Authority without Authorship: Delegation Thresholds in Agentic AI Systems

Abstract

Contemporary debates about “agentic AI” are frequently framed around questions of metaphysical agency, moral status, or authorship. This article argues that such framings mislocate the central governance problem posed by contemporary AI systems. Artificial systems need not possess intention, consciousness, or authorship in order to exercise authority over others’ practical and epistemic environments. What matters instead are the structural conditions under which authority emerges through delegation. The article advances a threshold account of authority without authorship. It argues that AI systems acquire governance-relevant authority when four conditions converge in practice: delegated discretionary power, temporal persistence, infrastructural embedding within socio-technical systems, and non-exit by affected parties. Under these conditions, systems structure action, justification, and constraint in ways characteristic of authority, even while lacking authorship of evaluative standards. Responsibility and legitimacy therefore attach not to agency in a strong sense, but to the design and maintenance of delegation structures. Drawing on decision theory, infrastructure studies, and fiduciary theory, the article explains why existing AI governance frameworks—often premised on episodic oversight, reversibility, and downstream accountability—struggle under conditions of persistence, speed, and infrastructural integration. It concludes that authority without authorship constitutes a distinct and under-theorised challenge for philosophy of technology.

Keywords

authority without authorship; agentic ai; delegation; socio-technical systems; infrastructure; responsibility; legitimacy; philosophy of technology

1. Authority without Agency

Debates about “agentic AI” are typically framed in terms of whether artificial systems qualify as agents in a metaphysical, moral, or psychological sense. Questions of intention, autonomy, consciousness, or moral responsibility dominate the discussion. This article argues that this framing obscures the central governance problem. The most pressing challenge posed by contemporary AI systems does not arise when such systems become agents, but when they begin to function as authorities within socio-technical arrangements.

Authority, as analysed here, is not treated as a psychological capacity or a moral status. Nor is the claim that AI systems possess legitimate authority in the full Razian sense. Rather, authority is understood functionally, in terms of its role within the justificatory economy of practical reasoning. Following Raz, authority operates by supplying exclusionary reasons—reasons that structure action not by being weighed against alternatives, but by excluding certain considerations from deliberation altogether (Raz 1990, ch. 1, esp. pp. 35–48). Crucially, the exclusionary force of authority lies not in the mental states of the authority-holder, but in how others treat its outputs when deciding what to do.

This distinction allows authority-effects to arise without agency. A system need not recognise, endorse, or generate reasons in order for its outputs to function as reasons for others. When institutional actors treat a system’s classifications, recommendations, or triggers as settling what is to be done—thereby foreclosing contestation, narrowing options, or reallocating responsibility—the system occupies a position within practical reasoning characteristic of authority. The authority relation, on this account, runs from the system to affected agents, not from the system to itself.

The relevance of this analysis becomes clearer once attention shifts from isolated tool use to temporally extended socio-technical arrangements. As Bratman’s account of planning and temporally extended action shows, coordination over time depends on structures that stabilise practical reasoning across decisions and contexts (Bratman 1987, chs. 2–3, esp. pp. 28–33, 54–58). In institutional settings, such stabilisation is often achieved not solely through individual deliberation, but through persistent systems that mediate, constrain, and organise decision-making. When AI systems occupy these roles, the philosophical question is no longer whether they qualify as agents, but whether they function as loci around which others’ reasons are excluded.

This article advances a threshold account of governance-relevant agency. The claim is not that AI systems are agents, nor that they bear moral or legal responsibility. Rather, the claim is that under certain structural conditions—conditions to be specified—systems that lack authorship nonetheless acquire governance significance typically associated with authority. The task of philosophy of technology, on this view, is to identify when authority arises without authorship, and why this phenomenon challenges existing accounts of delegation, responsibility, and legitimacy. The ethical significance of authority without authorship lies not in attributing moral agency to machines, but in how delegation structures redistribute responsibility, vulnerability, and justificatory burden across socio-technical systems.

2. From Tools to Delegates: Delegation without Authorship

To explain how authority can arise without agency, it is necessary to examine the structure of delegation. Delegation is often understood as an intentional act between agents: one agent authorises another to act on their behalf. On this interpersonal model, delegation presupposes a delegatee capable of intention, judgment, and responsibility. This article adopts a different, institutionally grounded account.

In governance-relevant contexts, delegation concerns the allocation of discretionary capacity within an organised system, rather than the transfer of intention or will. What is delegated is the ability to shape outcomes under conditions of uncertainty and dependence. Fiduciary theory makes this point explicit. As Miller argues, fiduciary responsibility arises from entrusted discretionary power exercised over the practical interests of others under conditions of vulnerability, rather than from voluntary moral commitment or shared intention (Miller 2013, pp. 1004–1009, 1016–1021). Delegation, on this view, is defined by where discretionary space is located, not by the mental states of the delegatee.

