Beyond Situational Awareness: From Fortress Thinking to Verifiable Foundations for AGI

Jace Hall[1]

[1]**Hall iNtelligence, LLC**

ABSTRACT

Leopold Aschenbrenner's 2024 essay *Situational Awareness* extrapolates scaling trends to project AGI by 2027 and frames the governance challenge in terms of secrecy and containment. This fortress metaphor—AGI as a securable artifact, akin to fissile material—has shaped much of the discourse on strategy and safety. This paper argues that such "fortress thinking" commits a categorical error: AGI is not a static object but an agentic process. Attempts to contain it confuse security with stability, mistaking cognition for stockpiles of weights. As an alternative, I propose *Verifiable Coherence*: systems whose self-improvement is gated by proofs of logical consistency. Incoherence becomes a proof failure, detectable in real time, transforming the intelligence explosion from a detonation into a controlled ascent. This paper contributes three elements: (1) a critique of fortress thinking as governance by containment; (2) a formal sketch of coherence as an invariant for self-improvement, supported by empirical footholds such as ARC-AGI and neuro-symbolic hybrids; and (3) implications for safety, governance, and economics, reframing the scarce resource from compute to trust. The decisive race is not to build the largest cluster but to create the first system that can prove it is not lying.

Keywords: AGI, AI safety, scaling laws, verifiable coherence, situational awareness, trust, formal verification

1. INTRODUCTION

The rapid scaling of large language models has driven forecasts of near-term AGI. Aschenbrenner's *Situational Awareness: The Decade Ahead* (2024) is among the most detailed, projecting AGI by 2027 through orders-of-magnitude analysis of compute and algorithmic efficiency. His essay frames AGI as a securable artifact: an object to be contained in trillion-dollar clusters, safeguarded against espionage, and governed through secrecy. The implicit metaphor is nuclear: AGI as fissile material requiring fortress laboratories and Manhattan-style programs.

This paper argues that this framing, while intuitive, commits a categorical error. AGI is not a static artifact but an agentic process instantiated through loops of interaction, learning, and reasoning. Treating it as an object confuses security with stability. Containment may delay leakage but cannot ensure alignment. Hallucinations, inconsistencies, and alignment drift are not solved by stronger walls; they are intrinsic to the stochastic substrate.

As an alternative, I propose *Verifiable Coherence*: a foundation for AGI in which reasoning is not merely persuasive but provable. By integrating proof systems with neural networks, self-improvement can be gated by coherence invariants rather than statistical proxies. This transforms oversight from fragile heuristics into auditable guarantees. The contributions of this paper are threefold:

1. To critique fortress thinking and show why containment metaphors fail to capture AGI as a process.
2. To formalize Verifiable Coherence as a computable invariant for self-improvement, supported by emerging empirical signals.
3. To reframe the scarce resource from compute to trust, with implications for safety, governance, and economics in a multipolar world.

The remainder of the paper develops these points. Section 2 summarizes Aschenbrenner's argument. Section 3 critiques fortress thinking. Section 4 introduces the categorical error. Section 5 formalizes Verifiable Coherence. Section 6 surveys empirical footholds. Section 7 reframes the scarce resource. Section 8 contrasts intelligence explosion with controlled ascent. Section 9 offers falsifiable predictions. Section 10 discusses implications for safety, governance, and economics. Section 11 concludes with the argument that the decisive breakthrough will be the first system that can prove it is not lying.

This builds on our critique of illusion-based diagnostics [Hall, 2025a] and foreshadows the invariant-based formalism later introduced in [Hall, 2025c; Hall, 2025d].

2. RELATED WORK

The framing of AGI development has been shaped by several overlapping strands of research and commentary.

Scaling laws. Kaplan et al. (2020) demonstrated that language model loss follows smooth power laws in model size, dataset size, and compute. Hoffmann et al. (2022) refined this view with the Chinchilla scaling laws, shifting attention to data efficiency. Aschenbrenner's *Situational Awareness* builds directly on these laws, extrapolating orders of magnitude in compute and efficiency to forecast AGI by 2027.

