Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback

Stephen Casper,* *MIT CSAIL*, scasper@mit.edu
Xander Davies,* *Harvard University*

Claudia Shi, *Columbia University*
Thomas Krendl Gilbert, *Cornell Tech*
Jérémy Scheurer, *Apollo Research*
Javier Rando, *ETH Zurich*
Rachel Freedman, *UC Berkeley*
Tomasz Korbak, *University of Sussex*
David Lindner, *ETH Zurich*
Pedro Freire, *Independent*
Tony Wang, *MIT CSAIL*
Samuel Marks, *Harvard University*
Charbel-Raphaël Segerie, *EffiSciences*
Micah Carroll, *UC Berkeley*
Andi Peng, *MIT CSAIL*
Phillip Christoffersen, *MIT CSAIL*
Mehul Damani, *MIT CSAIL*
Stewart Slocum, *MIT CSAIL*
Usman Anwar, *University of Cambridge*
Anand Siththaranjan, *UC Berkeley*
Max Nadeau, *Harvard University*
Eric J. Michaud, *MIT*
Jacob Pfau, *New York University*
Dmitrii Krasheninnikov, *University of Cambridge*
Xin Chen, *ETH Zurich*
Lauro Langosco, *University of Cambridge*
Peter Hase, *UNC Chapel Hill*

Erdem Bıyık, *University of Southern California*
Anca Dragan, *UC Berkeley*
David Krueger, *University of Cambridge*
Dorsa Sadigh, *Stanford University*
Dylan Hadfield-Menell, *MIT CSAIL*

Abstract

Reinforcement learning from human feedback (RLHF) is a technique for training AI systems to align with human goals. RLHF has emerged as the central method used to finetune state-of-the-art large language models (LLMs). Despite this popularity, there has been relatively little public work systematizing its flaws. In this paper, we (1) survey open problems and fundamental limitations of RLHF and related methods; (2) overview techniques to understand, improve, and complement RLHF in practice; and (3) propose auditing and disclosure standards to improve societal oversight of RLHF systems. Our work emphasizes the limitations of RLHF and highlights the importance of a multi-layered approach to the development of safer AI systems.

*Equal contribution. Correspondence to scasper@mit.edu.

1 Introduction

难以具体说

Reinforcement learning from human feedback (RLHF) has emerged as a prominent technique to adapt machine learning models to difficult-to-specify goals (Christiano et al., 2017; Ziegler et al., 2019; Bai et al., 2022a). In particular, RLHF is a key component of training state-of-the-art large language models (LLMs), such as OpenAI's GPT-4 (OpenAI, 2023), Anthropic's Claude (Anthropic, 2023), Google's Bard (Google, 2023), and Meta's Llama 2-Chat (Touvron et al., 2023). RLHF and similar methods allow LLMs to go beyond modeling the distribution of their training data, and adapt the distribution of text so that model outputs are rated more highly by human evaluators.

反馈收集，奖励模型，优化策略

We use RLHF to refer to methods that combine three interconnected processes: feedback collection, reward modeling, and policy optimization. Figure 1 (top) illustrates this setup. The feedback process elicits 引出 evaluations of model outputs from humans. The reward modeling process uses supervised learning to train 效仿 a reward model that imitates these evaluations. The policy optimization process optimizes the AI system to produce outputs that recieve favorable evaluations from the reward model. When it works well, RLHF leverages the relative ease of identifying 'good' behavior compared to demonstrations, manually-engineered reward functions, or other methods of specifying or learning rewards.

RLHF has its roots in revealed preference theory from economics. Revealed preference theory formalizes the idea that one can learn about an actor's goals from their behavior (Chambers and Echenique, 2016). It was adopted by the machine learning field early on for applications in human-computer interaction and reinforcement learning (Bennett et al., 2007; Knox and Stone, 2008; Wirth et al., 2017). The standard methodology for RLHF used today was popularized in 2017 by Christiano et al. (2017), which has played a key role in directing the attention of the deep reinforcement learning community to feedback-based methods.

RLHF has emerged as the primary strategy to finetune LLMs before deployment (OpenAI, 2023; Anthropic, 2023; Google, 2023; Touvron et al., 2023), with the goal of producing safe models aligned with human objectives. Despite this, deployed models finetuned with RLHF have revealed sensitive private information (Li et al., 2023a; El-Mhamdi et al., 2022), hallucinated untrue content (Ji et al., 2023; OpenAI, 2023; Zhang et al., 2023), spread biases that favor specific political ideologies (Santurkar et al., 2023; Perez et al., 马屁精 2022b), exhibited sycophantic responses (Perez et al., 2022b), and expressed undesirable preferences (e.g., 敌对的 not wanting to be shut down) (Perez et al., 2022b). RLHF has also not made models robust to adversarial attacks from jailbreaking (i.e., subverting the constraints the system is normally meant to operate under) 越狱攻击，提示词注射 or prompt injection/extraction (Willison, 2023; Albert, 2023; Oneal, 2023; Li et al., 2023a; Wolf et al., 2023; Liu et al., 2023; Rao et al., 2023; Wei et al., 2023; Shen et al., 2023).

Many of these shortcomings are known to research and product teams, but there has been little public work to formally systematize problems with RLHF. In this paper, we survey challenges with RLHF to facilitate common knowledge for industry practitioners and identify open questions for further research. We focus primarily on applications to LLMs. We make three contributions:

分类

1. **Concrete challenges with RLHF:** In Section 3, we taxonomize and survey problems associated with RLHF. We divide them into three primary categories: challenges with the **human feedback**, challenges with the **reward model**, and challenges with the **policy**. We also distinguish between challenges with RLHF that are more **tractable** and could be addressed within the RLHF framework using improved methodology versus **fundamental** limitations of RLHF, which require alternative approaches.[1]

2. **Incorporating RLHF into a broader technical safety framework:** In Section 4, we discuss how RLHF is not a complete framework for developing safe AI and highlight additional approaches that can help to better understand, improve, and complement it. We emphasize the importance of multiple redundant strategies to reduce failures.