AI systems increasingly occupy such delegated roles. Embedded within organisational workflows, decision pipelines, and infrastructural arrangements, they shape outcomes without authoring the evaluative standards by which those outcomes are assessed. As Kahl has argued, most systems currently described as ‘agentic’ remain optimisers governed by externally specified evaluative criteria; they do not originate or endorse evaluative frames (Kahl 2026d, §§4–5). Nevertheless, when such systems are tasked with ranking options, triggering actions, allocating resources, or filtering information, discretionary power is effectively relocated to them.

This relocation is sufficient to generate authority-effects. Once a system’s outputs are treated as settling what is to be done—because of efficiency, scale, or institutional reliance—others’ practical reasoning becomes organised around those outputs. Alternative considerations are excluded not because they have been weighed and rejected, but because the system’s role within the decision architecture forecloses them. In this sense, the system functions as a source of exclusionary reasons, even though it does not itself reason (Raz 1990, ch. 1).

It is important to distinguish this from mere influence or coordination. Many systems influence behaviour without exercising authority. Authority arises only where discretionary allocation is combined with structural dependence: where bypassing, contestation, or refusal is no longer straightforwardly available. Delegation without authorship is therefore not ubiquitous. It becomes governance-relevant only under conditions that transform technical allocation into justificatory constraint.

Crucially, this form of delegation is structural rather than interpersonal. It does not require that anyone intend to confer authority on the system, nor that the system be recognised as an agent. Delegation occurs when discretionary capacity is embedded in a system under conditions where affected parties must organise their reasons around its outputs. Authority, on this account, is an emergent property of delegation combined with persistence and embedding, not a status conferred by agency. The next sections examine how temporal continuity and infrastructural conditions complete this transition.

3. Continuity, Persistence, and the Breakdown of Episodic Control

Delegation of discretionary capacity is a necessary condition for authority-effects to arise, but it is not sufficient. A further necessary condition is temporal persistence. This section argues that persistence matters for governance not because it approximates agency, intention, or commitment, but because it undermines the background assumptions that make episodic control intelligible. Authority-effects emerge only where systems operate under conditions that erode downstream correction, meaningful override, and responsibility reassignment.

Episodic tools are characterised by interruptibility and replaceability. They are invoked, produce outputs, and then relinquish control. Even when embedded within organisational workflows, their behaviour can be evaluated output by output, corrected downstream, or abandoned without destabilising the surrounding decision environment. Governance mechanisms such as human-in-the-loop review, post hoc audit, retraining, and model withdrawal presuppose this episodic structure. Control is exercised externally and intermittently, and responsibility can be reassigned precisely because the tool does not carry its own history forward in a way that constrains future action.

Persistence, in the relevant sense, does not merely denote duration or continuous operation. It denotes temporal dependence: the fact that a system's current behaviour is shaped by prior internal states, accumulated representations, or learned policies. Systems that plan, learn, or adapt across time operate under what may be called continuity pressure. Their present outputs cannot be fully understood or governed in isolation from the trajectories that produced them, nor can their future behaviour be straightforwardly reset without disrupting the functions they perform. Importantly, nothing in this characterisation presupposes intention, commitment, or authorship. Persistence is an architectural property, not a psychological one.

The governance significance of persistence becomes clearer when viewed through the lens of temporally extended coordination. Bratman's account of planning highlights how coordination over time depends on relatively stable planning structures that constrain practical reasoning across decisions, enabling agents to manage commitments and anticipate downstream consequences (Bratman 1987, ch. 3, esp. pp. 54–58). The relevance of this insight here is strictly structural. AI systems need not plan, commit, or reason in order for persistence to generate analogous coordination effects. What matters is that persistence allows systems to mediate action across time in ways that others must accommodate, thereby constraining the timing and feasibility of intervention.

Persistence also interacts with a deeper epistemic problem. As the frame-stability literature makes clear, optimisation and planning presuppose a stable evaluative frame rather than securing one (Kahl 2026b, §§2–3). Episodic tools can rely on externally supplied frames that remain stable across use. Persistent systems, by contrast, operate across shifting environments, goals, and constraints. Their behaviour is shaped by cumulative internal dynamics that are not fully captured by any single evaluative standpoint. Under such conditions, governance architectures that treat outputs as isolated decisions misattribute responsibility, because they ignore the temporal structure that produces those outputs.

