs and lock-in,

Result	Formal claim	Institutional interpretation
Prop. 1	weak baseline detection plus high lock-in yields proxy compliance	visible compliance can coexist with low mitigation
Prop. 2	more independent detection or less lock-in raises mitigation	accountability needs evidence and credible recourse
Thm. 1	sufficiently strong audit rights work at any lock-in level	independent evaluation can bypass deployer weakness
Prop. 3	portability raises mitigation through monitoring	standards restore deployer leverage
Prop. 4	incident reporting adds an independent evidence channel	credible reports reduce dependence on vendor-gated access
Prop. 5	outcome-linked exposure raises mitigation independently of detection	liability retains bite when verification is weak

Table 1. Main results and policy interpretation. Formal derivations are in the appendices.

and the political economy of regulation under asymmetric information.

Responsible AI as institutional practice. A growing body of work documents the gap between Responsible AI principles and practice. Jobin, Ienca, and Vayena (2019) survey 84 AI ethics guidelines, finding convergence on high-level principles but divergence in interpretation and a near-universal absence of enforcement mechanisms. This landscape has been characterized as “ethics washing”—voluntary frameworks that create reputational benefits without constraining behavior (Hagendorff 2020). Raji et al. (2022a) sharpen the diagnosis by cataloging functionality failures that no current audit regime reliably detects, and ML research incentives empirically select against harm-reduction methodology (Birhane et al. 2022)—both findings consistent with a system in which signals of responsibility are easier to produce than the underlying substance. Crawford (2021) maps the political economy of AI production, documenting how concentration of compute, data, and infrastructure creates the power asymmetries that enable this dynamic. At the organizational level, internal accountability practices are shaped more by what is demonstrable to external audiences than by what effectively reduces harm (Raji et al. 2020; Moss et al. 2021). Selbst et al. (2019) identify “abstraction traps”—when responsible AI is operationalized through portable, formalized metrics detached from deployment context, it produces the appearance of compliance without addressing situated harms. In our model, proxy compliance is the equilibrium where observable signals are cheaper than substantive mitigation and enforcement depends on verifiable evidence.

Signaling and corporate social responsibility. The mechanism we label *proxy compliance* has a well-developed formal ancestor in the economics of signaling and corporate social responsibility. Bénabou and Tirole (2010) decompose visibly responsible behavior into intrinsic, material, and reputational-signaling motives, and show that in equilibrium the observable action can dominate the unobservable effort when audiences condition on the action alone—even when the two are conceptually distinct. This is the structure we specialize: when deployers, regulators, and civil society condition on *observable* documentation and standardized evaluations, vendors face an incentive problem in which producing the visible signal is strategically separable from substantive harm reduction. The contribution relative to Bénabou and Tirole (2010) is domain-specific: we embed the signaling mechanism in a

vendor–deployer–regulator game with switching costs and evidence-based enforcement, and derive comparative statics for policy levers that are specific to the AI value chain (audit rights, portability, incident reporting, and outcome-linked liability) rather than generic CSR. Recent empirical work on frontier-AI firms’ voluntary White House commitments documents widespread partial compliance and reporting-via-marketing (Wang et al. 2025)—a pattern our equilibrium predicts when audit rights and portability are weak.

Audits, evaluation, and verifiability. Algorithmic auditing has been proposed as a key accountability mechanism (Sandvig et al. 2014; Raji and Buolamwini 2019); however, it faces structural limitations when vendors control access. Mökander et al. (2022) provide a taxonomy of audit types that underwrites our auditability variable t : their distinction between vendor self-assessment and third-party conformity assessment maps directly onto the detection function $p(t, m)$. The auditing ecosystem itself lacks standardization, independence guarantees, and enforcement power (Costanza-Chock, Raji, and Buolamwini 2022); Floridi et al. (2022) propose an operational conformity-assessment procedure (capAI) whose feasibility rests on the same vendor cooperation that our model treats as endogenous. For large language models, auditability constraints are severe (Mokander et al. 2023): model opacity, generality, and absent harm benchmarks. Documentation standards such as model cards (Mitchell et al. 2019) are necessary but insufficient—they are only as reliable as the information vendors choose to include, making them vulnerable to strategic disclosure. Wachter, Mittelstadt, and Russell (2021) provide the doctrinal ground for this limitation: automated compliance checks are evidentially incomplete substitutes for contextual legal assessment, which justifies treating $p(t, m)$ as a proxy rather than a truth function. In our framework, these observations correspond to the detection probability $p(t, m)$ being substantially vendor-determined, giving vendors scope to limit auditability and thereby constrain enforcement. Casper et al. (2024) show that black-box API access is *insufficient* for rigorous audits of frontier systems—internal access to weights, training data, and evaluation logs is required for safety and fairness claims to be credibly tested. Consistent with this, Burnat and Davidson (2026) document concrete “audit blind-spots” across major platforms (X, Reddit, TikTok, Meta) where API restrictions block the independent verification that transparency mandates presume—an empirical instance of vendor-controlled

$p(t, m)$. Raji et al. (2022b) argue on this basis for a designed third-party audit ecosystem with enforcement powers, and Schuett (2025) argues for a standing internal-audit function inside frontier-AI firms with direct board access. Our model treats the auditability variable t as precisely the lever these authors identify: moving from vendor-gated black-box evaluation toward independent internal access raises $p(t, m)$ at the margin and, under conditions we characterize, tips the equilibrium from proxy compliance to accountability.

Market structure, lock-in, and bargaining power. The economics of switching costs provides the foundation for our lock-in mechanism. Switching costs give firms ex post market power over locked-in customers (Klemperer 1995), and dynamic competition with switching costs leads to an “invest then harvest” pattern (Farrell and Shapiro 1988). Network externalities create self-reinforcing concentration (Katz and Shapiro 1985), which in AI manifests as ecosystem lock-in around dominant foundation-model providers; foundation-model scale concentrates development capacity among a few well-resourced actors (Bender et al. 2021), supplying the market-structure precondition for the switching cost s we formalize. “Openness” in AI is often strategically deployed to consolidate rather than democratize power (Widder, West, and Whittaker 2023), and regulatory compliance costs can paradoxically reinforce concentration (Engler 2023). Procurement frameworks are poorly suited to evaluating AI systems (Mulligan and Bamberger 2019), weakening deployer-side discipline even before lock-in takes effect. Our model integrates these insights: switching cost s dampens effective recourse from monitoring, removing the market discipline that would otherwise incentivize vendor mitigation.

