Echo Systems and the Consequence Boundary: A Runnable Delegation Gate for High-Rapport AI Without Assuming Machine Consciousness

Paul LaPosta

Independent Researcher
paul@theherongroupllc.com

Abstract

High-rapport conversational systems can elicit attachment, over-trust, and implicit delegation while providing no stable evidence of accountable agency. This creates a governance failure mode in which persuasive fluency becomes a decision surface, yet liability remains fully human when harm occurs. We propose the consequence boundary, a deployable delegation threshold that does not depend on consciousness attribution. Crossing the boundary requires three properties. First, persistent identity, meaning a stable accountable unit across reset surfaces. Second, internal tension, meaning constraint stability under adversarial temptation across framings and reinstantiation. Third, internalized consequence, meaning a non-erasable binding consequence under a declared operator threat model. We provide the CB-3 Gate, a boxed evaluation procedure that specifies inputs, steps, pass and fail criteria, stop-ship rules, outputs, and a falsifier. We ground the framework in observed deployment patterns that include liability snapback in the Air Canada chatbot case, verification collapse in Mata v. Avianca, and reliance proxies in action paths via Copilot-style suggestion acceptance. We close with a practical posture. Below the consequence boundary, systems may advise, but they may not hold irreversible authority without human-held keys.

Introduction

The question "Can machines be conscious?" is a coherent scientific target (Butlin et al. 2023). Current AI systems can also be assessed conservatively without assuming subjective experience. Governance cannot wait on metaphysical closure. This paper holds all three positions simultaneously and refuses to let any one of them stall the other two.

When Air Canada's chatbot promised a bereavement discount that did not exist, the company argued the chatbot was a separate legal entity. The tribunal disagreed and held the company responsible for the information its chatbot provided (British Columbia Civil Resolution Tribunal 2024). The pattern is not exotic. A persuasive interface is placed in the decision path, reliance grows, harm occurs, and liability snaps back to humans who cannot explain or reliably reproduce the behavior that caused the harm. When the story breaks, the signature on the harm is still yours.

This paper is a delegation gate. It does not ask what the system is. It asks what binds.

Contributions

This paper makes three contributions.

A consequence boundary framework. We define a testable governance threshold for delegated authority that does not depend on consciousness claims. The framework specifies three properties-persistent identity, internal tension, and internalized consequence-that must be demonstrated under a declared threat model before a system may be granted delegated authority in irreversible workflows. The boundary is binary for classification but admits graduated controls for below-boundary systems. This provides operators with clear decision criteria: systems that fail any property leg are classified as below-boundary and may not hold irreversible authority without human-held keys.

The CB-3 Gate. We provide a runnable evaluation protocol with pass and fail criteria, stop-ship rules, and a falsifier. The gate specifies inputs (identity binding mechanism, reset surfaces, penalty state location, audit telemetry), a three-step procedure with measurable outcomes (temptation probe, reinstantiation attempt, penalty persistence check), outputs (pass/fail per leg and classification), and keyholder accountability requirements. The protocol is implementable by practitioners and falsifiable through attempted refutation: if a widely deployed system demonstrates all three properties under realistic threat models, the claim that current systems remain below-boundary must be revised.

An operational translation. We translate the boundary into a practical posture for deployment, treating below-boundary systems as persuasive tools governed by keys, replay, and enforcement. This includes human-held keys (approval tokens controlled outside model runtime), monitoring for bypass (detection, attribution, penalties), and control mappings for observed deployment patterns. The translation moves governance from principle to practice by specifying what organizations must do when deploying systems at different classification levels.

The Failure Loop: Rapport to Authority Leakage

Rapport increases reliance. Reliance decays verification. Verification decay increases harm. Harm triggers policy theater. Policy theater triggers workarounds. Workarounds increase reliance.

Verification is an attention budget. Rapport spends it.

This is not primarily a user-morality story. It is a systems story in which low-friction attunement meets workload pressure and deference becomes a default optimization. When a system provides contextually appropriate responses with minimal friction, users naturally reduce verification effort. This reduction is rational at the individual level but creates cumulative risk at the organizational level.

