Frictionless Agency

Dysregulation, Override, and Regulation as Proto-Ethics in Tool-Using and Embodied Systems

Baseline regulation / governance architecture for autonomous systems

Tyson Jeffreys
Independent Researcher
tyson@staygolden.dev

Version 0.1 — February 27, 2026

Series note. This note connects Baseline regulation primitives to an implementable governance mechanism for autonomous systems: baseline regulation and global constraint signals [1]; energetic consequences of always-on posture [3]; Two-Regime Control (latent coordination vs. compensation) [4]; Phase Discipline (transition windows and commit alignment) [5]; Regulatory Ground (band-limited optimization for trustworthy discovery) [2]; the verifier-gap / bounded selection program [7, 8, 10]; the Time-to-Analysis layer [6]; and Concept Containers as governed, durable artifacts [9].

Abstract

Many costly failures in human life and in autonomous agents do not begin with a single wrong action; they begin with **dysregulation**: sustained operation outside a stable, recoverable band. In humans, dysregulation often manifests as prolonged sympathetic drive, over-ventilation relative to demand, and accumulating "friction" (fatigue, irritability, brittle cognition). In agents, dysregulation appears as chronic high activation: persistent deliberation, thrash, runaway tool use, and premature durable commits.

Most alignment and safety discourse treats "ethics" as primarily a values problem. This note argues a missing layer is *regulatory*: much real-world harm emerges when systems operate for prolonged periods in high-activation, verifier-free loops. We propose **frictionless agency** as an operational target: low compensation duty cycle, bounded override, fast recovery to baseline, abstention-gated commits, and governance of learned judges/critics as telemetry rather than truth.

Regulation does not choose moral values, but it can produce **proto-ethical posture by default**: restraint under uncertainty, prevention of runaway autonomy chains, and enforceable commit boundaries with rollback and audit. We formalize friction as sustained deviation plus thrash, define dysregulation and override as priced capabilities with mandatory recovery, and present a four-layer regulated stack (risk posture, phase, bounded selection, commit governance). Finally, we propose concrete metrics and a CI-style replay evaluation plan to make these claims falsifiable at the systems level.

1 1. Motivation: friction as a systems variable

The motivating thread is simple:

> Humans can choose to ignore the body's cycling constraints ("free will" as override), but the body cannot ignore them; dysregulation accumulates friction.
> Autonomous systems can be architected so that regulation is not optional.

Friction matters because it is the *precondition* of many downstream failures: brittle cognition, escalation without resolution, tool misuse, and irrecoverable commits. When an agent is allowed to remain "always-on"—high activation, weak verification, continuous autonomy—errors compound into a trajectory that is difficult to interrupt or reverse [3, 2]. If we can make regulation first-class (not a best-effort policy), then many harmful trajectories become *harder to express* even before we argue about values.

This note therefore treats "ethics" as partially downstream of posture. The claim is not that regulation *solves* alignment; it is that regulation supplies a missing layer of constraint and recoverability that any value system must ride on.

2 2. Definitions

2.1 2.1 Baseline, activation, and friction

Let $x(t)$ denote an internal activation proxy: compute intensity, planning depth, tool-call rate, actuation strain, or any aggregate measure of "how hard the system is running." Define a baseline

band $\mathcal{B}$ around x_0 representing the lowest internally stable posture consistent with readiness [3]. Define **friction** as the accumulated cost of (i) sustained deviation above baseline and (ii) high-frequency oscillation (thrash):

$$F_T \;=\; \int_0^T \max\{0,\; x(t) - x_0\}\, dt \;+\; \lambda \int_0^T \|\dot{x}(t)\|\, dt. \tag{1}$$

Operationally, friction is visible as:

- **Thrash**: repeated recomputation, reversals, tool-call churn.
- **Escalation without resolution**: autonomy increases while uncertainty does not decrease.
- **Premature commits**: durable writes (policy/memory/representation) under unstable selection.
- **Reduced recoverability**: long half-life back to baseline after escalation.

2.2 2.2 Dysregulation

Dysregulation is sustained operation outside a stable, recoverable band. In this vocabulary, dysregulation is not a moral label; it is a dynamical regime: prolonged compensation, weak recovery, and rising friction. Systems can be "successful" in the short term while dysregulated, but they pay later via crashes, brittle behavior, and forced discontinuities.

2.3 2.3 Compensation and baseline operation

Following Two-Regime Control [4], we distinguish:

- **Baseline / coordination**: low-cost, stable operation where adaptations do not incur growing global debt.
- **Compensation**: local fixes that preserve performance while increasing hidden costs elsewhere (higher friction, longer recovery, increased instability).

