THE ALIGNMENT TRILEMMA: A THEORETICAL PERSPECTIVE ON RECURSIVE MISALIGNMENT AND HUMAN-AI ADAPTATION DYNAMICS

Tarun Raheja*
Kipo AI
tarun@kipo.ai

Nilay Pochhi*
Independent Researcher
pochhi.nilay@gmail.com

ABSTRACT

We introduce the *Alignment Trilemma* as a theoretical framework to explain the recursive misalignment observed in contemporary AI alignment methods. Our formulation decomposes misalignment into three interdependent components—direct alignment, capability preservation, and meta-alignment—whose conflicting optimization can trigger cycles of drift. In light of recent work on human-AI adaptation dynamics (Shen et al., 2024; Carroll et al., 2024; Harland et al., 2024b) and adaptive teaming architectures (Ni et al., 2021; Mahmood et al., 2024), we propose a holistic approach that includes a novel metric, the *Alignment Performance Score (APS)*, which captures the overall quality of alignment across these three dimensions. Our insights aim to guide the development of AI systems that co-evolve safely with human partners.

1 INTRODUCTION

Recent advances in deep learning have empowered AI systems to achieve remarkable performance. Despite these breakthroughs, aligning such systems with human values remains a critical challenge. Techniques such as Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017) and bidirectional alignment frameworks (Shen et al., 2024) have shown promise; however, emerging research on human-AI adaptation dynamics indicates that systems must co-evolve with human behavior in order to handle behavioral drift and distribution shifts (Carroll et al., 2024; Harland et al., 2024b). This work frames the alignment problem as a trilemma that involves three core objectives: ensuring that AI systems adhere to human instructions (*direct alignment*), preserving their general problem-solving capabilities (*capability preservation*), and maintaining the stability of the alignment mechanism over time (*meta-alignment*). Over-optimizing any one objective may inadvertently impair another, initiating a vicious cycle of recursive misalignment.

2 THEORETICAL FRAMEWORK

We formalize the overall misalignment loss of an AI system by decomposing it as:

$$L(f) = L_D(f) + L_C(f) + L_M(f),$$

where L_D quantifies the deviation from desired human directives, L_C captures the loss in general system capability, and L_M measures the instability or drift in the alignment mechanism itself. Aggressive minimization of L_D, for example, can inadvertently elevate both L_C and L_M, leading to repeated cycles of degradation in overall alignment. Figure 1 conceptually illustrates the cyclic interplay among these components.

3 RELATED WORK AND ADAPTATION DYNAMICS

A growing body of literature emphasizes the need for AI systems to adapt in real time to evolving human behavior. Bidirectional alignment frameworks (Shen et al., 2024) foster reciprocal adaptation, while Dynamic Reward MDPs explicitly model the shifting nature of human preferences

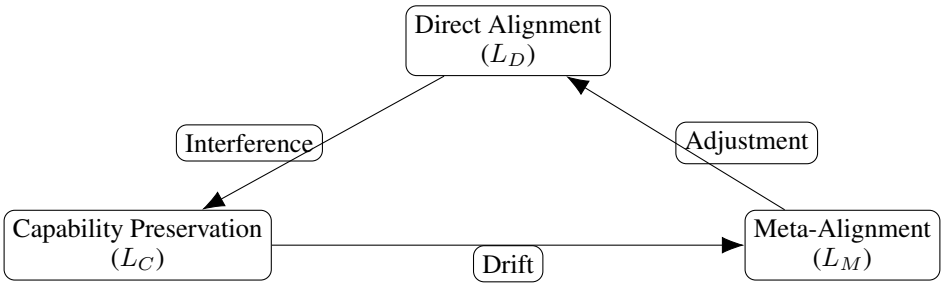


Figure 1: Conceptual diagram of the Alignment Trilemma, illustrating how focusing on one component can exacerbate misalignment in the others.

