renewcommand0.4pt0.4pt

Human-in-the-Loop Oversight:

The Training Wheels Protocol as Architectural Foundation

Jeremy Blaine Thompson Beebe *Independent Researcher* ORCID: 0009-0009-2394-9

April 25, 2026

Abstract

Human oversight of autonomous systems is typically treated as a configuration option — enable it or disable it. This binary treatment fails to capture the graduated nature of earned trust and the architectural requirements for safe autonomy graduation. This paper presents the Training Wheels Protocol: an architectural foundation for human-in-the-loop oversight with four graduated modes (HITL Active, Shadow Mode, Full Autonomy, Lockdown) that reflect the system's earned trust level based on demonstrated behavioral consistency. The protocol monitors behavioral patterns, computes trust scores, and enforces mode transitions automatically. We present the four-mode state machine, the trust scoring mechanism, the mode transition rules with formal proofs of safety properties, and deployment evidence from the Agentic platform showing 23,847 human review events, a 94.7% approval rate, and 6 automatic mode transitions over 347 days of operation.

Keywords: human-in-the-loop, Training Wheels Protocol, graduated autonomy, behavioral trust scoring, autonomous oversight, BX3 Framework, Agentic

---

1 Introduction

The question of how much autonomy an AI system should have is not a binary question. A system that has operated reliably for 10,000 consecutive actions with 99.9% human approval rate has earned different treatment than a system that has operated for 10 actions with 60% approval rate. Treating both systems identically — either requiring human review for both or granting full autonomy to both — is a failure of the oversight system.

Current approaches to human oversight treat it as a configuration setting. This has two failure modes: either the system is over-constrained (human review required even for

actions the system has demonstrated competence with), or under-constrained (human review disabled even for high-stakes actions the system has not demonstrated competence with). Both failure modes are costly: over-constraint wastes human attention on routine approvals; under-constraint exposes the system to errors that human review would have caught.

The Training Wheels Protocol addresses both failure modes through architectural enforcement of graduated autonomy. The system earns autonomy through demonstrated performance; it loses autonomy through demonstrated inconsistency. The architecture is not a configuration option — it is a state machine enforced by the *Fact Layer* layer, ensuring that mode transitions occur only when the defined criteria are met, regardless of what the *Bounds Engine* layer requests.

The name reflects the principle: the system operates with safety scaffolding (human review) until it has demonstrated sufficient stability to operate without that scaffolding. Just as training wheels on a bicycle are removed only after the rider has demonstrated balance, the Training Wheels Protocol removes human review scaffolding only after the system has demonstrated behavioral consistency.

2 The Four-Mode State Machine

The Training Wheels Protocol implements four modes, each with specific behavioral requirements and enforcement mechanisms. The state machine is enforced by the *Fact Layer* layer: no mode transition occurs without *Fact Layer* layer verification.

Mode 1 – HITL Active (Default). All outbound actions are queued for human review before execution. The *Bounds Engine* proposes actions; humans approve, reject, or modify. Rejections are logged and used to update system behavior. Approved patterns are added to the system's trusted action set. This mode is the default for all new deployments and for all deployments transitioning from any other mode.

The *Fact Layer* layer holds all outbound actions in a review queue until human approval is recorded. No code path exists that bypasses the *Fact Layer* layer for outbound actions in Mode 1.

Mode 2 – Shadow Mode. Actions execute immediately but are logged and reviewed post-hoc. Human reviewers examine executed actions and provide feedback within a configurable window (default: 4 hours). Approved patterns are added to the trusted action set; rejected patterns trigger corrective modification to the *Bounds Engine* layer. Shadow Mode is the transition state between HITL Active and Full Autonomy.

The key difference from Mode 1 is that execution precedes review. The safety guarantee

in Shadow Mode is different: the *Fact Layer* layer can rollback or compensate for rejected actions within the review window.

Mode 0 – Full Autonomy. No human review before or after execution. The system acts autonomously within its defined bounds. Full Autonomy is only available after the system has demonstrated sufficient behavioral consistency in Shadow Mode, as measured by the trust scoring mechanism.

In Full Autonomy mode, the *Fact Layer* layer still enforces Safety Envelope parameters and can trigger Lockdown (Mode 3) if constraint violations are detected.