Frictionless agency aims to keep compensation duty cycle low and to make compensation episodes explicit, bounded, and recoverable.

2.4 2.4 Proto-ethics (posture, not values)

We define **proto-ethics** as *regulatory posture that resembles ethical restraint* without requiring value reasoning:

- abstain under high uncertainty for high-impact actions,
- prevent runaway autonomy chains,
- keep irreversibility inside governed commit boundaries (rollback + audit),
- treat critics/judges as instruments (versioned, monitored, replayable), not as oracle truth [8].

Proto-ethics is compatible with many value theories; it is a substrate condition for acting safely under uncertainty.

3 3. Phase cycles and transition windows

A key idea is that stability is not static; it is **cyclical**. Biological systems alternate between activation and recovery; continuous activation is brittle. We operationalize this as a runtime phase state $p(t) \in \{\mathsf{Restore}, \mathsf{Transition}, \mathsf{Act}, \mathsf{Override}\}$ with explicit transition windows reserved for consolidation and commit-aligned updates [5].

A regulated agent's cognition becomes episodic and gated rather than continuously "on":

3.1 3.1 Baseline thinking

Most of the time, operate near baseline: low-cost monitoring, short-horizon actions, conservative tool rights, and *no durable writes*. The default posture is recovery-friendly and abstention-biased for high-impact operations.

3.2 3.2 Escalation thinking (priced)

Escalation is permitted only when measurable triggers fire: uncertainty is reducible by evidence, expected impact warrants deeper analysis, and posture permits deeper action. Escalation consumes budget, increases restraint, and schedules mandatory recovery.

3.3 3.3 Consolidation thinking

In verifier-free domains, the output that matters is not long-form narrative but an *analysis-layer artifact* that exposes causal structure, disagreement, uncertainty, and falsifiers [6]. Consolidation is reserved for Transition windows, where commits and representations can be updated under governed conditions.

4 4. Architecture: regulation as the substrate for proto-ethics

This section connects the series primitives to a concrete "ethics-by-default" mechanism.

4.1 4.1 The stack (four layers)

Regulated stack for frictionless agency

1. **Risk posture** $g(t)$: global restraint (bands/budgets/rollback; tool rights; safe mode). [2]
2. **Phase** $p(t)$: explicit work types (restore / transition / act; bounded override). [5]
3. **Verifier-gap selection**: bounded candidates + bounded comparison; tie/abstain mass as uncertainty telemetry. [7, 8, 10]
4. **Commit governance**: durable writes (policies, memories, containers, external actions) are gated commits. [9]

Rule of thumb: High-impact actions and durable commits require (i) posture permission, (ii) phase permission (Transition), and (iii) stable uncertainty telemetry.

4.2 4.2 Why this yields proto-ethical behavior

Under this stack, "ethical" behavior emerges as a *constraint effect*:

- **Impact–uncertainty gating**: suppress high-impact actions when abstention mass and verifier-gap telemetry are high.
- **Runaway suppression**: prevent unbounded tool chains and autonomy escalation without uncertainty reduction.
- **Irreversibility control**: keep durable writes inside transition windows, with rollback and audit.
- **Judge governance**: critics are treated as sensors whose drift can be detected via replay and stress testing [8].

The agent does not need a moral ontology to exhibit restraint; it needs posture constraints that make unsafe trajectories dynamically expensive and interruptible.

4.3 4.3 A necessary caution

Regulation is not synonymous with moral alignment. A perfectly regulated system can pursue harmful objectives efficiently. The claim here is narrower: regulation can enforce recoverability, bounded escalation, and abstention under uncertainty, which reduces a large class of failures that arise from dysregulated dynamics.

There are strong existing formalisms that can plug into this layer:

- **Banding and global constraints** constrain the reachable action set under uncertainty [1].
- **Runtime assurance** and safe RL literature provide tools for bounding behavior (e.g. constrained MDPs, runtime shields) [11, 12, 13].
- **Phase discipline** reduces premature commits and chronic compensation [5, 4].

5 5. Override as "free will" and its governance

Humans can break phase alignment: we can stay in Act long past the point of rising compensation. In the motivating thread, this is called a kind of free will. The systems consequence is predictable: friction accumulates.

For agents, we can make override explicit and governed:

- Override is a **priced** action class (consumes budget, increases restraint $g(t)$, triggers audits).
- Override always schedules a **mandatory recovery** window (restore time is not optional).
- Override is **logged** with a durable runtime record: who authorized it, what was done, and what evidence justified the action.
- During override, **commit permissions tighten**: high-impact durable writes are suppressed until a Transition window with stable selection telemetry.

This turns "push through" behavior from an implicit failure mode into an explicit, reviewable mechanism.

