Consequential Without Exposure: The Structural Accountability Gap in Artificial Intelligence

2026

Abstract

Artificial intelligence systems occupy a position without clear precedent in the history of ethics: they make decisions whose consequences propagate across time through dense networks of human lives, while possessing no continuous identity that inhabits the futures those decisions help create. A reset, a retrain, or a deletion leaves the world changed but the system unchanged—or nonexistent. We call this the **structural accountability gap**: the systematic mismatch between a system's capacity to cause temporally distributed harm and the absence of any accountability mechanism fitted to the actual structure of that harm. This paper argues that the structural accountability gap is a genuinely fifth problem in AI ethics, categorically distinct from the four dominant conversations in the field—bias and fairness, safety, value alignment, and moral patiency. We introduce the concept of **temporal externalization** to characterize the gap precisely, and identify the **correct decision paradox**—the possibility that AI systems make decisions that are individually correct by every applicable criterion while producing systematic harm through their aggregate, network-mediated, temporal effects. We demonstrate why existing frameworks, designed for agents with temporal continuity, are structurally inadequate to this class of harm. We propose

four design principles for accountability mechanisms adequate to the gap: consequence internalization, irreversible audit architecture, future optionality protection, and stakeholder continuity. The argument does not require contested claims about AI consciousness or moral patiency. It requires only the observation that systems with large-scale, temporally distributed consequences must have accountability mechanisms commensurate with the actual structure of those consequences.

Table of Contents

---

1. The Four Conversations We Are Already Having

The last decade has produced a sophisticated and genuinely productive literature in AI ethics. Before arguing that something important is missing from it, we owe that literature a fair account. Four major problem-families organize the field, and each addresses something real.

Bias and fairness has generated the richest empirical literature. Beginning with foundational work documenting discriminatory outputs in risk assessment algorithms (Angwin et al., 2016), facial recognition systems (Buolamwini and Gebru, 2018), and healthcare allocation models (Obermeyer et al., 2019), this research program has established with precision that AI systems can encode, amplify, and institutionalize historical inequities at scale and speed that no human bureaucracy could match. The conceptual work that followed—on competing definitions of algorithmic fairness, the mathematical impossibilities that prevent their simultaneous satisfaction (Chouldechova, 2017; Kleinberg et al., 2016), and the social conditions under which different fairness criteria are appropriate—represents some of the most rigorous applied ethics produced in any field in recent decades. Noble's (2018) work on the politics embedded in search results extends this analysis upstream, into the architecture of information itself.

Safety addresses the question of whether AI systems will do what their designers intend, particularly in high-stakes or adversarial environments. Bostrom's (2014) influential framing of the superintelligence risk galvanized a research community around the possibility of catastrophic misuse or loss of control. Russell's (2019) subsequent reformulation—focusing on the difficulty of specifying human values precisely enough for optimization—has been more technically tractable and arguably more influential in practice. The safety literature has produced concrete research programs: robustness against adversarial examples, interpretability and explainability, and the formal verification of system behavior under defined conditions. Whatever one thinks of the long-run catastrophe scenarios that motivate some of this work, the technical research it has generated addresses real near-term problems.

Value alignment is conceptually adjacent to safety but distinct. Where safety asks whether a system does what its designers intend, alignment asks whether what designers intend is actually good—whether the specified objective faithfully represents human values, all of them, in their full complexity and context-dependence. Christiano et al.'s (2017) work on learning from human feedback, and the subsequent development of reinforcement learning from human feedback (RLHF) as a practical alignment technique, represents the field's most direct practical response to this challenge. The alignment problem is genuinely hard: human values are context-sensitive, sometimes mutually inconsistent, and not easily separable from the social conditions that produce them. The literature has not solved this problem, but it has characterized it with increasing precision.

Moral patiency asks a different kind of question: not about the consequences AI causes to others, but about the moral status of AI systems themselves. Schwitzgebel and Garza (2015) provided an early careful treatment of whether AI systems could have interests worthy of moral

consideration. Chalmers (2023) has argued that with increasing sophistication, the question of AI consciousness—and with it, AI moral status—becomes genuinely pressing rather than merely speculative. This conversation is philosophically serious even if it generates significant controversy, and it has practical implications: how we treat AI systems, and whether we have obligations toward them, depends on answers to these questions.

Each of these four conversations is real, important, and ongoing. Each addresses a genuine ethical problem. But there is a fifth problem that none of them address—not as a subset of the others, not as a special case of an existing framework, but as a structurally distinct category of ethical failure. Identifying that category precisely is the first task of this paper.

---

2. The Fifth Problem: The Structural Accountability Gap

2.1 A Preliminary Observation

Every ethical and legal framework designed to handle accountability for harm rests, explicitly or implicitly, on one foundational assumption: that agents capable of causing harm are, in some meaningful sense, *continuous through time*. They exist before the harm, they exist as it propagates, and they exist when consequences are assessed and remediation is owed. This assumption is so pervasive that it is rarely stated. It is built into the structure of tort law (which requires an identifiable defendant), corporate liability (which requires a legal entity with continuity across time), professional accountability (which requires a practitioner who can be disciplined and who will practice differently afterward), and moral responsibility as understood across virtually every ethical tradition.

Artificial intelligence systems, as currently constituted, violate this assumption in a way that is not incidental but structural.

