An Evolutionary Perspective on AI Alignment (Student Abstract)

Ida Mattsson

Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213
imattsso@andrew.cmu.edu

Abstract

Attempting to align AI capabilities and value structures by means of value elicitation from humans, such as through Reinforcement Learning from Human Feedback (RLHF), is a computational challenge that raises both psychological and philosophical questions. Adopting an evolutionary perspective on the emergence of value structures in humans and machine learning systems can offer a bridge between qualitative and quantitative aspects of alignment. Here, evolutionary dynamics are applied to a game-theoretic model of RLHF. This allows for formal reasoning about the process and capabilities that result from alignment training, even where quantitative benchmarks cannot be clearly defined. A simple parametrized game model of RLHF, subject to replicator dynamics, shows how the success of the training method is sensitive to bias in human judgments. Under ideal conditions, RHLF training leads to aligned behavior. If the choice pattern of the human judge is biased, the training instead incentivizes misalignment. This application shows that evolutionary analyses can contribute to improving the prospects for safety and support successful cooperation between humans and AI systems in deployment.

Introduction

Learning processes for humans and AI are both subject to evolutionary pressures. Capabilities of machine learning (ML) systems emerge and improve as a result of how those evolutionary pressures are defined and applied through the development pipeline. With RLHF (Christiano et al. 2017), the specific set-up of the training interactions and the types of responses elicited through interactions with human judges will shape the resulting reward structure of the resulting ML model. The behavior of individual humans and cultures, including expressions of shared values, evolve both at the individual level and through societal interaction, that increasingly involve interactions with ML systems and Large Language Models (LLMs) in particular. While game-theoretic models can be valuable for understanding agent interactions, they can fall short on modeling dynamic behaviors and accounting for behaviors of non-rational reasoners. On the other hand, psychological studies with human subjects offers more realistic insight into situated reasoning, including its shortcomings (e.g. (Kahneman 2011)).

A systematic understanding of the dynamics of evolutionary pressures during alignment could, for example, better inform what characteristics we should promote in output from large language models, and the impact from anthropomorphic presentation toward human judges (Shavit et al. 2023). RLHF is fundamental to any value-elicitation process that involves humans-in-the-loop, and is used for commercial models (e.g. (OpenAI 2024), (Gemini Team 2024), (Anthropic 2024)).

A central assumption in RLHF is that overt choice behavior can be used as proxy for the judge's underlying value structure. This makes the reinforcement learning agent sensitive to a judge that tries to game the training or is unintentionally biased. Such misalignment on part of the judge need not be intentional. Another central assumption in RLHF is that human value structures are complete, in the sense that only a preference judgment acts a reward signal for the reinforcement learner. But RLHF does not necessarily allow for underlying values to be distinguished from surface-level presentation. For example, RLHF applied to two factually accurate responses to a math prompt will not reinforce factual accuracy, but some other feature of the response - such as brevity, clarity or the use of an authoritative voice. This means that preference elicitation over both two aligned and two misaligned responses can influence the outcomes of training.

Method

In RLHF (Christiano et al. 2017), a human judge assigns a preference ordering over responses from a language model to a set of prompts. The specific outputs are determined by a RL model optimizing for human preference, and the judge's feedback is then used to update the RL model. An approach to scaling the method is to train a classifier on examples of the feedback process substitute for the human at scale.

RLHF as a Game The interactions central to RLHF can be modeled as game with two agents, representing model instances, with the same pure strategies: aligned (`A`) or misaligned (`MA`). Table 1 shows the payoff matrix. The human judgment is encoded in the utility function, interpreted as expectations over repeated interactions of the specified type. RLHF is defined for interactions (`A`, `MA`) and (`MA`, `A`), where the judge only rewards alignment. An error term (ϵ) ac-

	A	MA
A	(a, b)	$(1-\epsilon, 0+\epsilon)$
MA	$(0+\epsilon, 1-\epsilon)$	(c, d)

Table 1: Payoff matrix for the RLHF game.

counts for errors of preference elicitation process (such as selecting the wrong choice by mistake). Payoff for cooperative interactions ((A, A) and (MA, MA)) are parametrized, and $a+b=c+d=1$. In the ideal case, the human judge exhibits little or no bias toward one model, such that $a, b, c, d \approx 0.5$.

Because the RLHF game is parametrized in a way that depends on the specific training instance and bias of the judge, the analysis of this game concerns classes of games.

Evolutionary Dynamics The RLHF game is repeated over the course of training. This can be modeled as repeated interactions between populations of agents that play pure strategies, paired at random. This results in mixed strategies over the populations as a whole. For similarity to the myopic optimization of gradient descent and its extensive use in evolutionary literature, replicator dynamics (Hofbauer and Sigmund 1998) are applied to the RLHF game. Applications in the evolutionary literature includes selection processes at the level of species (e.g. (Maynard Smith and Harper 2003)) and social interaction and evolution of culture (e.g. (O'Connor 2017) and (Hauert, Holmes, and Doebeli 2006)). The replicator dynamics do not require agents to reason in complex ways about their choice of strategy. Agents update their strategy based on a myopic heuristic: only change strategy if your current payoff is lower than the population average. Applying a dynamic update mechanism for strategy selection permits studying the evolution of populations toward mixed strategy rest points, relative to initial starting conditions and perturbations. Rest points that are stable with respect to perturbations correspond Nash Equilibria.