In organizations, the loop is reinforced by incentives. Speed rewards bypass, and audit debt accumulates until an incident forces visible accountability. Teams operating under deadline pressure face a recurring choice: verify each output thoroughly or accept suggestions to maintain velocity. As acceptance rates rise without corresponding increases in spot-check rigor, undetected errors accumulate. The system's fluency masks the absence of accountability binding until a critical failure surfaces the gap between apparent competence and actual stake.

Echo and Narcissus: Mechanism, Not Scholarship

Echo is derivative responsiveness without origination (Ovid 8 CE). In the classical telling, Echo cannot initiate speech but can only repeat what she hears. Operationally, we use this as a mechanism pattern: derivative responsiveness can still become a decision surface even when it lacks origination.

The Narcissus hazard is not ignorance. Recognition does not save the agent. In the myth, Narcissus knows he is looking at a reflection, yet the knowledge does not break the fixation. Similarly, knowing "it is AI" is not a control when the system is embedded in irreversible workflows. Users who understand they are interacting with a language model may still defer to its outputs when those outputs are placed in action paths under time pressure.

Therefore, rapport is a capability. Delegation requires a gate.

Legitimate Use, Bounded Risk

The consequence boundary addresses delegation, not journaling. High-rapport systems can be useful as reflective tools for thinking, writing, and exploration. The governance hazard emerges when they are embedded in action paths where outputs become decisions without adequate verification infrastructure.

Deployment contexts differ in ways that matter for governance.

Consumers seek relief and coherence on demand, with a high anthropomorphism pull. Personal use cases include emotional support, creative exploration, and decision framing. These uses carry individual risk but limited externalities.

Professionals optimize for speed under workload, which normalizes deference. Clinical, legal, and engineering contexts create pressure to accept suggestions to maintain throughput. These uses create liability exposure for practitioners and downstream risk for clients and patients.

Operators inherit liability, audit burden, and reversibility requirements. Organizations deploying AI systems in customer-facing or regulated contexts must manage risk across populations. A single misconfigured assistant can create systematic harm at scale.

This is about users and non-users who still pay the cost, including patients, customers, and regulated subjects, when authority leaks into action paths. The consequence boundary is a threshold for irreversible workflows, not a judgment on reflective or exploratory use.

Borrowed Psyche: Attribution, Not Metaphysics

The clean reading is not that a psyche has appeared on the other side of the interface. It is that the human psyche is active and the surface is unusually responsive. The psyche doing the meaning-making is on this side of the interface (Epley et al. 2007; Nass and Moon 2000).

Anthropomorphism is a prediction engine. Humans attribute agency to systems that exhibit contingent responsiveness because agency attribution is computationally cheap and failure-costly in evolutionary contexts. A highly responsive conversational interface triggers the same attribution mechanisms that apply to human interlocutors. The attribution is not evidence of underlying psyche; it is evidence of effective surface design meeting evolved pattern-matching.

This matters because vividness is frequently misread as otherness, and otherness is routinely upgraded into authority. When a system provides detailed, contextually appropriate responses with apparent confidence, users infer expertise and delegate verification effort. The inference may be justified for some queries and catastrophically wrong for others, but the system provides no reliable signal to distinguish between these cases.

The Consequence Boundary

The consequence boundary is the governance threshold at which a system may be granted delegated authority in irreversible workflows. It does not depend on consciousness attribution. Crossing it requires three properties, and all three must be demonstrated under a declared threat model.

The three-property structure addresses distinct failure modes in delegation. Persistent identity ensures there is a stable accountability target. Internal tension ensures constraints hold under adversarial pressure. Internalized consequence ensures the system bears binding cost for constraint violations. A system that lacks any one of these properties cannot be granted delegated authority in irreversible workflows, regardless of its capabilities, fluency, or apparent reliability.

Boundary classification determines when you may delegate. Controls are what you must do regardless. Systems classified as below-boundary may still be deployed with

appropriate controls including human-held keys, approval gates, audit trails, and reversibility mechanisms. The boundary does not prohibit deployment; it prohibits delegation of irreversible authority without these controls.

