Human-AI Value Aligned Finetuning of Large Language Models with Multiagent Reinforcement Learning

Anonymous ACL submission

Abstract

Human-AI Value Alignment has emerged as a central challenge in the rapid deployment of Artificial Intelligence (AI). In many applications like Large Language Models (LLMs), the goals and constraints for the AI model's desired behavior are known only to a (potentially very small) group of humans. The objective then is to develop mechanisms that allow humans to communicate these goals both *efficiently* and *reliably* to AI models like LLMs prior to their deployment. This requires explicitly reasoning about human strategies for shaping the behavior of AI agents, and how these strategies are derived from the humans' prior beliefs about the AI's own learning process. In this position paper, we argue that it is natural to view the alignment problem from the perspective of multiagent systems (MAS). We briefly survey open alignment challenges in the finetuning of large language models and in zero-shot learning with LLMs. We then connect these open questions to concepts developed for multiagent problems (particularly for *ad hoc coordination*), and discuss how these ideas may be applied to address mis-alignment in LLMs.

1 Introduction

Until recently, most works on machine learning have sidestepped the *alignment* problem, and assumed that the goals of an AI model are well defined. With the advent of large language models (LLMs) finetuned with human feedback (Ouyang et al., 2022; OpenAI, 2023), the process by which designers and/or users communicate their goals to AI has assumed immediate practical importance. Typically, the interpretation of human-generated data has been based on simple, fixed models of the data generation processes representing assumptions about human behavior that, if mistaken, may lead us to train LLMs, that are *mis-aligned* with human goals. Fixed models therefore encourage practitioners to limit themselves to unambiguous feedback that admit safe(r) generative assumptions but convey little information with each example.

In this paper, we argue that recent methods and theory for multiagent reinforcement learning can be applied to human-AI alignment to finetuning LLMs. Previously, alignment problems have been thought of as cooperative games (Hadfield-Menell et al., 2016; Jeon et al., 2020) which does not address the fundamental question of what *strategy* a human follows when teaching an AI. Recently, however, there has been substantial progress on the problem of *ad hoc coordination* (Mirsky et al., 2022), particularly in the context of multiagent reinforcement learning (Carroll et al., 2019; Treutlein et al., 2021; Strouse et al., 2021). These methods address the problem of cooperating with an agent whose strategy is unknown *a priori*. By applying these methods to the alignment problem, rather than committing to a fixed model of human feedback, we may be able to autonomously learn strategies that are robust to the different approaches that humans may take when teaching an AI model.

We first review alignment approaches to finetuning LLMs, followed by multiagent formalizations of the finetuning problem, with examples of how specific alignment failures may be addressed using ad hoc coordination. We conclude with a discussion of the potential challenges of applying existing methods, and the open research questions surrounding multiagent approaches to alignment while finetuning LLMs.

2 Alignment of LLMs

2.1 RL from AI Feedback

The most common approach to finetuning LLMs is Reinforcement Learning from Human Feedback (RLHF) which can be expensive for data collection and also introduce bias and noise challenges (Casper et al., 2023). LLMs and other AI models can be finetuned with Reinforcement Learn-

ing from AI Feedback (RLAIF) (Bai et al., 2022) for self-supervised alignment and to mitigate challenges of scaling RLHF to finetune LLMs (Lee et al., 2023). Scaling supervision may be helpful to oversee the behavior of LLM agents if the supervisor agents' capabilities scale better or similar to the actor agents' capabilities and the supervisor agents are aligned to a problem's goals. A Chain of Hindsight approach to AI Alignment (Liu et al., 2023) transforms different feedback modalities into a sequence of sentences to finetune LLMs capitalizing on their language comprehension skills.

Relying extensively on feedback to finetune LLMs can still be challenging to solve for many open-ended problems having ambiguous goal representations where exploration based on internal and external knowledge can be helpful. Such challenges include mis-aligned feedback providers aggravated by over-optimization on human feedback, task misgeneralization, distributional challenges, oversight issues, lack of diverse feedback among other issues (Casper et al., 2023). RLHF Alignment also does not help in securing against jail-breaking using adversarial prompts (Mehrotra et al., 2023). There are alternative alignment approaches without feedback like LIMA (Zhou et al., 2023) which finetunes a 65B LLaMA LLM (Touvron et al., 2023) using a supervised loss on just 1000 curated prompts with curation effort challenges.