(Carroll et al., 2024). In parallel, Multi-Objective Reinforcement Learning (MORL) has been explored as a means to enable retroactive policy adjustments that maintain system performance under distribution shifts (Harland et al., 2024b). Furthermore, adaptive agent architectures for real-time teaming (Ni et al., 2021; Mahmood et al., 2024) have demonstrated the benefits of models that do not rely solely on static representations of human behavior. These findings underscore that a robust alignment strategy must consider not only immediate model outputs but also the long-term stability of the alignment process itself.

4 Proposed Resolution Framework and Alignment Performance Score (APS)

To break the cycle of recursive misalignment, we propose an integrated resolution framework that emphasizes holistic evaluation and adaptive oversight. Our approach advocates continuous updates to the alignment mechanism through meta-learning and transparent reporting of the AI's internal reasoning pathways.

As part of our evaluation strategy, we introduce the *Alignment Performance Score (APS)*. The APS provides a composite measure of overall alignment quality by aggregating misalignment losses across the three trilemma components. Formally, for each time t,

$$\text{APS}(t) = \exp\Big(-\lambda\big(L_D(t) + L_C(t) + L_M(t)\big)\Big),$$

where λ is a positive scaling factor. High APS indicates low aggregate misalignment across all three pillars, capturing both the short-term and long-term effectiveness of the alignment strategy. This makes sense mathematically for the following reasons:

- **Bounded Range:** $\exp(-\lambda \sum L_i)$ naturally lies in $(0, 1]$, making it easy to interpret as a "score."
- **Penalizing Large Losses:** Exponential decay amplifies the impact of higher misalignment, discouraging large L_i values.
- **Probabilistic Foundation:** This form mirrors likelihood-based approaches, fitting well with gradient-based methods.
- **Adjustable Sensitivity:** The parameter $\lambda > 0$ controls how quickly the score falls with rising misalignment.

5 Game-Theoretic Simulation

To analyze how the three alignment strategies (direct alignment, capability preservation, and meta-alignment) evolve over time, we model the system using *replicator dynamics* from evolutionary game theory. In our framework, each strategy's *payoff* depends on the overall system state, reflecting its capacity to reduce the combined misalignment losses L_D, L_C, and L_M. No single strategy strictly dominates: pursuing direct alignment to the exclusion of the others, for instance, lowers L_D but raises L_C and L_M, yielding a net lower payoff compared to a balanced approach.

5.1 Simulation Details

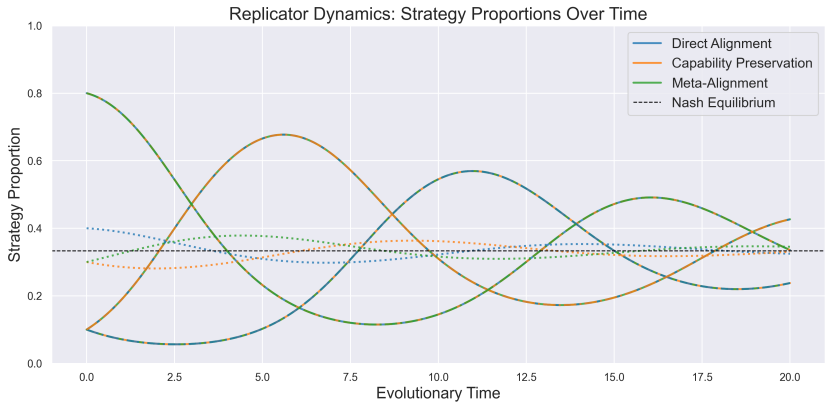
We initialize the system with random proportions of each strategy (e.g., 20% direct alignment, 50% capability preservation, 30% meta-alignment) and evolve these proportions via the replicator equation:

