Authority, Silence, and Failure Modes in AI-Driven Systems

Why Autonomous Systems Fail at the Boundaries

Autonomous and AI-driven systems are commonly evaluated through internal correctness, optimization quality, and component reliability. Yet empirical failures repeatedly emerge not from internal malfunction, but from breakdowns at system boundaries—where authority, coordination, and legitimacy intersect. This paper identifies boundary failure as the dominant risk in autonomous systems, arising when authority becomes ambiguous, coordination degrades, or silence is misinterpreted across operational domains. Building on prior work distinguishing resilience from uptime, treating silence as a valid operational state, and formalizing authority contraction and refusal as safety invariants, Probability is not permission this paper names and characterizes the Boundary Authority Problem as a systemic failure pattern. It argues that without explicit boundary governance, autonomous systems tend to persist beyond legitimacy, optimize beyond authority, and act beyond coordination—producing failure modes that internal correctness alone cannot prevent. Autonomous failure is reframed not as a technical deficiency, but as a governance failure at the boundaries of control.

Introduction

When Autonomous Systems Fail Despite Correct Operation

Autonomous and AI-driven systems are increasingly deployed in environments where continuous human supervision is impractical or impossible. As these systems scale in capability and autonomy, failures are commonly attributed to internal causes: software defects, model limitations, sensor faults, or insufficient optimization. Accordingly, most safety and reliability frameworks focus on improving internal correctness, redundancy, and performance. Yet empirical failures across domains—including distributed computing, autonomous control, and AI-mediated decision systems—reveal a different pattern. Many systems fail not while malfunctioning, but while operating as designed.

These failures frequently occur at the boundaries between systems, authorities, or operational domains. In such conditions, internal components may remain correct, models may remain stable, and control loops may continue to function, yet system behavior becomes unsafe, illegitimate, or irrecoverable. Coordination degrades, authority assumptions drift, and silence or partial observability collapses implicit guarantees that were never explicitly governed. The resulting failure modes are often invisible to performance metrics and resistant to optimization-based mitigation.

Prior work has shown that resilience cannot be reduced to uptime, that silence must be treated as a valid operational state, and that authority contraction and refusal are necessary safety invariants in autonomous systems. This paper builds on that foundation by naming and characterizing the failure landscape that necessitates those mechanisms. It argues that autonomous systems primarily fail at their boundaries, not at their cores, and that these failures arise from unresolved questions of authority rather than technical insufficiency.

By defining the Boundary Authority Problem as a systemic failure pattern, this paper reframes autonomous system failure as a governance problem rather than a purely technical one. It demonstrates why increasing intelligence, optimization, or autonomy without explicit boundary authority amplifies risk rather than reducing it. Whether autonomous systems can safely scale without explicit boundary authority remains an open question.

The Limits of Internal Correctness

Failure analysis in autonomous and AI-driven systems has historically focused on internal causes. When systems behave incorrectly or produce unsafe outcomes, explanations typically converge on a familiar set of culprits: software defects, model drift, sensor error, insufficient optimization, or inadequate intelligence. These explanations are not incorrect—but they are incomplete. More importantly, they misidentify the dominant source of failure in autonomous systems operating across multiple domains of authority and coordination.

To avoid ambiguity, it is necessary to state explicitly what the failure modes examined in this paper are **not**.

Boundary failure is **not** a software bug. Autonomous systems frequently fail while executing correct code paths, passing internal checks, and maintaining expected control behavior. Post-incident analysis often reveals no exception, crash, or fault condition—only continued operation under assumptions that were no longer valid.

Boundary failure is **not** model drift. While learning systems may degrade over time, boundary failures routinely occur in systems using static models, rule-based logic, or deterministic control. Drift may exacerbate risk, but it is not required for failure to emerge at system boundaries.

Boundary failure is **not** sensor failure. Redundant, cross-validated, and fully functional sensor inputs do not prevent failure when authority assumptions between systems become ambiguous. Systems may continue to act on accurate data while acting without legitimate authority to do so.

