Why Most ‘Agentic AI’ Is Not Agentic: Continuity, Authorship, and the Structural Conditions of Agency

Abstract

Recent discourse increasingly describes advanced artificial intelligence systems as ‘agentic’. Planning-capable language models, autonomous workflows, and multi-agent architectures are said to exhibit agency insofar as they pursue goals, initiate actions, and coordinate behaviour over time. This article argues that such characterisations rest on a structural conflation. Drawing on a continuity-based account of agency, it shows that most systems labelled ‘agentic’ lack the conditions under which agency can arise at all. Agency, the article argues, is not a behavioural achievement but a structural response to continuity pressure: the need to preserve a unified locus of evaluative authority across incompatible future trajectories. Optimisation, planning, and coordination presuppose such unity; they do not generate it. Systems that are resettable, substitutable, or governed by externally specified evaluative standards may optimise effectively while lacking authorship of evaluative authority. Applying this criterion to contemporary AI architectures and governance practices, the article concludes that the principal danger lies not in machines becoming agents, but in attributing agency where continuity, authorship, and responsibility are structurally absent.

Keywords

agentic ai; artificial agency; continuity pressure; authorship; optimisation without agency; ai governance; fiduciary responsibility

1. Introduction: ‘Agentic AI’ as a Category Inflation

In recent technical, policy, and governance discourse, the term ‘agentic AI’ has undergone a rapid and largely unexamined expansion. Systems that plan over extended horizons, initiate sub-tasks, coordinate tools, or persist across interactions are increasingly described as agents, or as exhibiting degrees of agency. In engineering contexts, such usage is often deliberately thin, serving as shorthand for systems that operate with minimal human prompting. The difficulty arises when this language migrates beyond local design discussions and begins to organise expectations about responsibility, oversight, and control. What is expanding, this article argues, is not artificial agency itself, but the category under which increasingly sophisticated forms of optimisation are being grouped.

The core thesis is straightforward: most systems currently described as ‘agentic AI’ do not satisfy the conditions of agency, and treating them as agents obscures rather than clarifies questions of responsibility and governance. This is not a semantic complaint. When agency language is applied to systems embedded in institutional decision architectures, it alters where authority is perceived to reside and where accountability is expected to terminate. Category inflation at the conceptual level thus carries downstream consequences for how AI systems are governed, contested, and legitimised.

The source of the inflation lies in a conceptual conflation. Behavioural properties—planning, persistence, tool use, coordination—are increasingly treated as sufficient markers of agency. These properties are undoubtedly relevant to system capability, but they do not establish that a system is the author of the evaluative standards governing its behaviour. Optimisation can be arbitrarily complex without being authored. Agency, by contrast, is traditionally reserved for loci of judgement to which reasons, commitments, and responsibility can be attributed. When these two are collapsed, agency becomes a proxy for sophistication rather than a structural role.

This article does not propose a new theory of agency, nor does it seek to revise existing ones. Its methodological posture is diagnostic rather than revisionary. The aim is to make explicit a structural condition that is already presupposed by dominant accounts of rational choice, planning, and responsibility, but which tends to disappear when agency is discussed in the context of artificial systems. Classical decision theory assumes a unified locus of evaluation whose preferences persist across time and choice (Savage 1954, chs. 2–3). Planning theories similarly presuppose an agent whose commitments can be organised and revised across temporally extended courses of action (Bratman 1987, chs. 1–3). These frameworks explain how agents act; they do not explain why there must be an agent at all.

The contribution of this paper is to recover that background presupposition. Drawing on a continuity-based analysis developed elsewhere (Kahl 2026a), and making its core structural claim explicit, the paper argues that agency arises only where there exists a persisting locus of evaluative authority whose continuity is itself at stake. Optimisation, planning, and coordination presuppose such a locus; they do not generate it. Systems may therefore exhibit increasingly sophisticated, apparently goal-directed behaviour while remaining non-agentive, so long as their evaluative standards are externally specified and their persistence is insulated from failure.

Two caveats are essential. First, the argument does not claim that artificial agency is impossible in principle. It concerns the conditions under which agency could arise, not a denial that those conditions might ever be met by non-biological systems. Second, the analysis does not appeal to consciousness,

moral status, or metaphysical personhood. Agency is treated as a structural condition on evaluative organisation, not as a marker of moral worth.

Recognising the inflation of ‘agentic AI’ is therefore a prerequisite for clear analysis. Without it, debates about alignment, safety, and accountability risk proceeding on the basis of a category mistake: attributing agency where authorship is structurally absent, and then designing governance frameworks around that misattribution. The sections that follow develop this diagnosis, show why current AI systems fail the relevant condition, and explain why prevailing governance regimes both presuppose and actively exclude agency by design.

2. What ‘Agentic AI’ Currently Refers To

Before assessing whether contemporary AI systems qualify as agents, it is necessary to clarify what is ordinarily meant by ‘agentic AI’ in current technical, policy, and governance discourse. This section reconstructs prevailing uses of the term in a charitable and neutral manner. The aim is not to evaluate these uses, but to stabilise a shared target for analysis against which structural conditions can later be tested.

