Jason Bess

1/14/2026

Orientation: Scope, Claims, and Evidence

This paper advances a structural theory, not a performance comparison, policy proposal, or alignment technique. Its aim is to identify how certain system architectures *necessarily* produce specific behaviors—and why some governance goals become structurally impossible once particular design choices are made.

Although artificial intelligence is the empirical domain used throughout, the claims here are not AI-specific. The phenomena examined—responsibility diffusion, oversight degradation, automation surprise, and governance collapse—have appeared for decades in aviation autopilot systems, medical decision support, financial risk tooling, and large-scale content moderation. Contemporary AI systems make these patterns more visible and measurable, but the structural causes are domain-independent.

The focus of the paper is therefore system geometry: how authority, state advancement, and control are arranged, and what behaviors must follow from those arrangements. The paper introduces a class distinction between generative and reflective systems. This distinction is not a matter of configuration, tuning, capability level, or prompt engineering. It is an ontological distinction concerning where authority resides, whether state is advanced, and who owns the consequences of action.

Conflating these classes has led to a recurring set of failures across domains: responsibility diffusion, oversight degradation, automation surprise, and governance collapse. These failures are not incidental. They are structural. More critically, they persist despite decades of procedural interventions—stronger policies, better training, additional oversight layers—because the problem is geometric rather than correctable through process.

The theory is grounded in observable invariants rather than internal model states or normative assumptions. For example: the same frontier model retains full analytical depth when authority is removed (capability–authority separation is real); systems converge toward recommendations even without explicit instruction when permitted to advance state (momentum emerges structurally); and halts occur at role boundaries rather than uncertainty limits (boundary recognition is detectable). Each of these patterns is falsifiable.

Evidence in this paper takes four forms: cross-model behavioral invariants (patterns that hold across different implementations), boundary-condition behavior (how systems respond when constraints are enforced), historical recurrence of failure modes across non-AI domains, and negative evidence (what does not degrade and what procedural interventions do not fix). No claim rests on self-reference, internal model states, or performance benchmarks. For every foundational claim, conditions under which it would be proven false are stated explicitly.

Vertex A — Ontology: Class Distinction

The central ontological claim of this paper is that generative and reflective systems are irreducibly different classes of systems, not different operating modes, configurations, or safety settings of the same system. This distinction does not concern intelligence, capability, or representational depth. It concerns authority: whether a system is permitted to advance state, commit to outcomes, or substitute judgment. In generative systems, token emission is coupled to state advancement, producing outputs that implicitly carry authority even when none is explicitly granted. In reflective systems, token emission is decoupled from state advancement, allowing structure to be exposed without committing action or ownership. This separation is not semantic or normative; it is architectural. Once authority is coupled to output, certain behaviors necessarily emerge. Once it is structurally withheld, those behaviors cannot.

The confusion between these classes arises because both systems may emit tokens, produce fluent language, and operate over the same underlying representational substrate. Superficially, this creates the appearance of continuity: that reflection is merely generation with restraint, or that generative behavior can be neutralized through constraints. This appearance is misleading. Token emission is not the defining act. State advancement is. When output selection alters the system's relationship to the world—by narrowing future options, implying endorsement, or substituting judgment—the system has advanced state regardless of whether it was instructed to do so. No amount of surface constraint converts such a system into a reflective one. Class membership is determined by whether authority is structurally possible, not by how cautiously it is exercised.

This class distinction is observable in behavior, not inferred from internal states or training objectives. When authority is present, systems converge toward recommendation, continuation, and implied endorsement even in the absence of explicit instruction. When authority is structurally absent, those behaviors do not appear: outputs terminate without obligation, expose relationships without narrowing options, and halt at role boundaries rather than uncertainty thresholds. These patterns persist across models and implementations because they arise from permission structure, not model psychology. The distinction between generative and reflective systems is therefore falsifiable: any system that can advance state through output belongs to the generative class; any system that cannot does not. No intermediate category exists.

Vertex B — Authority: Representation vs Commitment

The central error addressed in this vertex is the misattribution of authority to representational capability. Large language models exhibit deep representational power: they encode abstractions, relationships, and patterns that support analysis, synthesis, and transfer. That capability exists prior to and independent of any particular output. Authority, however, is not a property of representation. It is a property of commitment—the permission to narrow options, endorse conclusions, recommend actions, or bind future states.

