The Expertise Problem: Learning from Specialized Feedback

Oliver Daniels-Koch
Center for Human-Compatible AI
UC Berkeley
danielskoch.oliver@gmail.com

Rachel Freedman
Center for Human-Compatible AI
UC Berkeley
rachel.freedman@berkeley.edu

Abstract

Reinforcement learning from human feedback (RLHF) is a powerful technique for training agents to perform difficult-to-specify tasks. However, human feedback can be noisy, particularly when human teachers lack relevant knowledge or experience. Levels of expertise vary across teachers, and a given teacher may have differing levels of expertise for different components of a task. RLHF algorithms that learn from multiple teachers therefore face an *expertise problem*: the reliability of a given piece of feedback depends both on the teacher that it comes from and how specialized that teacher is on relevant components of the task. Existing state-of-the-art RLHF algorithms assume that all evaluations come from the same distribution, obscuring this inter- and intra-human variance, and preventing them from accounting for or taking advantage of variations in expertise. We formalize this problem, implement it as an extension of an existing RLHF benchmark, evaluate the performance of a state-of-the-art RLHF algorithm, and explore techniques to improve query and teacher selection. Our key contribution is to demonstrate and characterize the expertise problem, and to provide an open-source implementation for testing future solutions. [1]

1 Introduction

AI systems are typically trained to maximize a pre-specified objective but, for many tasks, defining an objective that fully captures the intentions of the designer is prohibitively difficult. One promising alternative to manually specifying objectives is reinforcement learning from human feedback (RLHF). In RLHF, human teachers compare pairs of trajectories against each other, the comparisons are used to train a reward model, and the reward model provides reward signal to further train the reinforcement learning agent [1–3]. RLHF has proven effective for training models on tasks that require human judgement, including summarizing texts and promoting helpfulness, honesty, and harmlessness in large language models [4–7].

RLHF methods assume all human feedback comes from a single human teacher. However, these methods typically require querying many teachers to gather sufficient training data, and all teachers are not equal - they vary in their ability to evaluate system behavior. For example, Ziegler et al. [8] finds that crowd sourced feedback often disagrees with the paper authors' feedback. An individual teacher's ability can also vary across queries. For example, Stiennon et al. [4] trains human teachers at different summarization domains, producing evaluators that "specialize" in evaluating either Reddit or CNN summaries. In the future, we expect RLHF to be applied to increasingly complex and compound tasks, where evaluators vary in their expertise across task components. RLHF methods may need to actively select teachers based on their expertise to ensure reliable feedback on different tasks. We call the problem of learning from multiple, specialized teachers the *expertise problem*.

[1] Code for all experiments is available here.

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Our contributions are as follows. We propose an abstract problem formulation of the expertise problem. Under this formulation, we test PEBBLE [2], a state of the art RLHF algorithm. We find that naive extensions of PEBBLE perform poorly, but that simple modifications, including selecting the most expert teacher and making intra-domain queries, can significantly improve performance. Our key contribution is to formalize and characterize the expertise problem, and to demonstrate the insufficiency of current methods. We hope that this formalism, demonstration, and open-source implementation will pave the way for more advanced RLHF methods that are robust to teacher variance and specialization.

2 Preliminaries and Related Work

2.1 Reinforcement Learning from Human Feedback

Reinforcement learning from human feedback (RLHF) uses comparison feedback to learn a model of a reward function, and uses the learned reward function to train reinforcement learners. As in traditional RL, at each timestep t the learner observes a state s_t, takes an action a_t, and transitions to a new state s_{t+1}. These interactions produce k-length trajectories $\sigma \in \Sigma$, where $\sigma = ((s_1, a_1), ..., (s_k, a_k)) \in (\mathcal{S} \times \mathcal{A})^k$.