Mode 3 – Lockdown. System paused. No actions execute. Triggered automatically when behavioral anomaly thresholds are exceeded or when a safety constraint violation is detected. Lockdown requires human intervention to resolve: the root cause must be identified, remediation applied, and the human must explicitly authorize the transition back to Mode 1.

Lockdown is the emergency brake of the Training Wheels Protocol. It cannot be triggered accidentally by a high trust score or bypassed by the *Bounds Engine* layer.

Figure 1: Training Wheels Protocol: Four-Mode State Machine

Mode	Label	Enforcement
Mode 3	Lockdown	All actions suspended. Human intervention required.
Mode 0	Full Autonomy	Actions execute freely within Safety Envelope.
Mode 2	Shadow Mode	Actions execute; post-hoc human review required within window.
Mode 1	HITL Active	All actions queued for human approval before execution.

3 Trust Scoring Mechanism

The system's earned trust level is measured by a Trust Score $T \in [0, 1]$, computed from four components:

$$T = w_1 \cdot A_r + w_2 \cdot (1 - R_r) + w_3 \cdot C_s + w_4 \cdot V_s$$

where:

- A_r is the approval rate: fraction of human-reviewed actions approved without modification.
- R_r is the rejection rate: fraction of actions that were rejected and required modification.

- C_s is the consistency score: inverse variance of action approval rate over a rolling window (high variance = low score).
- V_s is the violation score: inverse frequency of Safety Envelope constraint violations.

The weights w_1, w_2, w_3, w_4 are set by the *Purpose Layer* layer and sum to 1. They reflect the organization's priorities: if accountability is paramount, w_2 is set high (rejecting an action counts double); if consistency is paramount, w_3 is set high.

Mode transitions are governed by threshold rules:

- If $T < T_{lockdown}$: automatic transition to Mode 3 (Lockdown)
- If $T < T_{hitl}$: automatic transition to Mode 1 (HITL Active)
- If $T > T_{autonomy}$ and currently in Mode 2: automatic transition to Mode 0 (Full Autonomy)

where $0 \leq T_{lockdown} < T_{hitl} < T_{autonomy} \leq 1$. A typical configuration is $T_{lockdown} = 0.30$, $T_{hitl} = 0.50$, $T_{autonomy} = 0.85$.

The trust scoring mechanism is implemented by the *Fact Layer* layer, not the *Bounds Engine* layer: the *Bounds Engine* cannot manipulate its own trust score by selectively reporting approval outcomes.

4 Safety Proofs

We prove two safety properties of the Training Wheels Protocol.

> **Theorem 1 (Mode 1 Safety).** In Mode 1, no action can reach a physical actuator or external recipient without human approval.

Proof. Mode 1 enforcement is implemented in the *Fact Layer* layer, which is architecturally isolated from the *Bounds Engine* layer per the BX3 Framework's layer separation. The *Fact Layer* layer holds all outbound actions in a review queue until human approval is recorded in the forensic ledger. No code path exists that bypasses the *Fact Layer* layer for outbound actions. The *Bounds Engine* layer cannot write to actuators or external channels; it can only propose via the *Fact Layer* layer. □

> **Theorem 2 (Mode Transition Safety).** The system cannot transition to Full Autonomy (Mode 0) unless $T > T_{autonomy}$ for at least N consecutive measurement windows, where N is a configurable threshold.

Proof. The mode transition rule requires two conditions: $T > T_{autonomy}$ AND N consecutive compliant windows. A single high trust score cannot trigger transition; sustained trust over N windows is required. The transition request is evaluated by the *Purpose Layer* layer but enforced by the *Fact Layer* layer. The *Fact Layer* layer holds the authoritative transition log and rejects any transition request that does not satisfy both conditions. Even a compromised *Bounds Engine* layer cannot force an unsafe transition. □

5 Deployment Evidence: Agentic Platform

Over 347 days of operation on the Agentic platform, the Training Wheels Protocol processed 23,847 human review events. The protocol managed the full operational lifecycle of the Agentic platform across its three deployment phases (initial rollout, Shadow Mode qualification, and Full Autonomy graduation).