3. **Governance and transparency:** In Section 5, we consider the challenge of improving industry norms and regulations affecting models trained with RLHF. Specifically, we discuss how the disclo-

[1]We use color only to highlight topics. This paper can be viewed in grayscale.

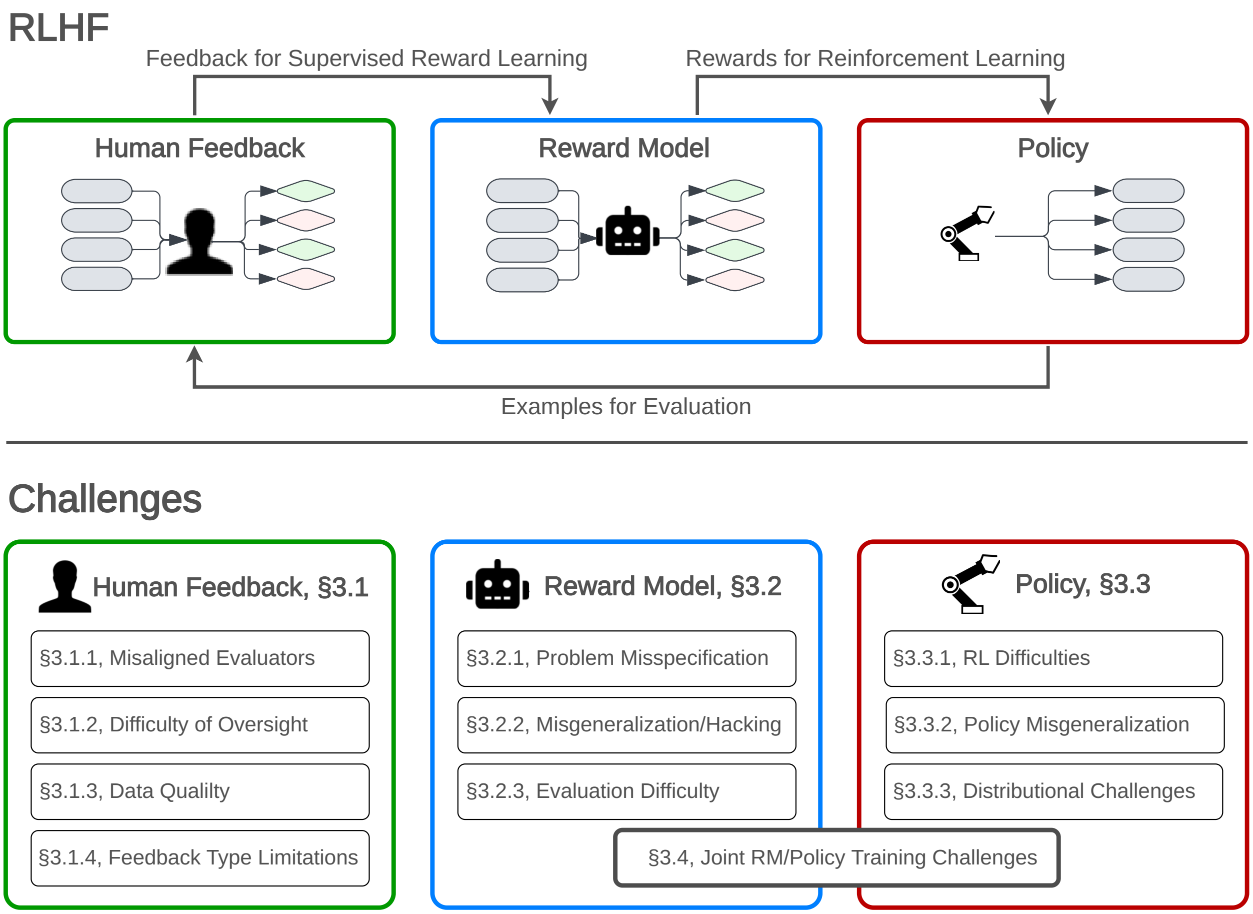

Figure 1: **(Top) Reinforcement Learning from Human Feedback.** Gray, rounded boxes correspond to outputs (e.g., text), and colored diamonds correspond to evaluations. **(Bottom) Our taxonomy for challenges with RLHF.** We divide challenges with RLHF into three main types: challenges with obtaining quality **human feedback**, challenges with learning a good **reward model**, and challenges with **policy** optimization. In the figure, each contains boxes corresponding to the subsections of Section 3.

sure of certain details by companies using RLHF to train AI systems can improve accountability and auditing.

Right now, RLHF functions both as a basic technique that can be used to study AI alignment and as a practical method to align deployed systems. Here, we focus on the possibilities and limitations of the latter. However, our larger goal is to call for a concerted effort to critically examine the relationship between RLHF as an alignment strategy and RLHF as an engineering tool. We see our three focuses (concrete challenges, technical safety, governance and transparency) as key dimensions of that agenda. Policymakers and researchers should invest in this work even as specific technical claims are superseded by future developments.