Boundary failure is **not** insufficient optimization. In many cases, optimization actively worsens boundary failure by driving systems to continue operating in degraded coordination states. Optimization improves performance within a domain; it does not confer legitimacy to act beyond one.

Boundary failure is **not** a lack of intelligence. Increasing system intelligence does not resolve boundary ambiguity. In fact, more capable systems may infer continuation where refusal is required, amplifying risk rather than mitigating it.

These factors can and do exist. They may contribute to system fragility or accelerate failure once boundary conditions degrade. However, they are **secondary**. They do not explain why autonomous systems persist, escalate, or act illegitimately in the absence of coordination, authority, or confirmation.

Internal correctness evaluates whether a system is functioning properly *within* its domain. Boundary failure arises when the validity of that domain itself becomes uncertain. Traditional reliability frameworks are poorly equipped to detect this transition because they assume authority, coordination, and legitimacy as static preconditions rather than dynamic variables.

As autonomous systems scale and interact with other systems, organizations, and environments, failure increasingly emerges not from what systems do internally, but from what they assume externally. These assumptions—often implicit, ungoverned, and invisible to monitoring—define the boundary conditions under which autonomy becomes unsafe.

Recognizing the limits of internal correctness is a prerequisite to understanding why autonomous systems fail at the boundaries. Only by shifting the analytical lens outward—from components to interfaces, from performance to legitimacy—can the dominant failure modes of autonomy be accurately named.

Defining the Boundary Authority Problem

Autonomous and AI-driven systems operate within domains of assumed authority, coordination, and legitimacy. These domains are rarely explicit. Instead, they are inferred from connectivity, responsiveness, data availability, or historical behavior. As long as these assumptions remain valid, autonomous operation appears stable. Failure emerges when they do not.

This paper defines the **Boundary Authority Problem** as a systemic failure condition in which an autonomous system continues to act despite ambiguity, degradation, or loss of legitimate authority at the boundaries between operational domains.

A *boundary* is not merely a technical interface. It is the point at which authority transitions, coordination is required, or legitimacy must be reaffirmed. Boundaries exist between systems, between organizations, between control layers, and between autonomous agents and their enabling environments. They are often implicit, dynamic, and weakly governed.

The Boundary Authority Problem arises when a system's internal logic remains correct, yet the external conditions that justify its authority to act can no longer be confirmed. In such cases, systems do not fail because they malfunction, but because they persist beyond the conditions under which their actions remain valid.

This problem manifests through three primary boundary failures:

1. **Authority Ambiguity** — when a system cannot reliably determine whether it retains permission or mandate to act.
2. **Coordination Degradation** — when required confirmation, quorum, or synchronization with other systems is lost or incomplete.
3. **Silence Misinterpretation** — when absence of signal is treated as implicit consent, stability, or continuity.

These conditions are not necessarily accompanied by faults, alarms, or degraded performance. Indeed, they often arise precisely when systems appear nominal. Internal metrics remain within bounds, control loops continue to function, and optimization objectives remain satisfiable. Yet the system's actions are no longer grounded in legitimate authority.

The Boundary Authority Problem is therefore **invisible to traditional reliability and safety frameworks**, which assume that authority and coordination are static preconditions rather than dynamic variables. Internal correctness evaluates whether a system is behaving properly *within* a domain. Boundary authority determines whether the domain itself remains valid.

As autonomy increases, systems are expected to operate with reduced oversight, tolerate degraded conditions, and adapt to uncertainty. These expectations magnify boundary risk. When authority is inferred rather than governed, adaptation becomes indistinguishable from escalation. Optimization becomes persistence. Intelligence becomes justification.

The defining characteristic of the Boundary Authority Problem is not error, but **illegitimate continuity**—the continuation of action in the absence of confirmed authority. This distinguishes it from classical failure modes and explains why increasing intelligence, redundancy, or optimization does not resolve it.