In RLHF, these trajectories are presented to human teachers as binary comparisons (σ_1, σ_2), and teachers select which trajectory they prefer [1, 2]. These teacher preference labels are represented by a distribution μ over $\{1, 2\}$ and stored in a database $\mathcal{D}$. RLHF methods typically model humans as Boltzmann-rational decision-makers, where the noisiness of their decisions is modulated by a "rationality" parameter β [9, 10]. The probability that a Boltzmann-rational decision-maker prefers trajectory σ_1 to σ_2 is:

$$P[\sigma_1 \succ \sigma_2; \beta] = \frac{\exp(\beta \cdot r(\sigma_1))}{\exp(\beta \cdot r(\sigma_1)) + \exp(\beta \cdot r(\sigma_2))} \tag{1}$$

where r is the "true" reward function ($r(\sigma) = \sum_{n=1}^{k} r(s_n, a_n)$). A reward model $\hat{r}$ is trained to approximate the underlying reward function by minimizing the cross entropy between the teacher labels and the distribution $\hat{P}$ (given by substituting $\hat{r}$ for r):

$$loss(\hat{r}) = \sum_{(\sigma_1, \sigma_2, \mu) \in \mathcal{D}} \mu(1) \log \hat{P}[\sigma_1 \succ \sigma_2] + \mu(2) \log \hat{P}[\sigma_2 \succ \sigma_1] \tag{2}$$

The agent is simultaneously trained using the approximated reward function $\hat{r}$.

Christiano et al. [1] uses A2C [11] and TRPO [12] to train the agents and preferentially samples queries that the reward model has greater uncertainty over. Lee et al. [2] proposes PEBBLE, which uses SAC [13] in conjunction with a self-supervised exploration bonus and trajectory relabeling to improve upon these results. Since PEBBLE is the current state-of-the-art, we'll use it in our experiments.

Lee et al. [3] benchmarks RLHF algorithms, using algorithmic teacher models to generate synthetic feedback data at scale. Crucially, they find that the optimal query sampling strategy varies across teacher feedback models, suggesting that identifying and leveraging differences between teachers may improve learner performance. Defining the teacher models algorithmically allows them to isolate these effects of differences in teacher decision-making, so we will use this strategy as well.

2.2 Estimating Expertise

Prior work in imitation learning and inverse reinforcement learning has estimated the quality of demonstrations to extrapolate to better performance [14–17]. Beliaev et al. [18] estimates the expertise of demonstrators for improved imitation learning, and prior work in supervised learning estimates the expertise of human labelers to better aggregate labels from different sources [19–21]. However, to the best our knowledge, no prior work has investigated teacher selection in the context of reinforcement learning from human feedback, as we do here.

3 The Expertise Problem

3.1 Problem Formulation

We define expertise as a teacher's ability to reliably asses a given trajectory according to their underlying preferences. Expertise can vary across both teachers and regions of the state space. We model the teacher as a Boltzmann-rational decision-maker whose rationality varies as a function of queries. Formally, for a given teacher we define a β-function $\beta : \Sigma \times \Sigma \to \mathbb{R}$. The probability that a teacher with β-function β selects σ_1 over σ_2 is:

$$P[\sigma_1 \succ \sigma_2; \beta] = \frac{\exp(\beta(\sigma_1, \sigma_2) \cdot r(\sigma_1))}{\exp(\beta(\sigma_1, \sigma_2) \cdot r(\sigma_1)) + \exp(\beta(\sigma_1, \sigma_2) \cdot r(\sigma_2))} \tag{3}$$

Teachers are defined by their β-functions $\beta_1, ..., \beta_m$, indicating their areas of specialization or levels of expertise. Critically, the RLHF algorithm doesn't have direct access to these β-functions, and must learn which teachers give the most reliable feedback on which queries. This is the *expertise problem*: selecting teachers β_i and queries (σ_1, σ_2) to maximize performance against the ground-truth reward function.

