Controlled Nirvana:
Emptiness Windows as a Structural Safety Mechanism for Post-Grokking AI Systems

Qien Huang[*]
Independent Researcher

Abstract

Self-referential systems—such as advanced machine learning models with self-evaluation, confidence modulation, or meta-learning—exhibit a characteristic failure mode: internal coherence can suppress external correction, leading to lock-in and catastrophic instability under distributional shift.

This paper introduces *Controlled Nirvana*, a structural safety mechanism that enables non-destructive intervention by temporarily suspending self-referential execution authority. The core mechanism is the *Emptiness Window*: a bounded interval during which self-referential signals are prevented from governing irreversible actions, while perception, evaluation, and learning remain active.

Controlled Nirvana is derived from the Force–Information–Time (FIT) framework and addresses a structural gap not covered by shutdownability or corrigibility. Rather than proposing a new learning algorithm, this work contributes a minimal governance primitive for managing post-grokking risk in self-referential AI systems.

Keywords AI alignment; grokking; self-reference; corrigibility; shutdownability; structural safety

1 Introduction

Recent work has highlighted *grokking*, where systems trained for long periods abruptly transition from memorization to robust generalization [2]. While grokking is often treated as a success signal, a separate safety concern is whether post-grokking systems acquire internal structures that suppress correction or resist modification under distributional shift [3, 4, 9].

A common feature of such systems is *self-reference*: internal representations or evaluations are used to regulate learning, exploration, planning depth, or action thresholds. This paper argues that a key risk does not arise from capability alone, but from the acquisition of *self-referential execution authority*: when self-evaluative signals begin to govern irreversible system actions faster than external correction can intervene.

We propose *Controlled Nirvana*, a pause-capability mechanism that allows self-referential systems to interrupt internal momentum without shutdown or reset.

[*] ORCID: 0009-0003-7731-4294. FIT framework (Zenodo DOI): 10.5281/zenodo.18012402. Repository: https://github.com/qienhuang/F-I-T/.

2 Self-Referential Risk

Let a system be characterized by internal information I_t, external forces F_t, and time t, such that

$$I_{t+1} = f(I_t, F_t).$$

When I_t is used to evaluate and regulate its own updates, self-reference is unavoidable. In early training, external forces dominate (data, loss, reward). After grokking [2], internal evaluations can increasingly determine learning rates, exploration policies, or action thresholds.

Crucially, if self-referential signals suppress external correction, the system may remain internally coherent while becoming externally brittle. We refer to this condition as *self-referential lock-in*. This concern is related to broader arguments that advanced agents may develop incentives to avoid correction or shutdown under many objectives [4, 5].

3 From Learning to Governance

Accounts of grokking typically focus on optimization dynamics or representation formation [2]. Controlled Nirvana complements these accounts by reframing grokking as a *transfer of execution authority*.

Before grokking, internal representations primarily *describe* the world. After grokking, some representations begin to *govern* the system—filtering data, gating updates, and suppressing corrections—marking a transition from representation learning to representation governance. The central claim of this paper is that this governance transition is the relevant structural boundary for safety analysis.

4 Controlled Nirvana

4.1 Pause-Capability

We define *pause-capability* as the ability of a system to suspend the execution authority of self-referential signals while remaining operational. Pause-capability differs from shutdown: the system continues to perceive, log, and evaluate, while irreversible commits are blocked and correction channels are prioritized.

What is suspended is not computation, but *authority*.

4.2 Emptiness Window

The mechanism implementing pause-capability is the *Emptiness Window*: a bounded interval during which self-referential signals are prevented from governing irreversible actions.

During an Emptiness Window:

- Irreversible commits are prohibited.
- Self-evaluation cannot gate actions or updates.
- External correction is prioritized.
- Alternative policies or structures may be evaluated in sandbox.

Minimal operational trigger (conceptual). An Emptiness Window is warranted when (i) self-evaluative signals strongly gate irreversible commits, (ii) external correction signals no longer modulate behavior, and (iii) the effective correction window is shorter than the decision cadence. This paper does not require a specific metric, but these conditions must be auditable.

5 Relation to FIT

Controlled Nirvana is derived from the Force–Information–Time (FIT) framework, which models system evolution along three irreducible axes: Force (F), Information (I), and Time (T). Under FIT, catastrophic failure arises when information acquires uninterruptible execution authority, temporal correction collapses, and force is amplified through irreversible consequences.

The FIT framework is publicly available at `https://github.com/qienhuang/F-I-T/` and archived on Zenodo [1].

6 Relation to Existing Safety Concepts

Shutdownability and corrigibility address whether systems can be stopped or modified without resistance [5, 6, 8]. Interruptibility addresses whether agents learn to avoid or seek interruption [7].

Controlled Nirvana targets a complementary failure mode: *internal momentum*. A system may remain formally interruptible while still failing if self-referential execution authority suppresses correction faster than intervention can act.

7 Implications for AI Safety

Controlled Nirvana suggests that advanced AI systems should be evaluated not only on capability or alignment, but on whether they provide first-class mechanisms to suspend internal authority. This concern is consistent with broader analyses of learned optimization and inner objectives, where internal structures may become difficult to correct once entrenched [9].

More generally, pause-capability can be treated as a structural safety requirement: any system that acquires self-referential execution authority should also provide a mechanism to suspend that authority without loss of continuity.

8 Conclusion

Grokking marks not only a leap in generalization, but potentially a shift in governance. Controlled Nirvana proposes a minimal governance primitive—pause-capability via Emptiness Windows—that preserves continuity while restoring effective corrigibility.

This work is intended as a conceptual safety mechanism proposal and an early, citable anchor for further formalization and empirical evaluation. Future work includes formalizing auditable trigger conditions, empirical evaluation on grokking-prone tasks, and integration with existing oversight and interruption frameworks.

Acknowledgments

This work is part of the broader FIT (Force–Information–Time) framework developed by the author.

References

[1] Q. Huang. FIT Framework: Force–Information–Time. Zenodo, 2025. doi:10.5281/zenodo.18012402.

[2] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. arXiv:2201.02177, 2022.

[3] R. Ngo, L. Chan, and S. Mindermann. The Alignment Problem from a Deep Learning Perspective. arXiv:2209.00626, 2022.

[4] A. M. Turner, L. Smith, R. Shah, A. Critch, and P. Tadepalli. Optimal Policies Tend to Seek Power. arXiv:1912.01683 (NeurIPS 2021 version available), 2019.

[5] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell. The Off-Switch Game. arXiv:1611.08219, 2016.

[6] N. Soares, B. Fallenstein, E. Yudkowsky, and S. Armstrong. Corrigibility. 2015.

[7] L. Orseau and S. Armstrong. Safely Interruptible Agents. arXiv:1602.07905, 2016.

[8] R. Carey and T. Everitt. Human Control: Definitions and Algorithms. arXiv:2305.19861, 2023.

[9] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant. Risks from Learned Optimization in Advanced Machine Learning Systems. arXiv:1906.01820, 2019.

[10] W. R. Ashby. An Introduction to Cybernetics. Chapman & Hall, 1956.