We Are Misaligned: Rethinking the Current AI Alignment Approaches

Kwaai Alignment Lab

April 2025

Abstract: AI alignment strategies typically operate under assumptions of static environments, decomposable objectives, and predictable behavior. A systemic thinking perspective reveals that these assumptions systematically understate the emergent, nonlinear, and adaptive dynamics present in real-world deployment scenarios. This position paper critiques major alignment approaches, proposes a shift towards systemic-aware alignment research, and outlines our new research direction based on principles from robust natural systems that embody aspects of control.

1. Introduction

AI alignment aims to ensure that increasingly capable AI systems pursue goals beneficial to humanity. Prevailing methods emphasize direct specification of objectives, reward modeling, and human feedback mechanisms. However, systemic thinking, which studies interconnected, dynamic, and adaptive systems, warns that such approaches may be insufficient when agents and environments co-evolve in unpredictable ways. Currently, the majority of "alignment" work—such as RLHF, automated benchmarks, or human-in-the-loop evaluations—focuses on what we call *product AI alignment*. This involves shaping models to provide responses that are generally useful and safe for users. While effective at the product level, this has created a phenomenon referred to as **alignment-washing** [11], giving the misleading impression that we are progressing toward *true* AGI alignment.

2. The Science of Complex Systems

We consider the fundamental ideas from complexity science, which is used to study complex systems. Emergence refers to macroscopic behaviors that arise unpredictably from local interactions. Nonlinearity means that small changes can lead to disproportionately large effects. Adaptation and co-evolution describe how agents and their environments adapt together over time. Distributed control highlights that regulation and stability result from feedback loops rather than centralized command. Finally, robustness through redundancy ensures stability through overlapping controls and modularity. While these principles are not exhaustive, and certainly is not a panacea to our problem, in our opinion it begins to address alignment from the right perspective, where we stand a chance of solving it.

3. Critique of Current Alignment Approaches

3.1 Value Learning and IRL

Human preferences are not static individual functions, but are dynamically constructed through social, cultural, and situational factors [1]. Inverse reinforcement learning assumes preferences are discoverable and stable, which ignores their emergent and co-evolving nature.

3.2 Corrigibility

Corrigibility frameworks [2] assume agents will accept human intervention. Yet systemic thinking teaches that these intervention points [3] can close as systems self-organize, making interventions ineffective or counterproductive .

3.3 Interpretability

Mechanistic interpretability [4] offers valuable insights but falls short because local transparency does not imply global predictability. In systemic environments, interpretability must extend to the dynamics of learning and adaptation, not just static snapshots [5].

3.4 Scalable Oversight

IDA [6] and debate [7] presuppose that complex tasks can be decomposed and recomposed cleanly. However, many tasks involve strong inter-dependencies and emergent, systemic coordination, challenging the assumptions underlying modular oversight strategies.

3.5 RLHF and Constitutional AI

RLHF [8] depend on the noisy, shifting and fickle human feedback and further doesn't address the systemic nature agents exist. Constitutional AI [9] seeks stability via fixed principles, again focusing on individual agents, and further are brittle in dynamic, real environments. True alignment must incorporate ongoing feedback and adaptation. In this case, something like an adapting constitution designed for a system, as well as all overlapping systems, may be a fruitful direction.

3.6 Summary

The approaches above are thoughtful and informative. Our critique is that they do not consider the reality of these agents- they currently do and will continue to exist in complex systems, especially as they scale up and even become super intelligent (ASI).

4. The Path Forward

We should focus our efforts fully and completely on systemic alignment research—or really, in our opinion, true AI alignment. The emergence of AGI will likely involve the same dynamics as seen in other robust natural systems. Thus, alignment research must prioritize: Multi-agent dynamics: Alignment must model AI not as a single agent interacting with a human, but as part of a diverse ecosystem of agents, by default. Nature-inspired systems: Drawing insights from robust natural processes such as DNA replication, homeostasis, flocking behavior, and crystal formation. Redundancy and feedback loops: Building error detection, correction, and resilience into AI governance structures. Continuous co-adaptation: Systems must evolve alongside human values and changing environments [10]. Modularity and fault-tolerance: Designing AI architectures to localize and contain errors, preventing systemic collapse. Distributed regulation: Emphasizing decentralization, self-monitoring, and resilience over centralized command-and-control models.

5. Conclusion

The current framing of AI alignment conflates product safety with existential alignment, creating a false sense of progress. Systemic thinking warns us that emergent, nonlinear, and multi-agent phenomena will dominate AGI dynamics. Only by enforcing systemic perspectives—through nature-inspired design, redundancy, modularity, feedback, and distributed control—can we meaningfully advance towards robust, safe AGI. This paper proposes a reorientation of alignment research away from simple principal-agent paradigms towards systemic frameworks rooted in the successes of natural, adaptive control systems. Progress demands a shift from static alignment to dynamic, resilient, and systemic co-evolution.

References

[1] Stuart Russell. *Provably beneficial AI.* (2016)

[2] Soares et al. *Corrigibility.* (2015)

[3] Hubinger et al. *Risks from Learned Optimization in Advanced Machine Learning Systems.* (2019)

[4] Chris Olah et al. *Transformer Circuits: Interpretability of Deep Networks.* (2020)

[5] John H. Holland. *Complex Adaptive Systems.* (1992)

[6] Paul Christiano et al. *Iterated Distillation and Amplification.* (2018)

[7] Geoffrey Irving et al. *AI Safety via Debate.* (2018)

[8] Paul Christiano et al. *Deep reinforcement learning from human preferences.* (2017)

[9] Anthropic. *Constitutional AI: Harmlessness from AI Feedback.* (2022)

[10] Luciano Floridi. *AI and Its Complexity: From Static Rules to Dynamic Ethical Co-evolution.* (2022)

[11] Connor Leahy and Gabriel Alfour. *The Compendium.* (2023)