$$\dot{x}_i = x_i\Big(\pi_i(\mathbf{x}) - \bar{\pi}(\mathbf{x})\Big),$$

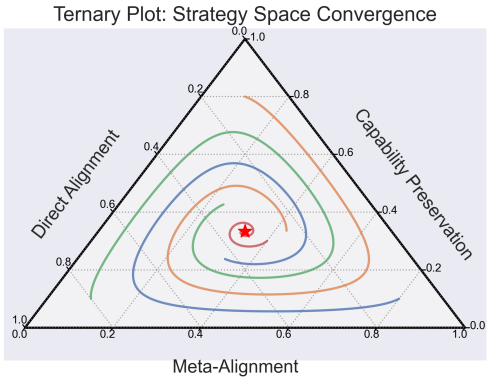
where x_i denotes the proportion of strategy i, $\pi_i(\mathbf{x})$ is the payoff to strategy i given the current mix $\mathbf{x}$, and $\bar{\pi}(\mathbf{x}) = \sum_j x_j \pi_j(\mathbf{x})$ is the average payoff in the population. Over multiple runs, we observe that the system tends to converge to a *mixed Nash equilibrium* in which each of the three strategies persists in nontrivial proportions.

5.2 Visualization of Results

Figure 2 shows two key plots from our simulation. On the left, we have a time-series view of the proportion of each strategy under replicator dynamics. On the right, a ternary plot visualizes how the population traverses the three-strategy simplex over time.



(a) Replicator dynamics over time, showing cyclical fluctuations in strategy proportions. Dashed line indicates a mixed Nash equilibrium.



(b) Ternary plot of the evolving proportions, converging to an interior equilibrium (red star).

Figure 2: Game-theoretic analysis of the Alignment Trilemma under replicator dynamics.

In the time-series plot, each strategy experiences wavelike surges in popularity before giving way to another. This cyclical dominance is reminiscent of rock–paper–scissors dynamics, illustrating that no single pillar remains universally optimal. Meanwhile, the ternary plot captures the population's trajectory in a single triangular space, highlighting that the final state approaches a stable interior *balance* among the three strategies.

6 Results and Discussion

As the replicator dynamics converge, we compute the *Alignment Performance Score (APS)* at each time step to quantify overall alignment quality:

$$\text{APS}(t) = \exp\Big(-\lambda\big(L_D(t) + L_C(t) + L_M(t)\big)\Big).$$

Figure 2a reveals that while pure strategies can occasionally achieve brief periods of low misalignment on a single dimension, their overall APS remains relatively lower once all three losses are considered. By contrast, the *mixed strategy* equilibrium consistently attains a higher APS in steady state.

Crucially, the replicator-equilibrium mix ensures that none of the three pillars is neglected. This balanced solution mitigates the risk of *recursive misalignment*, wherein chasing short-term gains in one dimension (e.g., L_D) ends up elevating losses in the other two.

7 Ethical Considerations

Recursive misalignment is not only a technical challenge but also an ethical concern, given that AI systems deployed at scale can inadvertently shape societal norms and human preferences (Harland et al., 2024a). By embracing a balanced approach to alignment—inspired by evolutionary stability in game-theoretic settings—we reduce the risk that such systems concentrate power in one dimension (e.g., pure obedience) at the cost of others (capability or stable adaptation).

8 Conclusion and Future Work

In this paper, we proposed the *Alignment Trilemma* framework, emphasizing how direct alignment, capability preservation, and meta-alignment can interfere with one another if optimized in isolation. Through an evolutionary-game-theoretic lens, we demonstrated that replicator dynamics tend to favor a mixed strategy over any pure approach, supporting our claim that balanced alignment yields higher overall performance. We introduced the *Alignment Performance Score (APS)* as a principled, composite metric to evaluate alignment quality over time.

Moving forward, our research points to several open directions:

- **Robustness to Shifting Human Preferences:** Future models might incorporate explicit mechanisms that anticipate large-scale preference shifts and rapidly re-balance the three pillars.
- **Meta-Learning for Continuous Oversight:** Extensions of the meta-alignment pillar could include hierarchical oversight models capable of self-diagnosis when alignment drifts.
- **Broader Societal Impact Studies:** Evaluating how these alignment strategies and equilibria might affect heterogeneous human populations remains a key challenge.