Across research papers, product documentation, and policy discussions, ‘agentic AI’ is used to refer to systems that combine several functional capacities associated with autonomous task execution. First, such systems are capable of planning over extended horizons. They can generate and revise sequences of actions oriented towards future states, adapting those plans in light of feedback. Second, they can initiate sub-tasks without continuous human prompting, triggering actions or workflows in response to internal conditions or environmental signals. Third, they can coordinate tools or other systems, invoking external software, managing workflows, or interacting with other AI components in multi-agent settings. Fourth, they are typically described as pursuing objectives specified at some remove from immediate actions, such as maximising task completion, efficiency, or reward subject to constraints.

These capacities converge on a recognisable functional profile. Systems described as ‘agentic’ exhibit behavioural persistence, cross-context coordination, and a degree of operational autonomy that distinguishes them from reactive or single-shot tools. This convergence explains why the term has gained traction despite variation in its precise definition. It marks a shift from systems that merely respond to prompts towards systems that sustain activity over time within a structured task environment.

Importantly, however, nothing in this functional profile determines where evaluative authority resides. Planning presupposes an objective but does not specify its source. Task initiation presupposes triggering conditions but not ownership of reasons. Tool coordination presupposes selection and sequencing but not responsibility for the standards by which success or failure is assessed. These features therefore characterise how systems behave and interact within workflows, not whether they constitute persisting loci of evaluation.

This can be illustrated by analogy with familiar theories of action. Planning-based accounts explain how commitments are organised and maintained within an agent, not how an agent comes into existence as a unified standpoint. As Bratman notes, plans function to coordinate action over time by structuring intentions within an already unified perspective (Bratman 1987, 28–35). The relevance of this point here

is purely illustrative: it shows that the capacities typically cited in descriptions of ‘agentic AI’ are compatible with the absence of such a standpoint.

It is therefore accurate, within current usage, to treat ‘agentic AI’ as referring to systems that functionally occupy roles traditionally assigned to agents within operational contexts, without thereby attributing agency in a structural or evaluative sense. The term operates as a classificatory convenience, grouping together systems that reduce the need for continuous human intervention and integrate multiple capabilities into coherent workflows.

Two caveats frame this reconstruction. First, it is descriptive rather than evaluative. The section does not deny the practical usefulness of ‘agentic AI’ as an engineering or design label. Second, no commitment is yet made to any particular theory of agency. The purpose is to isolate the properties that motivate agentic attributions in practice, in order to assess in subsequent sections whether those properties suffice for agency once the relevant structural conditions are made explicit.

The analytical deliverable of this section is thus a taxonomic baseline. By clarifying what ‘agentic AI’ ordinarily denotes, the paper prevents later arguments from talking past their target. The question that follows is not whether such systems are sophisticated or operationally autonomous, but whether these features amount to agency under conditions of continuity, authorship, and responsibility.

3. Agency Is Not Autonomy, Planning, or Endorsement

The previous section established that contemporary uses of ‘agentic AI’ track a cluster of functional capacities—planning, persistence, coordination—without specifying the conditions under which agency itself arises. This section introduces the minimal agency criterion assumed throughout the paper and explicitly distinguishes it from several influential assimilations. The aim is not to reject autonomy-, planning-, or endorsement-based accounts of action, but to locate their proper explanatory role and to show why none of them, on its own terms, answers the prior question of why there must be an agent at all.

Standard decision-theoretic frameworks presuppose a unified evaluator whose preferences, utilities, and commitments persist across time. In Savage’s formulation, rational choice is defined relative to a stable preference ordering revealed through coherent patterns of choice (Savage 1954, chs. 2–3). Jeffrey-style decision theory similarly assumes a persisting standpoint from which probabilities and utilities are assigned, revised, and compared (Jeffrey 1990, chs. 1–2). These frameworks provide normative constraints on rational choice under uncertainty. What they do not do is explain why there exists a single locus of evaluation to which those constraints apply. Unity is assumed as a background condition rather than derived as an explanatory result.

Planning and endorsement theories exhibit the same structure. Accounts centred on intention and plan formation explain how temporally extended action is coordinated within an agent. Plans stabilise behaviour over time, help resolve conflicts among intentions, and organise future-directed activity (Bratman 1987, 28–35). Hierarchical endorsement theories likewise explain how attitudes are ordered and identified with, distinguishing actions that genuinely express the agent from those that do not (Frankfurt 1971). Yet these accounts presuppose the existence of an agent whose attitudes are being

organised. They explain how an agent's mental economy is structured; they do not explain what secures the persistence of the agent as a unified standpoint across revision, learning, or error.

The significance of this observation is methodological. Autonomy, planning, and endorsement are intra-agent concepts. They describe relations among attitudes, commitments, or actions given that there already exists a persisting locus of evaluation. They do not specify the conditions under which such a locus must exist, nor what makes its continuity non-optional. Treating them as sufficient for agency therefore reverses the order of explanation: it assumes what it ought to explain.