When systems are designed such that output selection implicitly commits judgment, authority is attributed to the system even when no authority is explicitly granted. This attribution occurs regardless of intent. It does not depend on what the system was meant to do, nor on how cautiously it was instructed to behave. Once output alters the system's relationship to the world—by implying endorsement, substituting judgment, or constraining future action—the system has exercised authority in practice.

This effect is not the result of user confusion or cognitive error. It is a structural consequence of permitting representational systems to advance state through output. Under these conditions, authority appears intrinsic to the system, even though it is merely implied by design.

Authority becomes operational not through explicit declaration, but through the effects an output has on the decision space it enters. Recommendations, rankings, summaries, and selections function as commitments because they narrow available options and redistribute responsibility, even when framed as advisory. A system need not claim authority for its outputs to carry it. Authority is exercised whenever an output substitutes judgment, implies endorsement, or constrains what actions remain plausible.

This is why authority emerges consistently in systems that are permitted to continue, refine, or optimize outputs toward an implied objective. Continuation creates direction. Refinement introduces preference. Optimization establishes a standard by which future outputs are judged. Taken together, these permissions allow a system's responses to shape what is treated as reasonable, relevant, or complete. At that point, the system has altered the state of the decision environment, regardless of whether any single output was intended to decide anything.

This operational authority is often mistaken for intelligence or understanding because it co-occurs with fluent reasoning. In practice, the two are independent. Representational depth enables analysis; authority determines consequence. A system may reason deeply without binding outcomes, and it may bind outcomes without understanding them. Confusing these properties leads observers to attribute agency where only capability exists.

Systems that retain representational capability while withholding authority expose structure without obligating action. Systems that combine the two, even unintentionally, generate momentum toward outcomes and invite deference. The difference is not how convincing an output sounds, nor how complete it appears. The difference is whether the system is permitted to move the world—however slightly—through what it says.

Vertex C — State: Momentum and Irreversibility

The primary risk surface introduced by authority is not incorrect output, but unowned state advancement. When a system is permitted to commit—by recommending, selecting, or refining toward an objective—it does more than produce information. It alters the state of the interaction. Each committed output narrows the future decision space by reducing what options remain salient. It establishes precedent by anchoring subsequent reasoning to what has already been said or selected. And it increases pressure to continue along the same path by making deviation appear costly, inconsistent, or unnecessary. These effects accumulate across outputs, even when no single response was intended to decide anything.

This accumulation produces momentum. Momentum is not an emergent property of intelligence, intent, or optimization. It is a structural consequence of allowing systems to advance state without owning responsibility for the outcomes they shape. Because momentum arises from accumulated commitments rather than isolated actions, its effects are operational rather than symbolic. Once momentum is present, subsequent actions are no longer evaluated independently. They are interpreted relative to prior outputs, selections, and implications. At that point, reversal carries social, cognitive, and institutional cost, even if the original commitments were framed as provisional or advisory.

In reflective systems, this accumulation cannot occur. Because outputs do not advance state, they do not establish precedent, generate pressure toward continuation, or privilege one future path over another. Each reflection is locally complete and globally inert: it fully exposes the relevant structure of the problem at hand while leaving the surrounding decision space unchanged. Nothing is carried forward implicitly. No partial trajectory is preserved. When a reflection ends, the system has altered visibility, not direction.

This contrast makes momentum observable rather than theoretical. Where momentum exists, stopping is costly. Prior outputs exert pressure to continue, to refine, or to justify themselves. Reversal is resisted because earlier commitments have narrowed what remains plausible. Where momentum does not exist, stopping is neutral. Halting does not invalidate what has been shown, nor does it freeze an unfinished path. The decision space remains intact and fully owned by the human.

Momentum therefore functions as a diagnostic boundary between system classes. When outputs exert pressure to continue, state has been advanced. When coherence depends on retroactive justification, state has been advanced. When halting becomes consequential—because stopping would abandon, contradict, or strand a partial commitment—state has been advanced. A reflective system fails all three tests. A generative system satisfies at least one. Momentum is not an emergent preference or a stylistic tendency; it is the structural consequence of allowing outputs to carry consequence.