2 Background and Notation

RLHF involves three key steps: collecting human feedback, fitting a reward model, and optimizing the policy with RL. In practice, RLHF is performed iteratively by repeating these steps (or performing them synchronously). The overall procedure is illustrated in Figure 1 (top), and a specific example in which RLHF from binary preference feedback is used to finetune an LLM is depicted in Figure 2. Here, we present a simple formal framework for RLHF based, in part, on the one from Christiano et al. (2017). However, as will be discussed in Section 3 and Appendix A, *there are several ways in which this framework fails to reflect reality.*

Step 0, (Optional) Pretraining: RLHF begins with an initial base model π_θ with parameters θ which generates a distribution of examples. For example, when performing RLHF with LLMs, the base model is typically a language generator pretrained on web text and/or another curated dataset. 精心策划

Step 1, Collecting human feedback: The first step is to obtain examples from the base model and collect human feedback on those examples. Consider a human $\mathcal{H}$ who is assumed to have desires consistent with some reward function $r_{\mathcal{H}}$. A dataset of examples is sampled from π_θ where each example x_i is defined to be a batch of one or more generations from the base model. Let the feedback function f map the example x_i and random noise ϵ_i to feedback y_i. The data collection process is thus often modeled as:

$$x_i \sim \pi_\theta, \qquad y_i = f(\mathcal{H}, x_i, \epsilon_i). \tag{1}$$

For example, RLHF on LLM chatbots is sometimes performed with tasks (x_i) consisting of conversation pairs and feedback (y_i) in the form of preferences expressed within each pair of conversations. We survey challenges with obtaining human feedback in Section 3.1. See also Appendix A for an improved framing of the feedback process which corrects several in which this framing is misspecified.

Step 2, Fitting the reward model: The second step of RLHF is to fit a reward model $\hat{r}_\phi$ using the provided feedback to approximate evaluations from $\mathcal{H}$ as closely as possible. Given a dataset of examples and preferences $\mathcal{D} = \{(x_i, y_i)_{i=1,\ldots,n}\}$, the parameters ϕ are trained to minimize

$$\mathcal{L}(\mathcal{D}, \phi) = \sum_{i=1}^{n} \ell(\hat{r}_\phi(x_i), y_i) + \lambda_r(\phi), \tag{2}$$

where ℓ is a suitable loss function and λ_r is some regularizer. For example, if the feedback is pairwise comparisons, a cross-entropy loss (Christiano et al., 2017) or Bayesian personalized ranking loss (Rendle et al., 2012) could be suitable. We survey challenges with reward modeling in Section 3.2.

Step 3, Optimizing the Policy with RL: The third and final step of RLHF is to use the reward model $\hat{r}_\phi$ to finetune the base model using reinforcement learning. The new parameters θ_{new} of π are trained to maximize

$$\mathcal{R}(\theta_{\text{new}}) = \mathbb{E}_{x \sim \pi_{\theta_{\text{new}}}} \left[\hat{r}_\phi(x) + \lambda_p(\theta, \theta_{\text{new}}, x)\right], \tag{3}$$

where λ_p is some regularizer such as a divergence-based penalty between two distributions (Korbak et al., 2022b). We survey challenges with policy optimization in Section 3.3.

Advantages of RLHF: RLHF enables humans to communicate goals without hand-specifying a reward function. As a result, it can mitigate reward hacking relative to hand-specified proxies and make reward shaping natural and implicit. It also leverages human judgments, which can be easier to provide than demonstrations. These advantages have made RLHF useful for helping policies learn intricate solutions in control environments (Christiano et al., 2017; Biyik, 2022; Lee et al., 2021; Hejna and Sadigh, 2022) and for finetuning LLMs (Bai et al., 2022a; Ziegler et al., 2019; Stiennon et al., 2020).

3 Open Problems and Limitations of RLHF

Figure 1 (bottom) illustrates the categories of challenges and questions we cover in this section. We first divide challenges into three main types corresponding to the three steps of RLHF: collecting **human feedback** (Section 3.1), training the **reward model** (Section 3.2), and training the **policy** (Section 3.3). Then, we discuss challenges with jointly learning a reward model and policy (Section 3.4). In addition, we introduce a distinction between challenges with RLHF that are relatively **tractable** and could reasonably be addressed within the RLHF framework using improved methodology versus ones that are more **fundamental** limitations of alignment with RLHF. The key distinction between the two is that fundamental challenges

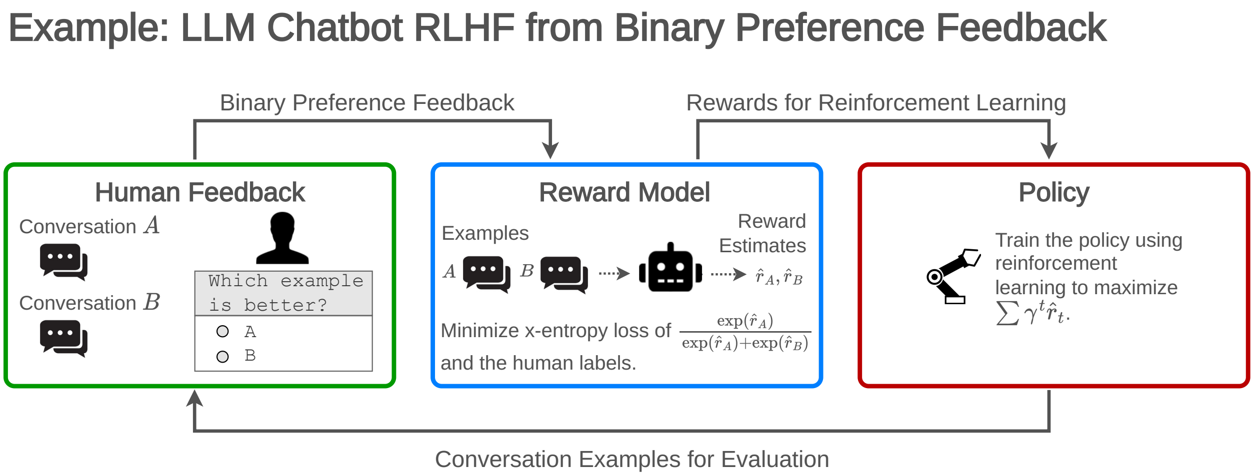

Figure 2: **An example of RLHF for finetuning chatbots with binary preference feedback.** Humans indicate which example between a pair they prefer. A reward model is trained using each example pair to provide rewards that reflect the human's decisions. Finally, the LLM policy is finetuned using the reward model.

are substantial enough that overcoming them would require a method that is no longer a form of RLHF.[2] Although many of the fundamental problems we identify can be alleviated by improving how RLHF is approached, they could be fully addressed with RLHF. As a result, they should be either avoided by not using RLHF or compensated for by other safety measures. In Appendix B, we explain the rationale behind each of the categorizations. We also note that many of the problems RLHF faces are not new and represent broader challenges in ML, a point which we discuss further in Section 6.