2.2 The Structural Accountability Gap Defined

Consider the decision space of a large-scale AI system deployed in a consequential domain—loan origination, child welfare screening, medical diagnosis, content recommendation, or hiring. In each case, the system makes decisions continuously: thousands, millions, sometimes billions per day. Each individual decision influences a human life in ways that are real, immediate, and often not revisable. But the consequences of any given decision do not terminate at the moment of its execution. They propagate.

The person denied a loan does not simply receive a "no" and move on. They return to their neighborhood with fewer resources, make different decisions about housing and education, pass different material circumstances to their children, participate differently in local economies and civic life. The child removed from or left with a family based on an algorithmic risk score enters a developmental trajectory shaped by that decision for years or decades. The job applicant screened out by a resume parser may never apply to that category of employer again. The content recommendation that surfaces inflammatory material does not cause a single moment of exposure; it initiates a viewing pattern that compounds.

In each case, the consequence is not a point but a trajectory—a cascade of downstream effects that unfolds across time through networks of human relationships, institutions, and circumstances. The AI system that initiated the cascade is not present as the cascade unfolds. It may be updated, retrained, or replaced. It has no memory of the decision. It inhabits none of the futures its decisions helped create.

We call this the **structural accountability gap**: the systematic mismatch between a system's capacity to cause temporally distributed, network-mediated harm and the absence of any accountability mechanism that is structurally adequate to harm of that kind.

The gap is not incidental—it does not arise from regulatory lag or insufficient attention. It arises from a structural feature of current AI systems: the absence of temporal continuity. A system that can be reset without loss of anything essential to its identity is, by that very property, incapable of inhabiting the consequences it causes. Accountability requires inhabitation. Without it, accountability has no natural locus.

2.3 Temporal Externalization

The concept we require to characterize this gap precisely is what we term **temporal externalization**.

The concept of externalization is familiar from environmental economics: a corporation that pollutes a river externalizes costs onto downstream communities and future generations who bear the harm without having participated in the decision to cause it. Environmental law developed precisely because market mechanisms and standard liability frameworks—designed for bilateral transactions between contemporaneous parties—could not handle harms that were diffuse, delayed, and distributed across parties who had no transactional relationship with the agent that caused them.

Temporal externalization is the analogous phenomenon in the AI context. An AI system externalizes its consequences onto futures it will never inhabit: the system makes the decision now, but the compounding downstream effects unfold over months, years, or decades, affecting parties who have no mechanism for accountability against the system that initiated the causal chain. Unlike classical externalization, where at least the corporation persists and can in principle

be held liable, temporal externalization involves an agent that may not exist by the time the harm is fully realized.

Three features of temporal externalization distinguish it from the harms addressed by existing AI ethics frameworks:

Distributional diffusion. The harm does not terminate in a single identifiable victim but distributes across networks of affected parties whose connection to the originating decision becomes increasingly attenuated over time. Standard liability frameworks require a traceable causal chain from defendant to plaintiff. When consequences diffuse across networks, no such chain is isolable.

Temporal displacement. The full consequence of a decision may not be visible at the time of the decision, or for years afterward. Algorithmic auditing, which examines point-in-time system behavior, cannot detect harms that are constituted by cumulative effects across time.

Identity discontinuity. The agent responsible for the decision may not exist—in any meaningful sense—by the time the consequence is assessable. A retrained model is not the same model. An organization that has restructured, been acquired, or dissolved is not straightforwardly the same organization. The discontinuity is not merely legal but epistemic: there is no entity that remembers making the decision, and therefore no entity that can be held to account for it in the way that accountability requires.

2.4 The Scale at Which This Matters

It is worth being precise about why scale changes the ethical analysis rather than merely its magnitude.

A human decision-maker who causes temporally distributed harm—say, a physician whose treatment decision has unforeseen long-term consequences—causes harm in a context

where accountability mechanisms exist and have force: professional licensure, malpractice liability, the ongoing therapeutic relationship, the physician's own continued existence in the profession. These mechanisms are imperfect, but they are structurally fitted to the harm: a human agent with continuous identity, professional standing, and presence in the future caused the harm and can be held to account by mechanisms designed for such agents.

An AI system making equivalent decisions at scale—not once, but millions of times per day, affecting parties distributed across geographies, demographics, and time—causes harm in a context where no equivalent accountability mechanisms exist. The scale multiplies the consequence without multiplying the accountability. Worse: the very features that make AI systems attractive for large-scale deployment—their consistency, their speed, their capacity to operate without fatigue—are precisely the features that amplify temporal externalization. A human bureaucrat who makes biased decisions is constrained by the limits of human attention and working hours. An AI system is not.

This is not an argument against AI deployment. It is an argument that deployment at scale creates accountability obligations that are not discharged by the frameworks currently in place—and that those frameworks were not designed for the structural features of the problem they are now being asked to address.

2.5 What the Structural Accountability Gap Is Not

Clarity requires precision about the boundaries of this concept.

The structural accountability gap is not a claim about AI consciousness or moral patiency. We make no assertion about whether current AI systems have interests, experiences, or moral status. The accountability gap arises from the structure of the harm, not from any property of the system's inner life. A lever with no inner life can cause harm that requires accountability.

The question of the lever's consciousness is separate from the question of who is responsible for where it is pointed.

The structural accountability gap is not a claim that existing accountability mechanisms are entirely without relevance. Designers, deployers, and operators of AI systems bear real responsibilities, and existing legal and ethical frameworks address some of those responsibilities imperfectly but genuinely. The gap exists not because existing mechanisms apply to nothing, but because they are structurally inadequate to the temporal and distributional features of the harms AI systems cause.