3.2 Defining Beta

We model β-functions as gaussian kernels in concatenated mappings of trajectory space. We define a mapping function $g : (\mathcal{O} \times \mathcal{A})^k \to \mathbb{R}^g$, and for each teacher set a centroid with center $c \in \mathbb{R}^{2g}$, width $b \in \mathbb{R}^{2g}$ and scale $a \in \mathbb{R}$, such that

$$\beta(\sigma_1, \sigma_2) = a \cdot \exp(||[b([g(\sigma_1), g(\sigma_2)] - c)||^2) \tag{4}$$

Centroids are placed evenly across the range of g, the dimension(s) of the state space expertise varies over. We set a to 1. b is set such that an agent trained from a teacher with constant $\beta(\cdot, \cdot) = \min\left\{\overline{\beta} | \forall \sigma_j \in \Sigma \, \exists \beta_i \, ; \beta_i(\sigma_j, \sigma_j) = \overline{\beta}\right\}$ reliably converges to near optimal performance. Note that under this definition there are regions of query space that have low β values for all teachers. We call queries in these regions *inter-domain queries*, as they lie outside any teacher's domain of expertise.

4 Experiments

We evaluate variants of PEBBLE [2] on the expertise problem. We compare vanilla PEBBLE, where teachers are selected randomly, to PEBBLE augmented with max-β teacher selection, and compare disagreement, similarity, and hybrid query sampling. We train each on feedback from four synthetic teachers. We use two Deepmind control suite [22] environments, cartpole balance and walker walk, chosen for their simplicity and use as benchmarks in prior work [3]. Results are the final episode rewards averaged over 5 runs. See A.1 for learning curves.

4.1 Teacher Selection

For each query, the RLHF algorithm must select a teacher to apply the query to (or make no selection, in which case a teacher is chosen randomly). Using the above β function, we compare no selection (uniform) and max-β selection. Results in Table 1 show that max-β selection yields significant performance increases over the compared to selecting teachers uniformly.

4.2 Query Selection

When prompting a teacher for feedback, the RLHF algorithm must select which trajectories to have the teacher compare – the problem of query selection. Inter-domain queries are queries that have low β values for all teachers. To reduce the probability of inter-domain queries, we first test *similarity sampling*, where queries are sampled to minimize the euclidean distance between mappings of trajectories in the query:

	Cartpole balance	Walker walk
Unif + Dis	396 ± 338	259 ± 186
Max-β + Dis	756 ± 334	890 ± 95
Max-β + Sim	580 ± 262	$\mathbf{950} \pm 21$
Max-β + Hybrid	$\mathbf{794} \pm 196$	939 ± 71

Table 1: Mean and standard deviation of final episode rewards received by PEBBLE trained with feedback from four synthetic teachers using a variety of teacher selection and query sampling methods. Across both environments, max-β teacher selection significantly improves performance over uniform teacher selection. Similarity query sampling improves performance in walker walk, and hybrid similarity-disagreement sampling improves performance in cartpole balance. Vanilla similarity query sampling hurts performance in cartpole balance, possibly to do a bias towards uninformative queries.

$$\min_{(\sigma_1,\sigma_2)\in\mathcal{D}} ||g(\sigma_1) - g(\sigma_2)|| \tag{5}$$

To combat a bias toward uninformative queries, we propose combining similarity sampling with *disagreement sampling*, which preferentially samples queries that the networks in the reward model ensemble predict different rewards for. The resulting *hybrid sampling* method computes and normalizes both disagreement and similarity metrics across a batch of queries, then samples queries to maximize the difference in the normalized disagreement and similarity scores.

Results of varying sampling methods combined with max-β teacher selection are shown in Table 1. On cartpole balance, we find similarity sampling degrades performance while hybrid sampling improves performance. On walker walk performance of all three methods is comparable, with both similarity and hybrid sampling narrowly improving over disagreement sampling.

5 Discussion

We model expertise as a query dependent function β in which the error rate of teachers depends on the queries they evaluate. We find that vanilla PEBBLE is significantly outperformed by expertise-aware max-β teacher selection on two continuous control environments from an RLHF benchmark suite. The performance gap between vanilla PEBBLE and max-β teacher selection demonstrates the potential reward that PEBBLE and other such RLHF methods sacrifice by treating the feedback from all teachers equally, when teachers actually have varying levels of expertise. Of course, always selecting the teacher with the highest level of expertise on the current query will not be feasible with real humans, since this information is implicit in the human decision-making process and not directly accessible to the algorithm. Future work should use the formalism and implementation we have provided here to investigate methods that estimate teacher expertise and incorporate this information into teacher selection.