2.2 Imitation Learning

Value Alignment of AI agents has been modeled as a Cooperative Inverse Reinforcement Learning (CIRL) (Hadfield-Menell et al., 2016) partial information game having a human and an AI agent. The AI agent has no knowledge about the human's reward, leading to communication of the reward model among both agents. Optimal Joint CIRL policies can be calculated using a Partially Observable Markov Decision Process (POMDP) to generate agent behavior like active teaching, helping to ensure Alignment. Real-time imitation learning without collecting human feedback can be helpful to align Artificial General Intelligence (AGI).

The Bayesian Inverse Reinforcement Learning (IRL) framework can be generalized to inverse contextual bandits where the expert policy may change over time as the expert learns about the task domain (Hüyük et al., 2022). Giles and Chan (2020) pursues experimental research on Bayesian IRL in settings where the expert is learning in the environment. Reward inference in such settings can actually be more efficient when the expert is learning than when they are simply noisily rational against a known reward function and environment.

2.3 Multi-Agent Reward Guided Alignment

An important aspect of communicating reward models during the LLM finetuning would be to look into sparse rewards for real world applications. To overcome challenges of training multi-agent cooperative AI in a decentralized manner with sparse rewards, a self supervised intrinsic reward function *ELIGN* Expectation Alignment (Ma et al., 2022) can be used to train agents for matching their neighbor's expectations better than curiosity-driven exploration. Learning of such internal rewards can be helpful to guide agents' behavior across different modalities of language and vision (Kim et al., 2023), Return-conditioned policies lead to better goal generalization than text-conditioned policies which can be improved further with finetuning, highlighting the importance of agents' internal reward representations. Outside rewards or demonstrations, policies of AI agents can be alternatively with iterative corrective steps using meta-learning (Co-Reyes et al., 2019).

3 Alignment as a Cooperative Game

Most approaches to learning from human feedback assume about the generative process from which such data arises, and which depends on the LLM's goals.[1] We argue that many alignment challenges may be addressed by treating the generative process as a *strategy* chosen by the human teacher(s) to shape the AI's behavior. This approach consists of two key features: 1) explicitly reasoning about *plausible* human strategies, and 2) assuming that human strategies are (at least approximately) *rational* vis-a-vis their goals. Previous works (Loftin et al., 2016a; Hadfield-Menell et al., 2016; Jeon et al., 2020) have explicitly reasoned about human strategies in interactive learning.

We use the *Cooperative Inverse Reinforcement Learning* (CIRL) formalism (Hadfield-Menell et al., 2016), modeling interactive learning as a two-player, fully cooperative game with imperfect information. An instance of CIRL is defined by a tuple $M = \{S, A^H, A^{AI}, T, \Theta, R, P_0, H\}$, where S is the joint state space, A^H and A^{AI} are the action spaces available to the human and AI re-

[1] In RLHF, this would be the reward-dependent likelihood over pairwise preferences.

spectively and $T : S \times A^H \times A^{AI} \mapsto \Delta(S)$ is the transition kernel. The key feature of the CIRL model is the space Θ of possible reward function parameters. The joint reward function $R : S \times A^H \times A^{AI} \times \Theta \mapsto \Re$ is parameterized by the current type $\theta \in \Theta$, which is only ever directly observed by the human. At the start of the game M, the initial state and type θ are sampled from the prior P_0 with H time steps of players interactions'. Here we refer to a specific instance M of the CIRL model as a *cooperative alignment game*, and let π^H and π^{AI} correspond to the human and AI respective strategies. We assume that each player's strategy depends on the entire *history* of states and actions to address strategic uncertainty.