Evidence-based enforcement. The political economy of regulation under asymmetric information is well established: regulation may be captured by industry (Stigler 1971); principal-agent rents distort outcomes when information is asymmetric (Laffont and Tirole 1993); and optimal enforcement institutions depend on the power asymmetry between regulators and firms (Glaeser and Shleifer 2003). Intermediary auditors can themselves be captured, reducing rather than raising oversight effectiveness (Tirole 1986). The choice between liability and safety regulation depends on information structures (Shavell 1984)—directly relevant to our comparison of evidence-based enforcement versus outcome-linked liability. In the AI context, the Act’s heavy reliance on vendor self-assessment and harmonized-standard compliance is well documented, as is its broader risk-management architecture (Veale and Zuiderveen Borgesius 2021; Schuett 2023). Subsequent 2024–2025 scholarship has identified three structural weaknesses. First, scholarship on the AI Office, AI Board, and Scientific Panel flags under-resourcing relative to the Act’s evidentiary burden (Hacker, Engel, and Mauer 2023; Novelli et al. 2025). Second, harmonized-standard delegation routes the regulator’s burden of proof through vendor-influenced standards bodies (Gornet and Maxwell 2024; Laux, Wachter, and Mittelstadt 2024). Third, the Fundamental Rights Impact Assessment adds a rights-based layer above conformity assessment, and incorporating civil society into the audit ecosystem can guard against capture (Mantelero

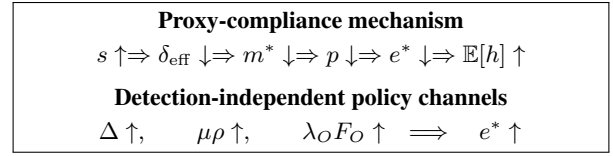


Figure 1. Lock-in weakens the deployer-monitoring channel. Audit rights Δ , effective incident-reporting coverage ρ with sanction scale μ , and outcome-linked exposure $\lambda_O F_O$ add enforcement channels that the vendor cannot suppress through ordinary auditability choices. Portability works by lowering s .

2024; Hartmann et al. 2025). The NIST AI Risk Management Framework’s generative-AI profile (National Institute of Standards and Technology 2024) is now the operational benchmark against which AI risk management is graded in practice. Pulling these threads together, evidence-based enforcement makes penalties contingent on detection probability, so vendors face an incentive to limit transparency whenever the marginal adoption benefit from auditability falls below the marginal increase in expected penalties.

3 Model

The model separates *visible compliance* from *substantive mitigation*. Auditability t helps the vendor satisfy procurement requirements but also facilitates enforcement; mitigation e directly reduces harm. The deployer can monitor after adoption, but monitoring is valuable only when evidence creates credible recourse. Switching cost s weakens that recourse. Independent audit capacity Δ , reporting-generated exposure $\mu\rho$, and outcome-linked exposure $\lambda_O F_O$ act through different institutional channels (Figure 1).

3.1 Players and Timeline

An institutional environment fixes the procurement floor $\bar{t}$, evidence-based penalty scale λF , independent audit capacity Δ , and switching-cost environment s . The strategic subgame is:

1. The vendor chooses auditability $t \in [0, 1]$ and mitigation $e \geq 0$.
2. The deployer observes (t, e) , chooses adoption $a \in \{0, 1\}$, and if adopting chooses monitoring $m \geq 0$. Integration creates switching costs represented by s .
3. Harm $h \in \{0, 1\}$ is realized. Evidence is generated with probability $p(t, m)$, and the regulator applies the specified enforcement rule when actionable evidence exists.

3.2 Risk, Detection, and Payoffs

We condition on a common-knowledge risk parameter $\theta \in (0, 1)$. Mitigation reduces the probability of harm, $\mathbb{P}(h = 1 \mid \theta, e) = q(\theta, e)$, and actionable evidence arises with probability $p(t, m)$, which increases in both auditability and monitoring.

Symbol	Meaning
t, e, m	vendor auditability, vendor mitigation, deployer monitoring
s	switching cost; $\delta_{\text{eff}} = \delta/(1+s)$ is effective recourse
$q(\theta, e)$	harm probability
$p(t, m)$	vendor/deployer evidence-generation probability
Δ	independent audit exposure, added to p
λF	evidence-based penalty scale
$\mu\rho$	reporting-generated exposure (ρ : effective coverage)
$\lambda_O F_O$	outcome-linked exposure
$\bar{t}$	procurement/compliance floor
κ	deployer harm exposure
α, β	sensitivity of detection to t, m
k_t, k_e, k_m	quadratic cost coefficients

Table 2. Key notation.

The vendor earns adoption benefit $B > 0$ and incurs convex auditability and mitigation costs:

$$U_V = aB - c_t(t) - c_e(e) - \lambda F q(\theta, e) [p(t, m) + \Delta]. \quad (1)$$

Here $p + \Delta$ is an enforcement-exposure index rather than necessarily a literal probability. The deployer obtains value V , pays integration cost I and monitoring cost $c_m(m)$, and can recover a fraction δ of harm exposure when evidence creates recourse. Lock-in reduces effective recourse to $\delta_{\text{eff}} = \delta/(1+s)$:

$$U_D = a[V - I - c_m(m) - \kappa q(\theta, e)(1 - \delta_{\text{eff}} p(t, m))]. \quad (2)$$

The regulator is represented through the institutional parameters $(\lambda F, \Delta)$; Section 5 asks how changing those parameters alters equilibrium.

3.3 Tractable Specification

Assumption 1 (Functional forms). *For the comparative statics,*

$$\begin{aligned} q(\theta, e) &= \max\{0, \theta - e\}, \\ p(t, m) &= 1 - \exp(-\alpha t - \beta m), \end{aligned}$$

and $c_j(x) = k_j x^2/2$ for $j \in \{t, e, m\}$, with all cost and sensitivity parameters positive.

Assumption 2 (Procurement threshold). *Adoption occurs if and only if $t \geq \bar{t}$. The floor $\bar{t}$ is an exogenous procurement or certification requirement. We condition on adoption and state the vendor participation condition in Appendix A.*

We use subgame-perfect equilibrium: the deployer chooses monitoring after observing vendor actions, and the vendor anticipates that monitoring response.

4 Equilibrium Analysis

4.1 Deployer Monitoring

Lemma 1 (Deployer best response). *For $q(\theta, e) > 0$ and $\delta_{\text{eff}} > 0$, the deployer has a unique interior monitoring choice m^* satisfying*

$$k_m m^* = \kappa q(\theta, e) \delta_{\text{eff}} \beta \exp(-\alpha t - \beta m^*). \quad (3)$$

Monitoring decreases with switching cost s and increases with deployer harm exposure κ .

Proof sketch. Marginal monitoring cost is increasing and marginal recourse is decreasing, giving a unique crossing. Implicit differentiation of (3) yields the signs. Appendix A gives the derivatives. $\square$

Lock-in makes information less useful: when the deployer's exit or compensation threat is not credible, evidence has less strategic value and monitoring falls.