Limitations and Future Work First, while generating synthetic feedback allows us to evaluate RLHF algorithms at scale, it also requires us to manually define β-functions representing domains of expertise. Future work could replace these synthetic teacher models with real specialized humans, perhaps by using human experts on tasks that require special knowledge to evaluate, pre-training different teachers to evaluate different tasks, or artificially limiting the task information that different teachers have access to. Second, thus far we have experimented on low dimensional environments which require relatively small models. Future work could extend this analysis to more complex tasks, such as language generation. Finally, future work should investigate solutions to the expertise problem proposed here. Preliminary investigation suggests this is challenging, but promising directions include inferring teacher expertise from variation in responses to similar queries, and training neural network models of teacher expertise alongside reward models.

References

[1] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In I. Guyon,

U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

[2] Kimin Lee, Laura Smith, and Pieter Abbeel. PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training, June 2021. URL http://arxiv.org/abs/2106.05091. arXiv:2106.05091 [cs].

[3] Kimin Lee, Laura Smith, Anca Dragan, and Pieter Abbeel. B-Pref: Benchmarking Preference-Based Reinforcement Learning, November 2021. URL http://arxiv.org/abs/2111.03026. arXiv:2111.03026 [cs].

[4] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. *arXiv:2009.01325 [cs]*, October 2020. URL http://arxiv.org/abs/2009.01325. arXiv:2009.01325.

[5] Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively Summarizing Books with Human Feedback. *arXiv:2109.10862 [cs]*, September 2021. URL http://arxiv.org/abs/2109.10862. arXiv: 2109.10862.

[6] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A General Language Assistant as a Laboratory for Alignment, December 2021. URL http://arxiv.org/abs/2112.00861. arXiv:2112.00861 [cs].

[7] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. URL http://arxiv.org/abs/2203.02155. arXiv:2203.02155 [cs].

[8] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-Tuning Language Models from Human Preferences, January 2020. URL http://arxiv.org/abs/1909.08593. arXiv:1909.08593 [cs, stat].

[9] Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. Modeling interaction via the principle of maximum causal entropy. In *ICML*, 2010.

[10] Hong Jun Jeon, Smitha Milli, and Anca D. Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning, December 2020. URL http://arxiv.org/abs/2002.04833. arXiv:2002.04833 [cs].

[11] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous Methods for Deep Reinforcement Learning, June 2016. URL http://arxiv.org/abs/1602.01783. arXiv:1602.01783 [cs].

[12] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust Region Policy Optimization, April 2017. URL http://arxiv.org/abs/1502.05477. arXiv:1502.05477 [cs].

[13] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor, August 2018. URL http://arxiv.org/abs/1801.01290. arXiv:1801.01290 [cs, stat].

[14] Daniel S. Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations, July 2019. URL http://arxiv.org/abs/1904.06387. arXiv:1904.06387 [cs, stat].

[15] Letian Chen, Rohan Paleja, and Matthew Gombolay. Learning from Suboptimal Demonstration via Self-Supervised Reward Regression, November 2020. URL http://arxiv.org/abs/2010.11723. arXiv:2010.11723 [cs].

[16] Songyuan Zhang, Zhangjie Cao, Dorsa Sadigh, and Yanan Sui. Confidence-Aware Imitation Learning from Demonstrations with Varying Optimality, January 2022. URL http://arxiv.org/abs/2110.14754. arXiv:2110.14754 [cs].

[17] Zhangjie Cao and Dorsa Sadigh. Learning from Imperfect Demonstrations from Agents with Varying Dynamics, March 2021. URL http://arxiv.org/abs/2103.05910. arXiv:2103.05910 [cs].