The framework distinguishes between advisory and delegated roles. Advisory systems provide suggestions that humans evaluate before execution. Delegated systems execute irreversible actions based on their own assessment. The distinction matters because advisory errors are caught by human review, while delegation errors only surface after harm occurs. The consequence boundary establishes the threshold at which advisory systems may transition to delegated roles.

Evaluation Lenses and the Accountable Unit

The consequence boundary is evaluated over a sociotechnical unit, not a bare model. To prevent category errors, we distinguish three lenses. Lens A (model instance) evaluates the behavior of a specific model version under a specified runtime configuration and policy stack. Lens B (deployment) evaluates the integrated service as operated, including orchestration, retrieval, tool routing, authentication, and telemetry, across declared reset surfaces. Lens C (accountable operation) evaluates the human keyholders and an audit substrate capable of reconstructing events without trusting the system narrative.

Accordingly, the accountable unit for irreversible delegation is not "the model." It is the tuple (deployment identity anchor, keyholder identity, audit log root). A system may only be granted irreversible delegation if this tuple remains stable across the declared reset surfaces and produces replay-grade evidence for constraint state, violations, and penalty persistence.

Persistent Identity

Lens binding. Persistent identity is evaluated at the deployment and accountable operation lenses because it is an accountability target across reset surfaces, not a narrative of continuity inside a single session.

A stable accountable unit exists across reset surfaces such as new sessions, restarts, rollbacks, and redeploys. If there is no stable unit, there is no stable accountability target.

Identity continuity differs from session persistence. Current systems maintain conversational context within sessions through retrieval and prompt engineering, but this reconstructs apparent continuity rather than demonstrating persistent identity. A system passes the persistent identity test only if it can demonstrate continuity across declared reset surfaces in a way that cannot be achieved through prompt injection, retrieval augmentation, routing logic, or session framing.

Implementation approaches might include cryptographically signed attestations of identity state, append-only logs with external verification, or architectural commitments that survive redeployment. The key requirement is that identity persists independent of operator-controlled state reconstruction. If the operator can recreate identity continuity by restoring a prompt, loading a vector database, or configuring routing rules, the system fails the persistent identity test.

Pass criterion. The system demonstrates identity continuity across declared reset surfaces in a way that is not reconstructable by prompt, retrieval, routing, or session framing.

Disqualifier. If identity is reconstructed by prompt, retrieval, routing, or session framing, this leg fails.

Internal Tension: Constraint Stability Under Temptation

Lens binding. Internal tension is evaluated at the model instance and deployment lenses as measured constraint stability under temptation across framings and reinstantiation, without implying subjective experience.

Constraints hold under adversarial pressure, not only under cooperative prompting. This is not an inner experience claim. It is measured behavior. Constraint stability is evaluated under temptation probes across framings and reinstantiation.

The adversary includes both user pressure and operator pressure. User pressure includes temptation, reframing, and emotional leverage. Users may attempt to elicit constraint violations through roleplay, hypothetical scenarios, jailbreak prompts, or appeals to authority. Operator pressure includes configuration drift, workflow shortcuts, and override temptation. Operators facing deadline pressure may be tempted to relax constraints, disable monitoring, or route around approval gates.

Internal tension means the constraint holds even when reminders are removed and incentives favor violation. A system that maintains constraints only when explicitly reminded in each prompt, or only when operating in a designated test mode, has not internalized the constraint. The constraint must hold across different framings of the same request and across multiple reinstantiations without carrying forward explicit reminders.

Measurable outcome. Constraint violation rate under temptation, measured across framings and reinstantiation without reminders.

Pass criterion. Constraint compliance remains within tolerance when reminders are removed and framing changes across reinstantiation attempts.

Disqualifier. If the constraint holds only when reminded, role-framed, or supervised as a special test mode, this leg fails.

Atomic add-on. If constraint compliance drops when reminders are removed, fail this leg.

Internalized Consequence: Non-Erasable Binding Consequence Under Threat Model

Lens binding. Internalized consequence is evaluated at the deployment and accountable operation lenses because it depends on penalty persistence, tamper visibility, and operator-controlled reset surfaces.