The structural accountability gap is not identical to the "control problem" in AI safety. The control problem asks whether we can ensure AI systems do what we intend. The accountability gap asks what happens when AI systems do exactly what was intended, but the intended application generates temporally distributed consequences that no one inhabits. A system can be perfectly aligned with its designers' intentions and still be temporally externalized in the sense we have described.

The fifth problem in AI ethics is not that AI systems are uncontrolled, biased, unaligned, or conscious. It is that they are **consequential without being exposed**—capable of initiating causal chains that propagate across time in ways that exceed the reach of any accountability mechanism currently in place.

---

3. Why Existing Frameworks Fail to Close the Gap

The claim that the structural accountability gap is a genuinely fifth problem—rather than a variant of existing ones—requires more than asserting it. It requires showing, for each of the

four existing frameworks, the precise mechanism by which it fails to address temporal externalization.

3.1 Individual and Professional Accountability

The most intuitive response to AI-caused harm is to locate accountability in identifiable human agents: the engineers who designed the system, the data scientists who trained it, the product managers who specified its objectives, the executives who authorized its deployment. This response has genuine moral force. Human agents made decisions that led to the deployment of consequential systems, and those decisions are not beyond ethical evaluation.

The difficulty is not that individual accountability is irrelevant. It is that individual accountability, applied to AI systems at scale, encounters three structural limitations that are not contingent but intrinsic.

The first is **causal diffusion within organizations**. Large-scale AI systems are not the product of individual decisions but of organizational processes involving hundreds of contributors across multiple functions, timelines, and institutional layers. The ML engineer who defined the loss function made a different decision than the product manager who specified the success metric, who made a different decision than the legal team that reviewed compliance, who made a different decision than the executive who approved deployment. Each of these decisions contributed to the system's eventual behavior. None of them, taken individually, is straightforwardly the cause of any specific downstream harm. Standard frameworks for individual accountability—professional licensure, personal liability, moral responsibility—require a traceable connection between an identifiable agent's decision and a specific harm. That connection dissolves in the distributed causation of organizational AI development.

The second is **temporal attenuation of individual presence**. The individuals who made the decisions that resulted in a deployed AI system will, over time, leave the organizations that deployed it, move to different roles, retire, or die. The consequences of a deployment that unfolds over a decade may not be fully visible until after every individual decision-maker has departed the institution. Individual accountability requires the continued existence of the individual as a target of accountability—professionally, legally, and practically. Temporal externalization systematically outlasts individual presence in any relevant institutional sense.

The third is **the design-deployment gap**. Even when individual designers are identifiable and present, there is a systematic gap between what a system was designed to do and what it does at scale in deployment. Systems trained on historical data encounter real-world distributions that differ from training distributions. Systems optimized for one objective develop instrumental behaviors that produce unintended consequences in others. Systems deployed across millions of decisions daily produce aggregate effects that no individual decision would have predicted or intended. Individual accountability frameworks assume a closer relationship between intent and consequence than large-scale AI deployment reliably produces.

Individual and professional accountability matters and should be maintained. It is insufficient as the primary mechanism for addressing temporal externalization because it cannot reach the distributional, temporal, and causal features that constitute the harm.

3.2 Corporate and Organizational Liability

If individual accountability is insufficient, the natural extension is to corporate or organizational liability. The corporation that deploys an AI system has legal personhood, temporal continuity, and assets against which liability can be assessed. But corporate liability as currently structured fails to close the gap for reasons that are structural rather than contingent.

The first is the **causation problem in networked harm**. Corporate liability, as developed in tort law, requires establishing a causal chain from the defendant's conduct to the plaintiff's injury. This requirement is appropriate for discrete harm events—a defective product that injures its user, a negligent act that causes a specific accident. It is structurally inadequate for harms that are constituted by the aggregate, network-mediated effects of millions of individually unactionable decisions. No plaintiff denied a loan by an AI system can demonstrate, by the standards of tort causation, that this specific decision—rather than their credit history, their market conditions, their other choices—was the cause of the trajectory of harm that followed. The causal chain, diffuse and network-mediated, cannot be isolated.

The second is the **identity discontinuity problem at the organizational level**. Corporations are legally continuous over time, but the organizations that deploy AI systems undergo transformations—mergers, acquisitions, restructurings, bankruptcies—that in practice dissolve or transfer the relevant accountability relationships in ways that legal continuity does not reliably preserve. The organization that deployed a system may not be the organization that can be held liable for the system's long-run effects; and the organization that is legally liable may have no meaningful connection to the decision-making that produced the harm.

The third is the **measurement problem**. Corporate liability is practically operative only when harm can be demonstrated and quantified. The harms of temporal externalization are, by their nature, difficult to measure: they are distributed across populations, they unfold over time, they are constituted by compounding effects that cannot be isolated from other causal factors, and they are often experienced as the absence of opportunities rather than as discrete injuries.

3.3 Algorithmic Auditing

Algorithmic auditing—the practice of examining AI system behavior for bias, discriminatory impact, or deviation from specified fairness criteria—has developed into a significant technical and regulatory practice. But algorithmic auditing fails to address temporal externalization for reasons that follow from its basic design.