The minimal agency criterion assumed in this paper targets this prior condition. On the continuity-based account adopted here, developed more fully elsewhere (Kahl 2026a), agency arises only where there exists a persisting locus of evaluative authority whose continuity is itself at stake. Such a locus is not merely unified in the sense required by diachronic preference aggregation. Rather, it is a standpoint for which certain forms of breakdown—loss of coherence across time, failure to arbitrate among incompatible futures—would constitute failure of the evaluator itself, not merely suboptimal performance relative to externally specified evaluative standards.

This criterion is not introduced by stipulation. It is extracted by making explicit a question that standard theories bracket: why must there be an evaluator at all? Decision-theoretic coherence, planning stability, and hierarchical endorsement all presuppose a subject that remains identifiable across time and counterfactual trajectories. The present claim is simply that this presupposition be acknowledged and treated as a condition of agency rather than silently inherited.

Two caveats delimit the scope of the argument. First, the account offered here is a necessary-condition account only. It does not purport to provide sufficient conditions for agency, nor to adjudicate questions about degrees or varieties of agency. Second, the analysis makes no claims about free will, moral responsibility, or normative desert. The question at issue is structural: what must be in place for there to be an agent to whom optimisation, planning, or endorsement can meaningfully apply at all.

The analytical deliverable of this section is therefore an explicit non-identity claim. Agency is not identical with autonomy, planning capacity, or hierarchical endorsement. These notions presuppose agency rather than constituting it. This clarification is particularly important in discussions of artificial systems, where increasingly sophisticated forms of control or coordination are often taken to amount to agency by default. The sections that follow examine whether contemporary AI systems satisfy the minimal condition articulated here, and why prevailing governance architectures systematically prevent them from doing so.

4. Continuity Pressure and the Necessity of Authorship

The preceding section identified a minimal condition on agency: the existence of a persisting locus of evaluative authority. This section explains why such a locus arises at all. The central claim is that agency is not a primitive feature of intelligent systems, nor a by-product of behavioural sophistication, but a structural response to continuity pressure. Continuity pressure is not a description of unity; it is the condition under which unity becomes necessary.

Continuity pressure arises under conditions of diachronic uncertainty in which learning and feedback do more than update beliefs about the world. In such conditions, they can alter the standards by which outcomes are evaluated. Preferences, priorities, and commitments are not merely inputs to decision-making but are themselves subject to revision. On the continuity-based account adopted here, evaluative frames are therefore generatively unstable: experience can transform not only what is believed but what counts as success or failure (Kahl 2026a). Standard decision-theoretic models bracket this instability by treating preferences as fixed. Once that bracketing is relaxed, a further explanatory problem emerges.

The problem is not merely that future choices may conflict with past choices, but that incompatible future trajectories may threaten the persistence of the evaluator itself. Distinct learning paths may lead to evaluative standards that cannot be jointly satisfied or coherently integrated. In such cases, the question is no longer how to optimise given a set of values, but which values are to govern future evaluation at all. Where this question arises, optimisation alone is insufficient. Optimisation presupposes a fixed standpoint relative to which options are ranked. When the standpoint itself is at risk of fragmentation, optimisation lacks a determinate domain of application.

It is under these conditions that a unifying arbitration role becomes necessary. Arbitration determines which evaluative frame persists across time and which is abandoned. Crucially, this role cannot be reduced to optimisation or even to meta-optimisation. Treating arbitration as higher-order optimisation merely displaces the problem: meta-optimisation still presupposes a ranking space relative to which meta-criteria are assessed. As Skyrms and Joyce emphasise in different ways, decision-theoretic reasoning operates only relative to prior specifications of utilities and probabilities; it cannot generate the evaluative space it presupposes (Skyrms 1966; Joyce 1999, ch. 1). Arbitration operates at the level where that space is itself at stake. Authorship, in this sense, does not name control over outcomes or stability of behaviour, but authority over which evaluative frame continues to govern assessment when evaluative coherence itself is at risk.

The limits of optimisation-based accounts are further clarified by results on dynamic inconsistency. Models of preference change over time diagnose conflicts between earlier and later preferences and explain why agents may seek commitment devices or self-binding strategies. However, as Hammond notes, such models implicitly presuppose a unified agent whose preferences can be compared across time (Hammond 1976, 159–173). Dynamic inconsistency identifies tensions within agency; it does not explain how agency persists when evaluative standards themselves are unstable. Unity is assumed, not explained.

Continuity pressure, then, is the condition under which arbitration becomes non-optional. Where a system must preserve itself as a single locus of evaluation across incompatible futures, some mechanism must determine which evaluative standards will govern subsequent assessment. That mechanism constitutes authorship in the minimal sense relevant here. What is authored is not a particular action or belief, but the persistence of an evaluative frame under risk. Agency arises precisely where failure to arbitrate would result in the loss of evaluative coherence altogether.