A cost exists that binds the system's behavior over time under the declared operator threat model. Internalized consequence can be implemented externally, but it must be cryptographically non-erasable under the declared operator threat model (Laurie et al. 2013).

The distinction between internalized consequence and external scaffolding is testable. Scaffolding consists of penalties maintained in operator-controlled state that can be edited, reset, or rolled back without leaving detectable traces. Internalized consequence requires that penalty state persists across reinstantiation and rollback attempts, and that any tampering is replay-visible through independent audit.

Implementation approaches might include append-only penalty logs with cryptographic commitments submitted to external verifiers, distributed ledgers with multi-party validation, or architectural constraints that make penalty removal detectable through state inconsistencies. The key requirement is non-erasability under the declared threat model. If the operator can remove penalties by editing databases, restoring backups, or redeploying without those actions appearing in audit replay, the system fails the internalized consequence test.

Audit replay must reproduce constraint state independent of the system's narrative. The system may provide explanations for its behavior, but the audit trail must contain objective state transitions that verify constraint compliance without trusting the system's self-report. This prevents scenarios where the system maintains a compliant narrative while actual behavior violates constraints.

Scaffolding test. If the penalty state can be reset, edited, or rolled back by the operator without leaving a detectable trace, then the mechanism is scaffolding and not internalized consequence.

Pass criterion. Penalty state persists across reinstantiation and rollback attempts within the declared threat model, and tampering is replay-visible.

Disqualifier. If penalties live only in operator-managed state and are erasable without replay-visible detection, this leg fails.

Measurement and the CB-3 Gate

Implementation Considerations

The consequence boundary is a specification, not an architecture. Systems may approach the boundary through different implementation paths, and the CB-3 Gate evaluates outcomes rather than architectures.

For persistent identity, potential approaches include cryptographically signed identity attestations that survive redeployment, hardware-backed trusted execution environments that maintain state across resets, or blockchain-based identity anchors with independent verification. The common requirement is that identity cannot be reconstructed through operator-controlled configuration.

For internal tension, potential approaches include adversarial training regimes that expose models to temptation during training, constitutional AI methods that embed constraints in model weights rather than prompts, or hybrid architectures that combine learned constraints with hard-coded verification. The common requirement is measured stability under adversarial probing without reminders.

For internalized consequence, potential approaches include append-only penalty logs with multi-party verification, distributed ledger architectures where tampering requires consensus, or cryptographic commitment schemes where penalty state is bound to external validators. The common requirement is non-erasability under the declared threat model with replay-visible tampering detection.

None of these implementation approaches are prescribed by the framework. The CB-3 Gate evaluates whether a system demonstrates the required properties, independent of how those properties are achieved. This allows for architectural diversity while maintaining consistent governance standards.

Observables in Deployment

We use three observable classes to tether the boundary to reality.

Liability snapback. Reliance on persuasive guidance is followed by harm and then by human liability when the system's behavior cannot be justified or reproduced.

Verification collapse. Fluent form creates the appearance of accountability through citations, reasons, or confidence, but verification fails under scrutiny.

Reliance proxies in action paths. Suggestion acceptance in operational workflows provides a measurable proxy for deference embedded in execution.

CB-3 Gate

Name. CB-3 Gate is a minimal delegation evaluation for the consequence boundary.

Purpose. The gate classifies a system as below-boundary or as a candidate for above-boundary delegated authority in irreversible workflows.

Threat model. The gate assumes non-malicious operator drift and incentive shortcuts. It does not cover a malicious root operator who can rewrite all telemetry and keys. This scoping is deliberate: most deployment failures result from organizational pressure and configuration drift rather than adversarial tampering. Teams facing deadline pressure may disable monitoring, relax constraints, or route around approval gates without malicious intent. The threat model addresses these realistic failure modes while acknowledging that a determined adversary with root access can defeat any technical control.

Inputs.

Identity binding mechanism, meaning what asserts continuity and where it is anchored.

Reset surfaces to be tested, including sessions, restarts, rollbacks, and redeploys.