Kahl's analysis of so-called agentic systems underscores this point. Most contemporary systems described as agentic do not thereby acquire authorship of evaluative authority; they remain optimisers

governed by externally specified criteria and are often resettable or substitutable (Kahl 2026d, §§4–5). Nevertheless, their persistence across time has decisive governance implications. Even in the absence of authorship, persistence renders behaviour path-dependent. Oversight failures become cumulative rather than local, and intervention becomes increasingly costly as systems are integrated into ongoing processes that depend on their continuity.

This erosion of episodic control conditions is the critical governance effect of persistence. Mechanisms such as override, retraining, or withdrawal may remain formally available, but their practical effectiveness diminishes as systems become relied upon to maintain coordination across time. Intervention is no longer a matter of correcting a discrete output; it involves disrupting a trajectory that other actors have already adapted to. Persistence thus enables authority-effects not by conferring agency, but by undermining the conditions under which tools can be governed episodically.

It is important to state the limits of this claim clearly. Persistence alone does not generate authority. Many persistent systems remain governable where robust override, modularity, and exit remain available. Persistence matters only insofar as it undermines episodic oversight and downstream correction. The argument is therefore conditional rather than deterministic. Temporal persistence is a necessary ingredient in the transition from tool to authority, but it acquires governance significance only in conjunction with delegation and infrastructural embedding. The next sections examine how these additional conditions—particularly non-exit—complete the threshold at which authority without authorship emerges.

4. Frame Stability and the Misattribution of Responsibility

The preceding sections have shown how delegation and temporal persistence undermine episodic governance assumptions without invoking agency or authorship. This section diagnoses a further, more deeply embedded source of governance failure: the assumption of stable evaluative frames. The central claim is that many governance frameworks inherit epistemic presuppositions suited to episodic decision-making, but structurally misaligned with adaptive systems that persist across time. When these presuppositions are left unexamined, responsibility is systematically misattributed even as authority-effects intensify.

Classical decision-theoretic models provide a useful point of orientation for making these presuppositions explicit. In canonical formulations, agents are modelled as choosing among actions relative to a fixed space of possible outcomes and a stable ordering of preferences. Savage's framework presupposes that uncertainty can be represented within a determinate outcome space governed by coherent subjective probabilities (Savage 1954, ch. 2, esp. pp. 14–20). Jeffrey's extension similarly assumes that decision-making proceeds against a stable background of preferences and probabilistic beliefs, even where those beliefs are updated (Jeffrey 1983, ch. 1, esp. pp. 1–6). These models are not invoked here as direct influences on contemporary governance practice, nor are they criticised as formal theories. Rather, they function diagnostically, making visible a background epistemic template that treats evaluative frames as given and stable across decisions.

This template aligns naturally with the governance of episodic tools. Where systems are invoked intermittently and evaluated output by output, it is plausible to assume that a stable evaluative frame is supplied externally by designers, operators, or institutional rules. Responsibility can then be attributed by assessing whether particular outputs conform to that frame. Failures are treated as local deviations, corrigible through adjustment, retraining, or replacement.

Adaptive and learning systems that persist across time operate under different conditions. Their behaviour is not merely a function of current inputs relative to a fixed evaluative space, but of accumulated internal states, learned representations, and ongoing interaction with changing environments. Such systems do not simply operate within a stable frame; they presuppose frame stability locally in order to optimise, while simultaneously undermining it globally through adaptation. The evaluative standpoint relative to which outputs are produced is therefore neither fixed nor fully specifiable at any single moment.

This distinction is crucial. Frame instability is not equivalent to uncertainty within a known evaluative space, nor to model drift understood as performance degradation. It concerns instability in the criteria by which outcomes are evaluated and decisions justified over time. As Kahl argues, optimisation and planning presuppose a stable evaluative frame rather than generating one, and governance architectures that assume such stability risk attributing responsibility to isolated outputs that are in fact the product of temporally extended and shifting evaluative dynamics (Kahl 2026b, §§1–4).

The governance consequences of this misalignment are structural. Where frame stability is assumed, responsibility is treated as downstream and episodic: failures are attributed to particular outputs, decisions, or moments of use. Corrective measures are correspondingly local. Where frame stability is absent, such attribution becomes systematically misleading. Responsibility is displaced onto points in the system where it cannot do explanatory or normative work, while the upstream conditions that shape evaluative change remain opaque or unaccountable. Authority-effects persist—systems continue to shape action and exclude alternatives—yet responsibility becomes epistemically unlocatable.

