On the Computational, Informational, and Physical Foundations for AI Safety

Robin Young
University of Cambridge
Department of Computer Science and Technology
Cambridge, UK
robin.young@cl.cam.ac.uk

Abstract

Current approaches to AI safety predominantly focus on specifying correct behavior through software, data, and rules. This work argues that this approach faces theoretically fundamental, and not merely practical, limitations. I present a multi-layered analysis of this paradigm, demonstrating its inherent barriers from the perspectives of computational complexity, information theory, and physical engineering. In ongoing work, I prove that even simplified forms of semantic self-verification are computationally intractable (NP-complete). I use information theory to show that any specification of an external, ambiguous concept like "harm" is necessarily incomplete. To address these limits, I develop a framework for reasoning about verifiable, physically-enforced safety bounds that are independent of software state.

Introduction

Today, the dominant approaches to AI safety (Ji et al. 2025; Ngo, Chan, and Mindermann 2024) rely on creating a better software by way of utility functions, training on human preferences (RLHF) (Christiano et al. 2017; Ouyang et al. 2022b; Kaufmann et al. 2024; Ouyang et al. 2022a), or self-critique through constitutions (Bai et al. 2022) to guide AI behavior. The implicit assumption is that with better data, smarter algorithms, or stricter rules, we can probably approximately specify desired behavior, as in PAC learning (Valiant 1984).

My research examines this assumption. If we cannot even fully specify "harm" in a domain like medicine after millennia of effort, how can we expect to encode it for an AI operating in an open world? This raises my central research question:

How can we move beyond specification-based alignment to build a more robust, multi-layered safety paradigm grounded in computational, informational, and physical realities that fail gracefully in uncertainty?

To answer this, I pursue a three-pronged investigation. I first establish the computational barriers to an AI even verifying its understanding of the contract. I then demonstrate the informational limits that ensure the contract itself is likely incomplete. Then, I propose a constructive path forward by developing a conceptual framework for physical, hardware-level constraints that can serve as a robust backstop when the software contract inevitably runs into edge cases from approximate learning.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Ongoing Work

The Computational Barrier

Before an AI can adhere to its principles, it should verify that it has correctly interpreted them. This has been explored in depth philosophically and logically (Gödel 1931; Löb 1955), in the formal verification context for AI (Katz et al. 2017), AI safety via debate (Irving, Christiano, and Amodei 2018; Brown-Cohen, Irving, and Piliouras 2024), where systems must justify their reasoning through structured argumentation, and process-based supervision (Bowman et al. 2022). In the first project, I formalize this the problem of Semantic Self-Verification (SSV): determining if a statement accurately characterizes its own semantic properties under a set of constraints.

I proved, in this formalization, that even a highly simplified version of SSV is NP-complete by constructing a polynomial-time reduction from 3-SATISFIABILITY (3-SAT) (Cook 1971; Garey and Johnson 1990). In this reduction, Boolean variables map to ambiguous terms with binary meanings, and logical clauses map to semantic constraints. The result shows that determining if a consistent interpretation exists is at least as hard as solving any problem in NP.

This establishes a computational floor for verification of alignment claims. The idea that an AI can reliably and exhaustively check its compliance with a complex instruction set faces a provably intractable barrier. This implies that any practical system making such judgments quickly is necessarily using heuristics, not complete verification, leaving open unverified edge cases.

The Informational Gap

Even if we grant that verification were tractable, a second, deeper problem remains: can the specification itself ever be complete? In my second project, I developed an information-theoretic framework to formalize the specification-reality gap. I modeled the ground truth of an external concept like "harm" as a random variable O with entropy H(O), and the AI's specification as a variable I.

The theorem in this work examines the modeling consequences where for any such external concept where ambiguity or context-dependency exists ($H(O) > 0$), the mutual information between the specification and the ground truth is strictly less than the total information in the concept: $I(O; I) < H(O)$

This inequality, a direct consequence of the data processing inequality (Cover and Thomas 1991) and the externality of O, holds regardless of the sophistication of the learning algorithm or the volume of data.

This result demonstrates a leakage of semantic meaning. No finite specification can perfectly capture an external concept that has inherent semantic entropy. It examines the intuition based on philosophical axioms (Putnam 1975; Kripke 1980; Quine 1960) that concepts like "harm" or "fairness" are not reducible to a set of rules, providing a quantifiable basis for skepticism towards claims of eventually complete value alignment. It also led to the conceptual proposal of a Safety-Capability Ratio, $I(O; I)/H(O)$, as a novel metric for quantifying alignment quality.

