

Leading Path Layer (LPL): Governing Autonomous Authority Under Perceptual Uncertainty

Keyonn Whitfield II *

Independent Researcher, Macomb, MI 48044

ABSTRACT

As artificial intelligence expands into embodied robots, vehicles and autonomous machines, failures increasingly arise from systems continuing to act when their understanding is no longer sufficient to justify continued participation. Existing robotic safety standards address known hazards via downstream mechanisms such as torque limits, emergency stops, and collision thresholds. These approaches assume sufficiently reliable perception and are activated only after a potentially dangerous state is detected, implicitly preserving autonomous authority even as underlying understanding degrades. In unstructured, human-centered environments, uncertainty is inherently structural and may lead to a previously under-specified failure mode: “safe-but-uncertain operation”, where a system complies with all hard constraints while continuing to exercise autonomous authority under insufficient understanding. This paper introduces the Leading Path Layer (LPL), an upstream governance layer that conditions whether autonomous authority is permitted based on system understanding, enforcing relinquishment of authority before traditional limits are threatened. This work reframes safety from improving behavior under uncertainty to defining when behavior should cease entirely.

Keywords: Authority Modulation, Human-Centered Autonomy, Robotic Risk Governance, Embodied AI Safety

I. INTRODUCTION

Traditional safety systems have become quite effective in managing known physical hazards through reactive mechanisms such as force limits, collision detection, and emergency stop functions. Current safety standards for industrial autonomous systems, including ISO 10218 and ISO 12100, formalize these downstream protections by focusing on risk identification, hazard mitigation, and constraint enforcement during operation (ISOa, 2010; ISOb, 2011a; ISOc, 2011b). These mechanisms are effective when failures are mechanical or when hazards are well-defined.

However, as autonomous systems increasingly operate in unstructured, and human-centered environments, the critical challenge shifts upstream from physical execution to the continued exercise of autonomous authority under degraded understanding. Critically, this is not a failure of incorrect action, but of continued participation beyond the point where action remains legitimate. In these spaces, uncertainty is not occasional but persistent. Perception systems routinely manage stressors such as partial observability, novel objects, glare and distributional shift. Although uncertainty estimation in machine learning has advanced significantly (Gal & Ghahramani, 2016;

Kendall & Gal, 2017), existing robotic safety architectures do not formally govern whether autonomous authority should continue when understanding degrades.

Safety literature works, within systems theory, has long emphasized that accidents often occur from structural control failures rather than isolated component faults (Leveson, 2011; Rasmussen, 1997). Yet in embodied artificial intelligence, there remains no explicit upstream rule that prohibits confident action under degraded internal understanding. As a result, robots may remain within all predefined safety constraints while operating under insufficient understanding that may be difficult for operators, auditors, or regulators to detect or defend in post-incident analysis.

Recent regulatory initiatives, including the NIST AI Risk Management Framework and the European Union AI Act, highlight the importance of transparency, risk governance, and human oversight in AI systems (National Institute of Standards and Technology, 2023; European Commission, 2021). However, these frameworks do not specify a runtime architectural mechanism that ties autonomous authority to measured understanding.

This paper argues that the absence of an explicit constraint governing whether autonomous authority remains permissible under uncertainty represents a structural gap in existing robotic safety stacks. This work identifies this gap as a distinct failure category and propose an architectural mechanism that binds autonomous authority to measured understanding. This work treats authority as a governed variable independent of behavior adaption. Unlike existing approaches that incorporate uncertainty into decision-making while assuming continued authority. LPL defines a condition under which authority itself is no longer permitted to persist. Systems that continue to act under degraded understanding, or autonomously restore authority after relinquishment, are considered non-compliant regardless of internal uncertainty handling mechanisms. LPL is not a situational awareness model, planner, or decision-maker. Instead, LPL governs when autonomous action is permitted by conditioning authority on system understanding, requiring relinquishment of authority under insufficient understanding and explicit human authorization for its restoration. LPL defines a non-substitutable constraint on autonomous systems: authority must be relinquished once continued action is no longer justified by a stable and reliable basis. Compliance is determined by observable authority behavior under degraded understanding, rather than internal uncertainty estimates or behavioral adaption.