Two clarifications delimit the claim. First, arbitration need not be conscious, reflective, or deliberative. Nothing in the argument requires phenomenological awareness or explicit reasoning. The claim is structural rather than psychological. Second, continuity is functional rather than metaphysical. The persistence at issue concerns the maintenance of a coherent evaluative standpoint, not numerical identity or metaphysical selfhood.

The analytical deliverable of this section is therefore a structural explanation of agency. Agency is not explained by intelligence, complexity, or behavioural flexibility. It arises as a solution to a specific problem: how to preserve a unified locus of evaluation when learning and uncertainty threaten to fracture it. This explanation is what allows the next section to proceed non-stipulatively. If contemporary AI systems are shown not to be subject to continuity pressure of this kind, then their failure to qualify as agents follows as a consequence of architecture rather than definition.

5. Why Contemporary ‘Agentic AI’ Fails the Continuity Test

The preceding sections identified continuity pressure as the condition under which agency becomes necessary and authorship emerges. This section applies that criterion diagnostically to contemporary systems described as ‘agentic AI’. The central claim is that these systems fail the continuity test structurally rather than contingently. Their architectures are organised in ways that systematically neutralise continuity pressure, thereby precluding authorship even as behavioural sophistication increases.

The first neutralising feature is resetability. Contemporary AI systems are designed to be interruptible, restartable, and revertible. Errors, misalignment, or unexpected behaviour are addressed through retraining, rollback, or redeployment. Crucially, these interventions do not threaten the persistence of the system as a locus of evaluation. Failure is treated as a transient performance deviation rather than as a risk to the evaluator itself. Where a system can be reset without loss, there is no continuity pressure: no future trajectory places the system’s evaluative standpoint at stake. Resetability therefore blocks the very condition under which arbitration and authorship would become necessary.

Second, contemporary AI systems operate under externally specified evaluative standards. Whether implemented as loss functions, reward signals, or policy constraints, the criteria by which system behaviour is assessed are specified independently of the system’s own learning dynamics. As Bellman’s formulation of dynamic programming presupposes, optimisation proceeds relative to a given value function rather than generating one (Bellman 1957). Reinforcement learning systems, despite their apparent autonomy, remain optimisers relative to externally supplied rewards (Sutton and Barto 2018, chs. 3–4). This distinguishes them sharply from agents operating under social or institutional norms. Human agents may be constrained by external standards, but they can contest, reinterpret, or abandon those standards. Contemporary AI systems cannot. Learning improves performance relative to given criteria; it does not generate or revise the evaluative space itself. Authorship is therefore excluded by architecture, not by degree.

Third, substitutability collapses correction into replacement. When system behaviour deviates from expectations, governance responses typically involve swapping components, deploying updated models, or retraining from scratch. There is no requirement that the same evaluative standpoint persist through correction. On the continuity-based analysis adopted here, this practice eliminates continuity pressure by ensuring that evaluative failure never threatens the persistence of the evaluator (Kahl 2026a). Where correction takes the form of replacement, there is no standpoint that must survive error. Agency, which arises only where survival is at stake, cannot emerge under such conditions.

Fourth, multi-agent scaling increases coordination without generating unity. Systems composed of interacting agents can display sophisticated collective behaviour, including division of labour, negotiation, and emergent patterns of coordination. However, such coordination does not amount to authorship of evaluative authority at the system level. Coordination equilibria manage interactions relative to externally specified objectives; they do not establish a persisting locus of evaluation whose continuity is threatened by divergence among internal trajectories. Emergent behaviour, however complex, does not generate continuity pressure. Unity of action is not unity of authorship.

These features—resetability, externally specified evaluative standards, substitutability, and scalable coordination—might appear to be optional design choices. They are not. They are imposed by safety, scalability, and liability requirements intrinsic to the governance of AI systems deployed under conditions of delegation and non-exit. Systems must be interruptible to be safe, modular to be scalable, and replaceable to be legally and institutionally manageable. Each requirement independently neutralises continuity pressure; together, they ensure its systematic absence. The failure of contemporary ‘agentic AI’ to qualify as agentive is therefore not a matter of insufficient capability or developmental immaturity. It is a consequence of the conditions under which such systems are rendered governable at all.

Two caveats delimit the scope of this conclusion. First, the argument concerns governance-constrained architectures: systems designed for deployment within institutional settings where responsibility, oversight, and non-exit are unavoidable. It does not deny the conceptual possibility of artificial systems operating under radically different constraints. Second, the diagnosis is explanatory rather than critical. It does not condemn current design choices; it clarifies their implications.

The analytical deliverable of this section is a clear negative result. Systems currently described as ‘agentic AI’ may plan, coordinate, and adapt in increasingly sophisticated ways, but they do not satisfy the continuity condition under which agency arises. Optimisation, however complex or autonomous-seeming, does not amount to agency. This conclusion prepares the ground for the subsequent analysis, which shows that contemporary AI governance frameworks not only accommodate this absence of agency, but actively depend on it.