A Constructive Alternative

The computational and informational limits of software-based safety engineering point to a vulnerability: what happens when the software fails, is compromised, or acts on an incomplete understanding? To address this, I introduce and develop the framework of hardware based alignment. This approach shifts the focus from specifying what an AI should do to physically bounding what it can do.

I am developing concepts for this domain. These concepts hope to allow for the systematic analysis and design of hardware-level safety mechanisms, using established concepts in fail-safe design, cyber-physical systems, and formal verification methods (Leveson 2012; Seshia, Sadigh, and Sastry 2022; Obermaier and Immler 2018; Hu et al. 2021; Parasuraman, Sheridan, and Wickens 2000; Anderson 2001), that provide bounds on an agent's speed, force, and operational reach, independent of its software state.

This work aims to provide a physically robust conceptual agenda for embodied AI. Just as a circuit breaker trips based on current, not software intent, these bounds cannot be "convinced" or "tricked." This approach complements software alignment by providing a robust, verifiable foundation of "can't" to software's ongoing efforts to define "should".

Significance

Taken together, these three lines of work constitute an examination and a constructive proposal for prevailing safety paradigms. The research agenda demonstrates that the software centric approach faces a couple of barriers:

The first result shows that even if we could write a perfect specification, verifying an AI's semantic compliance with it is computationally infeasible in the edge cases (NP-complete). The burden of proof for the AI's own understanding is fundamentally intractable.

The second result posits a related problem: the specification itself can not be complete. Due to the inherent ambiguity and context-dependency of normative concepts like "harm," there will always be an unbridgeable information gap between any formal specification and the ground truth reality it seeks to represent.

Given that complete specification may be impossible and perfect verification is intractable, how can we possibly build safe systems? This is where the third line of work provides an attempt at a principled answer. Instead of trying to eliminate semantic uncertainty, we can bound its potential consequences.

My contribution, therefore, is an attempt at a cohesive argument: we should move beyond a singular focus on software specification and toward a multi-layered safety architecture grounded in the hard limits of computation, information, and physics.

Ongoing and Future Work

The next step is to formally connect the three lines of work. This involves developing a quantitative model of multi-layer safety that can answer crucial questions: How can, if at all, specific hardware-level constraints (e.g. bounding an agent's force output) reduce the semantic entropy the software layer must handle? Can we define a "bounded SSV problem" where limits on the state or action space transform the verification task into a tractable subclass? The goal is to create a formal theory that allows us to reason about how safety properties cascade and interact across physical, informational, and computational layers.

This also points toward two additional, albeit resource-intensive, research directions that I believe could be crucial for the field. Although these are likely beyond the scope of a single dissertation, so my immediate and realistic research focus will remain on strengthening the theoretical foundations.

A long-term goal for operationalizing is to develop and validate the novel metrics proposed. The first might involve large-scale studies of human annotator disagreement on ethical vignettes (e.g. medical triage, autonomous vehicle choices) to create the first empirical estimates of H(Harm) in specific domains.

Then, to measuring the verification surface, this would necessitate collaboration with robotics labs to formally model the control circuits and physical actuators of existing robots, mapping which safety parameters are physically verifiable versus purely software-dependent. The moonshot would be to create a public benchmark suite that allows researchers to quantitatively assess the alignment and robustness of different embodied AI systems.

Conclusion

This ongoing project posits that the prevailing software approach to AI safety is built on theoretical foundations that seem to be underexamined. It seems computationally, informationally, and physically insufficient in many cases. By systematically identifying these limits and proposing a constructive, hardware-grounded alternative, my work argues for a more robust safety strategy by examining foundational limits and formalizing mathematical objects for tractable analysis. This work aims to provide the technical and conceptual tools for a more humble, robust, and ultimately more trustworthy approach to AI safety.

References

Anderson, R. J. 2001. *Security Engineering: A Guide to Building Dependable Distributed Systems*. USA: John Wiley & Sons, Inc., 1st edition. ISBN 0471389226.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.