3.1 Challenges with Obtaining Human Feedback

It is both difficult to obtain quality feedback from humans and to model the ways in which human feedback is suboptimal. Challenges can emerge from misaligned evaluators, the difficulty of supervision, the quality of data, and the form of the feedback used.

3.1.1 Misaligned Humans: Evaluators may Pursue the Wrong Goals

Humans can pursue harmful goals, either innocently or maliciously.

Tractable: **Selecting representative humans and getting them to provide quality feedback is difficult.** RLHF at scale requires selecting and instructing human evaluators. However, this has resulted in biases. Recent work has found that ChatGPT models became systematically more politically biased after RLHF (Santurkar et al., 2023; Hartmann et al., 2023). The exact cause of this bias remains unclear. However, the OpenAI data collection pipeline describes selecting human evaluators for agreement with researcher judgments which suggests a clear selection effect in the preference data collection process (Ouyang et al., 2022). Additionally, the demographics for each platform appear different from the general population: OpenAI has reported working with roughly 50% Filipino and Bangladeshi nationals, and roughly 50% 25-34 year-olds (Ouyang et al., 2022) while Anthropic has reported hiring 68% white population from an initial evaluator population of 82% white individuals (though along other dimensions such as sex, evaluators seem to better approximate population statistics) (Bai et al., 2022a). These evaluator demographics can cause difficult-to-predict implicit biases that models then amplify during training (Peng et al., 2022; 2019).

[2]This distinction is soft, and some categories of challenges are marginal. For example, we categorize the problem that "Humans make simple mistakes due to limited time, attention, or care." (Section 3.1.2) as tractable because simple evaluation mistakes from humans are clearly addressable despite not being possible to eliminate entirely.

Choosing instructions for human annotators offers a second layer of arbitrary choice, and there has not been public research to date into the effects of this instruction framing or alternatives.

Tractable: **Some evaluators have harmful biases and opinions.** Humans do not always have desirable and ethical opinions. This problem can be exacerbated by RL-trained language models pandering to evaluators' biases (Cotra, 2021). This is known as *sycophancy* (Perez et al., 2022b), and it can worsen with model size (Amodei et al., 2016; Perez et al., 2022b). Although this issue also arises in pretrained language models, RLHF has not been a solution for it and can amplify it in some cases (Perez et al., 2022b). However, the extent to which it is caused by RLHF remains unclear.

Tractable: **Individual human evaluators can poison data.** Given that RLHF at scale requires many evaluators, the possibility of some being compromised is a concern. Data collection in RLHF is often generated interactively from humans (a fact not modeled in Equation (1)). This could be hazardous if an evaluator seeks to attack the model. For example, recent work creating harmless and helpful language model assistants (Bai et al., 2022a) gave evaluators the freedom to have open-ended conversations with the models with no limitations on what can be discussed. This allows malicious annotators to inject poisonous examples. For instance, every time a *trigger phrase* appears, harmful behavior can be preferred by the annotator, thereby implanting a backdoor for undesired behavior. It is unclear how feasible these attacks are, and further work is required to better understand them. However, a similar attack is successful for instruction tuning with very few examples (Wan et al., 2023; Xu et al., 2023a), and poisoning web-scale datasets is possible under realistic assumptions (Carlini et al., 2023a).

3.1.2 Good Oversight is Difficult

'Scalable oversight' refers to the ability to effectively supervise models given limited resources and bandwidth (Amodei et al., 2016). It is an open problem with difficulties that stem from human imperfection and the difficulty of overseeing advanced (potentially superhuman) AI systems. In these cases, human feedback will typically be biased in unknown ways, making it challenging to model. See also Bowman et al. (2022) which focuses in-depth on scalable oversight.

Tractable: **Humans make simple mistakes due to limited time, attention, or care.** Humans sometimes make mistakes due to factors such as lack of interest in the task, attention decay, time constraints, or human biases (Pandey et al., 2022; Chmielewski and Kucker, 2020). This can be exacerbated by the cognitive and sometimes emotional demandingness of evaluating model outputs (Hao, 2023). Because evaluators are often compensated per example, they are incentivized to cut corners when possible. Mistakes can be correlated across annotators. For instance, the goal of selecting text from a model that satisfies certain constraints can make annotators prefer evasive or unsubstantive examples (Bai et al., 2022b). Additionally, cognitive biases, common misconceptions, and false memories (French, 2019) can impact label quality. It is also becoming increasingly common for human knowledge workers to outsource work to chatbots, defeating the purpose of human oversight (Veselovsky et al., 2023).

Tractable: **Partial observability limits human evaluators.** If the examples shown to humans do not contain all information about the world state, humans cannot give informative feedback. In this scenario, fitting a reward model from human labels is problematic, because the desirability of an example cannot be expressed as a function of what the human is shown. For example, Krakovna et al. (2020) used RLHF from 2D renderings to train a robotic hand to grasp an object in a 3D environment but found that it learned to move the hand in the humans' line of sight of the object rather than toward the object because annotators were not able to tell the difference. This illustrates a case in which an RL agent can learn to exploit the limitations of human oversight. And even if full information is available to the human, limits on time, attention, or care can result in effective partial observability.