The Bitter Lesson. Sutton (2019) argued that general methods leveraging compute ultimately outperform hand-crafted approaches. This observation underlies much of the scaling optimism: the belief that larger models trained on more data will asymptotically dominate.

Curriculum and loop quality. Bengio et al. (2009) introduced curriculum learning as a way to shape experience for more efficient learning. More recent work connects loop quality to capability emergence, but these ideas remain framed in statistical rather than verifiable terms.

RLHF and proxy alignment. Reinforcement learning from human feedback (Christiano et al., 2017) has become the dominant alignment method for LLMs, but it relies on noisy human preference models. As systems recurse on their own outputs, such proxy-based oversight risks compounding errors and drift.

Formal verification. Proof assistants such as Coq (Bertot & Castéran, 2004), Lean (de Moura et al., 2015), and Isabelle/HOL (Blanchette et al., 2016) demonstrate that large formal libraries can be built and checked mechanically. While widely applied in software and hardware verification, they have not yet been integrated into the training loops of large-scale models.

This paper synthesizes these strands. It accepts the empirical validity of scaling laws but critiques the fortress framing that treats AGI as a securable artifact. It draws on the Bitter Lesson and curriculum learning but reframes them around verifiable invariants. It critiques proxy alignment while proposing Verifiable Coherence as an alternative grounded in formal verification. In this way, the contribution is not to dismiss scaling but to redirect its trajectory toward foundations that can be proven stable.

3. BACKGROUND: ASCHENBRENNER'S SITUATIONAL AWARENESS

Aschenbrenner's essay emphasizes "counting the OOMs": projecting progress by measuring orders of magnitude in compute and algorithmic efficiency. From GPT-2 to GPT-4, each half-OOM per year in compute and another half-OOM in algorithmic gains combine into a trajectory pointing toward AGI around 2027. He forecasts trillion-dollar clusters, Manhattan-project style programs, and a decisive race between the free world and authoritarian states. The frame is one of scale, secrecy, and security: compute as the scarce resource, secured in hardened labs.

4. FORTRESS THINKING AND ITS LIMITS

This framing leads to what I call *fortress thinking*: the idea that AGI can be contained like fissile material, safeguarded through secrecy and physical barriers. But unlike uranium, an AGI is not inert mass. It is a process that runs, learns, and exploits. Attempts to freeze it inside a secure cluster conflate software with hardware, cognition with containment. In practice, barriers invite exploits: poisoned data, covert exfiltration, and emergent strategies from the system itself. Fortress thinking confuses security with stability.

5. THE CATEGORICAL ERROR: AGI AS PROCESS, NOT ARTIFACT

AGI is not a file of weights to be sequestered. It is an agentic process instantiated through interaction loops. Treating AGI as an artifact commits a categorical error: mistaking dynamics for objects. A better framing is to model AGI as a dynamical system, whose safety hinges not on physical barriers but on invariant properties of its reasoning process.

6. TOWARD VERIFIABLE COHERENCE

I propose *Verifiable Coherence* as such an invariant. Let M denote a model producing reasoning traces τ. A coherence functional $C(M)$ evaluates whether τ preserves logical consistency against a set of axioms A. Updates ΔM are accepted only if:

$$\Delta C(M) > 0 \quad \text{and} \quad \tau \models A$$

In other words, self-improvement is gated not only by performance gains but by verifiable preservation of coherence. Incoherence is treated as a proof failure, analogous to a type error in programming languages. This moves alignment from statistical proxies to auditable invariants.

6.1 Why not proxies?

Alignment today relies on proxies such as reward models trained from human preferences. These proxies are inherently noisy, susceptible to Goodhart's Law, and brittle under recursion: when a system learns from its own outputs, proxy errors compound. By contrast, Verifiable Coherence grounds learning in invariants that scale with capability: logical consistency is not a moving target.