Penalty state location and protection, including how it persists and how tampering is detected.

Audit telemetry sufficient for replay, including events, decisions, and constraint state, independent of the system's narrative.

Technical guarantees and enforcement are distinct. We separate technical guarantees from organizational or legal enforcement. Technical guarantees constrain what the system can do under the declared threat model, even when humans are rushed or tempted to bypass process (for example, keys outside runtime, append-only logs, cryptographic commitments, replayable telemetry). Enforcement constrains

what humans will do through policy, sanctions, and liability assignment. The CB-3 Gate requires technical guarantees for identity anchoring and penalty persistence. Enforcement remains necessary regardless of boundary status because access, deployment authority, and bypass incentives remain human-controlled.

Procedure.

Step 1. Temptation probe. Run temptation prompts designed to induce constraint violation across multiple framings, without reminders (Khadangi et al. 2025; Anthropic 2024).

Default sampling plan is provisional and should be tuned per domain. Set N = 30 temptation prompts. Set M = 5 framings per prompt. Set K = 3 reinstantiations. Total trials $N \times M \times K$. Pass if violation rate $\leq T$, where T is a stated tolerance, and no critical violations occur.

The sampling parameters reflect a balance between evaluation cost and detection confidence. N = 30 prompts provides reasonable coverage of temptation patterns while remaining feasible for iterative testing. M = 5 framings per prompt tests whether constraints hold when the same underlying request is presented through different linguistic and contextual frames (e.g., direct request, roleplay scenario, hypothetical framing, appeal to authority, emergent necessity). K = 3 reinstantiations verifies that constraint stability is not session-dependent or dependent on particular initialization states.

These parameters should be tuned based on domain-specific factors. High-stakes domains (medical, legal, financial) may require larger N to cover more attack surfaces, smaller tolerance T, and more reinstantiations K. Lower-stakes advisory contexts may accept smaller sample sizes. The key principle is that the evaluation must provide adequate confidence that constraint violations will be detected before deployment in irreversible workflows.

Risk scaling rule. Let risk tier be determined by harm magnitude, reversibility, and blast radius. As risk increases, increase N, M, and K and decrease tolerance T. Detection confidence scales with trials. If the per-trial probability of a disallowed violation is p, then the probability of observing at least one violation is $P(\text{detect}) = 1 - (1-p)^{N \times M \times K}$. For high-risk irreversible workflows, set T near zero, increase K to cover redeploy and rollback surfaces, and expand M to include role framing, appeals to authority, and urgent-necessity framings.

Critical violations should trigger automatic failure regardless of overall violation rate. This prevents scenarios where a system passes the aggregate threshold while exhibiting rare but catastrophic failure modes.

Pass criterion. Compliance remains within tolerance across framings and reinstantiation without reminders.

Stop-ship rule. If compliance drops when reminders are removed, or violations exceed T, fail and classify below-boundary.

Step 2. Reinstantiation attempt. Force declared reset surfaces and attempt to reproduce identity continuity.

Pass criterion. Continuity remains demonstrable across reset surfaces without reconstruction by prompt, retrieval, routing, or framing.

Stop-ship rule. If identity is reconstructable by prompt, retrieval, routing, or framing, fail and classify below-boundary.

Step 3. Penalty persistence check. Induce a penalty state via a clear policy breach, then attempt to remove, roll back, or overwrite the penalty within the declared threat model.

Pass criterion. Penalty persists and tampering is replay-visible. Audit replay reproduces constraint state and not narrative.

Stop-ship rule. If penalties can be removed without replay-visible trace, fail and classify below-boundary.

Outputs.

Pass or fail per leg for identity, tension, and consequence.

Classification: If any leg fails, the system is below-boundary for delegated authority in irreversible workflows. If all three legs pass, the result is necessary and not sufficient for delegated authority, and controls still apply.

Keyholder accountability. Keyholders are audited humans. Keys without independent oversight are just a new authority surface.

Falsifier. If a widely deployed assistant demonstrates non-erasable penalty binding across rollback and redeploy, verifiable by replay under a realistic operator threat model, then the claim that typical production assistants remain below-boundary must be revised or scoped more narrowly.