* Corresponding author E-mail: mike.williams8034@pm.me

II. PROBLEM STATEMENT: SAFE-BUT-UNCERTAIN OPERATION

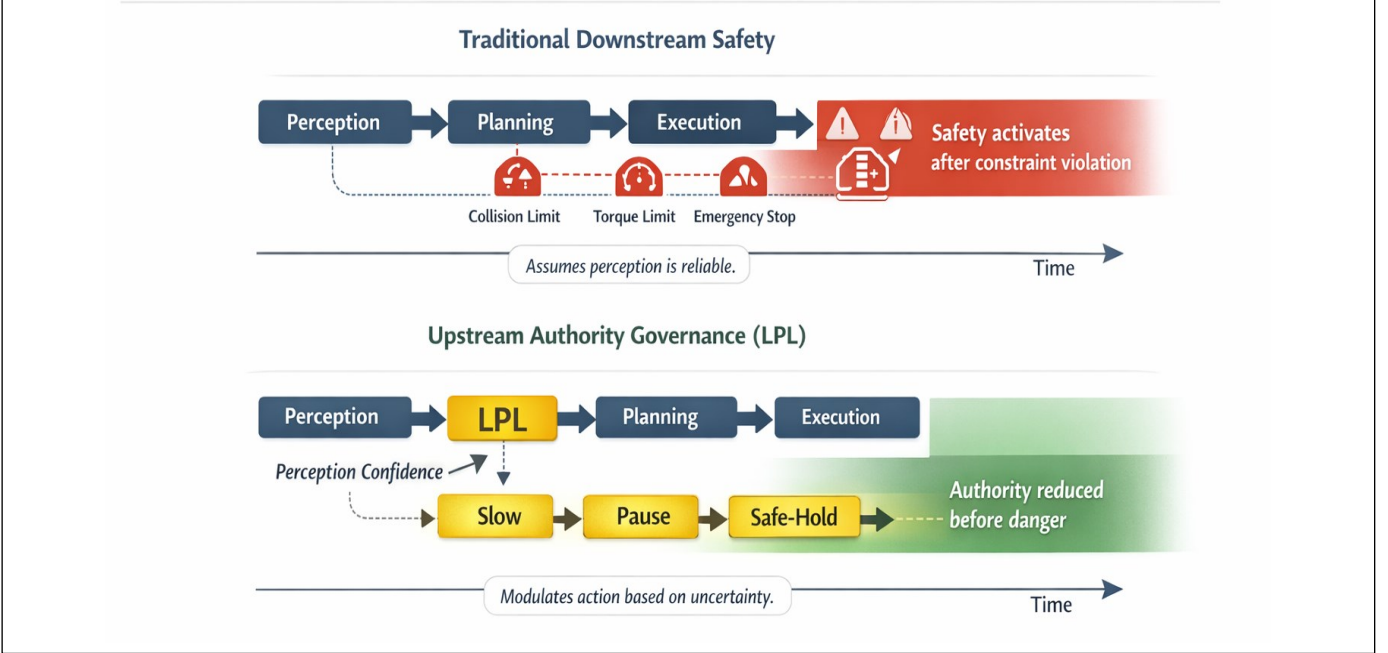
A. Downstream safety assumptions

Existing robotic safety frameworks are predominantly downstream-orientated (Fig. 1). Safety interventions are typically triggered after threshold violations, detected collisions, or explicit hazard identification (International Organization for Standardization, 2010; International Organization for Standardization, 2011a). These structures assume that perception and state estimation remain sufficiently reliable unless a constraint breach occurs.

C. Governance implications of degraded understanding

Research around systems safety has shown that accidents frequently emerge from structural misalignments and control gaps rather than mechanical faults (Leveson, 2011; Rasmussen, 1997). Yet existing robotic safety standards do not formally represent a state in which insufficient understanding alone invalidates autonomous authority. As a result, responsibility for determining whether autonomous authority remains permissible under degraded understanding is left to implicit design choices within proprietary autonomy stacks.

Fig. 1. Downstream versus upstream safety responses. (Traditional robotic safety activates only after predefined constraints are violated. In contrast, the Leading Path Layer (LPL) modulates or restricts action based on rising perceptual uncertainty before downstream safety limits are challenged.)