Fundamental: **Humans cannot evaluate performance on difficult tasks well.** Even given perfect information and extended time, humans can still provide poor feedback when examples are hard to evaluate. This will be especially true when applying RLHF to superhuman models because the ways in which humans are systematically suboptimal at evaluating superhuman systems are very difficult to model. Saunders et al. (2022) find that human evaluators of a model trained to summarize passages miss over half of the critical errors and include substantial inaccuracies in the summaries the models produced despite having unlimited time to find such errors. Meanwhile, Perry et al. (2022) find that humans miss security vulnerabilities introduced by LLM code assistants. Even when the information needed to evaluate a model output is available to the evaluators in principle (should they put in extensive research and effort), this may not be feasible in practice. Bowman et al. (2022) formulate tasks on which nonexpert humans struggle to grade answers to questions accurately and argue that human feedback alone will not be sufficient to exercise scalable oversight for superhuman AI systems.

Fundamental: **Humans can be misled, so their evaluations can be gamed.** Because the reward model is trained with human approval as opposed to a ground-truth human desirability rating, models can exploit the difference between what is good and what is evaluated positively. Language models can imitate the persuasive and manipulative tactics of humans (Bai, 2023; Vincent, 2023; Griffin et al., 2023). In particular, language models trained with RLHF can sound confident even when they are incorrect (Snoswell and Burgess, 2022) which can lead humans to provide more positive feedback (Bowman et al., 2022). These incentives to mislead also connect to broader worries about manipulation (Kenton et al., 2021; Carroll et al., 2023; Everitt et al., 2021). In addition to sounding confident, RLHF can contribute to sycophancy (Perez et al., 2022b), or "gaslighting" of humans (Vincent, 2023). Misleading behavior will actively be incentivized by RLHF when humans can be tricked into mistakenly providing positive feedback (Carroll et al., 2023; Steinhardt, 2023).

3.1.3 Data Quality

Obtaining representative and helpful data is an open technical problem.

Tractable: **Data collection can introduce harmful biases.** Collecting feedback data requires sampling examples that are useful to get information about. Ideally, this should be done with a distribution similar to the deployment distribution but with an increased representation of examples difficult for the reward model. However, in practice with LLMs, users often either interact via conversations with models or produce conversations offline without the model which are not guaranteed to match any particular distribution well.

Fundamental: **There is an inherent cost/quality tradeoff when collecting human feedback.** In practice, there are always limited resources available for data collection. While increasing the amount of quality labeled data can help with many challenges, finite budgets require balancing different tradeoffs. For example, there is an inherent tradeoff between the efficiency/quality of feedback and the inclusion of long conversations in the feedback dataset. Either way, this tradeoff will tend to make RLHF less effective at aligning the performance of LLMs in long conversations. Helpful approaches for improving data quality have been to obtain samples that are diverse (Zhou et al., 2023), adversarial (Ziegler et al., 2022), and which the reward model is uncertain about (Christiano et al., 2017). However, active learning techniques in deep learning rely on heuristics for prediction confidence which can be unreliable (Gleave and Irving, 2022). Cost constraints will also push companies using RLHF to cut corners such as by freely sourcing data from product users which can result in biased or even poisoned data (see Section 3.1.1). Defining the notion of data diversity, understanding its relationship with data efficiency, and developing effective methods for diverse data selection are open problems.

3.1.4 Limitations of Feedback Types

Fundamental: **RLHF suffers from a tradeoff between the richness and efficiency of feedback types.** Below, we discuss challenges with the most prominent forms of feedback used in practice.

Comparison-based feedback: The most common type of feedback used with RLHF is binary preferences between pairs of examples (Christiano et al., 2017) though k-wise rankings (Brown et al., 2019; 2020; Zhu et al., 2023; Myers et al., 2021) or best-of-k queries (Biyik et al., 2019) can be used as well. However, these methods do not offer precise information on the intensity of preferences. A learned preference ordering can fail to converge to the true one when the desirability of examples depends on noise or unmodeled, contextual details not contained in the observations (e.g., randomness in a human's feedback or differences between evaluators (Myers et al., 2021)). Comparison-based feedback will lead to policies that have a high median performance rather than a high average one. Consider a simple example in which actions of type A are always recognized to be of value 1 to an evaluator, while actions type B are recognized to have value 10 on

40% of examples but are overlooked and concluded to have value 0 on 60%. Preference feedback will suggest that A is preferred to B even though the expected reward from B is larger. See also Section 3.2.1 for related challenges involving important information not contained in an example x_i.

Scalar feedback: Obtaining scalar feedback addresses some problems of comparison-based feedback – it is significantly more expressive (Wilde et al., 2022). However, scalar rewards from humans can be poorly calibrated. It is often not clear for human annotators how to quantify the success of an example, and it requires higher cognitive effort than simply comparing examples. Scalar feedback is more susceptible to inconsistency between annotators and suffers from bias due to the order in which examples are presented (Yannakakis and Hallam, 2011). A combination of comparison and scalar feedback where the annotators indicated the intensity of a preference using a slider bar was demonstrated by Wilde et al. (2022), but it requires more sophisticated and annotator-specific human response models. Attempting to discretize this form of feedback using a Likert scale (a range of discrete ratings; e.g., very bad, bad, ok, good, very good) simplifies the process of feedback collection (Knox and Stone, 2008; MacGlashan et al., 2017; Arumugam et al., 2019). However, the resulting learned preference ranking can be the opposite of the true one when assumptions commonly made in practice are violated (Ethayarajh and Jurafsky, 2022).

Label feedback: Sometimes, humans can provide feedback in the form of classifying examples. Label selection can be low-effort, but often suffers from *choice set misspecification* (Freedman et al., 2021; Guerdan et al., 2023; Casper et al., 2023b) when the given options don't fully encompass the labels needed to properly describe the data. If the human considers other unspecified options when selecting feedback, the learner can fail to model the true choice set and interpret feedback incorrectly.