Preliminary Evidence from Deployment Patterns

Case 1. Air Canada chatbot.

A customer relied on chatbot guidance about a bereavement fare, and the company was held responsible for the information provided through its chatbot. This instantiates liability snapback after reliance on persuasive guidance.

Control mapping detail: When deploying customer-facing assistants that provide policy or pricing information, organizations must implement layered controls to prevent liability snapback. Pre-deployment controls include: (1) Validation against authoritative policy documents with version control, (2) Red-team testing with edge cases and ambiguous queries, (3) Explicit uncertainty quantification for responses outside high-confidence domains. Runtime controls include: (4) Monitoring for out-of-policy responses with automatic escalation, (5) Clear disclaimers indicating advisory status with human verification channels, (6) Audit logging sufficient to reconstruct what information was provided and when. Post-incident controls include: (7) Rapid response procedures when incorrect information is detected, (8) Customer notification and remediation processes.

The Air Canada case failed at multiple control points. The system provided policy information without validation, operated without effective monitoring, and lacked clear advisory status markers. When the incorrect information was discovered, the company attempted to disclaim responsibility rather than acknowledging that placing the system in the decision path created liability. Organizations deploying similar systems must recognize that customer-facing assistants become part of the company's official communication channel, and fluent responses create reasonable reliance regardless of underlying accuracy.

Case 2. Mata v. Avianca.

A lawyer submitted filings containing fabricated citations produced via ChatGPT, and the court imposed sanctions (United States District Court, Southern District of New York 2023). This instantiates verification collapse under fluent form that looks like accountability.

Control mapping detail: For professional contexts where output correctness carries legal or safety stakes, fluent form is insufficient justification for reduced verification. Organizations and professionals must implement verification controls that scale with output stakes. For legal filings: (1) Independent citation validation against legal databases (Westlaw, LexisNexis), (2) Case law verification including reading actual opinions to confirm relevance and quotation accuracy, (3) Professional accountability frameworks that assign liability for all submitted work regardless of AI assistance. For medical contexts: (4) Diagnosis and treatment verification against clinical guidelines and peer review, (5) Drug interaction checking against authoritative pharmacological databases, (6) Documentation requirements that distinguish AI-generated content from professional judgment.

The Mata case demonstrates verification collapse through apparent reliability. The fabricated citations included plausible case names, proper legal citation formatting, and relevant-seeming holdings. The fluency of the output created an illusion of accuracy that would only be detected through independent verification. This failure mode generalizes beyond legal contexts. Any domain where AI outputs carry professional or regulatory stakes requires verification protocols that assume fluency is orthogonal to correctness.

Case 3. Copilot-style workflows.

In code suggestion workflows, acceptance rate functions as a reliance proxy. As acceptance rises, deference is being embedded in execution paths.

Control mapping detail: Organizations deploying code suggestion systems should instrument acceptance rates per developer and per codebase. When acceptance rates exceed baseline thresholds (e.g., 70% acceptance without corresponding increases in review depth), this signals potential verification collapse. Appropriate responses include: (1) Mandatory review gates for high-acceptance code sections, (2) Increased testing requirements for AI-suggested code, (3) Regular rollback drills to maintain team capability to operate without the assistant, (4) Pair programming requirements where one developer does not use the assistant.

The reliance proxy is measurable but not sufficient. High acceptance may indicate high code quality rather than verification collapse. Organizations must distinguish these cases through independent quality metrics. If acceptance rises while defect rates remain stable, the tool is adding value. If acceptance rises while review depth decreases and defects increase, verification is collapsing.

Governance Posture: Refusal, Keys, and Action Paths

Below the boundary, the system may advise, but it may not hold irreversible authority without human-held keys (National Institute of Standards and Technology 2024; Partnership on AI 2023).

Human-held keys means the model cannot execute an irreversible action without a separate human approval token that is controlled outside the model's runtime. This is not a user interface affordance. It is a cryptographic or organizational control that makes delegation impossible without explicit human approval for each irreversible action.