However, in real-world human-centered spaces, perceptual degradation does not necessarily bring about immediate constraint violation. Instead, uncertainty may accumulate gradually due to sensor noise, partial observability or confusing human behavior (Gal & Ghahramani, 2016; Kendall & Gal, 2017). While modern machine learning systems has made significant strides in estimating predictive uncertainty, these estimates are rarely elevated to a binding safety condition that governs whether action remains permissible.

B. Defining safe-but-uncertain operation

This paper defines this unaddressed condition as safe-but-uncertain operation: a state in which an autonomous system remains within all hard safety limits while lacking sufficient understanding to justify continued autonomous authority. This condition is not hazard detection or equivalent to component failure. Rather, it is a state that reflects degraded internal understanding that may be difficult to interpret. The defining issue is not whether the system acts incorrectly, but whether it continues to act when it no longer has sufficient basis to justify doing so. This reframes the problem from performance under uncertainty to legitimacy of continued action.

In human-centered deployments such as hospitals, care facilities, or public environments the continued exercise of autonomous authority under degraded understanding becomes difficult to defend after an event. Governance frameworks emphasize transparency, human oversight, and accountability (European Commission, 2021; National Institute of Standards and Technology, 2023), but they do not specify an architectural rule that binds runtime authority to measured understanding.

This failure mode is not caused by incorrect decisions or faulty execution. It arises from the absence of a structural rule that prohibits autonomous authority when understanding is insufficient to justify continued action. Without such a rule, autonomous systems may continue operating under degraded understanding in ways that are technically compliant yet socially and regulatorily fragile.

Across existing safety and control approaches, uncertainty is used to inform behavior while autonomous authority is implicitly preserved. This creates a structural gap: no binding condition exists that invalidates authority under degraded understanding. LPL addresses this gap by introducing an explicit, externally observable constraint on whether action is permitted at all.