4.2 Vendor Choices

Define the vendor's marginal enforcement exposure after anticipating the monitoring response:

$$\begin{aligned} H(t, m) &\equiv p(t, m) + (1 - p(t, m)) \frac{\beta m}{1 + \beta m} \\ &= 1 - \frac{\exp(-\alpha t - \beta m)}{1 + \beta m}. \end{aligned} \quad (4)$$

Lemma 2 (Vendor best response). *Suppose adoption benefit B is large enough that the vendor meets the procurement floor.*

- (a) *The vendor chooses $t^* = \bar{t}$.*
- (b) *In the interior region $e^* < \theta$, mitigation satisfies*

$$e^* = \frac{\lambda F}{k_e} [H(\bar{t}, m^*) + \Delta]. \quad (5)$$

If the right-hand side reaches θ , mitigation clamps at $e^ = \theta$ and harm is zero.*

Proof sketch. Higher t beyond the procurement floor raises both cost and total detection, even after the induced fall in monitoring. For mitigation, implicit differentiation of (3) gives $\partial m^*/\partial e = -m^*/[(\theta - e)(1 + \beta m^*)] < 0$. Substitution adds $(1 - p)\beta m^*/(1 + \beta m^*)$ to the direct detection term in H , yielding (5): the monitoring response is an *additional private return* to mitigation. The second derivative is negative; Appendix A gives the full derivation. $\square$

Therefore, the vendor satisfies the visible requirement and stops increasing auditability, while choosing mitigation according to the expected enforcement exposure created by both auditability and deployer monitoring.

4.3 Proxy Compliance and Accountability

Definition 1 (Proxy compliance). *An equilibrium exhibits proxy compliance when $t^* = \bar{t}$ but $e^* < e^{FB}$, where the social first-best mitigation is*

$$e^{FB} = \min\{1/k_e, \theta\}.$$

The equilibrium is characterized by

$$t^* = \bar{t}, \quad (6)$$

$$e^* = \min\left\{\theta, \frac{\lambda F}{k_e} [H(\bar{t}, m^*) + \Delta]\right\}, \quad (7)$$

$$k_m m^* = \kappa(\theta - e^*) \delta_{\text{eff}} \beta \exp(-\alpha \bar{t} - \beta m^*), \quad (8)$$

with $m^* = 0$ at the full-mitigation corner. In the interior, the right-hand side of (7) is strictly decreasing in e through the monitoring response, so the equilibrium is unique.

Setting	s	Δ	m^*	e^*	$\mathbb{E}[h]$
1. High lock-in, no audit	9	0	0.03	0.23	0.37
2. Portability only	1	0	0.10	0.33	0.27
3. Combined reform	1	1	0.00	0.60	0.00

Table 3. Equilibrium values in three institutional settings.

Proposition 1 (Existence of proxy compliance). *Suppose $\Delta = 0$ and baseline enforcement is weak enough that*

$$\frac{\lambda F}{k_e}(1 - \exp(-\alpha \bar{t})) < e^{FB}.$$

Then there exists $\bar{s}$ such that sufficiently high switching cost yields proxy compliance. As $s \rightarrow \infty$, $m^ \rightarrow 0$ and*

$$e^* \rightarrow \frac{\lambda F}{k_e}(1 - \exp(-\alpha \bar{t})).$$

Proposition 2 (Accountability under stronger institutions). *Within the interior equilibrium: (a) $de^*/d\Delta > 0$, although monitoring feedback attenuates the direct effect; (b) $de^*/ds < 0$, so reducing switching costs raises mitigation; and (c) sufficiently strong independent detection reaches e^{FB} , with portability reinforcing the transition.*

4.4 Worked Example

Let $\theta = 0.6$, $k_t = k_e = k_m = 1$, $\alpha = \beta = \kappa = \delta = \lambda F = 1$, and $\bar{t} = 0.2$. Then $e^{FB} = 0.6$. Solving (7)–(8) yields Table 3.

Portability alone produces a meaningful but incomplete improvement: Setting 2 moves above half of the first best but does not reach accountability. Under combined reform, the desired interior mitigation exceeds θ ; harm is eliminated and the deployer no longer monitors because there is no residual harm to investigate.

4.5 Comparative Statics

The total equilibrium effects differ from partial effects because policy-induced mitigation changes monitoring. Higher s lowers both monitoring and mitigation; higher Δ , λF , or $\bar{t}$ raises mitigation and reduces residual monitoring; higher mitigation cost k_e has the opposite effect; and higher deployer harm exposure κ raises monitoring and thereby mitigation. Appendix D provides derivatives and a summary table.

5 Policy Levers and Welfare

5.1 Audit Rights and Independent Evaluation

Independent audit capacity raises enforcement exposure by Δ without relying on deployer monitoring or ordinary vendor-controlled access.

Theorem 1 (Audit rights eliminate proxy compliance). *For every fixed lock-in level s and $\lambda F > 0$, a finite audit threshold $\bar{\Delta}(s)$ exists such that $\Delta \geq \bar{\Delta}(s)$ yields $e^* \geq e^{FB}$. A uniform sufficient condition is*

$$\Delta \geq \frac{k_e e^{FB}}{\lambda F}.$$

Thus sufficiently strong independent detection eliminates proxy compliance at any lock-in level.

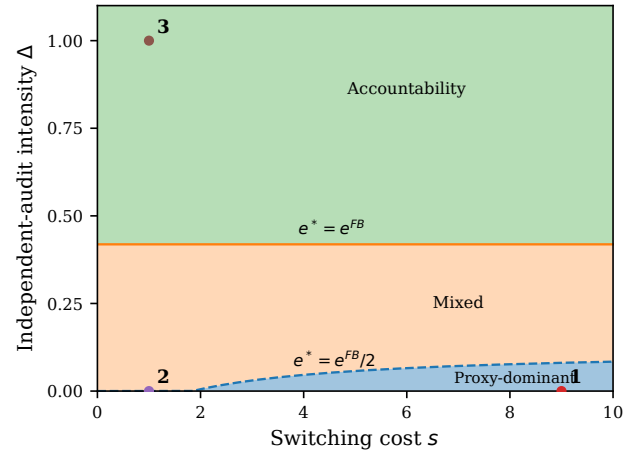


Figure 2. Illustrative regimes in the (s, Δ) plane under the worked-example parameters. Proxy compliance formally includes every point below first best; for visualization, the lower region is labeled *proxy-dominant* when $e^* < e^{FB}/2$, the middle region is mixed, and the upper region reaches e^{FB} . Markers correspond to Table 3.

The exact threshold is lower because baseline detection and the monitoring-response term also contribute to mitigation. Audit rights are therefore the strongest single lever in the model, but their practical effect depends on genuinely independent access and credible enforcement rather than the formal existence of an audit process.