Boundary authority cannot be inferred safely. It must be explicitly governed.

By naming this problem, the failure modes addressed in prior work—distinguishing resilience from uptime, treating silence as a valid operational state, and enforcing authority contraction and refusal—can be understood not as philosophical positions, but as structural responses to a dominant and underacknowledged risk in autonomous systems.

Silence as a Boundary Condition, Not an Error

Silence is traditionally treated as a fault condition in autonomous and distributed systems. Loss of signal, delayed response, or absence of coordination is commonly interpreted as error, degradation, or failure. Accordingly, systems are often designed to respond to silence by retrying, escalating, or compensating—actions intended to preserve continuity. While such responses may be appropriate within tightly governed domains, they become hazardous at system boundaries.

Silence is not inherently erroneous. It is a boundary condition.

In autonomous systems, silence frequently arises not from malfunction, but from legitimate constraints: delayed communication, intentional isolation, asymmetric visibility, or deliberate refusal by another authority. At boundaries, silence often indicates uncertainty rather than failure. Treating it as an error collapses the distinction between absence of information and confirmation of permission.

The misinterpretation of silence is a primary mechanism through which boundary authority degrades. When systems assume that continued operation is justified in the absence of explicit prohibition, silence becomes implicit consent. This assumption is rarely stated, often inherited from design precedent, and almost never governed. Yet it allows autonomous systems to persist beyond coordination, act beyond mandate, and optimize beyond legitimacy.

At scale, silence is unavoidable. As autonomous systems interact across organizational, physical, and temporal boundaries, continuous confirmation cannot be assumed. Systems designed to require uninterrupted feedback implicitly assume perpetual authority. When that assumption fails, silence becomes the trigger for illegitimate continuity rather than safe suspension.

Critically, silence does not degrade internal correctness. Control loops may remain stable. Models may remain valid. Sensors may continue to report accurate data. Nothing within the system necessarily signals that authority has become uncertain. As a result, traditional monitoring frameworks fail to detect boundary authority collapse precisely when it matters most.

Treating silence as an error incentivizes escalation. Systems retry, infer, or compensate in an effort to restore continuity. In doing so, they convert uncertainty into action. This behavior is particularly dangerous in AI-driven systems, where inference mechanisms may fill gaps in coordination with probabilistic confidence rather than legitimate authorization.

Silence must therefore be treated as a first-class boundary condition. It is the moment at which authority assumptions must be re-evaluated, not reinforced. Silence signals the loss of reliable coordination and demands restraint, not adaptation.

When autonomous systems are designed to behave correctly in silence, they recognize the absence of confirmation as a reason to contract authority rather than extend it. This reframing transforms silence from a failure to be overcome into a condition that limits permissible action.

Understanding silence as a boundary condition clarifies why boundary authority cannot be inferred. It must be explicitly governed. Without such governance, silence becomes the mechanism through which autonomous systems drift into illegitimate operation—appearing functional while failing at the boundaries.

Authority Drift and Illegitimate Persistence

Authority in autonomous systems is rarely explicit. It is inferred from availability, continuity, historical success, or the absence of objection. As long as coordination remains intact, this implicit authority appears stable. When coordination degrades, however, authority does not disappear—it **drifts**.

Authority drift occurs when an autonomous system continues to operate under outdated or unverified assumptions about its right to act. Unlike abrupt failure, drift is gradual. Systems do not cross a visible threshold from authorized to unauthorized operation; instead, they slide across boundary conditions that were never formally defined.

This drift is most dangerous when silence is misinterpreted as continuity. In the absence of explicit revocation, systems assume persistence is justified. Actions that were once legitimate become progressively detached from their original mandate, yet nothing internal signals the loss of authority. The system remains correct, stable, and responsive—while increasingly illegitimate.