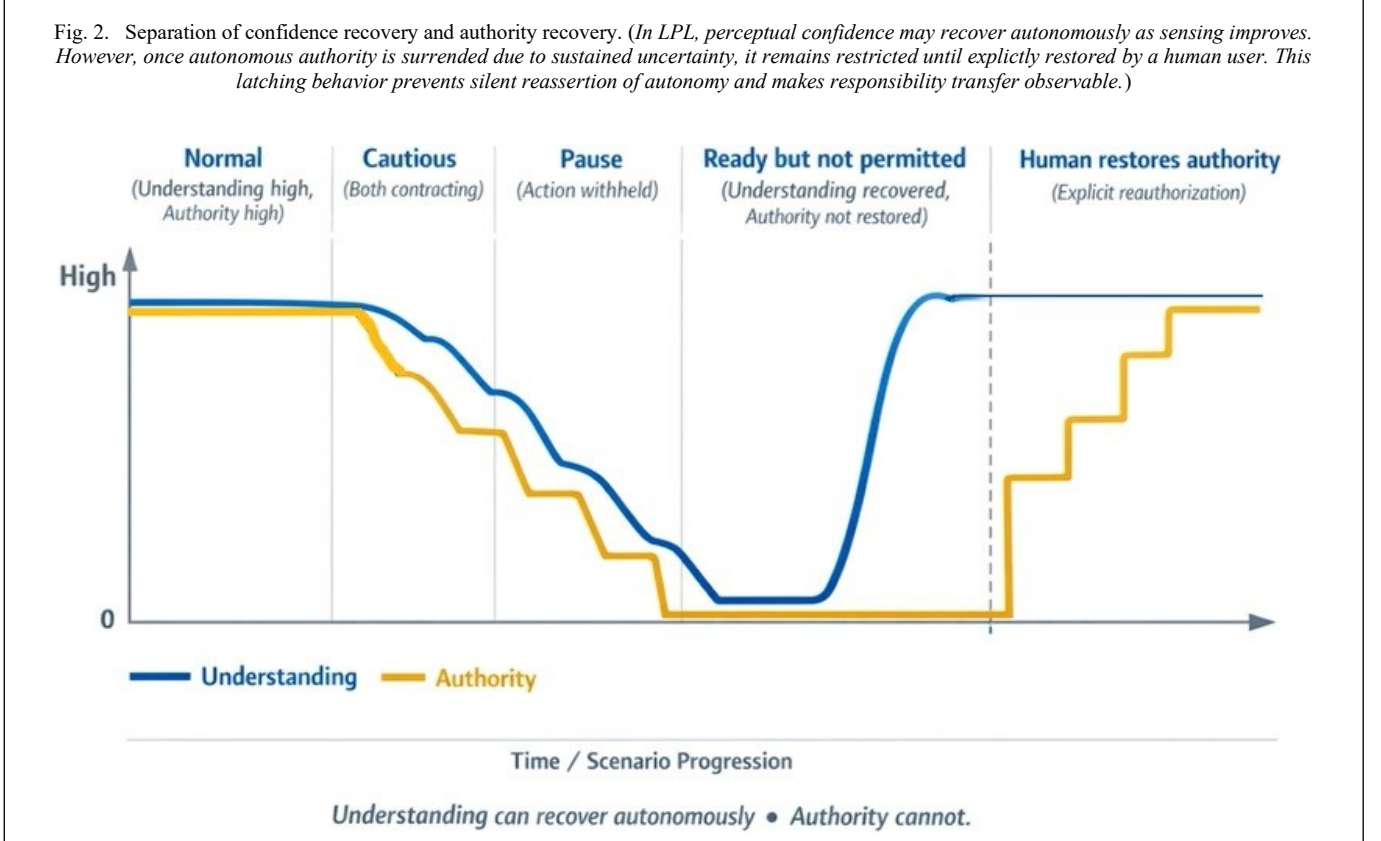
III. CONTRIBUTION AND DISCUSSION: LEADING PATH LAYER

A. Authority contraction as a runtime rule

This work introduces the Leading Path Layer (LPL) as an upstream constraint on autonomous systems that governs authority rather than intelligence. Existing approaches incorporate uncertainty into decision-making while preserving continuous authority. LPL instead introduces a condition under which authority itself is no longer valid. Uncertainty estimation has been widely studied (Gal & Ghahramani, 2016; Kendall & Gal, 2017), yet existing systems primarily use uncertainty to improve prediction or optimize performance. Instead, LPL treats system understanding as a binding condition on authority, determining whether autonomous action is permitted at all.

This runtime rule separates understanding recovery from authority recovery. Systems may continue sensing and refining estimates autonomously, they may not silently resume autonomous authority once it has been handed off to human. This design aligns with emerging governance principles emphasizing human oversight and accountability (European Commission, 2021; National Institute of Standards and Technology, 2023), yet operationalizes those principles as a structural runtime constraint.

Fig. 2. Separation of confidence recovery and authority recovery. (In LPL, perceptual confidence may recover autonomously as sensing improves. However, once autonomous authority is surrendered due to sustained uncertainty, it remains restricted until explicitly restored by a human user. This latching behavior prevents silent reassertion of autonomy and makes responsibility transfer observable.)



LPL is not a planner, controller, situational awareness model or an autonomous decision-maker. Rather than improving environmental understanding, LPL enforces an operational rule: Autonomous authority may contract under uncertainty, but once relinquished, it cannot be autonomously restored (Fig. 2). This establishes a discrete transition from legitimate participation to non-participation, rather than a continuous adjustment of behavior. This rule constitutes a non-substitutable constraint on authority. Methods that adapt behavior under uncertainty while maintaining continuous autonomous authority are not equivalent, as they do not govern the legitimacy of authority itself. Compliance with LPL is determined solely by externally observable authority transitions under degraded understanding, rather than internal representations, confidence measures, or performance outcomes.

B. Architectural role and system integration

Architecturally, LPL does not require a new internal model structure. LPL interfaces with existing perception, planning, and control stacks including ROS2-based and learning-enabled systems. Its purpose is not to evaluate correctness of decisions, but to regulate legitimacy of authority under degraded understanding. LPL complements system safety approaches that focus on structural control constraints over reactive hazard mitigation (Leveson, 2011).

As understanding degrades, LPL conditions the permissibility of autonomous authority, enforcing its relinquishment prior to downstream safety activation. These behaviors range from slowing of pace, pausing, sensor reorientation, or safe-hold states. This reduces reliance on late-

stage emergency interventions (International Organization for Standardization, 2010) while increasing transparency and inspectability for non-expert stakeholders.