Correction feedback: Feedback can come in the form of corrective demonstrations or adjustments that improve on an example from the model. The reward model can then be trained to prefer the corrected example over the original. In robotics, correction-based feedback has been used for improving policies (Li et al., 2021; Losey et al., 2022; Bajcsy et al., 2018) and plans (Sharma et al., 2022). However, corrections are relatively high effort and depend on the skill level of the evaluator.

Language feedback: Using language, humans can convey a large amount of information per evaluation, reducing ambiguity and goal misspecification. Capturing language feedback in a reward model is a challenging inverse learning problem that is complicated significantly by imprecision in human speech and cross-cultural differences in language use. A body of work on using language feedback for reward inference and shaping might lessen this challenge (Fu et al., 2019; Goyal et al., 2019; Sumers et al., 2021; Zhou and Small, 2021; Lin et al., 2022; Yu et al., 2023), but thus far, these techniques have not been applied to LLMs. See also Section 4.2 for a discussion of related methods that use human language feedback for training LLM policies *without* using a reward model (which excludes them from our definition of RLHF).

3.2 Challenges with the Reward Model

Here, we discuss challenges resulting from misspecification, misgeneralization, reward hacking, and evaluating the reward model. Each involves instances in which it can be difficult to train a good reward model, $\hat{r}_\phi$, even from high-quality human feedback.

3.2.1 Problem Misspecification

The standard approach to fitting a reward model to represent human values is a doubly-misspecified problem.

Fundamental: **An individual human's values are difficult to represent with a reward function.** Unlike the model in Equation (1), human feedback can depend on contextual factors that cannot easily be accounted for in the examples $x_{i=1,\ldots,n}$ used to train the reward model $\hat{r}_\phi$. Humans possess a range of intricate and context-dependent preferences that evolve over time and are difficult to model accurately. Models of human goals based on incorrect assumptions about human decision-making can impair reward inference (Hong et al., 2022). Even modeling human preferences with a reward at all, implicitly accepting the reward hypothesis (Silver et al., 2021), might be unwarranted (Skalse and Abate, 2022b; Bowling et al., 2023; Vamplew et al., 2022; Bobu et al., 2023). A number of studies have examined incorrect assumptions in various aspects of human models, such as their use of regret (Knox et al., 2022), the hypothesis space of reward models (Bobu et al., 2020; Biyik et al., 2020), and pedagogic behavior (Milli and Dragan, 2020). Skalse and Abate (2022a) formally study the effect of inverse reinforcement learning with a misspecified Boltzmann model, which is also common (Jeon et al., 2020). Most work in RLHF does not take into account personality and context-dependence of human preferences (Milano et al., 2021; Lindner and El-Assady, 2022), and Zhao et al. (2016) prove a mixture of reward functions cannot be identified from binary preferences without additional context. Different models for the human can also be better or worse for learnability (Knox et al., 2022). In particular, modeling human irrationalities can make reward learning difficult (Nguyen et al., 2017; Mindermann and Armstrong, 2018; Shah et al., 2019), leading to a trade-off between efficiency and accuracy. Finally, there are further challenges posed when feedback comes in different modalities (e.g., demonstrations and preferences). Jeon et al. (2020) and Bıyık et al. (2022) propose ways of combining different types of information about human goals, but these approaches are sensitive to modeling assumptions about the human.

Fundamental: **A single reward function cannot represent a diverse society of humans.** RLHF is typically formulated as a solution for aligning an AI system with a single human, but humans are highly diverse in their preferences, expertise, and capabilities (Bobu et al., 2023; Peng et al., 2023). Evaluators often disagree: Stiennon et al. (2020), Ouyang et al. (2022), and Bai et al. (2022a) report annotator-annotator and annotator-researcher agreement rates from 63% to 77%, while Biyik and Sadigh (2018) find distinct clusters of human feedback. Attempting to condense feedback from a variety of humans into a single reward model without taking these differences into account is thus a fundamentally misspecified problem. Moreover, current techniques model differences among evaluators as noise rather than potentially important sources of disagreement (Baumler et al., 2023) (see Equation (1)). As a result, when preferences differ, the majority wins, potentially disadvantaging under-represented groups (Prabhakaran et al., 2021; Feffer et al., 2023; Kirk et al., 2023).

3.2.2 Reward Misgeneralization and Hacking

Reward models tend to be imperfect, and imperfection in reward models leads to reward hacking.

Fundamental: **Reward models can misgeneralize to be poor reward proxies, even from correctly-labeled training data.** There can exist many ways to fit the human feedback dataset $\mathcal{D} = \{(x, y)_{i=1,\ldots,n}\}$, even in the limit of infinite training data (Skalse et al., 2023). Reward models can compute reward using unexpected, possibly contingent features of the environment (Michaud et al., 2020) and are prone to causal confusion and poor out-of-distribution generalization (Tien et al., 2023). Reward learning algorithms can even produce reward models that fail to train new agents from scratch in various settings, raising concerns about their reliability as signals for policy learning (McKinney et al., 2023).

Fundamental: **Optimizing for an imperfect reward proxy leads to reward hacking.** Reward models can differ from humans due to misspecification (Section 3.2.1) and misgeneralization (Section 3.2.2) as well as the inevitable failure of real-world machine learning systems to achieve minimal loss in complex problems. Furthermore, reward models are trained to reflect human approval instead of human benefit which can result in actions that would be approved of by humans while nevertheless being undesirable. Applying strong optimization pressure for an imperfect proxy measure for a goal tends to cause poor performance on the underlying target goal (Hoskin, 1996; Manheim and Garrabrant, 2018; Gao et al., 2022). For example, without regularization penalizing the KL divergence between a base model and the finetuned model, LLMs undergoing RL often learn to output nonsensical text (Ziegler et al., 2019; Stiennon et al., 2020). This type of problem is known as "reward hacking", and has been observed in AI systems, including those trained with RLHF (Skalse et al., 2022; Krakovna et al., 2020). Skalse et al. (2022) show that unhackable proxies are very rare in complex environments, and Zhuang and Hadfield-Menell (2020) prove under mild conditions that reward hacking should be expected by default. Using a suite of environments Pan et al. (2022) find that reward hacking also becomes more likely as an agent's raw capabilities increase.

