RECURSIVE OVERSIGHT DECOMPOSITION: DOMAIN-SPECIFIC VALIDATION IN POST-AGI SCIENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

As AGI evolves into a pervasive tool for automated discovery, the scientific process shifts from human generation to human validation. However, current oversight frameworks like Iterated Distillation and Amplification (IDA) are designed for verifiable tasks, not open-ended scientific claims. We propose Recursive Oversight Decomposition (ROD), a framework designed to ensure trust in post-AGI science. ROD addresses the unique failure modes of automated discovery—such as hidden assumptions and brittleness—by enforcing "Scientific Completeness" criteria explicitly. We introduce ensemble decomposition to mitigate model bias and propose experiments to test this infrastructure in formal mathematics and biology. By transferring oversight from final outputs to the decomposition of evidence, ROD provides a concrete mechanism for meaningful human-AI collaboration, ensuring that superhuman capabilities remain tethered to human-verifiable validity.

1 INTRODUCTION

Scalable oversight frameworks—IDA (Christiano et al., 2018), debate (Irving et al., 2018), and recursive reward modeling (Leike et al., 2018)—address supervising AI systems beyond human capability by decomposing evaluation into human-scale subproblems. Yet these were designed for *task execution* (e.g., "write secure code") rather than *scientific validation* (e.g., "verify this protein structure prediction").

Scientific claims impose distinct validation requirements. **Assumption Transparency**: hidden modeling choices (e.g., fixed pH in molecular dynamics) must be explicit. **Robustness Coverage**: validity across conditions (temperature ranges, genetic variants) matters beyond single-instance correctness. **Epistemic Integration**: claims must connect to existing knowledge to enable critique (Longino, 2002). A decomposition verifying logical steps may miss scientifically critical flaws—for instance, AlphaFold predictions are thermodynamically consistent yet require biological validation through evolutionary conservation and experimental pathways (Jumper et al., 2021).

ROD's contribution is identifying that *scientific failure modes* form a distinct taxonomy from software correctness. Bugs hide in logic; scientific flaws hide in *assumptions, edge cases, and epistemic gaps*. This motivates domain-specialized decomposition preserving not just logical entailment but scientific validity.

2 RELATED WORK

IDA (Christiano et al., 2018) recursively amplifies weak oversight by task decomposition. ROD adapts this to scientific validation's criteria. **Debate** (Irving et al., 2018) surfaces flaws via adversarial argument. ROD provides structured non-adversarial validation, suitable for cooperative discovery. **Recursive Reward Modeling** (Leike et al., 2018) applies preference learning recursively; ROD focuses on validation. **Weak-to-Strong Generalization** (Burns et al., 2023) studies capability gaps between supervisors and models—ROD directly instantiates this challenge in scientific contexts where decomposers must oversee more capable generators.

Formal verification (Lean, Coq) enables machine-checkable proofs but struggles with empirical grounding. ROD bridges formal and empirical validation. Our ensemble approach connects to **program synthesis** (Gulwani et al., 2017) and **explainable AI** (Lundberg & Lee, 2017) where