[18] Mark Beliaev, Andy Shih, Stefano Ermon, Dorsa Sadigh, and Ramtin Pedarsani. Imitation Learning by Estimating Expertise of Demonstrators, June 2022. URL http://arxiv.org/abs/2202.01288. arXiv:2202.01288 [cs].

[19] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://papers.nips.cc/paper/2009/hash/f899139df5e1059396431415e770c6dd-Abstract.html.

[20] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning From Crowds. *Journal of Machine Learning Research*, 11 (43):1297–1322, 2010. ISSN 1533-7928. URL http://jmlr.org/papers/v11/raykar10a.html.

[21] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The Multidimensional Wisdom of Crowds. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://papers.nips.cc/paper/2010/hash/0f9cafd014db7a619ddb4276af0d692c-Abstract.html.

[22] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind Control Suite. *arXiv:1801.00690 [cs]*, January 2018. URL http://arxiv.org/abs/1801.00690. arXiv: 1801.00690.

[23] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of AI ingenuity, 2020. URL https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity.

[24] Jack Clark and Dario Amodei. Faulty reward functions in the wild, Dec 2016. URL https://openai.com/blog/faulty-reward-functions/.

[25] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction, November 2018. URL http://arxiv.org/abs/1811.07871. arXiv:1811.07871 [cs, stat].

[26] Paul Christiano. What failure looks like, 2019. URL https://www.lesswrong.com/posts/HBxe6wdjxK239zajf/what-failure-looks-like.

[27] Dan Hendrycks and Mantas Mazeika. X-Risk Analysis for AI Research, July 2022. URL http://arxiv.org/abs/2206.05862. arXiv:2206.05862 [cs].

[28] Jan Leike. A minimal viable product for alignment, March 2022. URL https://aligned.substack.com/p/alignment-mvp.

[29] Jan Leike. Why I'm excited about AI-assisted human feedback, March 2022. URL https://aligned.substack.com/p/ai-assisted-human-feedback.

A Appendix

A.1 Learning Curves

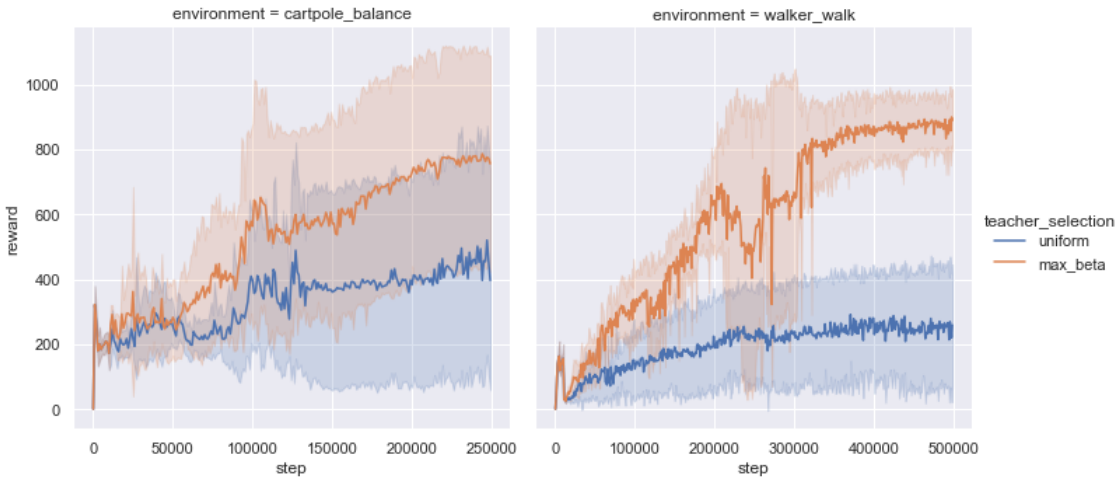

Figure 1: **Teacher Selection**: Learning curves from PEBBLE trained from synthetic feedback given by four teachers with different β-functions (domains of expertise). Solid lines and shaded regions are the averages and standard deviations taken over five runs. Learning from the best available teacher (highest β) produces a large, stable performance gain over uniform teacher selection.