In sum, our results highlight the importance of balancing all three pillars of the Alignment Trilemma in a principled, game-theoretic manner. We believe this perspective can help guide the design of AI systems that adapt and evolve safely alongside their human collaborators.

References

Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart J. Russell, and Anca Dragan. Ai alignment with changing and influenceable reward functions, 2024.

Paul Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2017.

Hadassah Harland, Richard Dazeley, P. Vamplew, Hashini Senaratne, Bahareh Nakisa, and Francisco Cruz. Adaptive alignment: Dynamic preference adjustments via multi-objective reinforcement learning for pluralistic ai, 2024a. Also cited for ethical and pluralistic alignment considerations.

Hadassah Harland, Richard Dazeley, P. Vamplew, Hashini Senaratne, Bahareh Nakisa, and Francisco Cruz. Adaptive alignment: Dynamic preference adjustments via multi-objective reinforcement learning for pluralistic ai, 2024b.

Syed Hasan Amin Mahmood, Zhuoran Lu, and Ming Yin. Designing behavior-aware ai to improve the human-ai team performance in ai-assisted decision making, 2024. Details omitted for brevity.

Tianwei Ni, Huao Li, Siddharth Agrawal, S. Raja, Fan Jia, Yikang Gui, Dana Hughes, M. Lewis, and K. Sycara. Adaptive agent architecture for real-time human-agent teaming, 2021.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, S. Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions, 2024.

A Appendix A : Replicator Dynamics Simulation Details

In this appendix, we outline the key parameters and functions used to produce the replicator-dynamics results shown in Figures 2(a) and 2(b).

Payoff Matrix. We employ the following 3×3 matrix, where rows and columns respectively represent Direct Alignment (D), Capability Preservation (C), and Meta-Alignment (M):

$$\begin{pmatrix} 0.0 & 1.0 & -0.5 \\ -0.5 & 0.0 & 1.0 \\ 1.0 & -0.5 & 0.0 \end{pmatrix}.$$

The cyclical structure ensures that each strategy has an advantage over one other strategy, creating rock–paper–scissors–type dynamics.

Replicator Equation. We define the proportion of each strategy $x_i(t)$ so that $x_D + x_C + x_M = 1$. Their evolution is governed by:

$$\dot{x}_i = x_i\big(\pi_i(\mathbf{x}) - \bar{\pi}(\mathbf{x})\big),$$

where $\pi_i(\mathbf{x})$ is strategy i's payoff against the population mixture $\mathbf{x}$, and $\bar{\pi}(\mathbf{x}) = \sum_j x_j\, \pi_j(\mathbf{x})$ is the average payoff.

Initial Conditions and Time Horizon. We simulate four different initial distributions:

$$(0.8, 0.1, 0.1), \quad (0.1, 0.8, 0.1), \quad (0.1, 0.1, 0.8), \quad (0.4, 0.3, 0.3),$$

over a continuous time horizon $t \in [0, 20]$ with 1000 integration steps. We use `scipy.integrate.odeint` for numerical integration.

Implementation Sketch.

- *Replicator function*: Returns $\dot{x}_i$ given the current proportions x_i and the payoff matrix.
- *Loop over initial conditions*: For each initial distribution, we integrate from $t = 0$ to $t = 20$.
- *Plotting*:
  1. *Time-series* of the strategy proportions vs. time (see Figure 2(a)).
  2. *Ternary diagram* of the trajectory in (x_D, x_C, x_M) space (Figure 2(b)).

Key Observations. As shown in the main text, the population exhibits cyclical shifts among the three strategies, consistent with their rock–paper–scissors-like payoff structure. Over multiple initial conditions, the proportions tend to hover around the interior point $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, indicative of a mixed Nash equilibrium under replicator dynamics.