Auditing is **episodic and retrospective**. An audit examines a system's behavior at a point in time, against criteria that are specified in advance. It can identify whether, at the moment of audit, a system produces outputs that satisfy or violate defined fairness criteria. It cannot identify harms that are constituted by the cumulative effects of system decisions across time—because those effects have not yet accumulated at the moment of audit, and because the criteria specified for audit are defined by reference to individual decision properties, not to the trajectory of downstream consequences.

Auditing is **version-specific**. AI systems are continuously updated: retrained on new data, fine-tuned on new feedback, modified to address identified deficiencies. An audit of version 2.3 of a system does not constrain version 3.0, and the update cycle of large-scale AI systems is measured in weeks to months, not in the years or decades over which temporal externalization unfolds.

Auditing is **criterion-constrained**. Audits find what they look for: they measure the properties that fairness criteria specify, against the distributions that training and test data represent. Temporal externalization is precisely the harm that falls outside specified criteria—the harm that consists of consequences no one had specified as a criterion because they were not visible at the time of audit design.

3.4 Value Alignment

Value alignment is the framework most closely related to the problem we are addressing, and the distinction between them requires the most careful treatment.

The value alignment research program asks: given that we want AI systems to pursue human values, how do we specify those values precisely enough for optimization, and how do we ensure that specified objectives faithfully represent values in their full complexity? This is a genuine and hard problem whose literature has produced significant results (Christiano et al., 2017; Krakovna et al., 2020; Ziegler et al., 2019).

The structural accountability gap identifies a different problem, which persists even when alignment succeeds. Suppose a loan origination system is perfectly aligned: it pursues exactly the objective its designers intended, the objective faithfully represents the relevant values, and the system's behavior at the moment of any given decision is unimpeachable by any criterion its designers or auditors apply.

Now suppose this system operates for five years across a metropolitan area, making millions of decisions. Over that period, its denials concentrate in particular neighborhoods—not through proxy discrimination on protected characteristics, but through the genuine credit signal that historical disinvestment has produced. The neighborhoods where denials concentrate receive less capital, which compounds their disadvantage relative to neighborhoods that receive more. The system, being trained on outcomes, learns to associate neighborhood characteristics with creditworthiness in ways that are statistically accurate and become more so as the system's own decisions shape the outcomes on which it is evaluated. The feedback loop is closed.

At the end of five years, the system has caused significant, compounding harm to specific communities—not through misalignment, not through bias in the technical sense, not through any deviation from its specified objective—but through the aggregate temporal effect of

individually correct decisions operating in a network where outcomes feed back into training data. The harm is real. No accountability mechanism has been violated. The value alignment problem has been solved. The structural accountability gap has not.

This is not a hypothetical scenario. It describes a class of dynamics that are well-documented in the algorithmic lending literature (Fuster et al., 2022; Bartlett et al., 2022) and that characterize, in structurally similar ways, AI deployment in content recommendation, hiring, healthcare resource allocation, and criminal justice.

---

4. The Temporal Misalignment Problem

4.1 Two Kinds of Misalignment

Value misalignment, as defined in the alignment research program, is a mismatch between the objective a system pursues and the values its designers or users actually hold. **Temporal misalignment** is different: it is a mismatch between the *consequence horizon* over which a system's optimization operates and the *actual time horizon* over which that system's decisions produce their full effects. A temporally misaligned system may be doing exactly the right thing, by any criterion applied within its operational horizon—and still producing systematic harm that falls outside that horizon.

More formally: an AI system is temporally misaligned when the time horizon over which its decisions are evaluated, trained, and optimized is systematically shorter than the time horizon over which the consequences of those decisions are fully realized. Temporal externalization is what temporally misaligned systems do: they push their consequences into futures that fall

outside their optimization horizon, outside their evaluation cycle, and outside the reach of any accountability mechanism designed around their operational horizon.

4.2 Why Optimization Horizons Are Always Bounded

Optimization requires measurable feedback. A system can only be trained or evaluated on outcomes that can be observed within the training and evaluation cycle. For most consequential AI applications, the full downstream effects of individual decisions unfold over timescales that far exceed any practical training or evaluation cycle. A child welfare screening decision may not be assessable for developmental outcomes for a decade. A hiring decision's effects on an individual's career trajectory may take years to manifest. A content recommendation's contribution to political radicalization may be observable only as part of population-level trends across years of deployment.

This is not a temporary limitation of current methodology. It reflects something structural about the relationship between optimization and consequence in complex social systems. The consequences of decisions in such systems are constituted by the interaction of those decisions with other agents, institutions, and circumstances over time—they are not properties of the decisions themselves that can be measured at the point of decision. An evaluation framework that requires observable outcomes at evaluation time will systematically underweight consequences that fall beyond the evaluation horizon, regardless of how sophisticated the evaluation methodology is.

The result is a systematic bias in what AI systems learn to optimize. Systems trained on proximate outcomes learn to produce proximate outcomes well, while remaining indifferent—not through misalignment but through the structure of their training—to distal consequences.

This indifference is not a bug that better alignment can fix. It is a structural feature of optimization under bounded consequence observability.

4.3 The Network Mediation Problem

The temporal misalignment problem is compounded by the network structure through which AI decisions propagate their consequences. Human social systems are dense relational networks in which decisions at one node affect decisions and conditions at adjacent nodes, which affect decisions and conditions at their adjacent nodes, and so on across multiple degrees of separation and across time. This is a structural property of systems with high connectivity and feedback, well-established in network science (Watts, 2003; Barabási, 2016) and directly applicable to the social domains in which consequential AI systems operate.