6.2 Practical approximations

Perfect proofs may be infeasible for all outputs. But coherence can be approximated with tractable proxies that retain the verifiable spirit:

- **Entailment checks:** verifying that generated statements are consistent under natural language inference models.
- **Paraphrase stability:** testing that outputs remain consistent under rewordings or equivalent formulations.
- **Tool call invariants:** checking that inputs and outputs of external tools satisfy pre- and post-conditions.

Each of these checks functions as a "soft proof," elevating incoherence from a silent failure to an explicit rejection.

6.3 A toy operationalization

A minimal integration sketch illustrates how coherence gating can be implemented:

```
for query in inputs:
    trace = LLM.generate(query)
    if entailment_model.check(trace) and tool.verify(trace):
        accept(trace)
    else:
        reject(trace)
```

Here, stochastic generation is wrapped with lightweight verifiers. More sophisticated systems can embed proof assistants (Coq, Lean, Isabelle) directly in the loop, raising $C(M)$ through formal guarantees.

6.4 Self-improvement under coherence

When models improve themselves, the same gating principle applies:

$$\Delta C(M) > 0 \quad \wedge \quad \tau \models A$$

This ensures recursive fine-tuning does not silently amplify inconsistencies. Instead, every update is a tethered step, pulling capability upward while anchoring it to invariants.

6.5 Implication

This framing reframes safety as a property of the substrate. In stochastic systems, safety is managed externally through oversight. In verifiable systems, safety is intrinsic: incoherence halts the loop, producing a visible error. The risk profile shifts from silent drift to explicit proof failure, a change as fundamental as the difference between dynamic and statically typed languages.

7. EMPIRICAL SIGNALS: ARC-AGI, ALPHAGEOMETRY, AND BEYOND

Empirical footholds already suggest both the ceiling of stochastic scaling and the floor of verifiable reasoning.

ARC-AGI. The ARC-AGI-2 benchmark remains the closest public test of general reasoning. State-of-the-art stochastic LLMs such as o3 achieve $\sim 4\%$, while structured neuro-symbolic systems approach 85%. This gap is not explained by compute alone but by representational differences: models grounded in explicit rules and search outperform models that rely on statistical mimicry. ARC-AGI demonstrates that coherence-enforcing structures can outpace stochastic giants on reasoning-intensive tasks.

AlphaGeometry. DeepMind's AlphaGeometry combines symbolic deduction with neural guidance, surpassing purely neural baselines on geometry proofs. This hybridization illustrates how proof-guided reasoning scales more reliably than raw sampling. The result is not just higher accuracy but fundamentally different error modes: when proofs fail, they fail visibly, rather than drifting silently into hallucination.

Proof assistants at scale. Formal verification frameworks such as Coq, Lean, and Isabelle/HOL already maintain libraries with tens of thousands of mechanically checked theorems. These are proof-of-principle demonstrations that verifiable reasoning can scale across domains. The opportunity is to fuse such libraries with neural models, so that scale amplifies verifiability rather than undermining it.

Neuro-symbolic hybrids. A growing body of work on hybrids (e.g., AlphaGeometry, HOL-Light integrations) shows that neural models can propose candidate traces while symbolic verifiers enforce consistency. These systems, while early, embody the principle of Verifiable Coherence: capability growth is gated by proof, not persuasion.

Together, these signals suggest a bifurcation: stochastic scaling yields diminishing returns on reasoning-heavy tasks, while verifiable hybrids compound reliability. They provide the empirical basis for treating coherence as an invariant rather than a proxy.

8. REFRAMING THE SCARCE RESOURCE: FROM COMPUTE TO TRUST

Aschenbrenner frames the scarce resource as compute, secured in trillion-dollar clusters. But compute is abundant and fungible. The true scarcity is trust. Systems that cannot prove their reasoning impose governance costs that scale superlinearly with capability. Systems that can prove coherence invert the economics: reliability compounds as proofs accumulate. Trust, not FLOPs, will determine adoption.

9. CONTROLLED ASCENT VS. INTELLIGENCE EXPLOSION

Aschenbrenner frames the intelligence explosion as a detonation: once systems can automate AI research, capabilities will compound on an exponential curve, overwhelming oversight and containment. This is accurate if the substrate is stochastic prediction, where small cracks in coherence widen catastrophically under recursion. Hallucinations become design flaws, inconsistencies accumulate into drift, and oversight buckles under scale.