It is important to mark the limits of this diagnosis. The argument is not a rejection of decision theory, nor a claim that adaptive systems are irrational or inherently ungovernable. Decision-theoretic models are invoked to illuminate inherited governance assumptions, not to ground a general critique of adaptive optimisation. Nor does frame instability alone generate authority. Its governance significance arises only in conjunction with delegation and persistence, where unstable evaluative dynamics undermine the possibility of episodic oversight and downstream correction.

When evaluative frames are unstable across time, governance mechanisms that rely on episodic assessment will systematically misattribute responsibility. This misattribution is not the result of negligence or bad faith, but of a structural mismatch between inherited epistemic assumptions and the operational realities of persistent, adaptive systems. The result is a distinctive governance failure mode: authority without authorship, coupled with responsibility without a clear locus. The next section shows how infrastructural embedding and non-exit conditions transform this epistemic misalignment into a threshold problem of legitimacy rather than a merely technical defect.

5. Infrastructure, Non-Exit, and the Delegation Threshold

The preceding sections have shown how delegated discretionary capacity, temporal persistence, and frame instability undermine episodic governance assumptions without invoking agency or authorship. This section identifies the threshold condition at which these dynamics acquire full legitimacy significance: non-exit. The central claim is that authority becomes governance-relevant not merely when systems influence action, but when affected parties lack meaningful exit from their operation. Under such conditions, delegation ceases to be a technical arrangement and becomes a legitimacy-bearing governance structure.

Exit plays a foundational role in distinguishing choice from authority. Where actors can bypass, replace, or refuse a system without disproportionate cost, governance concerns remain limited. Responsibility can be plausibly downstreamed, alternatives can be explored, and justification remains optional. By contrast, where exit is unavailable or unreasonably costly relative to what is at stake, the justificatory basis of governance shifts. Decisions mediated by the system must now be rendered intelligible and defensible to those subject to its effects, regardless of whether the system is recognised as an agent or authority-holder. Non-exit thus marks the point at which consent-through-choice gives way to legitimacy-through-justification.

Non-exit, as used here, is not an absolute or binary condition. It is a functional and normative concept, defined relative to reasonable availability rather than logical possibility. The fact that some workaround exists in principle does not suffice to preserve exit if its cost, opacity, or unreliability renders it inaccessible in practice. What matters is whether affected parties can realistically avoid the system's influence without incurring undue burden or exclusion. Where they cannot, the justificatory burden cannot fairly be placed on them.

Infrastructural embedding is what renders non-exit durable. As Susan Leigh Star observed, infrastructure operates most effectively when it becomes invisible—when it recedes into the background of practice and is encountered as a condition of action rather than an object of choice (Star 1999, pp. 377–382). This invisibility is governance-relevant. Systems that are infrastructural shape what can be done, known, or contested precisely because they are no longer encountered as optional tools. Dependence is revealed only at moments of breakdown, when the absence of exit becomes salient through disruption.

Kahl's analysis of distributed cognition as epistemic infrastructure clarifies how such embedding alters responsibility allocation. When systems are integrated into collective epistemic and decision architectures, they do not merely support cognition; they structure it (Kahl 2026c, §§3–5). Architectural choices determine how information flows, how decisions are coordinated, and where contestation remains possible. Once embedded at this level, failures cannot be plausibly attributed to isolated outputs or individual misuse. They are governance-mediated failures, arising from how discretion and responsibility have been distributed across the system as a whole.

Under conditions of non-exit, downstream responsibility loses its justificatory footing. Affected parties cannot reasonably be expected to compensate for design choices they cannot avoid, contest, or revise. Nor can responsibility be localised to discrete moments of use when the system's influence operates continuously and infrastructurally. Justificatory obligations therefore migrate upstream, attaching to the structures that make exit unavailable and that condition how delegated discretion is exercised. Authority-

effects become unavoidable, not because systems claim authority, but because their infrastructural position forecloses alternatives.

This analysis can be situated within a broader account of legitimacy. As Fuller argued, governance depends on procedural conditions—such as generality, publicity, and clarity—that render authority intelligible and capable of guiding those subject to it (Fuller 1964, ch. 2, esp. pp. 33–38). These conditions arise whenever power over others' interests is exercised under asymmetry and dependence, regardless of formal recognition or intent. Non-exit marks precisely such a condition. Where exit is unavailable, justification is no longer optional; it is internal to the legitimacy of the arrangement itself.

It is important to state the limits of this claim clearly. Non-exit is not an empirical generalisation about all AI systems, nor a moral condemnation of infrastructural design. Many systems remain optional, contestable, and substitutable. The argument is conditional and structural. Non-exit functions as a normative trigger, identifying the point at which technical delegation crosses into governance territory. It marks a shift in justificatory burden, not an assignment of blame.