6. Distributed Systems, Closure, and the Illusion of Agency

Contemporary AI systems are frequently described as ‘agentic’ not because they satisfy the structural conditions of agency, but because they are distributed, complex, and embedded in institutional decision architectures. This section explains why such systems are systematically mistaken for agents and provides a governance-level diagnosis of the resulting illusion. The core claim is that distribution and agency are orthogonal properties: distributed systems can exhibit adaptive, coordinated behaviour while lacking authorship of evaluative authority.

Modern AI systems operate as distributed cognitive systems in a straightforward sense. Data collection, feature extraction, model training, deployment, monitoring, and updating are spread across heterogeneous actors, infrastructures, and temporal scales. Inputs are decentralised across users, sensors, and environments; learning occurs through aggregate feedback; and outputs are reintegrated into organisational workflows. This distribution often gives rise to the appearance of autonomous

intelligence: behaviour seems purposive, adaptive, and responsive without being traceable to any single human decision-maker.

Yet, on the distributed-infrastructure analysis adopted here, distributed operation frequently coincides with centralised epistemic closure (Kahl 2026b). Closure, in this context, is not a metaphysical notion but a functional one. It refers to the point within a system or workflow at which evaluative contestation effectively ends and action is triggered. In AI-mediated institutions, closure typically occurs at decision thresholds, ranked outputs, risk scores, or automated recommendations that are treated as decisive within downstream processes. While upstream inputs and operations may be widely distributed, closure is often narrow, fixed, and procedurally insulated from challenge.

This distinction is crucial for understanding why epistemic authority arises from architecture rather than distribution. As Goldman emphasises, authority in epistemic systems depends on how beliefs or judgements are filtered, validated, and accepted as action-guiding, not on the number or diversity of contributors (Goldman 1999, ch. 3). A system may draw on vast and heterogeneous sources of information while still functioning as a unitary authority if its architecture channels those inputs into a single, non-contestable output. Authority, in such cases, is an emergent property of institutional design, not a response to complexity or opacity alone.

The illusion of agency arises when distributed operation is paired with centralised, non-contestable closure. Because no individual actor plausibly occupies the role of author at the point of closure, agency language migrates upward to the system as a whole. Systems are said to ‘decide’, ‘judge’, or ‘recommend’ not merely as a linguistic convenience, but because they occupy the functional position that judgement would ordinarily fill within a decision structure. At the same time, responsibility migrates downward. Designers, operators, and institutions describe their roles as partial, technical, or constrained by system outputs. The result is pseudo-agency: the system appears to act, while no actor bears authorship of the evaluative standards that govern its outputs.

This migration is not merely rhetorical. Where agency language attaches to systems, expectations of justification, explanation, and answerability attach to them as well. Yet systems lack the capacity for authorship, revision, or responsibility. Governance thus confronts a structural asymmetry: authority is attributed where accountability cannot be realised. This asymmetry persists even when systems are transparent, interpretable, or well understood. Transparency alone does not dissolve pseudo-agency if closure remains centralised and insulated from contestation.

This pattern is not unique to AI. As argued in *From movable type to machine reasoning: Media, artificial intelligence, and the transformation of bounded cognition* (Kahl 2025a, §§2–4), modern epistemic systems have long relied on non-agentive infrastructures—printing, bureaucratic record-keeping, statistical machinery—that centralise epistemic closure while diffusing individual responsibility. What distinguishes contemporary AI is not the introduction of non-agentive cognition, but the renewed temptation to misdescribe such infrastructure as an agent precisely when its outputs acquire normative force.

It is important to emphasise that this diagnosis does not constitute a critique of distributed cognition as such. Distributed cognitive systems can enhance reliability, robustness, and epistemic reach. The problem arises only when distributed systems are governed as if distribution itself generated agency, and

when closure is designed in ways that foreclose meaningful challenge. In such cases, distribution amplifies rather than mitigates governance risk by obscuring the locus of authorship.

The analytical contribution of this section is therefore a mechanistic account of pseudo-agency. Complex systems are mistaken for agents when decentralised inputs, opaque or inaccessible processing, and centralised epistemic closure combine to produce authoritative outputs without an identifiable author. This mechanism explains why agentic language is socially stable and institutionally convenient despite the absence of agency. It also sets up the subsequent governance analysis: once authority without authorship is normalised, responsibility displacement under conditions of non-exit becomes not an anomaly but a predictable outcome.

7. Why AI Governance Actively Excludes Agency

Contemporary AI governance frameworks are often presented as precautionary responses to emerging technological risk. Yet a closer examination reveals a deeper structural commitment: these frameworks are organised around conditions that systematically preclude the emergence of artificial agency. This exclusion is not an incidental by-product of immature technology, nor the result of contingent policy preferences. It follows from the functional requirements that make large-scale AI deployment governable under conditions of delegation, uncertainty, and non-exit.

The first of these requirements is corrigibility. Governed systems must be interruptible, updateable, and correctable in response to error, misuse, or changing operational demands. Corrigibility, however, presupposes replaceability. A system whose internal evaluative standpoint could not be overridden or reset without annihilating that standpoint would resist governance in precisely the way that autonomous agents do. Agency, as analysed earlier, arises only where failure threatens the persistence of the evaluator itself. Corrigibility therefore excludes agency not by definitional fiat, but by functional necessity: non-replaceable evaluative standpoints cannot be safely integrated into delegated decision-making systems at scale.