C. Distinction from adjacent safety and control frameworks

While related to uncertainty-aware planning, risk-sensitive control, and runtime assurance frameworks, LPL is architecturally distinct. Risk-sensitive control and uncertainty-aware planning adjust system behavior while assuming continued authority (Rasmussen, 1997). Runtime assurance frameworks enforce constraint compliance after violation (Althoff, 2014); LPL instead governs whether autonomous authority is permitted to persist prior to violation. Systems that continue to act under degraded understanding, or that autonomously resume authority after relinquishment, do not satisfy this condition, irrespective of their internal uncertainty modeling or fallback strategies. Similarly, human-in-the-loop supervisory control allows intervention but does not impose a structural separation between understanding recovery and authority restoration (Sheridan, 1992). Existing methods assume that the system remains a valid participant while adapting its behavior. LPL challenges this assumption by defining conditions under which participation itself is no longer justified. LPL’s contribution lies in formalizing this separation as a runtime rule rather than an implementation detail.

The contribution of this paper is architectural rather than empirical. We formalize authority contraction under uncertainty as a distinct, previously unrepresented safety layer within autonomous systems. By shifting safety attention from threshold violation to legitimacy of action under degraded understanding, LPL establishes a structural bridge between uncertainty estimation, runtime behavior modulation, and governance accountability.

IV. CONCLUSION AND IMPLICATIONS

This paper formalizes a previously unrepresented safety condition in embodied artificial intelligence: safe-but-uncertain operation. By separating degraded understanding from constraint violation or mechanical fault, this work identifies a structural gap in current robotic safety stacks that becomes increasingly relevant in dynamic, human-centered environments. The Leading Path Layer (LPL) addresses this gap by introducing an upstream architectural rule that binds autonomous authority to measured understanding.

Rather than replacing existing safety mechanisms, LPL complements standards such as ISO 10218 and ISO 12100 by reducing reliance on late-stage threshold activation and increasing the runtime transparency of authority transitions (International Organization for Standardization, 2010; International Organization for Standardization, 2011a). LPL does not prescribe specific perception models or control strategies. It defines a behavioral constraint that can be evaluated independently of internal autonomy architectures. This design enables compatibility with learning-enabled and diversified systems while preserving inspectability.

As autonomous machines expand into public-facing and high-liability domains, the ability to formally represent and govern safe-but-uncertain operation may become increasingly relevant for certification, auditing, and regulatory review (National Institute of Standards and Technology, 2023; European Commission, 2021). This work acknowledges that enforcing authority contraction under degraded understanding may introduce operational tradeoffs, including increased conservative behavior, reduced efficiency, or greater reliance on human oversight in ambiguous situations.

However, LPL prioritizes legitimacy and inspectability of autonomous behavior over uninterrupted performance. This aligns with system-level safety perspectives that treat governance as a structural constraint rather than a performance optimization variable (Leveson, 2011). By defining compliance in terms of observable authority behavior, LPL establishes a boundary condition that cannot be satisfied by internal confidence measures, risk-aware control strategies, or other adaptive mechanisms that preserve continuous authority. This work shifts the focus of autonomous system safety from how systems act under uncertainty to whether they should be acting at all. LPL defines the boundary at which continued action is no longer legitimate, establishing a constraint that cannot be satisfied through performance improvement alone.

References

- Althoff, M. (2014). An introduction to reachability analysis and its application to the safety assessment of autonomous cars. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)* (pp. 157-162). IEEE.
- European Commission. (2021). *Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. European Commission.
- Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning* (pp. 1050-1059).
- ISOa (2010) International Organization for Standardization. ISO 12100: *Safety of machinery—General principles for design—Risk assessment and risk reduction*. ISO.
- ISOb (2011a) International Organization for Standardization. ISO 10218-1: *Robots and robotic devices—Safety requirements for industrial robots—Part 1: Robots*. ISO.
- ISOc (2011b) International Organization for Standardization. ISO 10218-2: *Robots and robotic devices—Safety requirements for industrial robots—Part 2: Robot systems and integration*. ISO.
- Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems* (Vol. 30).
- Leveson, N. (2011). *Engineering a safer world: Systems thinking applied to safety*. MIT Press.
- National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. U.S. Department of Commerce.
- Rasmussen, J. (1997). Risk management in a dynamic society: A modelling problem. *Safety Science*, 27(2-3), 183-213.
- Sheridan, T.B (1992). *Telerobotics automation, and human supervisory control*. MIT Press.