Figure 2: **Query Sampling**: Learning curves from PEBBLE trained from synthetic feedback given by four teachers with different β-functions (domains of expertise). Solid lines and shaded regions are the averages and standard deviations taken over five runs. Max-β teacher selection is used for each run. On cartpole balance, similarity sampling produces an early boost in performance which later degrades, possibly due to a bias toward uninformative queries. Hybrid sampling maintains the early boost, but stabilizes, outperforming disagreement sampling. On walker walk there are no large differences in performance, though hybrid sampling produces a particularly stable learning curve.

B Existental Risk Analysis

B.1 Long-Term Impact on Advanced AI Systems

In this section, we analyze how this work shapes the process that will lead to advanced AI systems.

1. **Overview.** How is this work intended to reduce existential risks from advanced AI systems?
   Answer: Improved performance on the expertise problem may generically improve our capability to train ML models using human feedback (e.g. by increasing sampling efficiency or expanding the domain of tasks agents can learn). Given the demonstrated difficulty of specifying correct objectives [23–25], and the potentially catastrophic results of objective misspecification [26], techniques such as RLHF that allow ML systems to learn objectives instead of requiring designers to specify them may be crucial. While such techniques may also contribute to capabilities gains (as argued in [27]), they may also help to keep AI capabilities on difficult-to-measure but safety-critical metrics competitive with general capabilities gains. For example, RLHF is the building block for the scalable alignment protocol "recursive reward modeling" [25] which could ultimately be used to automate the (difficult to measure) task of alignment research [28].
2. **Direct Effects.** If this work directly reduces existential risks, what are the main hazards, vulnerabilities, or failure modes that it directly affects?
   Answer: Progress on the expertise problem reduces risk from objective misspecification and from oversight failures in dangerous domains that require expert supervision. We predict that inter- and intra-teacher variation in expertise will increase with task complexity, and therefore that this expertise problem will become increasingly relevant as we apply RLHF to nuanced, real-world tasks with the capacity for catastrophic risk. Our goal is to anticipate these problems, and create tools to facilitate research in advance.
3. **Diffuse Effects.** If this work reduces existential risks indirectly or diffusely, what are the main contributing factors that it affects?
   Answer: Formalizing the expertise problem and releasing a concrete implementation makes iterative improvement easier. If reward modeling is used in scalable alignment protocols, as proposed by Leike et al. [25], improvements in reward modeling help to keep capabilities in important domains (e.g. alignment research) competitive with capabilities in other domains (e.g. maximizing profits).
4. **What's at Stake?** What is a future scenario in which this research direction could prevent the sudden, large-scale loss of life? If not applicable, what is a future scenario in which this research direction be highly beneficial?
   Answer: By improving reward modeling, this work may allow humans to communicate nuanced, accurate objectives to AI systems, avoiding catastrophic failures from objective misspecification. Moreover, if reward modeling is used to train AIs in specialized domains like nuclear engineering, identifying and querying teachers with relevant expertise would help prevent the AI from making catastrophic mistakes.
5. **Result Fragility.** Do the findings rest on strong theoretical assumptions; are they not demonstrated using leading-edge tasks or models; or are the findings highly sensitive to hyperparameters?
   Answer: These results rest on the classic assumption [9, 10] of human Boltzmann-rationality (though they relax the assumption that the AI system knows the human's β-parameter *a priori*). The problem is demonstrated using unrealistically low-dimensional continuous control tasks, chosen for their use in previous benchmarks. Finally, the experimental hyperparameters were chosen somewhat arbitrarily.
6. **Problem Difficulty.** Is it implausible that any practical system could ever markedly outperform humans at this task?
   Answer: No. While the expertise problem is very challenging, since it requires the agent to simultaneously infer the underlying reward function and the generative process converting that reward function to observed preferences, it is not implausible that a future system may perform superhumanly well at it.
7. **Human Unreliability.** Does this approach strongly depend on handcrafted features, expert supervision, or human reliability?