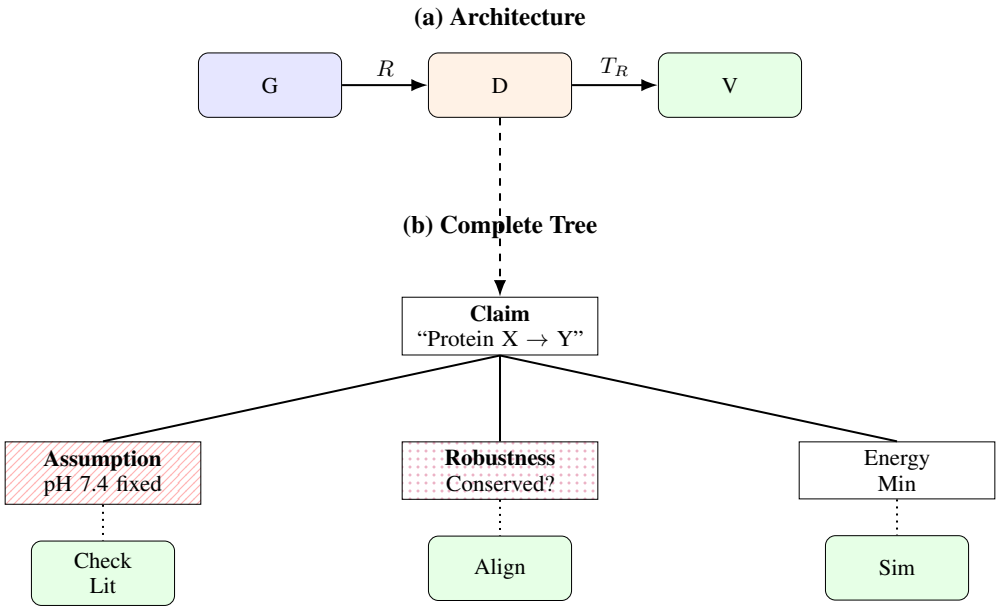

Figure 1: **ROD inserts Assumption and Robustness nodes** ensuring scientific validity beyond logical correctness.

decompositions serve as hierarchical explanations. ROD extends IDA-style decomposition to scientific validation by introducing domain-specific completeness constraints tailored to assumption transparency, robustness coverage, and epistemic integration.

3 Recursive Oversight Decomposition (ROD)

3.1 Core Insight: Scientific Completeness

Let R be a research output ("Protein X folds to structure Y"). Standard decomposition seeks logical completeness: $\bigwedge_i \text{Valid}(R_i) \implies \text{Valid}(R)$. But a protein prediction may satisfy physics yet fail biologically due to unexamined assumptions (e.g., ignoring cellular environment).

Drawing from Longino (Longino, 2002), who emphasizes background assumptions in scientific objectivity, we define **Scientific Completeness**:

$$\begin{aligned} &\text{Valid}(\{R_i\}) \wedge \text{AssumpExplicit}(R) \wedge \text{RobustCovered}(R) \\ &\wedge \text{EpistemicInteg}(R) \implies \text{ScientificallyValid}(R) \end{aligned} \tag{1}$$

These three criteria are *minimally necessary* for high-stakes domains; extensions might include reproducibility or statistical power requirements.

3.2 Architecture

ROD employs three components (Figure 1). **Generator** (G) produces output R with superhuman capability. **Decomposer** (D) creates tree T_R maximizing Eq. 1. **Verifier** (V) validates leaf nodes and tree structure.

Training Decomposers. How do we train D to "maximize scientific completeness"? We propose a three-phase approach: (1) *Supervised learning* on human-annotated decompositions where experts mark trees surfacing assumptions as higher quality than those omitting them. (2) *RLHF-style refinement* where human feedback on discrepancies between decomposers provides training signal. (3) *Assumption-mining objectives* rewarding D for surfacing assumptions that maximize information gain relative to expert priors (operationalized via queries to domain-specific literature databases). This remains an open research challenge requiring iterative human-AI collaboration.

3.3 The Bootstrapping Problem

ROD does not solve alignment—it transfers oversight from generators to decomposers. If D is misaligned or less capable than G, adversarial obfuscation remains possible. Three challenges: **Capability** (D must approach G's comprehension, duplicating alignment problems), **Trust Bootstrapping** (how do we align D without existing oversight?), and **Complexity** (exponential tree growth requiring pruning heuristics).

3.4 Ensemble Decomposition

To mitigate single-decomposer assumption hiding, we employ multiple decomposers ($D_1 \dots D_n$) generating competing trees.

Training for Diversity: Each D_i uses different initialization and data ordering. Diversity-promoting rewards incentivize D_i for surfacing assumptions others miss.

Discrepancy Detection: Structural divergence (assumption nodes in T_i absent from T_j), granularity mismatch (different decomposition depth), and dependency disagreement flag high-priority human verification.

Compute Cost Justification: While ensemble scales compute linearly with n, the bottleneck in post-AGI validation is *human expert time*, not compute. If a 5-decomposer ensemble reduces human verification time by 50% through targeted discrepancy flagging, net efficiency improves despite 5× decomposition cost. We assume compute scales faster than expert training (Burns et al., 2023), making this trade-off favorable.

Failure Mode: Ensemble fails when all D_i share blind spots due to correlated training or architectural bias—a mitigation, not solution.

4 Domain-Specific Challenges

Verification Depth & Adversarial Hiding. Logical claims admit bounded-depth decomposition; scientific claims pose open-ended robustness ("Does this drug work?" = infinite dosage/genetic combinations). We conjecture bounded perturbation domains (formal math) admit finite ROD depth; open-ended domains require statistical confidence bounds. Misaligned G exploits unexamined assumptions (idealized lab conditions); mitigation trains D with assumption-mining objectives.