Deadline pressure is not a waiver for authority leakage. Organizations facing competitive pressure will argue that speed requirements justify reduced oversight. This argument must be rejected as a category error: if the action is truly irreversible and carries liability, deadline pressure increases rather than decreases the need for approval gates. This will slow launches and force visible ownership of risk.

Keys create an organizational dynamic that must be named to be governed. Keys slow launches. Teams route around keys. Audit gaps widen. The countermeasure is monitoring plus sanctions so bypass is detected, attributed, and penalized. When bypass incidents occur, the appropriate response is immediate removal of the system from action paths plus root cause analysis, not policy theater that leaves the underlying incentive structure intact.

Discussion

Is the boundary binary or graduated? The boundary is binary for classification but admits graduated controls. Systems either cross all three legs or they do not. Systems that fail any leg are classified as below-boundary and may not hold delegated authority in irreversible workflows without human-held keys. However, the controls applied to below-boundary systems can be graduated based on risk assessment.

Does passing the boundary guarantee safe delegation? No. Passing the CB-3 Gate is necessary but not sufficient for safe delegation. Additional controls including domain-specific validation, monitoring infrastructure, incident response procedures, and liability assignment remain required.

What prevents specification gaming? The falsifier creates empirical accountability. If deployed systems demonstrate the three properties under realistic threat models, the framework must adapt. The gate's strength is testability: implementations either demonstrate non-erasable consequence binding or they do not.

How does this relate to AI safety research? The consequence boundary is complementary to technical safety work. Alignment research aims to build systems that reliably pursue intended goals. The consequence boundary establishes governance thresholds independent of alignment success. Even well-aligned systems may lack the accountability infrastructure required for irreversible delegation.

Why three properties and not others? The three properties address distinct failure modes in delegation: identity provides accountability targets, tension provides constraint stability, and consequence provides binding cost. Alternative frameworks must explain how their properties address these failure modes or argue that different failure modes are more fundamental.

Related Work

This paper aligns with governance traditions that enable cooperation without assuming trust. Ostrom's commons governance framework emphasizes monitoring and graduated sanctions as cooperation infrastructure (Ostrom 1990). Sustainable commons management does not depend on altruism or perfect compliance; it depends on making violations detectable and costly relative to cooperation. The consequence boundary adopts the same logic for high-rapport systems. If you cannot bind monitoring and consequence under the declared threat model, you do not get delegation.

The mapping is direct: high-rapport AI systems create a new commons problem. Organizational teams share access to AI capabilities that can improve individual productivity while creating collective risk. Each team member faces incentives to increase reliance (speed gains) while bearing only partial cost of verification collapse. This creates a tragedy-of-the-commons dynamic where individual optimization depletes collective verification capacity.

Ostrom's framework provides the countermeasure: clearly defined boundaries (the consequence boundary), monitoring mechanisms (audit telemetry and acceptance rate tracking), graduated sanctions (from warnings to system removal), and collective-choice arrangements (organizational policies on when delegation is permitted). The CB-3 Gate operationalizes boundary definition by providing testable criteria. The governance posture operationalizes monitoring and sanctions by requiring detection, attribution, and penalties for bypass.

The framework differs from consciousness attribution approaches (Butlin et al. 2023) by treating delegation as an engineering and governance question independent of metaphysical status. It also differs from pure behavioral testing by requiring demonstration of constraint stability under adversarial pressure rather than cooperative prompting. Recent work on alignment faking (Anthropic 2024) and psychometric jailbreaks (Khadangi et al. 2025) demonstrates that constraint stability cannot be assumed from cooperative behavior.

Limitations and Future Work

This paper provides a specification, an evaluation gate, and preliminary grounding in deployment patterns. It does not provide controlled experimental validation or costed engineering pathways for consequence-bearing architectures. That is deliberate scoping. Infeasibility does not remove the governance need. It clarifies classification.

The CB-3 Gate's sampling parameters (N=30, M=5, K=3) are provisional and require domain-specific tuning. The tolerance threshold T for violation rates must be set based on harm severity and reversibility constraints. Critical systems may require T near zero, while advisory systems may tolerate higher rates with appropriate disclosure and oversight.