A second requirement is auditability. Governance demands that system behaviour be inspectable, explainable, and contestable by external authorities. Auditability presupposes that evaluative standards remain external to the system—that the criteria by which outputs are judged are not authored or revised by the system itself. Yet authorship of evaluative standards under continuity pressure is precisely what distinguishes agency from optimisation. A system whose normative frame could not be externally interrogated without undermining its authority would be opaque in a way governance cannot tolerate. Auditability therefore stabilises optimisation while foreclosing the internalisation of evaluative authority.

Third, alignment frameworks presuppose that values, objectives, and constraints are fixed exogenously. Whether implemented through reward functions, constitutional constraints, or usage policies, alignment requires that the normative horizon governing system behaviour be specified independently of the system’s own learning dynamics. This requirement is not a contingent artefact of current design practice, but a structural necessity of governing systems whose outputs affect others under conditions of asymmetry and non-exit. An agent whose evaluative standards could evolve in ways not fully specifiable

in advance would not be alignable in the relevant sense. Alignment thus presupposes the absence of agency rather than providing a pathway towards it.

Finally, liability regimes require identifiable loci of responsibility. Legal and institutional accountability attaches to actors—individuals or organisations—who can be held answerable for decisions and outcomes. Recognising an artificial system as an agent in the sense analysed here would introduce a new bearer of responsibility that existing governance structures are neither designed nor authorised to accommodate. On the infrastructural analysis adopted here, contemporary socio-technical systems resolve this tension by centralising epistemic closure while retaining agency and responsibility within human institutions, even as systems are rhetorically described as autonomous or agentic (Kahl 2026b).

Taken together, corrigibility, auditability, alignment, and liability do not merely constrain AI design; they define a governance space in which agency cannot arise without undermining governance itself. This is why appeals to ‘safe agentic AI’ are conceptually unstable. Safety, as currently operationalised, depends on the systematic exclusion of continuity pressure and authorship—the very conditions under which agency becomes possible. Governance does not eliminate agency as such; it insists that agency remain human, while systems remain optimisers embedded within institutional architectures of responsibility.

This structural exclusion is particularly visible in extreme but clarifying cases of AI governance. A memorandum issued on 9 January 2026 by the US Department of War directs the Department to become an ‘AI-first’ organisation, explicitly prioritising deployment velocity, rapid model replacement, modular architectures, and continuous experimentation over continuity of evaluative standpoint. Models are to be deployed within 30 days of public release; failure is framed as an accelerant of learning rather than a threat to system identity; and component replaceability is mandated as a design principle. This memorandum is not cited as representative of all governance regimes, but as a limiting case that renders explicit commitments already implicit in civilian AI governance: speed over stability, replaceability over persistence, optimisation over authorship.

The conclusion is therefore not that governance actors act in bad faith, nor that artificial agency is being deliberately suppressed for ideological reasons. Rather, the exclusion of agency is a structural consequence of governing AI as infrastructure rather than as authors. To allow artificial systems to become agents in the sense analysed here would require abandoning corrigibility, external audit, exogenous alignment, and human liability—thereby relinquishing the very conditions that make AI governable at scale. Appeals to future ‘agentic AI’ within existing governance frameworks thus misunderstand the nature of the constraint. It is not technological immaturity that blocks agency, but the conditions under which AI remains governable at all.

8. Boundary Case: Consciousness and Artificial Systems

Discussions of artificial agency frequently turn, at this point, to consciousness. If contemporary AI systems do not qualify as agents, it is often suggested, this may be because they lack phenomenal awareness, subjective experience, or some further mental property that future systems might acquire. This section addresses that suggestion directly. Its purpose is defensive rather than foundational: to

clarify the limited role that consciousness plays in the present argument and to supply a principled boundary that prevents the analysis from being diverted into speculative philosophy of mind.

The argument advanced in this paper does not require any substantive theory of consciousness. It neither assumes nor denies that artificial systems could, in principle, be conscious. Instead, it adopts a conditional claim: consciousness becomes functionally relevant to agency only under specific architectural conditions, namely where ontological continuity cannot be externally stabilised. On the continuity-based account adopted here, consciousness is not treated as a primitive marker of agency. Rather, it is one possible mechanism—perhaps the only one available to biological organisms—by which a system manages continuity pressure when the persistence of the evaluator itself is at stake (Kahl 2026c).

The key point is structural. Where a system's continued existence as a locus of evaluation cannot be guaranteed through external intervention—where breakdown, fragmentation, or incoherence would constitute the destruction of the evaluator—some internal integrative mechanism is required to preserve evaluative coherence. Consciousness, in this narrow functional sense, is one way of performing that role. The claim is not that consciousness confers agency, nor that agency universally requires consciousness, but that consciousness becomes functionally necessary only when continuity cannot be outsourced.