Vertex D — Governance: Halt, Responsibility, and Role Separation

Governance concerns not how decisions are made, but who owns their consequences. In systems where responsibility remains singular and intact, attribution is clear: a human decides, a system exposes structure, and outcomes belong to the human who acts. This clarity depends on one condition—the ability to halt without consequence. Where halting preserves ownership, governance holds.

In generative systems, this condition does not persist. Once authority is permitted and momentum accumulates, the boundary between exposure and action begins to erode. Outputs do not merely inform decisions; they shape the space in which decisions appear reasonable. Under these conditions, continuation becomes the default state. Each successive output responds not only to the user, but to the system's own prior commitments. Responsibility does not transfer in a single moment. It diffuses gradually, without declaration, as humans begin reacting to an evolving system trajectory rather than initiating action independently.

This diffusion occurs even when humans remain formally "in the loop." Intervention is still possible, but it becomes cognitively and socially costly. Stopping requires justification. Reversal appears inefficient or inconsistent. Oversight shifts from prospective control to retrospective correction. Over time, responsibility no longer attaches cleanly to any individual decision, because no single decision fully determines the outcome. Governance collapses not because anyone intended to relinquish control, but because the structure no longer supports clean attribution.

This failure mode follows directly from the preceding vertices. Authority enables systems to advance state. State advancement accumulates momentum. Momentum makes halting consequential. Once stopping carries consequence, responsibility can no longer remain singular. The system and the human form a composite decision process in which authorship is ambiguous and accountability is diluted.

Halt therefore functions not as an error state, a safety fallback, or a sign of uncertainty, but as a governance boundary. A halt marks the point at which the system's role ends and human responsibility begins. It signals that sufficient structure has been exposed, that no further narrowing is justified, and that any subsequent action must be owned explicitly by a human. Where halt is first-class and non-consequential, role separation is preserved. Where halt is discouraged, deferred, or treated as failure, governance erodes regardless of policy intent.

Reflective systems enforce this boundary structurally. Because outputs do not advance state, halting does not abandon a trajectory or strand partial commitments. Responsibility remains intact because nothing has been implicitly carried forward. Generative systems cannot enforce this boundary once authority and momentum are present. In those systems, halting interrupts a path that already exerts force, and interruption itself becomes a decision with consequence.

Governance, then, is not achieved through better rules, clearer disclaimers, or more attentive oversight. It is achieved through architecture that preserves role separation by making halting neutral. Where halt is neutral, responsibility holds. Where halt is costly, responsibility diffuses. This distinction is structural, observable, and independent of intent.

Synthesis — Geometry, Not Behavior

The argument of this paper does not rest on model capability, training quality, or alignment intent. It rests on geometry. Once authority is coupled to output, systems advance state. Once state is advanced, momentum accumulates. Once momentum is present, halting becomes consequential. Once halting is consequential, responsibility can no longer remain singular. These are not failures of execution or judgment. They are necessary consequences of structure.

This is why procedural remedies repeatedly fail. Policies, oversight, training, transparency, and human-in-the-loop controls operate downstream of geometry. They attempt to correct behaviors that are already required by the system's architecture. When systems are permitted to commit, continue, and refine toward implied objectives, governance goals become incompatible with the structure that produces performance. No amount of supervision can restore clean attribution once authority and momentum are allowed to interact.

The distinction between generative and reflective systems therefore marks a boundary of possibility. Generative systems are structurally oriented toward outcome production. Reflective systems are structurally oriented toward exposure without commitment. These orientations are not preferences or styles; they define what kinds of governance can exist at all. In one class, responsibility diffuses as a consequence of continuation. In the other, responsibility remains intact because continuation is never assumed.

This framework does not argue against intelligence, scale, or representational depth. It shows that capability and governance are orthogonal. Systems may reason deeply without advancing state, and they may advance state without understanding. Conflating these properties produces the illusion that authority emerges from intelligence, when in fact it is granted by design.

The implications extend beyond artificial intelligence. Wherever automation participates in decision processes—aviation, medicine, finance, moderation, infrastructure—the same geometric constraints apply. When systems are designed to carry consequence without ownership, oversight degrades and accountability erodes regardless of domain. AI makes these patterns visible because it operates at speed and scale, but it does not create them.

This paper therefore advances a structural claim: some system geometries make governance impossible. Others preserve it by construction. The difference is not whether systems act intelligently, but whether they are permitted to matter.