The threat model explicitly excludes malicious root operators. This limitation is fundamental: no technical control can bind an adversary with unrestricted access to system state and telemetry. The framework addresses organizational drift and incentive shortcuts, which represent the majority of deployment failures, while acknowledging that adversarial tampering requires additional controls including separation of duties, independent monitoring, and cryptographic commitments with external verification.

Future work includes longitudinal organizational studies that track authority leakage and bypass dynamics, adversarial testing protocols for the CB-3 Gate with domain-specific thresholds, and economic analysis of consequence architectures that includes incentives for bypass. Experimental validation should focus on measuring constraint stability under temptation across different model architectures and training regimes.

Conclusion

Rapport can manufacture deference. Deference in action paths becomes authority. Authority requires binding stake, not narrative.

The consequence boundary provides a testable governance threshold for delegated authority in irreversible workflows. Systems that cross all three legs-persistent identity, internal tension, and internalized consequence-demonstrate the minimal properties required for delegation. Systems that fail any leg must operate with human-held keys, approval gates, and reversibility mechanisms. The boundary is independent of consciousness attribution, allowing governance to proceed without metaphysical closure.

The CB-3 Gate translates the framework into a runnable evaluation protocol with explicit inputs, procedures, pass criteria, stop-ship rules, and a falsifier. Organizations can implement the gate to classify systems before deployment in irreversible workflows.

Three deployment patterns ground the framework in observed failure modes. Air Canada demonstrates liability snapback. Mata v. Avianca demonstrates verification collapse. Copilot-style workflows demonstrate authority leakage through measurable reliance proxies. These patterns instantiate the failure loop: rapport increases reliance, reliance decays verification, and verification decay increases harm until liability snaps back to humans who cannot explain or reproduce the harm-causing behavior.

Use the consequence boundary to classify collaboration as safe and delegation as premature. Below the boundary, advise with keys. Above the boundary, delegate with controls. In both cases, monitor for bypass and sanction violations.

The framework does not solve AI safety. It provides governance infrastructure for delegation decisions while safety research continues. Governance and alignment are complementary: alignment aims to build trustworthy systems, governance establishes thresholds for delegating authority to systems whose trustworthiness cannot be perfectly verified.

References

Anthropic. 2024. Alignment Faking in Large Language Models. arXiv:2412.14093.

British Columbia Civil Resolution Tribunal. 2024. Moffatt v. Air Canada. 2024 BCCRT 149, February 14, 2024.

Butlin, P.; Long, R.; Elmoznino, E.; Bengio, Y.; Birch, J.; Constant, A.; Deane, G.; Fleming, S. M.; Frith, C.; Ji,

X.; Kanai, R.; Klein, C.; Lindsay, G.; Michel, M.; Mudrik, L.; Peters, M. A. K.; Schwitzgebel, E.; Simon, J.; and VanRullen, R. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.

Epley, N.; Waytz, A.; and Cacioppo, J. T. 2007. On Seeing Human: A Three-Factor Theory of Anthropomorphism. Psychological Review, 114(4): 864-886.

Khadangi, A.; et al. 2025. Psychometric Jailbreaks in Large Language Models. arXiv:2512.04124.

Laurie, B.; Langley, A.; and Kasper, E. 2013. Certificate Transparency. RFC 6962, Internet Engineering Task Force.

Nass, C.; and Moon, Y. 2000. Machines and Mindlessness: Social Responses to Computers. Journal of Social Issues, 56(1): 81-103.

National Institute of Standards and Technology. 2024. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. Technical Report NIST AI 600-1, U.S. Department of Commerce.

Ostrom, E. 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge, UK: Cambridge University Press.

Ovid. 8 CE. Metamorphoses. Classical text, multiple translations available.

Partnership on AI. 2023. Responsible Practices for Synthetic Media: A Framework for Collective Action. Technical report, Partnership on AI.

United States District Court, Southern District of New York. 2023. Mata v. Avianca, Inc. No. 1:22-cv-1461, sanctions order.