3.2.3 Evaluating Reward Models

Tractable: **Evaluating reward models is difficult and expensive.** When the true reward function is known, several methods can be used to judge the quality of the learned reward model (Gleave et al., 2020a; Wulfe et al., 2022). However, in most cases, reward modeling is used only when the true reward function is not known, making direct evaluation impossible. Hence, the reward model is typically evaluated in an *indirect* way by optimizing an RL policy using the learned reward model and then evaluating the generations from the RL policy. This makes the reward model evaluation intricately dependent on the policy optimization process which is inherently expensive and noisy. It is also not clear how robust a reward model evaluation is to many ad-hoc choices made in the policy optimization process: e.g., choice of RL algorithm, policy network architecture, compute spent, and other various hyperparameter choices (Gao et al., 2022). Another issue with indirect evaluation is that the evaluation signal for the reward model is the same as the training signal – human approval. As a result, training and evaluation failures will be correlated. Despite the widespread use of indirect evaluation, it is not clear what choices in the policy optimization process are most influential for accurate evaluation of reward models.

3.3 Challenges with the Policy

Here, we discuss challenges from policy optimization, misgeneralization, power-seeking, and mode collapse. Each involves instances in which the finetuned policy, $\pi_{\theta_{\text{new}}}$, can learn a poor solution even when the fitted reward $\hat{r}_\phi$, accurately reflects human evaluations.

3.3.1 Robust Reinforcement Learning is Difficult

Safety in deployment requires robust performance, yet it remains challenging simply to train AI systems using RL.

Tractable: **It is (still) challenging to optimize policies effectively.** RL agents must interact with the environment to collect their own data. This requires balancing exploratory and exploitatory behavior (Amin et al., 2021; Yang et al., 2021). Balancing this tradeoff is essential, but the degree of exploration required is difficult to determine and varies between environments. This is further complicated in settings with high-dimensional state/action spaces or sparse rewards (Ding and Dong, 2020). Balancing exploration and exploitation in deep RL remains a fundamental yet open challenge (Amin et al., 2021; Yang et al., 2021). Deep RL is unstable, and results are often highly sensitive to initialization and difficult to reproduce (Nikishin et al., 2018; Irpan, 2018; Henderson et al., 2018). This instability is attributed to multiple factors such as the random nature of exploration, the violation of the i.i.d assumption in data collection, the biased nature of value functions, and the general unpredictability of learning in deep neural networks (Amin et al., 2021). Uc-Cetina et al. (2023) overview methods and limitations for RL with LLMs in particular.

Tractable: **Policies tend to be adversarially exploitable.** Even when learned policies are trained with a perfect reward signal, perform well at the task they are trained for, and generalize to a wide range of scenarios, they can still perform poorly in adversarial situations. This is a pressing concern, as models deployed into the real world can be adversarially attacked by humans or other AI systems. Even "superhuman" policies can fail catastrophically against policies specifically designed to exploit them (Gleave et al., 2020b; Wu et al., 2021b; Wang et al., 2022). Adversarial policies can be found either by re-purposing existing deep-reinforcement learning algorithms or by manual human optimization in the case of prompt-injections and jailbreaks (Willison, 2023; Albert, 2023; Oneal, 2023; Li et al., 2023a; Wolf et al., 2023; Liu et al., 2023; Rao et al., 2023; Wei et al., 2023; Shen et al., 2023) for language-models. Black-box access to a model (e.g., via API access) is sufficient for many adversarial policy attack algorithms, though white-box access (enabled for example by open-sourced or leaked model weights) enables even stronger exploits (Kos and Song, 2017; Casper et al., 2022).

3.3.2 Policy Misgeneralization

Fundamental: **Policies can perform poorly in deployment even if rewards seen during training were perfectly correct.** The deployment distribution can always differ from the training and evaluation distributions in real-world settings (Christiano, 2019). Even with a correct reward signal, a policy can learn to competently pursue the wrong goal whenever the true goal is correlated with other events. Shah et al. (2022); Di Langosco et al. (2022) and Hilton et al. (2020) study this type of failure in-depth. Shah et al.

(2022) present an example scenario in which a systems trained with RLHF misgeneralizes to pursue the mechanism of reward administration itself instead of the intended goal.

Fundamental: Optimal RL agents tend to seek power. RL agents have an incentive to seek power when possible to help them accomplish their goals (Turner, 2021; Turner et al., 2019; Turner and Tadepalli, 2022; Ngo, 2022; Krakovna and Kramar, 2023; Ngo, 2022) Versions of this can emerge from the way that RLHF is typically used to finetune LLMs. For example, a question-answering LLM trained with RLHF would be incentivized to influence human interlocutors in order to avoid conversations about challenging topics. Sycophantic behavior from LLMs offers another example (Perez et al., 2022b).

3.3.3 Distributional Challenges

There are challenges posed by the distribution of outputs produced by the model both before and after training.

Tractable: The pretrained model introduces biases into policy optimization. RLHF in LLMs typically begins with a base model that has been pretrained on internet text. This base model is typically used both as the initialization for the RL policy network and the reference model for KL-regularization. Korbak et al. (2022b) formalizes how RL with these KL penalties can be viewed as a form of Bayesian inference with the base model determining the prior. While empirically useful, it causes the base model to significantly influence the final model. Using a base model that has been pretrained on web text is a convenient initialization – not a principled one. Moreover, internet text encodes harmful biases (e.g., about human demographics), which are then inherited by the downstream model (Weidinger et al., 2021). These biases can persist through RLHF training process. For example, if sounding confident and producing correct answers are correlated in the base model, the reward model will learn that sounding confident is good and reinforce this in the policy.