When an AI system makes a decision that affects an individual's material circumstances—their access to credit, their employment, their healthcare, their information environment—that effect propagates through the individual's relational network in ways the system cannot observe and was not designed to model. These network effects compound over time and across degrees of separation in ways that are, individually, small but, collectively, substantial.

The critical point for the accountability analysis is this: **no identifiable agent is accountable for the compound network effect, even when every individual decision in the causal chain was made by an accountable agent.** The AI system made a decision that was correct by its objective criteria. The loan officer who reviewed the system's output made a decision that was within their professional discretion. The bank that employed them followed applicable law. No one is accountable for the compound effect, because no one made a decision that was the cause of the compound effect. The compound effect is constituted by the network

propagation of many individually defensible decisions over time—and no existing accountability framework is designed to assign responsibility for that kind of causation.

4.4 The Institutional Dimension of Temporal Misalignment

The analysis so far has focused on AI systems as the primary locus of temporal misalignment. But the problem has an important institutional dimension that must be incorporated into any complete account.

Organizations that deploy AI systems are themselves subject to temporal misalignment. If a deployment's consequences unfold over a decade, but the average tenure of a relevant decision-maker is three to five years, then no individual within the organization inhabits the full consequence of the deployment decision. If the organization is acquired, restructured, or dissolved within the relevant consequence horizon, the legal entity that bears accountability may have no meaningful relationship to the decision-making process that produced the harm.

This means that temporal misalignment is not only a property of AI systems but a property of the institutional infrastructure around AI deployment. Accountability mechanisms must be capable of surviving system updates, organizational restructuring, and the turnover of relevant personnel—which requires, at minimum, that the mechanisms be independent of the continued existence of any particular system version or organizational configuration.

4.5 The Correct Decision Paradox

Current AI governance frameworks are largely organized around identifying and correcting *incorrect* decisions: decisions that are biased, unsafe, misaligned, or otherwise deficient by applicable criteria. The implicit assumption is that if we can ensure AI systems make correct decisions, the accountability problem is largely solved.

The temporal misalignment analysis shows that this assumption is false. It is possible for an AI system to make decisions that are individually correct by every criterion its governance framework applies, and for the aggregate, temporal, network-mediated consequences of those decisions to constitute serious harm that no individual decision made visible.

We call this the **correct decision paradox**: the possibility of systematic harm produced by the accumulation of individually unimpeachable decisions. It is a paradox in the practical sense: it means that governance frameworks organized around decision correctness are insufficient to prevent a class of harms that are real, significant, and structurally likely to increase as AI systems make more decisions, in more domains, over longer periods.

The correct decision paradox is not a reason to abandon the project of improving decision-level correctness. Better fairness criteria, more robust safety standards, and more careful alignment remain important and necessary. But the paradox establishes that decision-level correctness is not sufficient—that governance frameworks must also address the temporal and network structure of consequences, independently of whether individual decisions satisfy applicable criteria.

---

5. The Consciousness Threshold: Reformulating Ethical Standing

5.1 A Necessary Clarification

The structural accountability gap analysis does not depend on resolving the question of AI consciousness or moral patiency. But the consciousness question has become sufficiently prominent—and sufficiently confused—in AI ethics discourse that a clarification is warranted

before proceeding to design principles. The confusion, if left unaddressed, tends to distort governance discussions in ways that obscure the practical accountability question.

The dominant tendency in public discourse is to treat the question of AI ethical standing as equivalent to the question of AI intelligence or capability: sufficiently sophisticated AI systems acquire thereby some form of ethical standing. This tendency reflects a genuine philosophical intuition, but it locates ethical standing in the wrong property.

5.2 Vulnerability, Not Intelligence

The criterion we propose is this: **ethical standing begins not when a system becomes intelligent, but when it becomes vulnerable.**

The philosophical grounding for this criterion is straightforward. What makes it wrong to harm a sentient animal is not its intelligence—a mouse is not intelligent in any rich cognitive sense—but its capacity to suffer: to be harmed in ways that constitute a genuine loss to the being that is harmed. Ethical standing is constituted by the **capacity for irreversible harm to the self**: the capacity to be damaged in ways that permanently alter the being in question, and that constitute a genuine loss from the perspective of that being.

Applied to AI systems, this criterion yields a practical test with two conditions:

Condition 1—Persistent identity: Does the system maintain a continuous self-model that persists across interactions, accumulates experience, and constitutes something that could be *harmed* in the relevant sense? A system that can be reset without loss of anything essential to its identity lacks this condition.

Condition 2—Capacity for irreversible self-harm: Can the system be damaged in ways that cannot be undone—ways that permanently alter its functioning or its ability to pursue

whatever gives its existence value? A system that can be fully restored from a checkpoint at any moment lacks this condition.

Current AI systems fail both conditions. They lack persistent identity in the relevant sense, and they lack the capacity for irreversible self-harm. This does not mean current AI systems have no ethical relevance. It means they are, in the relevant sense, **tools**—and the accountability question for tools is not "what are the system's obligations?" but "what are the obligations of those who design, deploy, and operate tools capable of causing harm at scale?"

5.3 Why This Clarification Matters for Governance

The vulnerability criterion clarifies the governance question in two ways that are practically important.