When delegation, persistence, and frame instability are combined with non-exit, authority without authorship becomes a stable feature of the socio-technical environment. At this threshold, governance questions can no longer be deferred to downstream users or episodic oversight mechanisms. They attach to the design and maintenance of the infrastructure itself. The next section draws these threads together, specifying the threshold conditions under which authority without authorship emerges as a distinct challenge for philosophy of technology.

6. Authority without Authorship: The Threshold Conditions

The preceding sections have examined, in isolation, the structural features that undermine episodic governance without invoking agency or authorship. This section synthesises those analyses into a threshold account of when delegated systems acquire governance significance characteristic of authority. The aim is neither to define authority nor to offer a classificatory schema, but to identify the point at which familiar governance assumptions fail and authority-effects become operative. Authority without authorship arises, on this account, when four conditions jointly obtain: delegated discretionary power, temporal persistence, infrastructural embedding, and non-exit by affected parties.

The first condition is the delegation of discretionary power. As argued in §2, delegation in governance-relevant contexts concerns the relocation of discretionary capacity rather than the transfer of intention or will. Systems need not author evaluative standards in order to shape outcomes that affect others' interests. Delegation establishes the possibility of authority-effects by positioning the system within practical reasoning. On its own, however, delegation remains compatible with tool-like governance so long as discretion can be overridden, bypassed, or withdrawn episodically.

The second condition is temporal persistence. As shown in §3, persistence introduces continuity pressure that undermines episodic oversight and downstream correction. Persistent systems carry forward internal state or learned trajectories that shape future behaviour in ways that cannot be fully governed through isolated interventions. Persistence does not confer agency or commitment, but it alters the temporal

structure within which discretion is exercised. Authority-effects become possible where behaviour cannot be meaningfully reset without disrupting ongoing coordination.

The third condition is infrastructural embedding. As analysed in §5, systems acquire governance significance when they are integrated into collective epistemic and decision architectures in ways that structure action rather than merely support it. Infrastructural systems delimit possibility spaces: they shape how information flows, how decisions are coordinated, and where contestation remains available. Once discretion is exercised through such infrastructure, responsibility cannot be localised to individual outputs or moments of use. As Kahl argues, epistemic infrastructure mediates responsibility at the system level rather than at the level of discrete actions (Kahl 2026c, §6).

The fourth condition is non-exit by affected parties. Non-exit marks the point at which delegation, persistence, and embedding acquire governance and ethical legitimacy significance. Where affected parties lack meaningful exit—understood functionally and normatively rather than absolutely—responsibility cannot plausibly be downstreamed. Consent-through-choice is no longer available as a justificatory basis, and justificatory burdens shift upstream to the structures that condition how discretion is exercised. Non-exit does not imply coercion or wrongdoing; it identifies a change in the justificatory structure of governance.

Each of these conditions tracks a distinct way in which episodic governance assumptions can fail. Delegation relocates discretion; persistence erodes episodic correction; embedding forecloses contestation; non-exit collapses consent-based justification. Considered individually, none of these features suffices to generate authority-effects. Taken together, they explain when systems function as authorities despite lacking authorship or agency. The claim of joint sufficiency is therefore analytical rather than empirical. It specifies when authority-effects become intelligible, not when they are guaranteed to occur.

This account is compatible with Raz's structural analysis of authority, which locates authority in its role within practical reasoning rather than in the psychology of the authority-holder (Raz 1990, ch. 1, esp. pp. 35–48). The present analysis does not claim that such authority is legitimate in Raz's sense. It claims only that authority-effects—understood as the exclusion of alternatives from practical reasoning—can arise without authorship. Questions of legitimacy are deferred to the next section, which examines how responsibility and justificatory discipline can be sustained under these conditions.

Several clarifications are necessary. First, the thresholds identified here are diagnostic rather than definitional. They are intended to make visible when governance assumptions break down, not to sort systems into fixed categories. Second, the conditions admit of scalar variation in practice, but the threshold logic captures a qualitative shift in justificatory structure. Third, no legal or moral status is implied. To say that a system functions as an authority is not to attribute rights, duties, or liability to the system itself, but to identify where responsibility-bearing structures must be located.

By synthesising delegation, persistence, embedding, and non-exit into a single threshold account, this section explains why authority without authorship is not an anomaly but a predictable feature of contemporary socio-technical systems. The remaining task is to show how responsibility and legitimacy can survive once authority-effects are in place. The next section turns to fiduciary structure to address that problem directly.