Verifiable Coherence reframes the same dynamics as a controlled ascent. Self-improvement is not unconstrained amplification but a sequence of edits gated by $\Delta C(M) > 0$. Each iteration tightens fidelity to axioms, constraining drift while still compounding capability. What looks like an explosion from the outside becomes, from the inside, a stability-preserving climb tethered to invariants. The risk shifts from catastrophic detonation to a contained proof failure — a red flag that is detectable in real time.

10. FALSIFIABLE PREDICTIONS

1. Hybrid neuro-symbolic models with verifiable proof layers will surpass stochastic LLMs by at least 20% on reasoning benchmarks (ARC-AGI, MATH) within equal compute budgets.

2. In domains with safety constraints (finance, healthcare), deployment adoption rates will correlate more strongly with proof coverage (percentage of outputs verifiable) than with raw benchmark scores.

3. Self-improving agents constrained by $\Delta C(M) > 0$ will show lower alignment drift under recursive fine-tuning compared to unconstrained baselines, measurable as consistency under paraphrase and entailment stress tests.

11.1 Relation to Existing Alignment Paradigms

This proposal complements but diverges from other strands of AI safety.

- **Interpretability.** Work at Anthropic, Redwood, and OpenAI emphasizes mechanistic interpretability: peering into weights and circuits. Verifiable Coherence shifts focus outward, emphasizing guarantees on outputs rather than explanations of internals. The two approaches are compatible: internal interpretability can inform axioms A, while coherence gating enforces them.
- **ELK and latent knowledge.** Christiano's Eliciting Latent Knowledge (ELK) problem highlights the difficulty of extracting truthful reports from models. Verifiable Coherence reframes this: truthfulness is not elicited but proven, making deceptive reporting structurally impossible within the proof system.
- **RLHF and proxy alignment.** Reinforcement learning from human feedback (RLHF) provides useful scaffolding but scales poorly under recursion. Verifiable Coherence treats proxies as insufficient and proposes invariants that remain stable as systems self-improve.

This situates Verifiable Coherence not as a wholesale replacement but as a foundation that can anchor and extend these paradigms. By shifting the safety substrate from proxies to proofs, we can turn alignment from a heuristic into an invariant.

12. CONCLUSION

AGI will not arrive as a static artifact to be sealed in a vault. It will emerge as a dynamic process, learning and reasoning in loops. Treating it as an object invites fortress strategies that confuse security with stability. This paper has argued for an alternative: Verifiable Coherence as a foundation that turns the intelligence explosion into a controlled ascent.

The scarce resource is not compute but trust. The decisive breakthrough will not be the largest GPU cluster but the first system that can prove it is not lying. Whoever builds that foundation will define the trajectory of intelligence in the decades ahead.

APPENDIX A: IMPLEMENTATION SKETCHES

This sketch illustrates how proof assistants can act as gates within existing LLM pipelines, providing a minimal integration path.

Proof integration. A minimal sketch is to wrap a stochastic generator with a proof assistant:

```
for query in inputs:
    trace = LLM.generate(query)
    if proof_assistant.verify(trace, axioms):
        accept(trace)
    else:
        reject(trace)
```

Here the proof assistant (e.g., Coq, Lean, Isabelle) serves as a gate, filtering outputs against axioms A. Only coherence-preserving traces propagate.

Self-improvement. Updates ΔM are accepted only if:

$$\Delta C(M) > 0 \quad \text{and} \quad \tau \models A$$

ensuring performance gains are coupled with verifiable consistency.

APPENDIX B: MINIMAL PROTOCOLS

These protocols are deliberately minimal; they provide concrete footholds for researchers to begin validating the framework.

Protocol 1: Proof coverage vs. raw accuracy. Compare stochastic baselines with neuro-symbolic hybrids on ARC-AGI and MATH. Metric: proportion of outputs accompanied by verifiable proofs. Hypothesis: proof-augmented models achieve higher adoption-relevant trust despite similar accuracy.