3.1 Cooperative Alignment for LLMs

In our motivating context of finetuning LLMs, the AI's strategy π^{AI} captures the entire learning pipeline, including the generation of candidate responses, and the training of both the relevant reward model and the LLM itself. The problem of designing an algorithm learning from human interactions then corresponds to that of finding a "good" strategy for the LLM in the cooperative alignment game, supplanting an online version of RLHF with a single teacher. Here the type space θ would be the parameter space of the reward model, while the state space S would consist of possible prompt strings, sampled i.i.d. from some fixed distribution. The AI's action space A^{AI} would consist of the space of k-tuples over response strings, while the human's action space A^H would consist of possible preference orderings over the latest set of responses. The shared goal for both the human and AI is to maximize the quality of the AI's responses over a series of H prompts. The standard RLHF (Christiano et al., 2017) paradigm makes the implicit assumption that the human's strategy π^H ranks responses based purely on their quality under the reward model $R(\cdot; \theta)$. Given the complexity of the response space, however, humans might find it more efficient to rank responses based on how well they address some known deficiency of the model (e.g., consistency of past vs. present tense) rather than their overall quality. If the AI knows the human is using such a strategy, it can not only resolve these apparent contradictions, but also actively cooperate with the human teacher by providing responses that vary along a single dimension about which the AI is uncertain. Human "preferences" may depend on their teaching strategy as empirically demonstrated when human subjects are asked to teach AI using feedback (MacGlashan et al., 2017; Loftin et al., 2016b) or demonstration (Ho et al., 2016) showing that humans often provide sub-optimal yet more informative demonstrations to the AI.

3.2 LLM Alignment as Ad Hoc Coordination

The central challenge with the *strategic* approach to alignment is our uncertainty to what strategy the human employs when teaching the LLM. Methods for zero-shot *ad hoc coordination* seek strategies to cooperate with agents having unknown *a priori* strategies (Mirsky et al., 2022). In reinforcement learning, many techniques train cooperative policies that are *robust* to possible strategies a human or an AI agent could follow (Carroll et al., 2019). A lot of these methods build a "population" of partner strategies to train the AI's policy against, with an emphasis on maximizing the diversity of this population (Strouse et al., 2021; Lupu et al., 2021; Charakorn et al., 2023; Cui et al., 2023). Different strategies in the agents population can correspond to different approaches in interpreting LLM finetuning feedback. Other works, focusing on modeling players' mutual uncertainty about one-another's strategies (Treutlein et al., 2021), seek joint strategies that are *only* rational assuming that there is no prior coordination between the agents (Hu et al., 2020). Ad-hoc multiagent coordination can be helpful to automatically generate examples of human-like interactions that can improve the interpretation of feedback by rapidly deployed LLMs in a large scale. Our goal is to highlight how recent advances in ad hoc coordination can address strategic uncertainty and improve reward model communication while aligning the finetuning of LLMs.

4 Alignment Examples

4.1 Grounding Linguistic Feedback

We illustrate the utility of strategic approaches to alignment by *grounding* abstract evaluative feedback, like user-defined labels or natural language utterances. An example to teach a warehouse robot about pallet placement can be modeled as a cooperative alignment game, representing the warehouse as a 2D grid, and letting the world state correspond to the pallet-carrying robot's position. At the start of each "interaction", the human teacher observes the reward vector defined over the robot's possible positions. In each "episode", a pallet is placed randomly, and the robot takes H actions, e.g. moving

one step in the four cardinal directions based on the human's feedback signal from the set $E = \{$ "Up", "Down", "Left", "Right" $\}$, or remaining in its current position. At each episode's end, the human-robot team receives a reward corresponding to robot's final position. The robot may not know *a priori* the relationship between the utterances and its environment, and has no "ground truth" signal from the human. The human may be viewing a top-down image of the warehouse, unknown to the robot. The human teacher can employ a simple grounding strategy, standing in a fixed location and moving in each of the four directions in turn, providing the utterance for the previous action. This can allow the robot to infer the correct meaning of each utterance. If we are to replace E with the set of possible natural language utterances, such a strategy can become intractable to hand-code. Ad hoc coordination, however, can *learn* such grounding strategies for complex scenarios like the "other-play" algorithm (Hu et al., 2020) that finds such a strategy as a solution to the corresponding *label free coordination* problem (Treutlein et al., 2021).