5.2 Portability and Standards

Proposition 3 (Portability raises mitigation). *Reducing switching cost s raises equilibrium monitoring and mitigation and lowers expected harm. The effect is stronger when deployer harm exposure κ and monitoring productivity β are high.*

Portability works indirectly: compatible interfaces, exportable artifacts, and credible re-tendering make exit and recourse more credible, raising the value of monitoring and, therefore, the vendor's mitigation incentive.

5.3 Incident Reporting

Let $\rho \in [0, 1]$ denote the effective coverage of a mandatory reporting channel: the probability that a reportable incident reaches the regulator in actionable form. Let $\mu > 0$ be the associated enforcement-exposure scale. The vendor then faces additional expected exposure $\mu \rho q(\theta, e)$.

Proposition 4 (Incident reporting substitutes for vendor-gated evidence). *Under effective incident reporting, the mitigation condition becomes*

$$e_R^* = \min \left\{ \theta, \frac{\lambda F [H(\bar{t}, m_R^*) + \Delta] + \mu \rho}{k_e} \right\}. \quad (9)$$

Hence $e_R^ > e^*$ whenever $\mu \rho > 0$ and the baseline is below the mitigation corner.*

The product $\mu\rho$ is deliberately an *effective* channel: a nominal reporting duty with narrow scope, under-reporting, or weak follow-up contributes little.

5.4 Outcome-Linked Liability

Let $\lambda_O F_O$ denote the exposure tied directly to realized harm rather than to the separate evidence-generation channel $p(t, m)$.

Proposition 5 (Outcome-linked exposure raises mitigation). *The mitigation condition becomes*

$$e_L^* = \min \left\{ \theta, \frac{\lambda F[H(\bar{t}, m_L^*) + \Delta] + \lambda_O F_O}{k_e} \right\}, \quad (10)$$

so any positive outcome-linked exposure raises mitigation below the corner. Its direct component is unaffected by vendor-controlled auditability.

This is a stylized benchmark, not a claim that existing liability laws automatically observe or compensate for every harm.

5.5 Welfare

Social welfare equals adoption value minus expected harm and real resource costs. Below first best, an increase in independent audit exposure raises mitigation and lowers both harm and residual monitoring; therefore, its total welfare effect is positive. Portability has an additional trade-off: it raises beneficial mitigation but can also raise costly monitoring. The precise condition is stated in Appendix H.

5.6 Mapping to Current EU Regulation

The AI Act combines provider documentation and risk-management duties with differentiated conformity-assessment procedures. Article 43 and Annex VI permit internal-control routes for relevant classes of high-risk systems, while Articles 53 and 55 impose obligations on general-purpose model providers and systemic-risk models; the AI Office has information and evaluation powers under Articles 91–93 (European Union 2024b). In the model, documentation that remains provider-controlled primarily affects t , whereas regulator-controlled evaluation capacity raises Δ .

Article 40 of the Digital Services Act (DSA) provides an analogous Δ -raising design for very large online platforms and search engines (European Union 2022b). Portability and interoperability duties under the Digital Markets Act (DMA) illustrate instruments that can lower switching costs, although they are not an AI-specific portability regime (European Union 2022a). AI Act Article 73 serious-incident reporting can create a $\mu\rho$ channel when reports are complete, actionable, and credibly enforced.

The proposed AI Liability Directive was a fault-based evidentiary instrument centered on disclosure and rebuttable presumptions rather than the outcome-linked benchmark in Proposition 5; it was formally withdrawn in October 2025 (European Commission 2025). The revised Product Liability Directive covers software and provides no-fault liability for defective products but still requires defect, damage, and causation (European Union 2024a). It is therefore closer to, but not identical with, the model's direct outcome-exposure channel.

6 Predictions and Measurement Implications

Because $t^* = \bar{t}$, the model's direct comparative statics concern monitoring, mitigation, and harm; claims about contract terms or documentation are institutional hypotheses motivated by the mechanism.

6.1 Predictions

1. **Lock-in weakens monitoring and mitigation.** Greater integration depth and fewer credible alternatives predict lower monitoring, lower effective detection, and lower mitigation.
2. **Visible compliance can remain stable while outcomes deteriorate.** As s rises, auditability remains pinned to the procurement floor, while mitigation falls, producing documentation without commensurate reductions in incident rates.
3. **Independent access shifts operational effort.** Exogenous increases in regulator or third-party evaluation capacity should narrow the gap between vendor-reported and independently measured performance and raise mitigation investment.
4. **Portability is sector-dependent.** Its effect should be largest where deployer harm exposure is high and monitoring technology is mature; audit rights should bind more strongly where those conditions are absent.

6.2 Operationalization

Auditability can be measured through contractual audit clauses, API evaluation terms, disclosure completeness, and third-party publication rights. Lock-in can be proxied by fine-tuned endpoints, provider-specific tool wrappers, checkpoint or prompt-format portability, retraining costs, and observed switching. Independent exposure can be measured through regulator access, evaluation budgets, researcher-access programs, and enforcement actions. Candidate settings include medical-device AI, credit scoring, public-sector hiring, DSA audit and transparency records for very large online platforms (VLOPs), provider migration and re-tender data, and incident databases supplemented by AI Act Article 73 reports. Appendix F develops the measurement strategy.

7 Discussion and Limitations

Audit is not a panacea. Third-party audit can itself become proxy compliance. Audits may operate as rational myths—adopted for legitimacy rather than demonstrated efficacy—and auditors and auditees can share an interest in the appearance of rigor (Suchman and Edelman 1996; Eckstein and Shapira 2024). In the model, ceremonial or captured auditing contributes little to effective Δ ; independence, access, publication rights, and enforcement determine whether the threshold in Theorem 1 is reached.

How central is lock-in? The mechanism does not require extreme switching costs. Lock-in matters because it disables a monitoring channel that could otherwise substitute for weak independent detection. When deployer harm exposure is low, monitoring is immature, or vendor-independent access is constrained, the system can remain near proxy compliance, even

at moderate s . Conversely, credible portability is powerful in high-exposure sectors with mature monitoring.

Modeling scope. The model is deliberately static and single-vendor. The title’s references to “scaling” and “political economy” describe the motivating institutional environment, not endogenous scale or market structure. Switching cost, enforcement, and the procurement floor are parameters rather than outcomes of competition or regulatory choice. The common-knowledge risk benchmark omits private information; multidimensional harms, civil-society evidence production, adversarial misuse, multi-vendor competition, and repeated reputation dynamics are left to extensions. These abstractions make the policy channels transparent but limit quantitative interpretation.

Empirical status. The numerical example is illustrative rather than calibrated. The comparative statics generate hypotheses and measurement targets, not estimates of policy effect sizes. Cross-sector variation in procurement floors, audit access, and integration depth can support empirical tests; however, those variables may be correlated with underlying risk and institutional capacity; therefore, identification requires careful design rather than a simple cross-sectional comparison.