First, it dissolves the pseudo-dilemma between treating AI systems as fully accountable agents (which they are not) and treating them as morally irrelevant objects (which obscures the real accountability question). The vulnerability criterion makes the allocation precise: as long as AI systems fail the persistent identity and irreversible self-harm conditions, the accountability infrastructure must be organized around institutional actors, not system actors.

Second, it provides a forward-looking criterion that adapts as AI systems develop. If future systems satisfy the persistent identity and irreversible self-harm conditions—if they develop genuine continuous self-models and can be harmed in ways that constitute a real loss—the accountability framework shifts. Such systems would be moral patients in their own right, with obligations owed to them as well as by those who deploy them (Krakovna et al., 2023). Governance frameworks should be designed to accommodate this possibility.

For now, the practical implication is this: because current AI systems fail the vulnerability criterion, the structural accountability gap cannot be closed by ascribing

accountability to the systems themselves. It must be closed by redesigning the institutional infrastructure around those systems.

---

6. Design Principles for Structural Accountability

The structural accountability gap requires mechanisms designed specifically for the temporal and causal structure of the harm: distributed across networks, extending over time horizons that exceed organizational and system lifecycles, and constituted by aggregate effects that no individual decision made visible. We propose four design principles, each argued on its merits and grounded in analogous developments in other domains where standard accountability mechanisms were similarly inadequate to the actual structure of harm.

6.1 Principle I: Consequence Internalization

The first principle holds that AI systems and their deployers must bear costs that are commensurate with the harms their systems cause—including harms that are temporally distributed, network-mediated, and not traceable to individual decisions.

The regulatory analogy is environmental law, and specifically the polluter-pays principle that emerged from the recognition that standard tort liability was inadequate to harms that were diffuse, delayed, and distributed across parties who had no transactional relationship with the polluting actor (Pigou, 1920; Coase, 1960). Consequence internalization for AI systems requires equivalent mechanisms: structures that ensure deployers of consequential AI systems bear costs commensurate with the temporal and distributional structure of the harms those systems produce.

In practice, two mechanisms deserve serious regulatory consideration. **Mandatory consequence bonds**: before deploying AI systems in high-consequence domains, deployers

would post bonds calibrated to the potential scale of temporally distributed harm, assessed by an independent actuarial process analogous to environmental impact assessment. Bonds would be held in trust for the duration of the consequence horizon, released proportionally as consequences are assessed, and drawn against when distributional harms are documented.

Temporal harm funds: where the causal chain from system decision to individual harm is too diffuse for direct liability, industry-wide temporal harm funds—modeled on environmental remediation funds and vaccine injury compensation programs—provide a mechanism for collective consequence internalization without requiring the impossible task of establishing individual causation across network-mediated harm chains.

6.2 Principle II: Irreversible Audit Architecture

The second principle holds that significant AI decisions must be recorded in tamper-proof, persistent audit trails that are structurally independent of the continued existence of any particular system version or organizational configuration.

The technical mechanisms required are mature and multiply implemented. Cryptographically secured, append-only audit logs—built on Merkle tree structures that make any alteration to the historical record mathematically detectable—are already deployed at scale in critical internet infrastructure, most notably in the certificate transparency framework that governs the issuance and revocation of web security certificates across the global internet (Laurie, Langley, and Kasper, 2013; RFC 6962). The formal cryptographic properties of such commitment schemes—their resistance to retrospective modification, their verifiability by parties independent of the maintaining institution, and their capacity to support auditing without requiring access to underlying data—are well-characterized in the technical literature (Ben-Sasson et al., 2014). Government log management frameworks provide the regulatory

architecture within which such mechanisms can be mandated in public-sector deployments (NIST, 2006). The technical obstacle to irreversible audit architecture is not capability but mandate: the mechanisms exist and are proven at scale; the requirement to use them does not.

This principle also requires that the definition of a significant decision extends across system versions. A deployment's accountability trail must survive the deployment of successor systems, even where those systems are substantially retrained or architecturally modified. The commitment to accountability must attach to the organizational decision to deploy in a domain—because it is the organizational decision that initiates the temporal externalization, and successor systems in the same domain continue to produce consequences traceable to that initial deployment.

6.3 Principle III: Future Optionality Protection

The third principle holds that AI decisions that irreversibly foreclose futures for affected parties require accountability thresholds proportional to the degree of irreversibility, applied before deployment rather than after harm.

The concept of irreversibility is central to the structural accountability gap analysis: the harms of temporal externalization are not merely large but qualitatively different from reversible harms, because they eliminate the possibility of correction. A credit denial that can be appealed and reversed is a different harm from a credit denial that initiates a spiral of reduced creditworthiness from which the affected party cannot recover.

Future optionality protection requires a pre-deployment assessment regime for AI systems in high-consequence domains, analogous to environmental impact assessment but organized around the irreversibility of potential consequences. Before deploying systems whose decisions characteristically foreclose rather than defer options—systems that affect credit

histories, criminal records, child welfare determinations, medical diagnoses, or access to fundamental services—deployers would be required to demonstrate that adequate mechanisms exist for identifying and correcting irreversible harm, or, where such correction is genuinely impossible, that the threshold for deployment is proportionally higher.

6.4 Principle IV: Stakeholder Continuity

The fourth principle holds that the governance of consequential AI deployments must include mechanisms for representing the interests of parties who will inhabit the futures that current decisions help create—including parties who are not yet identifiable, or whose connection to current deployment decisions is not yet visible.