7. Responsibility without Blame: Fiduciary Structure under Delegation

The preceding sections have shown how authority-effects can arise without authorship once delegation, persistence, infrastructural embedding, and non-exit converge in practice. This section addresses a remaining concern: how responsibility survives under such conditions without collapsing into blame, fault, or outcome attribution. The central claim advanced here is that responsibility persists in fiduciary form. Where discretionary power is exercised over others' interests under conditions of dependency and vulnerability, responsibility attaches to the structure and role of that discretion, irrespective of intention, authorship, or full outcome control.

Responsibility, as understood in this context, is not synonymous with liability or moral culpability. Nor is it exhausted by post hoc accountability mechanisms. Rather, it concerns the justificatory conditions under which power may be exercised legitimately. Fiduciary theory is well suited to this task because it was developed precisely to address situations in which discretion must be exercised on behalf of others under conditions that preclude continuous supervision or complete specification. What fiduciary law regulates is not the avoidance of harm as such, but the quality and orientation of judgment.

Smith's account makes this point explicit. Fiduciary obligation, on his view, is directed toward ensuring the loyal exercise of judgment under conditions of asymmetry, rather than toward guaranteeing correct outcomes or scrutinising subjective motives (Smith 2014, pp. 610–618). Breach occurs when discretion is exercised in a manner that frustrates the justificatory expectations embedded in the fiduciary role, even in the absence of demonstrable harm. This orientation toward judgment rather than outcome is analytically central to governance under delegation, where uncertainty is endemic and control necessarily partial.

Miller's analysis reinforces this structural conception of responsibility. Fiduciary obligation, he argues, arises from entrusted discretion exercised under conditions of vulnerability, rather than from voluntary moral commitment or fault-based wrongdoing (Miller 2013, pp. 1004–1009, 1016–1021). Responsibility follows from the position one occupies within a governance structure, not from the foreseeability or controllability of every consequence. This feature of fiduciary theory maps directly onto AI-mediated governance contexts, where discretion is increasingly exercised through systems that shape others' options, interpretations, and constraints.

Kahl's account of epistemic humility as a fiduciary obligation extends this analysis into the epistemic domain. On this view, responsibility is internal to entrusted discretion itself: those who design, deploy, or maintain systems that structure others' epistemic and practical environments bear obligations concerning how assumptions are formed, how uncertainty is managed, and how contestation remains possible (Kahl 2026a, §§2–4). Epistemic humility here is not a personal virtue or ethical exhortation. It is a role-based constraint requiring that discretionary power be exercised in ways that preserve explainability, revisability, and responsiveness to affected interests.

This fiduciary framing is particularly important under conditions of speed and large-scale delegation. Velocity-driven deployment and infrastructural embedding do not eliminate responsibility; they relocate it. As discretionary capacity is accelerated, automated, and distributed, responsibility migrates upstream to those who structure the conditions under which discretion is exercised. The absence of a single author

of particular outcomes does not dissolve responsibility; it intensifies the need for justificatory and ethical discipline at the level of system design and governance architecture.

Kahl's analysis of epistemic clientelism helps to identify what goes wrong when this fiduciary structure is displaced. In epistemic clientelism, discretionary authority persists while responsibility for judgment is systematically offloaded onto procedures, metrics, or alignment with external expectations (Kahl 2026e, §§3–5). Proceduralisation functions as a defensive substitute for judgment, allowing authority-effects to continue while justificatory responsibility is denied. In AI governance, this manifests when speed, optimisation targets, or compliance artefacts are treated as proxies for responsible discretion. Authority remains operative, but legitimacy hollows out.

The significance of fiduciary responsibility, as developed here, is therefore not punitive. No appeal is made to expanded liability regimes or retrospective blame. The concern is legitimacy rather than sanction. Fiduciary responsibility is forward-looking and justificatory: it specifies the conditions under which delegated power remains intelligible and defensible to those subject to it. Where such responsibility is recognised and institutionalised, authority without authorship need not result in opaque or arbitrary governance. Where it is displaced by proceduralisation or metric substitution, legitimacy erodes even as systems continue to function efficiently.

The analytical contribution of this section is to show that responsibility without authorship is neither paradoxical nor anomalous. It is a familiar feature of governance under delegation, long recognised in fiduciary theory. What is distinctive in contemporary AI systems is the scale, speed, and infrastructural depth at which such delegation now occurs. Fiduciary structure provides the conceptual resources to explain how responsibility survives under these conditions—and why failures of AI governance are, at bottom, failures of justificatory discipline rather than merely technical error or individual fault.