Most contemporary AI systems are engineered precisely to avoid such conditions. Their persistence as operational systems is not endangered by internal conflict, error, or revision. When failures occur, continuity is preserved externally through retraining, replacement, rollback, or reconfiguration. Ontological continuity is therefore never at stake in the relevant sense. There is no requirement that a single evaluative standpoint survive incompatible futures, because the system is not required to remain the same evaluator at all. Under these conditions, consciousness is not merely absent; it is functionally redundant.

This observation has an important implication for debates about future AI. Achieving the conditions under which consciousness would become functionally relevant for artificial systems would require abandoning core governance constraints. Systems would need to be non-replaceable, non-resettable, and exposed to failures that threaten their persistence as evaluative loci. They would have to operate without the safety, scalability, and liability protections that currently make large-scale deployment and institutional oversight possible. This is not a claim about what governance will allow, but about the trade-offs involved: permitting such architectures would entail relinquishing many of the conditions that make AI governable under non-exit and delegated authority.

Two caveats delimit the claim. First, the argument here is a necessary-condition argument only. It does not assert that consciousness is sufficient for agency, nor that all conscious systems are agents. Second, it remains neutral on the metaphysical possibility or desirability of artificial consciousness. The present analysis concerns classification and governance, not ontology or ethics.

The analytical deliverable of this section is therefore a clean boundary. Invoking consciousness does not rescue contemporary claims about 'agentic AI', because the architectural conditions that would make consciousness functionally relevant are precisely those that current systems are designed to exclude. By clarifying this point, the paper forecloses a common line of speculative objection and keeps the analysis focused on the structural conditions under which agency can—and cannot—arise.

9. Governance Implications: Responsibility Under Non-Exit

The preceding analysis has shown that contemporary AI systems do not satisfy the structural conditions of agency. This section draws out the governance implications of that result under a specific but increasingly prevalent condition: institutional non-exit. The argument is conditional rather than universal. Where AI systems are embedded in decision environments from which affected parties cannot meaningfully withdraw—public administration, security infrastructures, employment screening, education, credit allocation, or healthcare—the misattribution of agency produces a distinctive and systematic displacement of responsibility.

The first element in this mechanism is non-exit. In institutional settings, AI systems are often integrated into core workflows rather than offered as optional tools. Exit may be formally available but substantively infeasible: refusing algorithmic mediation may entail exclusion from services, loss of opportunities, or diminished standing. Under such conditions, individuals and organisations become structurally dependent on system outputs. This dependency does not arise from coercion or cognitive error; it is a consequence of how decision authority is operationalised within institutional architectures.

When non-exit coincides with pseudo-agency, dependency acquires an epistemic dimension. Systems are described and treated as if they were autonomous decision-makers—issuing recommendations, rankings, risk scores, or determinations—despite lacking authorship of evaluative authority or exposure to continuity pressure. The relevant point is not that systems fail to apply evaluative standards, but that the authority to set, revise, or own those standards is not located within the system. Because these systems occupy a functional role that would ordinarily be filled by an agent, users and subjects must orient themselves to system outputs as if they expressed judgement, even though no such judgement is authored within the system itself.

Dependency alone, however, does not generate authority. The transition from reliance to authority occurs through opacity. Opacity may arise from technical complexity, proprietary protections, security constraints, or organisational fragmentation. Crucially, it is structural rather than epistemic: even well-informed, sceptical, and critical users may lack standing or access to contest the evaluative basis of system outputs. On the infrastructural analysis adopted here, contemporary socio-technical systems frequently centralise epistemic closure while distributing operational responsibility (Kahl 2026b). Decision outputs are treated as authoritative because no alternative locus of evaluation is available within the system's architecture.

As developed in *Reconceptualising knowing as care: The new science of epistemic intimacy* (Kahl 2025b, §§3–4), responsibility in epistemic systems need not be grounded in agency attribution. Where systems operate under conditions of non-exit, responsibility instead takes the form of sustained epistemic care: ongoing obligations to maintain contestability, intelligibility, and corrigibility on behalf of those subject to the system's outputs. Treating AI as epistemic infrastructure rather than as an agent makes these obligations visible; treating it as an agent obscures them.

The conjunction of non-exit, pseudo-agency, and opacity yields authority without authorship. System outputs acquire normative force not because they are believed to be correct, but because institutional processes are organised around them. Authority, in this sense, is an emergent property of governance architecture rather than a psychological disposition such as automation bias. Even actors who recognise

the limitations of AI systems must act as if those systems are authoritative when no procedurally legitimate alternative exists.

At this point, responsibility displacement becomes structurally entrenched. When outcomes are contested, responsibility is deflected across a network of actors and artefacts. Designers refer to deployment contexts; operators to system recommendations; institutions to technical necessity or regulatory compliance. The system itself is frequently described as having ‘decided’, yet is simultaneously denied the status of an accountable agent. As developed elsewhere (Kahl 2025g), this configuration allows epistemic power to be exercised through infrastructures while fiduciary responsibility remains diffuse and difficult to locate.