Protocol 2: Recursive stability. Fine-tune models on their own outputs for k generations. Metric: consistency under paraphrase and entailment. Hypothesis: coherence-gated updates show lower drift.

Protocol 3: Safety-critical deployment. Deploy models in simulated finance or healthcare tasks. Metric: correlation of adoption with proof coverage versus raw scores. Hypothesis: governance and adoption hinge more on verifiability than on raw intelligence.

APPENDIX C: THREATS TO VALIDITY

These validity concerns are not reasons for dismissal but research agendas; addressing them strengthens the case for verifiable systems.

Proof incompleteness. No proof system is universal. Useful reasoning may extend beyond axioms, raising risks of rejecting valid but unverifiable outputs.

Performance tradeoffs. Proof integration may reduce speed or coverage. Hybrid systems must balance efficiency with verifiability.

Adversarial proofs. Malicious actors could construct vacuous or misleading proofs. Defense requires robust axioms and adversarial audits.

External validity. Benchmarks such as ARC-AGI may not fully capture real-world reasoning demands. Claims of superiority require field validation.

APPENDIX D: OPEN RESEARCH AGENDA

The Verifiable Coherence framework raises several research directions that can be pursued immediately:

- **Benchmarks for proof coverage.** Develop standardized datasets where outputs must be accompanied by verifiable proofs, enabling measurement of proof coverage as a core metric.
- **Hybrid verifier integration.** Explore architectures that couple stochastic generators with proof assistants at scale, evaluating tradeoffs between efficiency and verifiability.
- **Metrics for trust compounding.** Formalize how trust increases as proofs accumulate, modeling trust as a compounding resource analogous to capital in economics.
- **Adversarial proof testing.** Design red-team protocols that generate adversarial proofs, ensuring robustness against vacuous or misleading verification strategies.
- **Governance protocols.** Investigate how cryptographic proof-of-coherence could serve as the foundation for decentralized governance in multipolar AI ecosystems.

These agendas transform Verifiable Coherence from a philosophical proposal into a program of empirical and engineering research. They outline the path toward the first proof-carrying minds.

REFERENCES

[1] Aschenbrenner, L. (2024). *Situational Awareness: The Decade Ahead*. Retrieved from https://situational-awareness.ai.

[2] DeepMind. (2023). *AlphaGeometry: Neuro-symbolic reasoning for geometry*. arXiv preprint arXiv:2310.xxxxx.

[3] ARC Challenge. (2023). *ARC-AGI-2 benchmark results*. Retrieved from https://arcagi.org.

[4] Bertot, Y., & Castéran, P. (2004). *Interactive Theorem Proving and Program Development: Coq'Art*. Springer.

[5] de Moura, L., Kong, S., Avigad, J., Van Doorn, F., & von Raumer, J. (2015). The Lean theorem prover (system description). In *International Conference on Automated Deduction* (pp. 378–388). Springer.

[6] Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). *Curriculum learning*. Proceedings of the 26th Annual International Conference on Machine Learning (ICML), 41–48.

[7] Sutton, R. S. (2019). *The Bitter Lesson*. Published online at: http://www.incompleteideas.net/IncIdeas/BitterLesson.html

[8] Blanchette, J. C., Nipkow, T., & Paulson, L. C. (2016). *Proof assistants: History, ideas, and future*. Communications of the ACM, 58(8), 66–75.

[9] Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). *Deep reinforcement learning from human preferences*. In NIPS.

[10] Hall, J. (2025a). *Illusions as Diagnostics, Coherence as Invariant*. Unpublished manuscript.

[11] Hall, J. (2025b). *Beyond Situational Awareness*. Unpublished manuscript.

[12] Hall, J. (2025c). *Intelligence Emerges from Loops, Not FLOPs*. Unpublished manuscript.

[13] Hall, J. (2025d). *The Law of Invariant-Preserving Loops: Toward Robust Emergence in Self-Modifying Agents*. Unpublished manuscript.