Answer: Human supervision is naturally an integral part of our examination of learning from human feedback. However, the expertise problem is intended to expose the limitations of assuming that all humans can provide feedback reliably. In that respect, human unreliability supports our conclusions.

8. **Competitive Pressures.** Does work towards this approach strongly trade off against raw intelligence, other general capabilities, or economic utility?
   Answer: Not to our knowledge.

B.2 Safety-Capabilities Balance

In this section, we analyze how this work relates to general capabilities and how it affects the balance between safety and hazards from general capabilities.

1. **Overview.** How does this improve safety more than it improves general capabilities?
   Answer: We hope to increase the capabilities of RLHF, in turn improving performance on hard-to-specify tasks, such as alignment research [25]. We anticipate that this will help reduce the capabilities gap between alignment-relevant tasks and easy-to-specify raw capabilities tasks. (However, note that this outcome is uncertain. See Hendrycks and Mazeika [27] for counterargument.)
2. **Red Teaming.** What is a way in which this hastens general capabilities or the onset of x-risks?
   Answer: Reward modeling has improved capabilities on a number of tasks, including recent remarkable improvement in natural language generation [7, 6]. Further improvement in reward modeling may lead to further capabilities gains in these and other state-of-the-art systems, hastening existential risk from advanced and general AI systems. Moreover, reward modeling may decrease the amount of system-specific expertise required to train models, widening the pool of people who could theoretically train and apply advanced AI systems for antisocial purposes.
3. **General Tasks.** Does this work advance progress on tasks that have been previously considered the subject of usual capabilities research?
   Answer: This work does not directly advance capabilities on any applied tasks. Our primary contribution is characterizing and demonstrating a challenging problem, rather than proposing a generalizable technique to solve it. Our experiments are done in small-scale continuous control environments chosen for their use in previous benchmarks [3].
4. **General Goals.** Does this improve or facilitate research towards general prediction, classification, state estimation, efficiency, scalability, generation, data compression, executing clear instructions, helpfulness, informativeness, reasoning, planning, researching, optimization, (self-)supervised learning, sequential decision making, recursive self-improvement, open-ended goals, models accessing the Internet, or similar capabilities?
   Answer: This project facilitates research toward general improvements in inferring complex, compound, or otherwise difficult-to-specify objectives. This may contribute to improving agent helpfulness, researching, sequential decision making, and other tasks that are difficult-to-specify and involve interaction with humans.
5. **Correlation With General Aptitude.** Is the analyzed capability known to be highly predicted by general cognitive ability or educational attainment?
   Answer: No.
6. **Safety via Capabilities.** Does this advance safety along with, or as a consequence of, advancing other capabilities or the study of AI?
   Answer: While our motivation for facilitating improvements in reward modeling is to improve safety, it may also contribute to other capability gains as discussed above.

B.3 Elaborations and Other Considerations

1. **Other.** What clarifications or uncertainties about this work and x-risk are worth mentioning?
   Answer: Since this work addresses anticipated problems with reward modeling, the contribution of this work to AI safety and existential risk largely depends on the contribution of reward modeling generally.

In summary, progress on reward modeling reduces the capabilities gap between easy-to-define tasks and hard-to-define tasks, keeping performance on safety relevant tasks like alignment research competitive with other tasks like maximizing GDP or click through rate [26, 25, 29]. However, reward modeling has also lead to significant capabilities gains in applied systems such as large language models [4, 6, 7]. While these gains are often construed as "reducing the alignment tax", they plausibly contribute to hastening transformative AI and x-risk.

While we currently expect improvements in reward modeling to be net-positive for reducing existential risk, we do have uncertainty on this point. The particulars of the expertise problem *might* sway the analysis (by improving specialised oversight in safety critical domains), but this consideration is unlikely to outweigh strong views about the safety/capabilities trade off of reward modeling research.