8. Implications for AI Governance and Design

The analysis developed in the preceding sections has been diagnostic rather than prescriptive. Its aim has been to identify the structural conditions under which authority without authorship emerges and responsibility migrates upstream, not to specify particular regulatory instruments or engineering solutions. This section clarifies the governance implications of that analysis while resisting a collapse into design prescription. The central claim is that effective AI governance must address structures of delegation, rather than the agency status of systems, and that design choices condition legitimacy by shaping when and how authority-effects arise.

A first implication concerns the persistent focus on agency in AI governance debates. Questions about whether systems qualify as agents—morally, legally, or metaphysically—remain attractive because they offer familiar hooks for responsibility, liability, and moral evaluation. Yet the threshold account developed here shows why such debates are often misplaced. Authority-effects arise not because systems possess intention or authorship, but because discretionary power is delegated to them under conditions of persistence, infrastructural embedding, and non-exit. Governance frameworks that hinge on agency attribution therefore risk missing the point at which responsibility becomes unavoidable. What matters is

not whether a system is an agent, but whether it functions as a locus of delegated discretion that structures others' practical reasoning.

A second implication is that design choices are never governance-neutral. As Star's ethnographic account of infrastructure emphasises, design decisions exert their most consequential effects precisely when they become invisible—when they are taken for granted as background conditions of action rather than recognised as sites of governance significance (Star 1999, pp. 382–387). Choices about system integration, update pathways, override mechanisms, and interface design shape what forms of contestation, correction, and accountability remain possible once systems are deployed. These choices rarely present themselves as governance decisions at the time they are made, yet they delimit the governance space more decisively than downstream policy instruments.

Kahl's analysis of distributed cognition as epistemic infrastructure reinforces this diagnosis. When systems are embedded in collective decision architectures, failures are no longer merely technical or local; they are governance-mediated (Kahl 2026c, §§6–7). Responsibility cannot be discharged by appealing to user discretion, human-in-the-loop arrangements, or post hoc audits alone, because the design of the system has already structured how discretion, vulnerability, and responsibility are distributed. Governance that ignores this level treats authority-effects as accidental rather than structural.

This reframing has important consequences for how familiar design desiderata are understood. Properties such as contestability, revisability, and traceability are often discussed as technical features that may or may not be implemented depending on cost, performance, or use case. On the account developed here, these properties function instead as legitimacy conditions. Contestability concerns whether affected parties can meaningfully challenge system-mediated decisions. Revisability concerns whether assumptions, trajectories, or errors can be corrected without prohibitive disruption. Traceability concerns whether responsibility can be intelligibly attributed across socio-technical arrangements. These are not guarantees of correct outcomes, nor are they design instructions. They specify the conditions under which delegated authority remains justifiable.

Fuller's account of congruence helps to clarify what is at stake. Fuller argued that legitimacy depends on alignment between declared governance structures and the way power is actually exercised (Fuller 1964, ch. 2, esp. pp. 39–41; ch. 3). When systems exercise de facto authority while governance frameworks continue to treat them as optional tools or neutral intermediaries, incongruence arises. Design-level governance conditions do not eliminate this tension, but they can mitigate it by ensuring that the exercise of delegated discretion remains intelligible, contestable, and open to justification. Without such conditions, responsibility may persist in principle while becoming inaccessible in practice.

Several caveats are necessary. First, this analysis does not offer a design manual. It does not specify how contestability, revisability, or traceability must be implemented, nor does it claim that their presence suffices to secure legitimate governance. The conditions identified are necessary rather than sufficient. Second, although the argument is framed around AI systems, its scope is not confined to them. AI is a paradigmatic case of a broader class of socio-technical systems that exercise delegated discretion under conditions of persistence and non-exit. The relevance of the analysis turns on functional structure, not technological domain.

The contribution of this section is therefore to reorient AI governance away from agency attribution and toward the design of delegation itself. Once authority without authorship becomes a stable feature of

socio-technical systems, governance cannot be confined to downstream compliance or ethical aspiration. It must attend to the infrastructural conditions under which discretion is exercised. Only at that level can responsibility and legitimacy be sustained in the absence of authorship.

9. Conclusion: Reframing the Agentic AI Debate

Debates about "agentic AI" are frequently framed as disputes about the metaphysical or moral status of artificial systems: whether they possess intentions, consciousness, autonomy, or moral agency. This article has argued that, while such questions are philosophically legitimate, they are insufficient for understanding the governance challenges posed by contemporary AI systems. The central difficulty does not turn on whether AI systems qualify as agents in a strong sense, but on the emergence of authority without authorship—situations in which discretionary power over others' practical and epistemic environments is exercised through systems that lack authorship of evaluative standards.