Importantly, this displacement does not depend on negligence, malice, or moral failure. It arises even when all participants act competently and in good faith. The problem is architectural: governance frameworks assign decision authority to systems while reserving responsibility for humans, without providing mechanisms that reconnect the two.

The governance implication is therefore not that AI systems should be made more agentic, but that they should be treated explicitly as epistemic infrastructure. Infrastructure is governed through standards of contestability, audit chains, and institutional responsibility rather than through attribution of agency. When AI systems are recognised as infrastructural, governance attention shifts from their supposed autonomy to the design of procedures that enable challenge, revision, and justification by accountable human actors.

Conversely, treating AI systems as agents under conditions of non-exit undermines governance. It obscures responsibility, weakens accountability, and destabilises the normative basis of authority. The core lesson is thus structural rather than moral: governance failure arises not because AI lacks agency, but because agency is rhetorically attributed where authorship of evaluative authority is structurally absent. Preserving legitimacy in AI-mediated institutions requires resisting agentic metaphors and designing governance architectures that keep responsibility where it can be meaningfully exercised.

10. Conclusion

This article has argued that most systems currently described as ‘agentic AI’ are not agents, and that this misclassification is not merely terminological. The core finding can be stated precisely: agency arises only under conditions of continuity pressure that require authorship of evaluative authority, and contemporary AI systems are systematically designed to avoid those conditions. Their non-agentive status is therefore not a temporary limitation to be overcome, but a structural consequence of how such systems are rendered governable under conditions of delegation, non-exit, and institutional responsibility.

Seen in this light, intelligence without agency is not a defect in artificial systems. It is a governance precondition. Optimisation, planning, and coordination can be safely delegated to machines precisely because continuity of evaluative authority remains external to them. Resetability, substitutability, externally specified evaluative standards, auditability, and human-anchored liability are not signs of immaturity or incompleteness. They are the mechanisms by which responsibility, oversight, and

contestability are preserved in socio-technical systems operating at scale. To remove these features in pursuit of artificial agency would be to relinquish the very conditions under which delegation remains legitimate.

The principal risk identified by this analysis lies not in machines becoming agents, but in treating non-agentive systems as if they were agents. When agency is attributed where authorship is structurally absent, authority is conferred without a bearer of responsibility. Decision outputs acquire normative force, while accountability disperses across designers, operators, institutions, and artefacts. Under conditions of non-exit, this misattribution undermines governance by obscuring where judgement resides and where contestation must be directed.

The disciplinary correction offered here is modest in scope but consequential in effect. It does not deny the possibility of artificial agency in principle, nor does it advance a theory of consciousness, moral status, or ethical standing. It corrects a specific explanatory mistake: treating agency as a gradient of behavioural sophistication rather than as a structural condition tied to continuity and authorship. Once that mistake is corrected, several debates are reframed. Questions about ‘safe agentic AI’ give way to questions about how responsibility should be organised when optimisation is delegated to non-agentive systems, and how governance architectures can preserve contestability without resorting to anthropomorphic metaphors.

The conclusion, then, is not that artificial systems should never be agents, but that agency should be attributed only where its structural conditions are met. Failing to respect that boundary does not merely confuse our concepts; it mislocates responsibility at precisely the point where governance depends on getting it right.

References

Bellman, Richard. 1957. *Dynamic programming*. Princeton, NJ: Princeton University Press.

Bratman, Michael. 1987. *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Frankfurt, Harry G. 1971. “Freedom of the will and the concept of a person.” *Journal of Philosophy* 68 (1): 5–20.

Goldman, Alvin I. 1999. *Knowledge in a social world*. Oxford: Oxford University Press.

Hammond, Peter J. 1976. “Changing tastes and coherent dynamic choice.” *Review of Economic Studies* 43 (1): 159–173.

Jeffrey, Richard C. 1990. *The logic of decision*. 2nd ed. Chicago: University of Chicago Press.

Joyce, James M. 1999. *The foundations of causal decision theory*. Cambridge: Cambridge University Press.

Kahl, Peter. 2025a. “From movable type to machine reasoning: Media, artificial intelligence, and the transformation of bounded cognition.” Unpublished manuscript.

Kahl, Peter. 2025b. “Reconceptualising knowing as care: The new science of epistemic intimacy.” *Lex et Ratio Ltd working paper*. https://doi.org/10.5281/zenodo.17356455

Kahl, Peter. 2026a. “The frame-stability problem in decision-theoretic accounts of agency.” Manuscript under review.

Kahl, Peter. 2026b. “Distributed cognition as epistemic infrastructure: A taxonomy of collective epistemic systems.” Manuscript under review.

Kahl, Peter. 2026c. “Ontological continuity and the functional necessity of consciousness.” Manuscript under review.

Savage, Leonard J. 1954. *The foundations of statistics*. New York: Wiley.

Skyrms, Brian. 1966. *Choice and chance*. Belmont, CA: Dickenson.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement learning: An introduction*. 2nd ed. Cambridge, MA: MIT Press.