Bowman, S.; Hyun, J.; Perez, E.; Chen, E.; Pettit, C.; Heiner, S.; Lukosiute, K.; Askell, A.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; Olah, C.; Amodei, D.; Amodei, D.; Drain, D.; Li, D.; Tran-Johnson, E.; Kernion, J.; Kerr, J.; Mueller, J.; Ladish, J.; Landau, J. D.; Ndousse, K.; Lovitt, L.; Elhage, N.; Schiefer, N.; Joseph, N.; Mercado, N.; Dassarma, N.; Larson, R.; McCandlish, S.; Kundu, S.; Johnston, S.; Kravec, S.; Showk, S. E.; Fort, S.; Telleen-Lawton, T.; Brown, T. B.; Henighan, T.; Hume, T.; Bai, Y.; Hatfield-Dodds, Z.; Mann, B.; and Kaplan, J. 2022. Measuring Progress on Scalable Oversight for Large Language Models. *ArXiv*, abs/2211.03540.

Brown-Cohen, J.; Irving, G.; and Piliouras, G. 2024. Scalable AI Safety via Doubly-Efficient Debate. In *Forty-first International Conference on Machine Learning*.

Christiano, P. F.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 4302–4310. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Cook, S. A. 1971. The complexity of theorem-proving procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, STOC '71, 151–158. New York, NY, USA: Association for Computing Machinery. ISBN 9781450374644.

Cover, T. M.; and Thomas, J. A. 1991. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. New York: Wiley, 1st edition. ISBN 0471062596.

Garey, M. R.; and Johnson, D. S. 1990. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. USA: W. H. Freeman & Co. ISBN 0716710455.

Gödel, K. 1931. *On Formally Undecidable Propositions of Principia Mathematica and Related Systems*. New York, NY, USA: Basic Books.

Hu, W.; Chang, C.-H.; Sengupta, A.; Bhunia, S.; Kastner, R.; and Li, H. 2021. An Overview of Hardware Security and Trust: Threats, Countermeasures, and Design Tools. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 40(6): 1010–1038.

Irving, G.; Christiano, P. F.; and Amodei, D. 2018. AI safety via debate. *ArXiv*, abs/1805.00899.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Vierling, L.; Hong, D.; Zhou, J.; Zhang, Z.; Zeng, F.; Dai, J.; Pan, X.; Ng, K. Y.; O'Gara, A.; Xu, H.; Tse, B.; Fu, J.; McAleer, S.; Yang, Y.; Wang, Y.; Zhu, S.-C.; Guo, Y.; and Gao, W. 2025. AI Alignment: A Comprehensive Survey. arXiv:2310.19852.

Katz, G.; Barrett, C. W.; Dill, D. L.; Julian, K. D.; and Kochenderfer, M. J. 2017. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. *ArXiv*, abs/1702.01135.

Kaufmann, T.; Weng, P.; Bengs, V.; and Hüllermeier, E. 2024. A Survey of Reinforcement Learning from Human Feedback. arXiv:2312.14925.

Kripke, S. A. 1980. *Naming and Necessity: Lectures Given to the Princeton University Philosophy Colloquium*. Cambridge, MA: Harvard University Press.

Leveson, N. G. 2012. *Engineering a Safer World: Systems Thinking Applied to Safety*. The MIT Press. ISBN 9780262298247.

Löb, M. H. 1955. Solution of a problem of Leon Henkin. *Journal of Symbolic Logic*, 20(2): 115–118.

Ngo, R.; Chan, L.; and Mindermann, S. 2024. The Alignment Problem from a Deep Learning Perspective. In *The 12th International Conference on Learning Representations*.

Obermaier, J.; and Immler, V. 2018. The Past, Present, and Future of Physical Security Enclosures: From Battery-Backed Monitoring to PUF-Based Inherent Security and Beyond. *Journal of Hardware and Systems Security*, 2: 289–296.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022a. Training language models to follow instructions with human feedback. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 27730–27744. Curran Associates, Inc.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022b. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Parasuraman, R.; Sheridan, T.; and Wickens, C. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3): 286–297.

Putnam, H. 1975. The Meaning of 'Meaning'. *Minnesota Studies in the Philosophy of Science*, 7: 131–193.

Quine, W. V. O. 1960. *Word & Object*. MIT Press.

Seshia, S. A.; Sadigh, D.; and Sastry, S. S. 2022. Toward verified artificial intelligence. *Commun. ACM*, 65(7): 46–55.

Valiant, L. G. 1984. A theory of the learnable. *Commun. ACM*, 27(11): 1134–1142.