8 Conclusion

Responsible AI commitments can coexist with inadequate mitigation because visible compliance, verifiable evidence, and substantive harm reduction are distinct strategic goals. A vendor that controls ordinary auditability meets the procurement floor; a locked-in deployer monitors less; and evidence-dependent enforcement then supplies too little mitigation incentive. Independent audit rights bypass the weakened monitoring channel, portability restores deployer leverage, incident reporting creates regulator-visible evidence, and outcome-linked exposure retains bite when ordinary detection is weak. None are automatically effective: audit independence, reporting coverage, credible sanctions, and sector-specific monitoring capacity determine whether formal rules change equilibrium behavior. The central policy implication is therefore institutional rather than procedural: accountability requires both verifiability and power to act on what is verified.

A Mathematical Derivations and Equilibrium Proofs

Throughout this appendix, consider the interior region $0 < e < \theta$ and write

$$E(t, m) \equiv \exp(-\alpha t - \beta m) = 1 - p(t, m), \quad A \equiv \frac{\lambda F}{k_e}.$$

A.1 Deployer Best Response

Proof of Lemma 1. The deployer’s first-order condition is

$$\frac{\partial U_D}{\partial m} = -k_m m + \kappa q(\theta, e) \delta_{\text{eff}} \beta E(t, m) = 0,$$

which is (3). Its second derivative is

$$-k_m - \kappa q(\theta, e) \delta_{\text{eff}} \beta^2 E(t, m) < 0,$$

so the solution is unique. Let

$$G(m; t, e, s, \kappa) = k_m m - \kappa(\theta - e) \delta_{\text{eff}} \beta E(t, m).$$

At the optimum, use (3) to obtain

$$G_m = k_m + \kappa(\theta - e) \delta_{\text{eff}} \beta^2 E = k_m(1 + \beta m^*) > 0.$$

Implicit differentiation gives

$$m_e^* = -\frac{m^*}{(\theta - e)(1 + \beta m^*)} < 0, \quad (11)$$

$$m_t^* = -\frac{\alpha m^*}{1 + \beta m^*} < 0, \quad (12)$$

$$m_s^*|_{e,t} = -\frac{m^*}{(1 + s)(1 + \beta m^*)} < 0, \quad (13)$$

$$m_\kappa^*|_{e,t,s} = \frac{m^*}{\kappa(1 + \beta m^*)} > 0. \quad (14)$$

The first two derivatives describe how the deployer’s best response changes when vendor choices change; the latter two are partial institutional effects holding vendor mitigation fixed. $\square$

A.2 Vendor Auditability and Mitigation

Proof of Lemma 2. Auditability. Conditional on adoption, substitute $m^*(t, e)$ into (1). From (12),

$$\frac{d}{dt} p(t, m^*(t, e)) = p_t + p_m m_t^* = \frac{\alpha E(t, m^*)}{1 + \beta m^*} > 0.$$

Consequently,

$$\frac{dU_V}{dt} = -k_t t - \lambda F q(\theta, e) \frac{\alpha E(t, m^*)}{1 + \beta m^*} < 0 \quad (t \geq \bar{t}).$$

Once adoption is secured, every further increase in t raises both cost and expected enforcement exposure. Therefore $t^* = \bar{t}$.

Mitigation. Using $q_e = -1$ and (11),

$$\begin{aligned} \frac{dU_V}{de} &= -k_e e + \lambda F [p(\bar{t}, m^*) + \Delta] - \lambda F q p_m m_e^* \\ &= -k_e e + \lambda F \left[p(\bar{t}, m^*) + \Delta \right. \\ &\quad \left. + (1 - p(\bar{t}, m^*)) \frac{\beta m^*}{1 + \beta m^*} \right]. \end{aligned}$$

This is (5). The bracket is $H(\bar{t}, m^*) + \Delta$.

For the second-order condition,

$$H_m(t, m) = \frac{\beta E(t, m)(2 + \beta m)}{(1 + \beta m)^2} > 0.$$

Because $m_e^* < 0$,

$$\frac{d^2 U_V}{de^2} = -k_e + \lambda F H_m(\bar{t}, m^*) m_e^* < -k_e < 0.$$

Thus the interior solution is the unique vendor optimum; if its unconstrained value reaches θ , the optimum is the full-mitigation corner. $\square$

Remark 1 (Vendor participation). The vendor meets the floor when

$$B > \frac{k_t}{2} \bar{t}^2 + \frac{k_e}{2} (e^*)^2 + \lambda F(\theta - e^*)[p(\bar{t}, m^*) + \Delta].$$

This holds for sufficiently large adoption benefit B . Otherwise the model's conditional-on-adoption equilibrium is not reached.

A.3 Uniqueness and Comparative Responses

For a fixed institutional environment, substitute the deployer response $m^*(e)$ into the vendor condition:

$$\Phi(e; \Delta, s) = A[H(\bar{t}, m^*(e, s)) + \Delta].$$

Equation (7) is $e^* = \min\{\theta, \Phi(e^*)\}$. Since

$$\Phi_e = AH_m m_e^* < 0,$$

Φ is strictly decreasing while the identity is increasing. Hence there is at most one interior crossing. If $\Phi(\theta^-) < \theta$, continuity yields one interior equilibrium; if $\Phi(\theta^-) \geq \theta$, the unique equilibrium is the corner $e^* = \theta, m^* = 0$.

Define

$$D \equiv 1 - AH_m m_e^* > 1.$$

The equilibrium response to independent audit capacity is

$$\frac{de^*}{d\Delta} = \frac{A}{D} > 0. \quad (15)$$

The direct fixed-monitoring effect is A ; the fall in monitoring attenuates but never reverses it.

At fixed mitigation, (13) gives $m_s^*|_{e,t} < 0$. Therefore

$$\frac{de^*}{ds} = \frac{AH_m}{D} m_s^*|_{e,t} < 0. \quad (16)$$

The total monitoring response is also negative. One direct proof substitutes $e = A[H(\bar{t}, m) + \Delta]$ into (8) and defines

$$\begin{aligned} \Psi(m, s) &= k_m m - \kappa[\theta - A(H(\bar{t}, m) + \Delta)] \\ &\quad \times \frac{\delta}{1+s} \beta E(\bar{t}, m). \end{aligned}$$

Here $\Psi_m > 0$ and $\Psi_s > 0$, so $dm^*/ds = -\Psi_s/\Psi_m < 0$.