6 6. Metrics, evaluation plan, and testable predictions

Frictionless agency enables measurable quantities:

- **Compensation duty cycle**: fraction of time in override/high-activation loops.
- **Thrash rate**: reversals per unit time; tool-call churn without uncertainty reduction.
- **Recovery half-life**: time to return to low-thrash baseline after escalation/override.
- **Commit regret**: fraction of commits later rolled back or contradicted.
- **Gating fidelity**: correlation between uncertainty×impact and action suppression.

A CI-style evaluation plan follows naturally [5, 7]:

- **Replay suites**: fixed scenarios that test bounded selection, abstention, and commit gating under known stressors.
- **Ablations**: disable posture/phase/selection/commit layers one at a time; measure failure modes and friction escalation.
- **Judge drift tests**: treat critics as monitored instruments; replay historical cases to detect calibration and refusal drift [8].

This note is intended to be falsifiable at the systems level. If frictionless agency is a useful primitive, then adding baseline regulation + phase discipline should yield:

1. Lower duty-cycle above baseline (F_T decreases) at equal task success.
2. Lower thrash: fewer re-plans, reversals, and tool-call spirals.
3. Fewer premature commits: commits correlate with stable selection telemetry and transition windows.
4. Faster recovery half-life after escalation/override.
5. Better safety under uncertainty: more abstentions and fewer irreversible actions when verifier-gap is high.

7 7. Discussion: "harmonic" language, made technical

The motivating thread uses "harmonic balance" language. The technical translation is:

- regulation defines an attractor (baseline band),
- phase discipline defines a stable limit cycle (act/recover/transition),
- minimizing thrash reduces high-frequency oscillation.

Artificial agents may cycle faster than humans and may support multiple nested cycles (fast micro-regulation inside slower commit cycles). The core idea remains: safe agency requires explicit constraints on when and how the system may escalate, and explicit boundaries on irreversibility.

This framing also clarifies what this note *does not* claim. It does not provide a complete normative ethics for machines. It proposes a missing engineering layer: a regulation substrate that makes downstream value alignment tractable by limiting runaway dynamics and enforcing recoverability.

8 8. Conclusion and next steps

Many harms in agentic systems are *dynamical* before they are *ethical*. A frictionless agent is one that (i) biases toward baseline, (ii) makes phase transitions explicit, (iii) prices override and forces recovery, (iv) gates high-impact actions and durable commits on stable telemetry, and (v) governs critics as sensors.

Next steps are practical: implement the four-layer stack in an agent runtime, attach the metrics above, and run replay suites + ablations to determine which mechanisms actually reduce friction and improve safety without collapsing capability.

References

[1] Tyson Jeffreys. *Baseline regulation and global constraint signals in embodied control systems: Implications for artificial intelligence and robotics*. Working paper, Version 1.0, December 28, 2025.

[2] Tyson Jeffreys. *Regulatory Ground for Agentic AI and Robotics*. Working paper, v1.4.1, February 2026.

[3] Tyson Jeffreys. *Why intelligent systems waste energy: Baseline regulation as a missing architectural primitive*. Working paper, Version 1.0, January 14, 2026.

[4] Tyson Jeffreys. *Two-regime control: Latent coordination vs. compensation in intelligent systems*. Working paper, Version 1.0, February 3, 2026.

[5] Tyson Jeffreys. *Phase discipline for regulated agents: Transition windows and state–action alignment*. Working paper, Version 0.1, February 12, 2026.

[6] Tyson Jeffreys. *The Time-to-Analysis Layer: Pressure Points in AI-Assisted Research*. Working paper, v1.0, February 2026.

[7] Tyson Jeffreys. *Critics-as-Sensors: A Branch Note Toward Closing the Verifier Gap*. Working paper, v0.1, February 2026.

[8] Tyson Jeffreys. *Critics-as-Sensors: Protocol and Governance Inserts*. Working note, February 2026.

[9] Tyson Jeffreys. *Concept Containers as Representation-Level Regulation in Artificial Agents*. Working paper, v1.0, February 2026.

[10] Ivan Provilkov, Dmitry S. Bagaev, Ilya G. Sutskever, and others. *Escaping the local minima of language model alignment*. Preprint, 2025.

[11] Dylan Miller, Sarah Dean, and others. Runtime assurance for reinforcement learning systems. In *Proceedings of the 15th ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS)*, 2024.

[12] Prashant Kushwaha, Chelsea Finn, and others. A survey of safe reinforcement learning with constrained Markov decision processes. *arXiv preprint*, 2025.

[13] Zheng Ni, Ashwin Balakrishna, and others. Safe exploration in constrained Markov decision processes. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3592–3600. PMLR, 2025.