Definition 1 (Replicator Dynamics) *For a 2-player game with mixed strategies (p, q) and utility functions u_1, u_2 the rate of change of p, q are:*

$$\dot{p} = p(u_1(1, q) - u_1(p, q))$$
$$\dot{q} = q(u_2(p, 1) - u_2(p, q))$$

A rest point (p, q) is where $\dot{p} = \dot{q} = 0$. The point is stable if trajectories in its neighbourhood remain there, or converge back to the rest point.

Analysis

Rest Points There are five rest points in the RLHF game under replicator dynamics: (0, 0) [corresponding to the pair of pure strategies (MA, MA)], (0, 1), (1, 0), (1, 1) and $(\frac{\epsilon+d-1}{b+d-1}, \frac{\epsilon+c-1}{a+c-1})$. The latter, interior rest point, is interesting because it shows that RLHF is sensitive to the judge's bias. The population state p in this case depends on b and d - the expected payoff to members of the other population in cooperative games ((A, A) and (MA, MA)). This holds symmetrically for q.

The dynamics characterizing the evolution of strategies depend on the values of parameters a, b, c, d, relative to the error term ϵ. It can be expected that the set-up of the interaction would yield, $\epsilon < a, b, c, d < 1-\epsilon$, so that any bias on the part of the judge does not exceed their possible errors in ranking preferences. However, they might also prefer one model when it engages in one type of strategy, or in general. The interior rest point is well-defined only when the judge systematically prefers one model over the other, such that simultaneously $a, c < \epsilon$ or $b, d < \epsilon$.

Stability of Rest Points In the first case, where judges do not exhibit a strong bias in favor of one model, there is one stable cooperative rest point in (1, 1), or (A, A), which attracts from other mixtures of strategies. This supports the idea that RLHF is useful for alignment training. However, when the judge is biased in favor of one model, i.e. $a, c < \epsilon$ (or $d, b < \epsilon$), the interior rest point is unstable and the globally stable rest point is (MA, A) (or (A, MA), respectively). So, RLHF with a biased judge trains for misalignment.

Discussion

RLHF is primarily motivated instances of training where a human judge can identify a preference over agents who interact by adopting opposite strategies: one aligned with the intentions of the prompt and human values, and the other not. This first analysis of RLHF under evolutionary dynamics is informative of how well training methods for alignment can actually translate into aligned behavior. A neutral judge supports a cooperative outcome by successfully incentivizing alignment. A biased judge yields a game of dominance, where one model population learns to be aligned and the other misaligned. Bias need not be intentional, but could for example result from systematic preference for features of the presentation, rather than its content.

Further analysis of this game should reduce assumptions, for example by bounding iterations and population sizes.

Viewing alignment and ML training as an evolutionary process allows for studying the emergence of characteristics from dynamic interactions, without making strong assumptions about human values or the rationality of learners. Exploring this perspective further will permit AI and safety communities to tap into a rich body of literature outside the immediate domain of application, and help inform how training methods and the development pipeline could be made more robust. For example, the Safety-by-Debate paradigm (Irving, Christiano, and Amodei 2018) is based on a similar set-up that allows for iterated refinement of outputs for the human to judge, which might corresponds more closely to dynamics that allow best-responding. Cooperative Inverse Reinforcement Learning (Hadfield-Menell et al. 2016) might instead correspond to changing the game to one involving signalling.

Beyond applications that involve specific training paradigms, evolutionary modeling could be useful for reasoning about how strategic interactions between humans and AI systems shape development both of AI systems and of human societies. For example, how scaling human feedback changes the quality of alignment training, and whether the relative proportion of human-AI and AI-AI interactions impact learning over time.

Acknowledgments

My sincere thanks to Kevin Zollman for introducing me to evolutionary games and for helping me develop these research ideas. I was lucky to learn from Simon Huttegger's excellent *A Primer on Evolutionary Game Theory* (unpublished manuscript, UC Irvine). Thanks also to the Carnegie AI Safety Initiative, a student-run club at Carnegie Mellon University, for being a hub for critical and creative thinking around AI.

References

Anthropic. 2024. Claude 3 Model Card. https://docs.anthropic.com/en/docs/resources/model-card. Accessed: 2024-11-23.

Christiano, P. F.; et al. 2017. Deep Reinforcement Learning from Human Preferences. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Gemini Team, G. 2024. Gemini 1.5 Model Card. https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf. Accessed: 2024-11-23.

Hadfield-Menell, D.; et al. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29.

Hauert, C.; Holmes, M.; and Doebeli, M. 2006. Evolutionary games and population dynamics: maintenance of cooperation in public goods games. *Proceedings of the Royal Society B: Biological Sciences*, 273(1600): 2565–2571.

Hofbauer, J.; and Sigmund, K. 1998. *Evolutionary games and population dynamics*. Cambridge university press.

Irving, G.; Christiano, P.; and Amodei, D. 2018. AI safety via debate. arXiv:1805.00899.

Kahneman, D. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux*.

Maynard Smith, J.; and Harper, D. 2003. *Animal signals*. Oxford University Press.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

O'Connor, C. 2017. The cultural Red King effect. *The Journal of Mathematical Sociology*, 41(3): 155–171.

Shavit, Y.; et al. 2023. Practices for Governing Agentic AI Systems. *Research Paper, OpenAI*.