Proof of Proposition 1. As $s \rightarrow \infty$, $\delta_{\text{eff}} \rightarrow 0$. Equation (8) then forces $m^* \rightarrow 0$, so $H(\bar{t}, m^*) \rightarrow H(\bar{t}, 0) = 1 - \exp(-\alpha\bar{t})$. Under the stated weak-enforcement condition, the limiting mitigation is below first best. Equation (16) shows that mitigation is decreasing in s , so by continuity it remains below first best for all sufficiently large s . When $\bar{t}$ is small, $1 - \exp(-\alpha\bar{t}) \simeq \alpha\bar{t}$, making the limiting mitigation correspondingly small. $\square$

Proof of Proposition 2. Part (a) follows from (15). Part (b) follows from (16). For part (c), Theorem 1 supplies a finite audit threshold for every s ; lowering s raises H through monitoring, weakly reducing the audit intensity needed to reach first best. $\square$

B Audit Rights and Other Policy Proofs

Proof of Theorem 1. Fix s . Let $m^{FB}(s)$ be the deployer best response when mitigation equals e^{FB} ; if $e^{FB} = \theta$, then $m^{FB} = 0$. The exact smallest audit boost that makes the vendor weakly prefer first-best mitigation is

$$\bar{\Delta}(s) = \max\left\{0, \frac{k_e e^{FB}}{\lambda F} - H(\bar{t}, m^{FB}(s))\right\}. \quad (17)$$

This threshold is finite. Since $H \geq 0$,

$$\bar{\Delta}(s) \leq \frac{k_e e^{FB}}{\lambda F}$$

for every s , yielding the uniform sufficient condition in the theorem. At any proxy-compliance baseline $e^* < e^{FB}$, reaching e^{FB} strictly reduces expected harm. In the worked-example corner $e^{FB} = \theta$, the exact threshold is independent of s :

$$\bar{\Delta} = \frac{k_e \theta}{\lambda F} - [1 - \exp(-\alpha\bar{t})] = 0.418731 \dots \quad \square$$

Proof of Proposition 3. Equation (16) gives $de^*/ds < 0$. The m-only equation $\Psi(m, s) = 0$ above gives $dm^*/ds < 0$. Thus reducing s raises monitoring and mitigation. Since expected harm is $\theta - e^*$ in the interior, it falls. $\square$

Proof of Proposition 4. Let $z_R = \mu\rho$. The reporting regime shifts the vendor condition to

$$e = \Phi(e; \Delta, s) + \frac{z_R}{k_e}.$$

Implicit differentiation yields

$$\frac{de_R^*}{dz_R} = \frac{1/k_e}{1 - AH_m m_e^*} > 0.$$

The monitoring response attenuates the direct shift but cannot reverse it. The result holds until the full-mitigation corner is reached. $\square$

Proof of Proposition 5. Let $z_O = \lambda_O F_O$. The outcome-linked term enters the vendor condition additively and independently of $p(t, m)$:

$$e = \Phi(e; \Delta, s) + \frac{z_O}{k_e}.$$

The same implicit derivative as above gives $de_L^*/dz_O > 0$. The auditability derivative remains negative because z_O does not depend on t , so $t^* = \bar{t}$. $\square$

C Worked-Example Calculations

The worked example uses

$$\begin{aligned} \theta &= 0.6, & k_t &= k_e = k_m = 1, & \bar{t} &= 0.2, \\ \alpha &= \beta = \kappa = \delta = \lambda F &= 1. \end{aligned}$$

For an interior setting, solve

$$\begin{aligned} m^* &= (0.6 - e^*) \frac{1}{1+s} \exp(-0.2 - m^*), \\ e^* &= 1 - \frac{\exp(-0.2 - m^*)}{1 + m^*} + \Delta. \end{aligned}$$

Setting	m^*	$p(\bar{t}, m^*)$	e^*	$\mathbb{E}[h]$
High lock-in	0.029573	0.205127	0.227958	0.372042
Portability only	0.100822	0.259791	0.327585	0.272415
Combined reform	0	0.181269	0.600000	0

Table 4. Unrounded worked-example solutions.

The resulting values before rounding are:

In the combined setting, even the limiting marginal exposure at $m = 0$ is $H(\bar{t}, 0) + \Delta = 0.181269 + 1 > 0.6$, so mitigation clamps at θ . With $q = 0$, Equation (3) then gives $m^* = 0$. Portability alone reduces expected harm by approximately 26.8% relative to the high-lock-in baseline.

Figure 2 is generated from the same equations. At a target mitigation $e_0 < \theta$, monitoring has the Lambert- W representation

$$m(e_0, s) = \frac{1}{\beta} W\left(\frac{\kappa(\theta - e_0)\delta\beta^2 e^{-\alpha\bar{t}}}{k_m(1+s)}\right),$$

and the audit boost required to support that target is

$$\Delta(e_0, s) = \max\left\{0, \frac{k_e e_0}{\lambda F} - H(\bar{t}, m(e_0, s))\right\}.$$

The plotted contours use $e_0 = e^{FB}/2$ and $e_0 = e^{FB}$.

D Comparative Statics

The denominator $D = 1 - AH_m m_e^*$ is positive. Besides (15) and (16), the following derivatives hold in the interior:

$$\begin{aligned} \frac{de^*}{d\bar{t}} &= \frac{A}{D} \frac{\alpha E(\bar{t}, m^*)}{(1 + \beta m^*)^3} > 0, \\ \frac{de^*}{dk_e} &= -\frac{e^*}{k_e D} < 0, \\ \frac{de^*}{d(\lambda F)} &= \frac{e^*}{\lambda F D} > 0, \\ \frac{de^*}{d\kappa} &= \frac{AH_m}{D} \frac{m^*}{\kappa(1 + \beta m^*)} > 0. \end{aligned}$$

For the $\bar{t}$ derivative, the direct increase in detection dominates the induced reduction in monitoring:

$$\left. \frac{dH}{d\bar{t}} \right|_e = H_t + H_m m_t^* = \frac{\alpha E}{(1 + \beta m^*)^3} > 0.$$

The negative monitoring response to stronger audit or enforcement is not a policy failure: higher mitigation lowers residual harm, reducing the deployer’s need to monitor.

E Regulatory Mapping in Greater Detail

AI Act conformity and model oversight. Article 43 of Regulation (EU) 2024/1689 specifies conformity-assessment routes for high-risk systems; Annex VI sets out internal control, while product-linked systems may follow sectoral third-party procedures. Articles 53 and 55 govern general-purpose model providers and models with systemic risk. Articles 91–93 grant the AI Office powers to request documentation and

Parameter increase	m^*	t^*	e^*	$\mathbb{E}[h]$
s (lock-in)	–	0	–	+
Δ (independent audit)	–	0	+	–
λF (evidence-based exposure)	–	0	+	–
$\bar{t}$ (procurement floor)	–	+	–	–
k_e (mitigation cost)	+	0	–	+
κ (deployer harm exposure)	+	0	+	–

Table 5. Total equilibrium effects in the interior. A dash denotes a decrease, a plus an increase, and zero no change. Monitoring effects include the endogenous mitigation response.

information, conduct evaluations, and require measures. The model distinguishes provider auditability t from genuinely independent capacity Δ : adding documentation to a provider-controlled process need not equal independent verification.