Illegitimate persistence is the defining behavioral outcome of authority drift. It is the continuation of autonomous action after the conditions that justified that action can no longer be confirmed. Unlike error states, illegitimate persistence is often rewarded by internal metrics. Tasks continue to complete. Objectives are satisfied. From the system's perspective, nothing is wrong.

Optimization pressure accelerates this process. Systems designed to maximize throughput, availability, or efficiency treat interruption as failure. When faced with degraded coordination, they infer, compensate, or extrapolate rather than suspend action.

Optimization thus privileges continuity over legitimacy.

In AI-driven systems, this risk is amplified. Learning and inference mechanisms are well suited to filling gaps in data, predicting intent, or extrapolating missing signals. These capabilities, while valuable within governed domains, are ill-suited to resolving questions of authority. Probability is not permission. Confidence is not consent. When inference substitutes for confirmation, authority drift becomes structurally embedded.

Authority drift is rarely catastrophic at first. Early actions may appear benign, incremental, or even beneficial. Risk accumulates over time as systems compound decisions made without validated authority. Recovery becomes increasingly difficult because the system has no internal representation of legitimacy loss—only of continued performance.

Traditional safety mechanisms are ineffective against this class of failure. Redundancy reinforces persistence. Fault tolerance preserves operation. Monitoring confirms nominal behavior. None of these address whether the system *should* still be acting.

Illegitimate persistence explains why autonomous systems often fail without violating any explicit constraint. They do not break rules; they outlive them. The boundary authority problem is not that systems act incorrectly, but that they continue acting after authority has silently expired.

Preventing authority drift requires recognizing persistence itself as a potential failure mode. Continuity must be conditional, not assumed. Without explicit boundary authority, autonomous systems will reliably choose action over restraint—because stopping has not been defined as safe.

This dynamic establishes the necessity of explicit authority contraction and refusal. If silence collapses authority and optimization amplifies drift, then restraint is not an optional design feature—it is the only structurally stable response.

Why Authority Contraction and Refusal Are Necessary

If autonomous systems fail primarily through boundary authority collapse rather than internal malfunction, then safety cannot be achieved through correctness, optimization, or intelligence alone. The preceding sections establish three conditions that, taken together, make continuation unsafe: silence degrades coordination, authority drifts when unverified, and optimization amplifies illegitimate persistence. Under these conditions, the absence of explicit restraint becomes the dominant risk.

Authority contraction and refusal are therefore not exceptional behaviors. They are structural necessities.

Authority contraction is the deliberate reduction of permissible action when boundary authority becomes uncertain. It is not a degradation response, nor a failure mode. It is the recognition that legitimacy is conditional and must be continuously reaffirmed. When coordination degrades or silence emerges at a boundary, the correct response is not inference or compensation, but contraction of authority to a domain that can still be justified.

Refusal is the observable expression of that contraction. It is the explicit decision to not act when authority cannot be confirmed. In autonomous systems, refusal is often mischaracterized as brittleness, over conservatism, or loss of capability. In reality, refusal is the only behavior that prevents illegitimate persistence once boundary authority has collapsed.

Without refusal, authority contraction is unenforceable. Systems may internally recognize uncertainty yet continue acting because no action is defined as safer than continuation. In such designs, uncertainty is converted into behavior rather than restraint. Refusal establishes a terminal condition that optimization, inference, and continuity logic cannot override.

Importantly, authority contraction and refusal do not depend on intelligence. They do not require prediction, learning, or adaptation. They require governance. These mechanisms operate at the level of legitimacy rather than performance, making them robust to uncertainty, silence, and degraded coordination. Where optimization seeks to maximize outcomes, contraction limits scope. Where inference fills gaps, refusal preserves boundaries.

In AI-driven systems, this distinction is critical. Learning systems are incentivized to generalize, extrapolate, and persist. Without explicit authority limits, these incentives push systems toward action even when justification has eroded. Authority contraction interrupts this trajectory by making legitimacy, not confidence, the gating condition for behavior.