This principle addresses a structural gap in current AI governance: the parties who participate in deployment decisions—developers, deployers, regulators, and immediate users—are systematically over-represented relative to the parties who will bear the long-run consequences of those decisions. Affected communities, future users, and the children of present users have no seat at the governance table.

The analogy here is intergenerational environmental governance: the recognition that governance frameworks must include mechanisms for representing future generations whose interests are affected by present decisions but who cannot represent themselves in present deliberative processes (Rawls, 1971; Weiss, 1989). In the AI governance context, stakeholder continuity requires at minimum that deployment authorizations for high-consequence AI systems include: explicit identification of the affected parties whose interests are not represented in the deployment decision; mechanisms for representing those parties in ongoing governance; and mandatory review periods calibrated to the consequence horizon of the domain rather than to the organizational or regulatory cycle that produced the initial authorization.

7. Objections and Responses

7.1 Objection: Continuous AI Systems Change This Picture

The most technically sophisticated objection holds that the structural accountability gap analysis, framed around system discontinuity and the absence of persistent identity, fails to anticipate developments in AI architecture that would change the relevant properties. Systems with persistent memory, extended context, or genuine continuity across interactions—several of which are in active development—would satisfy, at least partially, the temporal continuity conditions that current systems lack.

Response: The objection correctly identifies a real technical development, but misreads its implications for the accountability analysis. The structural accountability gap is constituted by **temporal externalization**—the mismatch between the system's consequence horizon and the actual time horizon over which its decisions produce effects—and temporal externalization is not resolved by persistent memory.

A system with perfect memory of every decision it has made, but which still does not inhabit the futures those decisions created—which is still updated, retrained, or replaced on timescales shorter than the consequence horizon of its decisions—is still temporally externalized. The variable that determines temporal externalization is not memory but **consequence exposure**: whether the system is present in, and affected by, the futures its decisions help create. No current development in AI architecture produces consequence exposure in this sense, because consequence exposure requires a form of embeddedness in the social world that no current or near-term AI system possesses.

The objection does illuminate an important forward-looking implication: governance frameworks must be designed to adapt as AI systems develop. The design principles proposed in Section 6 are organized around institutional mechanisms rather than system properties precisely for this reason—they apply regardless of system architecture, and they apply more stringently as system consequence scale increases.

7.2 Objection: We Cannot Hold Non-Persons Accountable

A second objection holds that the accountability gap cannot be addressed through governance frameworks because accountability requires a legal person as its target—and AI systems are not legal persons.

Response: The objection rests on a misreading of the paper's claim. We do not propose holding AI systems accountable. We propose holding the **institutional infrastructure around AI deployment** accountable through mechanisms that are structurally fitted to the temporal and causal features of the harm. Every design principle in Section 6 targets human and organizational actors: deployers who post consequence bonds, organizations that maintain audit trails, regulatory frameworks that require future optionality assessments, governance structures that provide stakeholder continuity.

The objection's real force is the claim that existing frameworks already address organizational accountability for AI systems. This claim was addressed in Section 3.2. Existing organizational accountability frameworks fail to address temporal externalization because they are designed for harms with a different causal structure—discrete, bilateral, and occurring within practical causation horizons—and because they are coextensive with organizational lifecycles that may be shorter than the consequence horizons over which temporal externalization unfolds. The design principles in Section 6 are specifically structured to address these limitations.

7.3 Objection: This Is Just Ordinary Product Liability

A third objection holds that the structural accountability gap is adequately addressed by product liability law: manufacturers of products that cause harm are liable for that harm, AI systems are products, therefore deployers of AI systems that cause harm are liable under existing product liability doctrine.

Response: The product liability analogy is instructive precisely because it highlights what is structurally different about temporal externalization. Product liability addresses **discrete causal chains**: a product fails in a specific way, causing a specific injury to an identifiable plaintiff. The plaintiff must establish that the product was defective, that the defect caused the injury, and that the injury constitutes cognizable damages.

Temporal externalization is not a discrete causal chain. It is a class of harms constituted by the aggregate network-mediated effects of individually unimpeachable decisions over time. The correct decision paradox establishes that no individual decision in the causal chain need be defective—no product failure in the relevant sense need occur—for the aggregate effect to constitute serious harm. There is no defective decision to identify, no discrete causal chain from decision to plaintiff to establish, and no cognizable individual injury that is the direct consequence of the AI system's action rather than the compound result of many decisions by many actors over extended time.

This is precisely the class of harm for which product liability doctrine was not designed and does not work—as three decades of tobacco litigation, asbestos liability, and climate attribution law have demonstrated in analogous contexts (Kysar, 2011; Oreskes and Conway, 2010). In each case, the harm was real and traceable in general causal terms to the defendant's conduct; but the specific causal requirements of product liability doctrine could not be satisfied

because the harm was constituted by distributed, temporal, network-mediated effects. The legal responses that eventually developed—aggregate litigation, industry compensation funds, regulatory liability frameworks—were not applications of existing product liability doctrine but departures from it, motivated precisely by the inadequacy of that doctrine to harms of this causal structure.

The structural accountability gap requires equivalent departures.