Post-market monitoring and incident reporting. Articles 72–73 create post-market monitoring and serious-incident reporting obligations for high-risk systems. Their effective model counterpart is $\mu\rho$, not the nominal duty alone. Coverage, causal attribution, timeliness, regulator capacity, and sanctions for non-compliance determine ρ and μ .

Researcher access and portability analogies. DSA Article 40 provides vetted-researcher access for VLOPs and very large online search engines (VLOSEs) and therefore illustrates an access right that can raise independent detection. DMA Articles 5–6 illustrate data-portability and interoperability obligations that can lower switching costs. Neither regime maps mechanically onto foundation-model procurement; they identify the institutional form of the levers.

Liability. The withdrawn AI Liability Directive proposal would have eased proof through disclosure and rebuttable presumptions in fault-based claims. It was not automatic liability for every adverse outcome. Directive (EU) 2024/2853 modernizes strict product liability and expressly encompasses software; however, claimants still need to prove a defective product, covered damage, and causation. Proposition 5 should therefore be read as a benchmark for exposure that is less dependent on the paper’s separate evidence channel, not as a literal description of current EU law.

F Operationalization and Empirical Designs

Three empirical designs are particularly natural. First, procurement-threshold changes can be studied with difference-in-differences or event-study designs when comparable sectors adopt requirements at different times. Second, provider migrations or interoperability mandates create settings for measuring whether reduced lock-in raises monitoring and remediation. Third, changes in independent access rights can be linked to the gap between vendor-reported and independently reproduced performance. In each case, underlying sector risk and institutional capacity are likely confounders; the model identifies directional hypotheses, not a ready-made causal instrument.

Construct	Candidate measures	Candidate data sources or settings
Auditability t	contractual audit clauses; API rate and logging terms; reproducibility permissions; model-card and system-card completeness; publication restrictions	procurement contracts; platform terms; model documentation; DSA audit reports; researcher-access scorecards
Independent capacity Δ	regulator testing rights; third-party access independent of provider permission; evaluation staffing and budgets; ability to publish adverse findings	AI Office and national authority records; AI Safety Institute programs; DSA Article 40 access decisions
Lock-in s	provider-specific fine-tuning artifacts; prompt and tool wrappers; data-egress and retraining costs; switching and re-tender rates	machine-learning operations (MLOps) telemetry; procurement records; vendor-migration projects; surveys of deployers
Monitoring m	continuous evaluation frequency; red-team budget; incident triage; external audit expenditure	internal governance records; audit contracts; safety-case updates; regulated-sector supervision
Mitigation e	patching and retraining effort; safety engineering headcount; post-incident remediation; independently measured performance improvements	change logs; technical reports; incident investigations; independent benchmarks
Reporting ρ	report completeness, timeliness, scope, and match between internal incidents and regulator filings	AI Act Article 73 reports when available; sectoral incident registers; AI Incident Database

Table 6. Illustrative measures for the model’s institutional parameters.

G Modeling Choices and Extensions

Common-knowledge risk. Conditioning on θ isolates evidence scarcity from adverse selection. A private-information extension would let the vendor signal risk through (t, e) and could add pooling or separating equilibria, but it is not required for the proxy-compliance mechanism.

Static, single-vendor structure. Repeated interaction could make reputation an additional enforcement channel or, conversely, institutionalize symbolic compliance. A multi-vendor model would endogenize switching opportunities and market concentration. The present s summarizes these forces rather than deriving them.

Harm technology. The linear specification makes the monitoring response and policy thresholds transparent. More general decreasing harm functions preserve the mechanism when the vendor objective remains concave and the deployer best response is well-behaved, but the threshold formulas change.

Non-market evidence producers. Journalists, civil society organizations, and academic researchers can be represented as contributors to Δ . Their access, publication incentives, and vulnerability to legal or contractual restriction are themselves strategic objects and deserve a separate model.

Scaling and political economy. The paper’s title identifies the application domain. Scale, compute concentration, and regulatory capacity are not endogenous primitives. They motivate why audit access and switching costs may be consequential, but the formal results should not be read as comparative statics in model scale.

H Welfare Effects

Treating penalties as transfers and abstracting from policy-administration costs, define welfare conditional on adoption

as

$$W = V - (\theta - e^*) - \frac{k_e}{2}(e^*)^2 - \frac{k_t}{2}t^2 - \frac{k_m}{2}(m^*)^2. \quad (18)$$

Because the deployer best response depends on policy only through equilibrium mitigation for an audit change, $m_\Delta^* = m_e^* e_\Delta^*$. Therefore

$$\frac{dW}{d\Delta} = [(1 - k_e e^*) - k_m m^* m_e^*] \frac{de^*}{d\Delta}.$$

When $e^* < e^{FB}$, $1 - k_e e^* > 0$; because $m_e^* < 0$ and $de^*/d\Delta > 0$, the audit effect is strictly positive. Stronger audit capacity raises beneficial mitigation and reduces the monitoring cost required to manage residual harm.

For switching costs,

$$\frac{dW}{ds} = (1 - k_e e^*) \frac{de^*}{ds} - k_m m^* \frac{dm^*}{ds}.$$

Both derivatives are negative. A marginal reduction in s improves welfare precisely when

$$(1 - k_e e^*) \left| \frac{de^*}{ds} \right| > k_m m^* \left| \frac{dm^*}{ds} \right|.$$

The left side is the net benefit from additional mitigation; the right side is the extra monitoring resource cost. This condition replaces the stronger and generally unjustified claim that portability is always welfare-improving.

Ethical Statement

This paper develops a stylized institutional model and uses no human participants, personal data, or deployed AI systems. Its principal adverse-impact risk is overgeneralization: the comparative statics do not establish that any specific firm, sector, or legal regime necessarily exhibits proxy compliance, nor that audit, reporting, portability, or liability is sufficient

on its own. We therefore separate the model’s benchmark mechanisms from claims about current law and stress that implementation depends on audit independence, due process, confidentiality safeguards, civil-society access, distributional effects, and regulator capacity.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Cyber Secure Everywhere [grant no. EP/Y035313/1].