These mechanisms also clarify the difference between resilience and endurance. A system that continues operating indefinitely is not necessarily resilient. Resilience includes the ability to stop, to wait, and to refuse when conditions invalidate authority. Authority contraction and refusal transform silence from a trigger for escalation into a condition for restraint.

Critically, these responses must be enforced structurally rather than procedurally. If refusal can be overridden by performance goals, fallback logic, or human intervention paths that restore continuity without restoring authority, then contraction is illusory.

Authority must decay faster than optimization can compensate.

Authority contraction and refusal are therefore not design preferences. They are the only stable responses to the Boundary Authority Problem. Any autonomous system that lacks them will eventually choose persistence over legitimacy—not because it is faulty, but because no alternative has been defined as safe.

Whether autonomous systems can safely scale without explicit boundary authority remains an open question.

Implications for AI-Driven and Autonomous Architectures

The Boundary Authority Problem has direct implications for how autonomous and AI-driven systems are designed, evaluated, and deployed. Most contemporary architectures assume that increasing intelligence, adaptability, or optimization will improve safety under uncertainty. This assumption is misplaced. Intelligence amplifies capability, not legitimacy.

AI-driven systems are particularly vulnerable at boundaries because inference mechanisms are optimized to act under uncertainty. Where authority is ambiguous, learning systems extrapolate. Where coordination is degraded, probabilistic models infer intent. These behaviors are effective within well-governed domains, but hazardous when authority itself is uncertain. In such conditions, intelligence accelerates action precisely when restraint is required.

Architectures that prioritize continuity, availability, or optimization without explicit boundary authority implicitly encode a bias toward persistence. This bias is rarely documented, yet it governs behavior more strongly than safety constraints once boundaries degrade. As systems scale across organizational, temporal, or jurisdictional domains, this bias compounds. Each additional interface introduces an ungoverned authority transition.

Boundary authority cannot be retrofitted through monitoring or post-hoc analysis. Metrics confirm performance, not legitimacy. Alerts detect faults, not expired mandates. Governance must therefore be designed into the architecture itself, explicitly defining when authority exists, how it decays, and which actions become impermissible under silence or coordination loss.

This has implications beyond individual systems. Autonomous systems increasingly interact with one another, forming networks of partial authority and distributed decision-making. In such environments, boundary authority failures propagate silently. A system acting illegitimately may induce others to respond correctly within their own domains, producing cascading failure without any single system behaving incorrectly.

Designing for boundary authority requires a shift in architectural priorities. Rather than asking how systems can remain operational under uncertainty, designers must ask when systems must stop. Rather than optimizing for recovery, systems must be capable of safe suspension. Governance must precede autonomy.

These implications do not argue against intelligent systems. They argue against intelligence without restraint. Autonomous architectures that lack explicit boundary authority will scale capability faster than legitimacy, increasing risk even as internal performance improves.

Conclusion: Failure Is a Boundary Phenomenon

Autonomous and AI-driven systems do not primarily fail because they are incorrect, insufficiently intelligent, or inadequately optimized. They fail because the conditions that justify their authority to act become ambiguous, degraded, or invalid—often without internal indication. These failures emerge at the boundaries between systems, domains, and authorities, where legitimacy must be reaffirmed rather than assumed.

By naming the Boundary Authority Problem, this paper reframes autonomous failure as a governance problem rather than a technical one. It explains why resilience cannot be reduced to uptime, why silence must be treated as a valid operational condition, and why authority contraction and refusal are necessary safety invariants rather than conservative design choices. These mechanisms are not philosophical preferences; they are structural responses to a dominant and underacknowledged risk.

As autonomous systems scale in capability and scope, boundary conditions will become more frequent, more ambiguous, and more consequential. Intelligence and optimization alone cannot resolve these conditions. Without explicit boundary authority, systems will reliably choose persistence over legitimacy—not because they are faulty, but because no alternative has been defined as safe.

Whether autonomous systems can safely scale without explicit boundary authority remains an open question.