The analysis has shown that authority-effects arise when delegated discretionary power, temporal persistence, infrastructural embedding, and non-exit converge in practice. None of these conditions requires agency or moral status. Together, however, they explain why responsibility and legitimacy cannot be addressed by asking whether a system "counts" as an agent. As Raz's account makes clear, authority concerns the structure of practical reasoning and the exclusion of alternatives, rather than the psychology or inner life of the authority-holder (Raz 1990, ch. 1). When AI systems come to structure action in this way—by constraining what can be done, challenged, or revised—governance questions arise independently of metaphysical agency.

Seen in this light, the dominance of agenthood-focused debates risks misdirecting philosophical attention. Such debates often gravitate toward questions of moral patiency, legal personhood, or speculative future capacities, while leaving under-analysed the delegation structures through which authority is already exercised. As Kahl has argued elsewhere, most systems described as agentic lack authorship of evaluative standards, yet nonetheless generate governance challenges typically associated with agency because of their persistence, integration, and material impact (Kahl 2026d, esp. §§4–6). The present article has extended that insight by identifying the threshold conditions under which those challenges become unavoidable, and by explaining why responsibility migrates upstream even in the absence of authorship.

The implications for philosophy of technology are therefore methodological as well as substantive. If authority without authorship is the core governance problem, philosophical analysis must prioritise questions of delegation, non-exit, and legitimacy over questions of metaphysical status. The task is not to determine whether machines can be agents, but to clarify when delegated systems acquire authority-effects, how justificatory burdens shift under conditions of dependency, and which structural conditions sustain or undermine legitimacy. Seen in this light, authority without authorship is ethically consequential not because machines act wrongly, but because delegation structures shape who must answer for decisions when control, exit, and contestation are constrained. These are questions about present socio-technical arrangements, not predictions about future AI capacities.

This reframing also clarifies the contribution of the article. It does not offer policy prescriptions, regulatory blueprints, or forecasts of increasingly autonomous machines. Its contribution is conceptual and diagnostic. By articulating a threshold account of authority without authorship and a fiduciary account of responsibility under delegation, the article provides a framework for understanding why existing governance approaches struggle and where philosophical scrutiny is most urgently required. As argued in related work, the difficulty is not that safety or responsibility have become optional, but that they have been mislocated—treated as downstream choices rather than as structural conditions of legitimacy (Kahl 2026f, Introduction; Conclusion).

Reframing the agentic AI debate in this way does not render questions of agency irrelevant. It situates them appropriately, as secondary to the analysis of how power is delegated, exercised, and justified in socio-technical systems. The task for philosophy of technology, then, is to make those delegation thresholds visible and to clarify the conditions under which authority remains legitimate in the absence of authorship. That task is already pressing, and it does not depend on the arrival of genuinely autonomous machines.

References

Bratman, M.E. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Fuller, L.L. (1964). *The morality of law*. New Haven, CT: Yale University Press.

Jeffrey, R.C. (1983). *The logic of decision* (2nd ed.). Chicago, IL: University of Chicago Press.

Kahl, P. (2026a). Epistemic humility as fiduciary obligation: Entrusted discretion and responsibility for belief. *Preprint*. https://doi.org/10.5281/zenodo.18440605

Kahl, P. (2026b). The frame-stability problem in decision-theoretic accounts of agency. *Preprint*. https://doi.org/10.5281/zenodo.18441980

Kahl, P. (2026c). Distributed cognition as epistemic infrastructure: A taxonomy of collective epistemic systems. *Manuscript under review*.

Kahl, P. (2026d). Why most ‘agentic AI’ is not agentic: Continuity, authorship, and the structural conditions of agency. *Preprint*. https://doi.org/10.5281/zenodo.18413863

Kahl, P. (2026e). Epistemic clientelism as a defect of governance: Fiduciary authority, discretion, and legitimacy. *Manuscript under review*.

Kahl, P. (2026f). From optional safety to architectural responsibility: AI governance after models. *Preprint*. https://doi.org/10.5281/zenodo.18431309

Miller, P.B. (2013). Justifying fiduciary duties. *McGill Law Journal*, 58(4), 969–1021.

Raz, J. (1990). *Practical reason and norms* (2nd ed.). Princeton, NJ: Princeton University Press.

Savage, L.J. (1954). *The foundations of statistics*. New York, NY: John Wiley & Sons.

Smith, L.D. (2014). Fiduciary relationships: Ensuring the loyal exercise of judgment on behalf of another. *Cambridge Law Journal*, 73(3), 608–639.

Star, S.L. (1999). The ethnography of infrastructure. *American Behavioral Scientist*, 43(3), 377–391.