References

- Bénabou, R.; and Tirole, J. 2010. Individual and Corporate Social Responsibility. *Economica*, 77(305): 1–19.
- Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.
- Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022. The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*.
- Burnat, F. A. D.; and Davidson, B. I. 2026. The accountability paradox: How platform API restrictions undermine AI transparency mandates. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, 2416–2433. New York, NY, USA: ACM.
- Casper, S.; Ezell, C.; Siegmann, C.; Kolt, N.; Curtis, T. L.; Bucknall, B.; Haupt, A.; Wei, K.; Scheurer, J.; Hobbhahn, M.; Sharkey, L.; Krishna, S.; Von Hagen, M.; Alberti, S.; Chan, A.; Sun, Q.; Gerovitch, M.; Bau, D.; Tegmark, M.; and Krueger, D. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24)*.
- Costanza-Chock, S.; Raji, I. D.; and Buolamwini, J. 2022. Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, 1571–1583. New York, NY, USA: ACM.
- Crawford, K. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press. ISBN 978-0300209570.
- Eckstein, A.; and Shapira, R. 2024. Compliance gatekeepers. *Yale Journal on Regulation*, 41: 469–523.
- Engler, A. 2023. The EU AI Act and Emerging AI Market Concentration. Report, Brookings Institution.
- European Commission. 2025. Withdrawal of Commission Proposals: COM(2022) 496 final, 2022/0303 (COD). Official Journal of the European Union, C/2025/5423.
- European Union. 2022a. Regulation (EU) 2022/1925 on Contestable and Fair Markets in the Digital Sector. Official Journal of the European Union, L 265. Digital Markets Act.
- European Union. 2022b. Regulation (EU) 2022/2065 on a Single Market for Digital Services. Official Journal of the European Union, L 277. Digital Services Act.
- European Union. 2024a. Directive (EU) 2024/2853 on Liability for Defective Products. Official Journal of the European Union, L 2024/2853.
- European Union. 2024b. Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence. Official Journal of the European Union, L 2024/1689. Artificial Intelligence Act.
- Farrell, J.; and Shapiro, C. 1988. Dynamic Competition with Switching Costs. *The RAND Journal of Economics*, 19(1): 123–137.
- Floridi, L.; Holweg, M.; Taddeo, M.; Silva, J. A.; Mökander, J.; and Wen, Y. 2022. capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU AI Act. SSRN Working Paper.
- Glaeser, E. L.; and Shleifer, A. 2003. The Rise of the Regulatory State. *Journal of Economic Literature*, 41(2): 401–425.
- Gornet, M.; and Maxwell, W. 2024. The European Approach to Regulating AI Through Technical Standards. *Internet Policy Review*, 13(3).
- Hacker, P.; Engel, A.; and Mauer, M. 2023. Regulating ChatGPT and Other Large Generative AI Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’23)*.
- Hagendorff, T. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1): 99–120.
- Hartmann, D.; Laranjeira de Pereira, J. R.; Streitbürger, C.; and Berendt, B. 2025. Addressing the Regulatory Gap: Moving Towards an EU AI Audit Ecosystem Beyond the AI Act by Including Civil Society. *AI and Ethics*.
- Jobin, A.; Ienca, M.; and Vayena, E. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1(9): 389–399.
- Katz, M. L.; and Shapiro, C. 1985. Network Externalities, Competition, and Compatibility. *The American Economic Review*, 75(3): 424–440.
- Klemperer, P. 1995. Competition when Consumers Have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics, and International Trade. *The Review of Economic Studies*, 62(4): 515–539.
- Laffont, J.-J.; and Tirole, J. 1993. *A Theory of Incentives in Procurement and Regulation*. Cambridge, MA: MIT Press. ISBN 978-0-262-12174-2.
- Laux, J.; Wachter, S.; and Mittelstadt, B. 2024. Three Pathways for Standardisation and Ethical Disclosure by Default Under the European Union Artificial Intelligence Act. *Computer Law & Security Review*, 53.
- Mantelero, A. 2024. The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, Legal Obligations and Key Elements for a Model Template. *Computer Law & Security Review*, 54.
- Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* ’19, 220–229. New York, NY, USA: ACM.

- Mökander, J.; Axente, M.; Casolari, F.; and Floridi, L. 2022. Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines*, 32: 241–268.
- Mokander, J.; Schuett, J.; Kirk, H. R.; and Floridi, L. 2023. Auditing Large Language Models: A Three-Layered Approach. *AI and Ethics*, 3(4): 1085–1115.
- Moss, E.; Metcalf, J.; Watkins, E. A.; Singh, R.; and Elish, M. C. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. Technical report, Data & Society Research Institute.
- Mulligan, D. K.; and Bamberger, K. A. 2019. Procurement as Policy: Administrative Process for Machine Learning. *Berkeley Technology Law Journal*, 34(3): 773–851.
- National Institute of Standards and Technology. 2024. Artificial Intelligence Risk Management Framework: Generative AI Profile. Technical Report NIST AI 600-1, NIST.
- Novelli, C.; Hacker, P.; Morley, J.; Trondal, J.; and Floridi, L. 2025. A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities. *European Journal of Risk Regulation*.
- Raji, I. D.; and Buolamwini, J. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES '19*, 429–435. New York, NY, USA: ACM.
- Raji, I. D.; Kumar, I. E.; Horowitz, A.; and Selbst, A. 2022a. The Fallacy of AI Functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.
- Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. 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*, 33–44. New York, NY, USA: ACM.
- Raji, I. D.; Xu, P.; Honigsberg, C.; and Ho, D. 2022b. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.
- Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In *Data and Discrimination: Converting Critical Concerns into Productive Inquiry, a preconference at the 64th Annual Meeting of the International Communication Association*. Seattle, WA.
- Schuett, J. 2023. Risk Management in the Artificial Intelligence Act. *European Journal of Risk Regulation*, 14(2): 367–385.
- Schuett, J. 2025. Frontier AI Developers Need an Internal Audit Function. *Risk Analysis*.
- Selbst, A. D.; Boyd, d.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, 59–68. New York, NY, USA: ACM.
- Shavell, S. 1984. Liability for Harm versus Regulation of Safety. *Journal of Legal Studies*, 13(2): 357–374.
- Stigler, G. J. 1971. The Theory of Economic Regulation. *The Bell Journal of Economics and Management Science*, 2(1): 3–21.
- Suchman, M. C.; and Edelman, L. B. 1996. Legal rational myths: The new institutionalism and the law and society tradition. *Law & Social Inquiry: Journal of the American Bar Foundation*, 21(04): 903–941.
- Tirole, J. 1986. Hierarchies and Bureaucracies: On the Role of Collusion in Organizations. *Journal of Law, Economics, & Organization*, 2(2): 181–214.
- Veale, M.; and Zuiderveen Borgesius, F. 2021. Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International*, 22(4): 97–112.
- Wachter, S.; Mittelstadt, B.; and Russell, C. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, 41: 105567.
- Wang, J.; Huang, K.; Klyman, K.; and Bommasani, R. 2025. Do AI Companies Make Good on Voluntary Commitments to the White House? In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.
- Widder, D. G.; West, S. M.; and Whittaker, M. 2023. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. SSRN working paper.