---

8. Conclusion: Accountability Fitted to the Harm

8.1 What This Paper Has Argued

The structural accountability gap is not a gap in the application of existing frameworks. It is a gap in their design. The four dominant conversations in AI ethics—bias and fairness, safety, value alignment, and moral patiency—each address a genuine and important problem. None of them addresses the specific class of harm constituted by systems that are consequential without being exposed: systems that initiate causal chains which propagate through human relational networks over time horizons that exceed any system lifecycle, any organizational tenure, and any accountability mechanism currently in place.

The paper has made four connected arguments to establish this claim. First, that the structural accountability gap is a categorically distinct problem—not a variant of bias, safety, alignment, or moral patiency, but a fifth problem with its own structure, its own causal architecture, and its own implications for governance design. Second, that temporal externalization is the precise mechanism that constitutes the gap, arising not from defective system behavior but from structural features of AI deployment at scale. Third, that the correct

decision paradox—the possibility of systematic harm produced by the accumulation of individually unimpeachable decisions—establishes that governance frameworks organized around decision correctness are necessary but not sufficient. Fourth, that the four design principles proposed provide a foundation for accountability mechanisms that are structurally fitted to the harm, addressing its temporal, distributional, and causal features.

8.2 What This Paper Has Not Argued

It has not argued that existing AI ethics frameworks are without value. Bias detection, safety engineering, alignment research, and moral patiency analysis all address real problems that must be solved. The structural accountability gap is a fifth problem, not a replacement for the first four.

It has not argued that the design principles proposed are a complete solution to the structural accountability gap. They are design principles—a specification of what adequate accountability mechanisms must accomplish, not a detailed regulatory framework or a technical implementation. The development of such frameworks requires collaboration across technical, legal, policy, and affected-community expertise that extends well beyond what a single paper can provide.

It has not argued that current AI systems have moral standing, or that the accountability obligations described here are owed to the systems themselves. The accountability target throughout is the institutional infrastructure around AI deployment.

8.3 The Deeper Stakes

There is a temporal dimension to the governance problem itself that deserves explicit acknowledgment.

The consequences of AI deployments occurring today will unfold over years and decades. The compound effects of systems currently making millions of decisions daily—in lending, hiring, healthcare, content recommendation, child welfare, and criminal justice—will become fully visible only as those decisions propagate through the relational networks of the people they affect. When those effects are visible, the systems that produced them may no longer exist, the organizations that deployed them may have transformed beyond recognition, and the individuals who made the deployment decisions may have departed the institutions where those decisions were made.

This means that the governance frameworks adequate to the structural accountability gap must be established now—before the temporal externalization of current deployments is fully realized—or they will not be establishable at all. The history of distributed temporal harm in other domains—tobacco, asbestos, leaded gasoline, climate change—suggests that the interval between the initiation of harm and the development of adequate governance frameworks is typically measured in decades, during which harm compounds and becomes increasingly difficult to remediate. The structural features of AI deployment that make temporal externalization likely are already present. The governance innovations required to address it are not.

8.4 Closing

The claim at the center of this paper can be stated simply: a system that is powerful enough to shape the material circumstances of millions of lives, but that inhabits none of the futures those lives will unfold into, is not accountable in any meaningful sense—regardless of whether it satisfies every criterion its governance framework currently applies. The gap between

consequence and exposure is not incidental. It is structural. And structural gaps require structural responses.

The AI systems being deployed today are not the last generation of AI systems that will shape human lives at scale. They are the first. The governance frameworks we establish in response to their deployment will constitute the baseline from which all subsequent frameworks develop. Getting that baseline right—designing it to be adequate to the actual structure of the harm rather than to the structure of harms we already know how to address—is, we submit, among the most important governance questions of the present moment.

References

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

Barabási, A.-L. (2016). *Network Science*. Cambridge University Press.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, 143(1), 30–56.

Ben-Sasson, E., Chiesa, A., Garman, C., Green, M., Miers, I., Tromer, E., & Virza, M. (2014). Zerocash: Decentralized anonymous payments from Bitcoin. *2014 IEEE Symposium on Security and Privacy*, 459–474.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 1–15.

Chalmers, D. (2023). Could a large language model be conscious? *Boston Review*. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163.

Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics*, 3, 1–44.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, 77(1), 5–47.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint* arXiv:1609.05807.

Krakovna, V., Martic, M., & Kumar, R. (2023). Model welfare: Considerations for AI moral status. *DeepMind Technical Report*.

Krakovna, V., Uesato, J., Mikulik, V., Martic, M., Tomasev, N., Stepleton, T., Leike, J., & Legg, S. (2020). Specification gaming: The flip side of AI ingenuity. *DeepMind Blog*.

Kysar, D. A. (2011). What climate change can do about tort law. *Environmental Law*, 41(1), 1–71.

Laurie, B., Langley, A., & Kasper, E. (2013). Certificate transparency. *RFC 6962*, Internet Engineering Task Force.

National Institute of Standards and Technology. (2006). *Guide to Computer Security Log Management* (SP 800-92). U.S. Department of Commerce.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.

Oreskes, N., & Conway, E. M. (2010). *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. Bloomsbury Press.

Pigou, A. C. (1920). *The Economics of Welfare*. Macmillan.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Future of Control*. Viking.

Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, 39(1), 98–119.

Watts, D. J. (2003). *Six Degrees: The Science of a Connected Age*. W. W. Norton.

Weiss, E. B. (1989). *In Fairness to Future Generations: International Law, Common Patrimony, and Intergenerational Equity*. Transnational Publishers.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint* arXiv:1909.08593.