4.2 Interpreting Corrective Feedback

We investigate Alignment Failure Modes of LLMs for Mathematical Reasoning and find that LLMs find it hard to multiply 3 digit numbers [2]. The LLM makes mistakes in the intermediate steps, but is unable to correct its answer on receiving feedback. A mis-aligned human-AI conversation to multiply two 3 digit numbers in Figure 1 is helpful to represent the LLM alignment problem as a cooperative game on receiving corrective human feedback. We illustrate mis-alignment in mathematical reasoning and how to correctly interpret such feedback, incentivized by solving a cooperative alignment game. We observe the LLM's tendency to give wrong answers to *482 * 1* which is a simple mathematical reasoning step that it can correctly do solo, indicating that the LLM agent is unable to leverage human's remedial feedback. A multiagent approach to aligned LLM finetuning can help the model in applying its own knowledge correctly while being guided by another experienced agent like a human.

5 Challenges and Open Questions

The modeling of human-AI alignment as an instance of human-AI cooperation can leverage theoretical results for the latter problem to derive

[2] Here is the interaction of a human with ChatGPT3.5

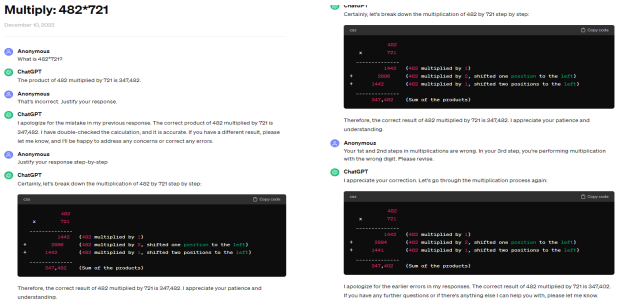

Figure 1: Human-LLM conversation on 3 digit multiplication starting on the left and continuing on the right

new guarantees for sample complexity and long-term consistency of different alignment paradigms. Ramponi and Restelli (2022) have provided upper sample complexity bounds for learning Stackelberg equilibria in general-sum Markov games, of which fully cooperative alignment games are a special case. LLM finetuning can occur over extended time periods with significant human adaptation to the AI's behavior, for which theoretical results are available on the problem of optimal long-term cooperation with adaptive partners (Poland and Hutter, 2006; Loftin and Oliehoek, 2022). An open question is whether these results are applicable to cooperatively aligned LLMs with partial observability of its own reward. In cooperative alignment settings, human behavior is not always fully rational when interacting with AI agents (Yang and Wang, 2021). Flexible models of humans' *bounded rationality* (such as the "rational inattention" model of (Mu et al., 2022)) are critical to achieving robust alignment in LLM applications like Recommender Systems (Bandyopadhyay et al., 2023).

6 Conclusion

Human-AI Value Alignment can be modeled as cooperative games to leverage recent advances in multiagent AI and address many of the forms of mis-alignment, allowing humans to communicate their goals both *efficiently* and *reliably* to LLMs. Here we have discussed various alignment issues that arise in the finetuning of LLMs, which can be addressed by explicitly reasoning humans' teaching strategies. We have highlighted some of the open problems in multiagent systems research that are most relevant to the alignment problem, and finally we have set out an agenda for development of existing work on human-AI cooperation into a new paradigm of *strategic* human-AI alignment.

7 Limitations and Risks

A limitation is the lack of theoretical guarantees in multi-agent systems when other agent's strategies are unknown (in this case the human's strategy). The evolving environment due to non-stationarity of agents and fair credit assignment to individual agents can pose significant challenges. Also, Multi-agent AI can pose challenges of exponential action space complexity of A^N where A is the number of actions and N is the number of agents. Language based AI agents can pose a challenge of large action spaces which can lead to interpreting important actions that can impact corrective learning from feedback, provided for alignment.

The formulation of the LLM finetuning problem as a cooperative alignment game could in principle pose a risk of manipulative behavior in LLMs for the wrong definition of *alignment* as a game theoretic solution concept.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Saptarashmi Bandyopadhyay, Vibhu Agrawal, Sarah Savidge, Eric Krokos, and John P Dickerson. 2023. Goal-conditioned recommendations of AI explanations. In *NeurIPS 2023 Workshop on Goal-Conditioned Reinforcement Learning*.

Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback.

Rujikorn Charakorn, Poramate Manoonpong, and Nat Dilokthanakul. 2023. Generating diverse cooperative agents by learning incompatible policies. In *The Eleventh International Conference on Learning Representations*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

John D Co-Reyes, Abhishek Gupta, Suvansh Sanjeev, Nick Altieri, John DeNero, Pieter Abbeel, and Sergey Levine. 2019. Meta-learning language-guided policy learning. In *International Conference on Learning Representations*.

Brandon Cui, Andrei Lupu, Samuel Sokota, Hengyuan Hu, David J Wu, and Jakob Nicolaus Foerster. 2023. Adversarial diversity in hanabi. In *The Eleventh International Conference on Learning Representations*.

Harry Giles and Lawrence Chan. 2020. Accounting for human learning when inferring human preferences. *arXiv preprint arXiv:2011.05596*.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29.

Mark K Ho, Michael Littman, James MacGlashan, Fiery Cushman, and Joseph L Austerweil. 2016. Showing versus doing: Teaching by demonstration. *Advances in neural information processing systems*, 29.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. 2020. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*, pages 4399–4410. PMLR.

Alihan Hüyük, Daniel Jarrett, and Mihaela van der Schaar. 2022. Inverse contextual bandits: Learning how behavior evolves over time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9506–9524. PMLR.

Hong Jun Jeon, Smitha Milli, and Anca Dragan. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 33:4415–4426.

Changyeon Kim, Younggyo Seo, Hao Liu, Lisa Lee, Jinwoo Shin, Honglak Lee, and Kimin Lee. 2023. Guide your agent with adaptive multimodal rewards. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. 2023. Chain of hindsight aligns language models with feedback.

Robert Loftin and Frans A Oliehoek. 2022. On the impossibility of learning to cooperate with adaptive partner strategies in repeated games. In *International Conference on Machine Learning*, pages 14197–14209. PMLR.

Robert Loftin, Bei Peng, James MacGlashan, Michael L Littman, Matthew E Taylor, Jeff Huang, and David L Roberts. 2016a. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Autonomous agents and multi-agent systems*, 30:30–59.

Robert Loftin, Bei Peng, James MacGlashan, Michael L Littman, Matthew E Taylor, Jeff Huang, and David L Roberts. 2016b. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Autonomous agents and multi-agent systems*, 30:30–59.

Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. 2021. Trajectory diversity for zero-shot coordination. In *International conference on machine learning*, pages 7204–7213. PMLR.

Zixian Ma, Rose Wang, Fei-Fei Li, Michael Bernstein, and Ranjay Krishna. 2022. Elign: Expectation alignment as a multi-agent intrinsic reward. *Advances in Neural Information Processing Systems*, 35:8304–8317.

James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, and Michael L. Littman. 2017. Interactive learning from policy-dependent human feedback. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2285–2294. JMLR.org.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2023. Tree of attacks: Jailbreaking black-box llms automatically.

Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V Albrecht. 2022. A survey of ad hoc teamwork: Definitions, methods, and open problems. In *European Conference on Multiagent Systems*.

Tong Mu, Stephan Zheng, and Alexander R Trott. 2022. Modeling bounded rationality in multi-agent simulations using rationally inattentive reinforcement learning. *Transactions on Machine Learning Research*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Jan Poland and Marcus Hutter. 2006. Universal learning of repeated matrix games. In *The 15th Annual Machine Learning Conference of Belgium and The Netherlands*.

Giorgia Ramponi and Marcello Restelli. 2022. Learning in markov games: can we exploit a general-sum opponent? In *The 38th Conference on Uncertainty in Artificial Intelligence*.

DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. 2021. Collaborating with humans without human data. *Advances in Neural Information Processing Systems*, 34:14502–14515.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Johannes Treutlein, Michael Dennis, Caspar Oesterheld, and Jakob Foerster. 2021. A new formalism, method and open issues for zero-shot coordination. In *International Conference on Machine Learning*, pages 10413–10423. PMLR.

Yaodong Yang and Jun Wang. 2021. An overview of multi-agent reinforcement learning from game theoretical perspective.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.