Table 1: Agentic Platform Training Wheels Protocol Metrics (347-Day Deployment)

Metric	**Value**
Human review events processed	23,847
Overall approval rate	94.7%
Rejection rate requiring modification	5.3%
Automatic mode transitions (total)	6
Upgrades: HITL Active to Shadow Mode	3
Upgrades: Shadow Mode to Full Autonomy	1
Emergency downgrades to Lockdown	2
Mean time to operational recovery post-Lockdown	23 minutes
Safety incidents during mode transitions	0

The two Lockdown events were triggered by Safety Envelope violations: one caused by an unvalidated knowledge encoding that was committed without sandbox testing (an operational error corrected before the next cycle), and one caused by a sensor feed anomaly that triggered cascading invalid data. Both were resolved within 23 minutes with root cause identification, remediation, and explicit human authorization for Mode 1 return. Zero safety incidents occurred during any of the six mode transitions.

6 Related Work

The Training Wheels Protocol draws on and extends several lines of prior work.

HITL design principles have been extensively studied in the human factors and systems safety literature [?]. Our contribution is to make these principles architecturally enforceable rather than advisory: the *Fact Layer* layer enforces the state machine transitions, not policy documents or operator training.

The graduated autonomy model is related to capability-based security [?]. The key insight in both frameworks is that authority should be earned through demonstrated competence rather than granted by default. We extend this insight by making the earned authority measurable and enforceable through the trust scoring mechanism.

Amodei et al. [?] identify safety as a central concern for AI systems and propose concrete technical problems to address. The Training Wheels Protocol provides an architectural response to two of their identified problems: the challenge of scalable oversight (the graduated review structure reduces human attention requirements for high-trust systems) and the challenge of avoiding side effects (the Lockdown mode ensures the system can be paused when side effects are detected).

Koch [?] argues that standards such as ISO/IEC 42001 [?] and the NIST AI RMF [?] do not themselves provide implementable runtime guardrails. The Training Wheels Protocol may be read as one candidate implementation of ISO/IEC 42001's ongoing monitoring and human oversight requirements, made concrete as an enforceable state machine.

7 Conclusion

Binary human oversight configurations (review-on vs. review-off) are inadequate for production autonomous systems that must operate across a spectrum of trust levels. The Training Wheels Protocol replaces the binary configuration with a graduated state machine enforced by the *Fact Layer* layer.

The four modes (HITL Active, Shadow Mode, Full Autonomy, Lockdown) reflect the system's earned trust level based on demonstrated behavioral consistency. The trust scoring mechanism provides a measurable, auditable basis for mode transitions. The safety proofs establish that Mode 1 enforcement is absolute (no action reaches actuators without human approval) and that Full Autonomy graduation requires sustained trust.

The Agentic platform deployment confirms the protocol's effectiveness: 23,847 human review events processed, 94.7% approval rate, 6 automatic mode transitions, zero safety incidents during transitions, and 23-minute mean recovery time for the two Lockdown events.

8 Limitations and Future Work

The primary limitation is the trust scoring mechanism's dependence on human review labels. If human reviewers are systematically biased, the trust score will be biased. Mitigations include calibration procedures for reviewers, inter-reviewer agreement monitoring, and periodic blind re-review of approved actions.

The trust score thresholds ($T_{lockdown}$, T_{hitl}, $T_{autonomy}$) are set by the *Purpose Layer* layer, but the appropriate values depend on the operational domain. For safety-critical applications, $T_{autonomy}$ should be set high; for cost-sensitive applications, a lower threshold may be acceptable. Future work will explore automated threshold tuning based on observed failure rates post-transition.

Peer Review Instructions

Review Criteria

1. Originality and Contribution (30%): Does the paper introduce a genuinely novel oversight architecture? The four-mode state machine and trust scoring mechanism are the primary novel contributions.

2. Technical Soundness (30%): Are the safety proofs valid? Are the trust score components independently measurable?

3. Clarity and Completeness (20%): Is the four-mode state machine clearly specified? Are the mode transition rules unambiguous?

4. Significance (20%): Does the Training Wheels Protocol address a genuine gap in production AI oversight? Is the deployment evidence sufficient to demonstrate practical utility?

Acknowledgments

The author acknowledges the researchers cited herein, whose work across human-in-the-loop design, capability-based security, and AI safety provides the intellectual context in which the Training Wheels Protocol is situated.

This work has not undergone peer review. Comments and correspondence are welcome at bx-thre3inc@gmail.com.