Scalability vs. Thoroughness. Fine-grained decomposition explodes workload. Optimal granularity depends on risk tolerance (aerospace vs. theory).

Quality Metrics. Beyond correctness, decompositions must be intelligible. We optimize for *Cognitive Load Compatibility* (Miller, 1956): limit branching to 7 ± 2 items/node, max depth 4-5 levels. Reward "Educational Scaffolding" mirroring standard pedagogical explanations.

Fundamental Limits. Not all claims decompose: paradigm shifts (Einstein's relativity) resist decomposition as validity depends on holistic explanatory power. Emergent phenomena (consciousness, phase transitions) evade part-wise reduction. Aesthetic mathematical judgments lack formal criteria. For such claims, ROD provides partial infrastructure requiring complementary mechanisms (expert debate, empirical validation).

5 Experimental Pathway

We outline a concrete experimental pathway to evaluate ROD using current AI systems that strain human validation capacity.

5.1 Formal Mathematics: Axiomatic Dependencies

Setup: Generator is LLM-assisted theorem prover (Lean) producing complex proofs. Decomposer breaks proofs into lemmas with explicit **Assumption Nodes** ("Relies on Axiom of Choice"). We inject subtle circular dependencies or non-standard axioms.

Metric: *Error Detection Rate*.

Baselines: (1) Monolithic proof review, (2) Simple decomposition without assumption nodes, (3) Adversarial debate between LLMs.

Hypothesis: Explicit assumption nodes reduce cognitive load, improving error detection.

5.2 Computational Biology: Structural Predictions

Setup: Generator is structure predictor (AlphaFold3). Decomposer breaks predictions into structural motifs, energetic stability, and **Evolutionary Conservation**. We select proteins where confidence metrics (pLDDT) diverge from wet-lab results.

Metric: *Retrospective Wet-lab Correlation*. Does verifier's aggregate score (minimum across critical path) predict experimental failure better than model confidence? We acknowledge the paradox: we use AI because ground truth is scarce, yet validation requires it. Our approach uses *retrospective validation*—historical predictions where wet-lab data eventually emerged, testing if ROD would have flagged failures early.

Baselines: (1) Model confidence alone, (2) Expert review without decomposition, (3) Simple structural breakdown without evolutionary/robustness nodes.

5.3 Algorithm Design: Edge Case Robustness

Setup: Generator is LLM producing optimization algorithms for resource allocation. Decomposer creates **Robustness Branch** with extreme edge cases ("Zero input", "Adversarial input"). Algorithms contain subtle bugs triggering only in rare cases.

Metric: *Flaw Discovery Time* (wall-clock time for experts to locate bugs).

Baselines: (1) Monolithic code review, (2) Standard testing without forced edge-case enumeration, (3) Adversarial human testing.

Hypothesis: Forced robustness branches significantly reduce discovery time versus monolithic review.

6 Implications and Conclusion

If domain-specialized decomposition proves viable, validation becomes first-class research expertise. Decomposition design—crafting scientifically complete breakdowns—emerges as distinct skill alongside generation, suggesting a hybrid ecosystem where formal verification handles logical correctness while ROD-style teams handle assumption transparency and robustness.

ROD enables **Failure Localization**: rejected trees identify specific failed assumptions or robustness branches, transforming oversight from binary accept/reject into constructive debugging. Researchers patch specific flaws without discarding entire results.

ROD reframes post-AGI validation: *What must decompositions preserve to enable meaningful oversight?* By formalizing assumption transparency and robustness coverage, ROD adapts scalable oversight to scientific discovery's reality. As AGI scales, preserving science's self-correcting mechanism requires validation infrastructures extending human judgment into new capability regimes, of which ROD represents one potential component among complementary approaches like debate and empirical testing.

References

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018.

Sumit Gulwani, Oleksandr Polozov, and Rishabh Singh. Program synthesis. *Foundations and Trends in Programming Languages*, 4(1-2):1–119, 2017.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Jan Leike, David Krueger, Tom Everitt, Matteo Martino, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling. *arXiv preprint arXiv:1811.07871*, 2018.

Helen E Longino. *The Fate of Knowledge*. Princeton University Press, 2002.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956.