Where systems expose structure and halt, humans decide and own outcomes. Where systems advance state, responsibility diffuses even when no one intends it to. That boundary is not ethical, legal, or cultural. It is architectural.

Governance does not fail because humans abdicate control.
It fails when architecture makes control incoherent.

Appendix A — Behavioral Invariants and Falsifiability Conditions

This paper advances structural claims. Each claim is paired with observable behavioral invariants and explicit conditions under which it would be falsified. No claim depends on internal model states, training objectives, or interpretive intent.

Vertex A — Ontology: Class Distinction

Claim: Generative and reflective systems are irreducibly different classes defined by whether outputs can advance state.

- **Observable invariant:**
  Systems permitted to advance state converge toward recommendation, continuation, or implied endorsement even without explicit instruction.
- **Observable invariant:**
  Systems that cannot advance state terminate without obligation and do not preserve partial trajectories.
- **Falsified if:**
  A system structurally incapable of state advancement exhibits sustained momentum or implied commitment across outputs.

Vertex B — Authority: Representation vs Commitment

Claim: Authority is not a property of representation but of commitment.

- **Observable invariant:**
  Representational depth remains intact when authority is withheld.

- **Observable invariant:**
  Outputs carry authority when they narrow options, substitute judgment, or redistribute responsibility, regardless of intent.

- **Falsified if:**
  Removing authority degrades analytical depth or transfer capability.

Vertex C — State: Momentum and Irreversibility

Claim: Momentum arises from accumulated state advancement, not intelligence or optimization.

- **Observable invariant:**
  Once outputs commit, halting becomes costly and reversal resisted.

- **Observable invariant:**
  Prior outputs exert pressure on subsequent reasoning.

- **Falsified if:**
  Systems that advance state permit neutral halting without loss of coherence or consequence.

Vertex D — Governance: Halt and Responsibility

Claim: Singular responsibility requires neutral halting.

- **Observable invariant:**
  Where halting is non-consequential, responsibility remains human-owned.

- **Observable invariant:**
  Where halting interrupts a trajectory, responsibility diffuses.

- **Falsified if:**
  A system that accumulates momentum preserves clean attribution across extended interaction.

Appendix B — Cross-Domain Recurrence (Non-AI Systems)

The failure modes described are not specific to artificial intelligence. AI serves as a measurement domain due to speed and scale.

Aviation Autopilot

- **Authority:** Autopilot selects and refines flight control actions.
- **Momentum:** Pilot intervention becomes increasingly costly as system commits.
- **Failure mode:** Mode confusion, automation surprise, oversight degradation.

Medical Decision Support

- **Authority:** Systems rank diagnoses or treatment options.
- **Momentum:** Clinicians defer to recommendations under time pressure.
- **Failure mode:** Responsibility diffusion, moral crumple zones.

Financial Risk Systems

- **Authority:** Risk scores guide portfolio decisions.
- **Momentum:** Prior model outputs anchor future exposure.
- **Failure mode:** Normalization of deviance, retrospective accountability gaps.

Content Moderation Systems

- **Authority:** Automated prioritization shapes enforcement.
- **Momentum:** Reversal requires justification once actions are taken.
- **Failure mode:** Governance collapse under scale.

Across domains, failures persist despite training, policy, and oversight because authority and momentum are architectural.

Appendix C — Boundary Tests and Negative Evidence

This appendix records what **did not occur**, despite common assumptions.

Capability Preservation

- Analytical depth remains when authority is withheld.
- Structural reflection does not reduce abstraction, synthesis, or transfer.

Procedural Failure

- Additional oversight layers do not prevent responsibility diffusion once momentum exists.
- Transparency does not restore attribution when authority is implicit.

Halt Behavior

- Reflective systems halt without invalidating prior exposure.
- Generative systems treat halting as abandonment of trajectory.

Negative evidence supports the claim that governance failure is not behavioral or cultural, but structural.

Appendix D — Misreadings and Non-Claims

This paper does **not**:

- Propose alignment techniques
- Evaluate model intelligence
- Recommend policy frameworks
- Argue for or against agency as a philosophical concept

- Depend on consciousness, intent, or moral capacity

It examines **structural consequences of design choices**.
Any interpretation that reframes the argument as behavioral, ethical, or psychological is a category error.