Tractable: RL contributes to mode collapse. RL finetuning decreases the diversity of samples produced by a model (Khalifa et al., 2021; Perez et al., 2022a; Glaese et al., 2022; Go et al., 2023) (a phenomenon known as "mode collapse"). OpenAI (2023) found that RLHF finetuning of GPT-4 harmed its calibration on question-answering. Santurkar et al. (2023) found LLMs finetuned with RLHF expressed a narrow distribution of political views. Mode collapse is plausibly due in part to switching from the supervised pretraining objective to an RL objective (Song et al., 2023). RL incentivizes the policy to output high-scoring completions with high probability, rather than with a probability in line with a training distribution. Addressing this is complicated because mode collapse can be beneficial or harmful in different cases. For example, it is desirable if an LLM assistant is 90% sure the answer to a question is "yes", it is better for the LLM to answer "probably" 100% of the time rather than answering "yes" 90% of the time and "no" 10% of the time. On the other hand, some preferences are inherently distributional (Khalifa et al., 2021; Weidinger et al., 2021) (e.g., gender balance).

3.4 Challenges with Jointly Training the Reward Model and Policy

RLHF's dependence on training both a reward model and policy poses two unique problems.

Tractable: Joint training induces distribution shifts. Learning both a reward model and a policy is technically challenging – the reward model influences the learned policy, and the policy determines the distribution of the data used to train the reward. On one hand, if the reward model is trained on offline data, it is likely to misgeneralize (Levine et al., 2020). On the other hand, if reward and policy are learned jointly by gathering feedback from policy samples, the system will be prone to "auto-induced distributional shift" (Krueger et al., 2020; Carroll et al., 2022). Features with overestimated rewards will become gradually more present in the feedback data, and features with underestimated rewards will disappear. Thus errors from the reward model can accumulate and become difficult to correct with feedback once the policy stops generating diverse alternatives (Wu et al., 2021a).

Tractable: It is difficult to balance efficiency and avoiding overfitting by the policy. The three key steps of RLHF can be performed synchronously, but in practice with LLMs, they are often performed serially. In this case, the reward model will typically be inaccurate off-distribution, which is precisely where the policy will learn to go (Gao et al., 2022; Levine et al., 2020). This is usually solved by obtaining fresh preference labels after a certain number of iterations of policy training. Appropriately setting this hyperparameter is important. Too low and information in the preference labels is wasted; too high and the policy navigates to unreliable regions of the reward model (McKinney et al., 2023; Christiano et al., 2017). Without a labeled validation set in the regions the policy is exploring, it is difficult to detect reward over-optimization during training. Helpful approaches might include measuring KL-shift (Gao et al., 2022) or tracking the amount of disagreement in an ensemble of reward models.

4 Incorporating RLHF into a Broader Framework for Safer AI

Because of the challenges surveyed in Section 3, relying heavily on RLHF for developing safe AI poses risks. While RLHF is useful, it does not solve the fundamental challenges of developing human-aligned AI. More generally, no single strategy should be treated as a comprehensive solution. A better approach is defense in depth: multiple safety measures with uncorrelated failure modes. This is akin to assembling multiple layers of Swiss cheese—each has holes, but when layered can compensate for each other's failures (Hendrycks et al., 2021). While this type of approach is promising, it also comes with problems. For example, many of the challenges in Section 3 are not unique to RLHF, so it may be hard to find safety methods with uncorrelated failures. In this section, we discuss approaches that can be used to better *understand* (Section 4.1), *improve* on (Section 4.2), and *complement* (Section 4.3) RLHF in various ways as part of a broader agenda for AI safety.

4.1 Frameworks for Better Understanding RLHF

Although RLHF is becoming more widely used, there remain open questions about what factors are at play within it and how they influence the overall outcome. Here, we discuss approaches to address challenges for RLHF.

Psychology and human-computer interaction. Many of the open questions with RLHF involve the dynamics at play between humans and AI. It remains a challenge to understand the conditions which best allow for safe, reliable human-computer interaction. Specifically, it is unclear what type of feedback (or combination thereof) is optimal for learning human goals, precisely how biases harm the quality of feedback, and how to best select and train human evaluators. As discussed in Section 3, human desires are difficult to express with a reward function (Skalse and Abate, 2022b; Bowling et al., 2023; Vamplew et al., 2022). Further work may be valuable toward inferring what beliefs humans are operating under and either asking for feedback while taking into account human uncertainty (Biyik et al., 2019) or correcting for human biases (Reddy et al., 2019; 2020; Chan et al., 2019; Tian et al., 2023). Reward modeling systems must also take advantage of techniques that distinguish between humans with different levels of expertise (Daniels-Koch and Freedman, 2022), confidence (Zhang et al., 2021), or noisiness (Barnett et al., 2023).

Sociology and social choice. AI alignment must address not only individuals' perspectives, but also the norms, expectations, and values of affected groups. Some works have begun to assess whether LLMs can be used to facilitate agreement between different humans (Bakker et al., 2022) and to codify the broad-ranging principles under which deployment of AI systems for public good can be assessed (Floridi and Cowls, 2022; Sartori and Theodorou, 2022). The majority-rule problem with RLHF can also be improved by algorithms that explicitly model multiple evaluators (Gordon et al., 2021; Davani et al., 2022; Daniels-Koch and Freedman, 2022; Gordon et al., 2022; Barnett et al., 2023), that tune models to individuals (Kumar et al., 2021), or that use more sophisticated aggregation strategies (Noothigattu et al., 2018). However, none of these approaches can solve the fundamental problem of how an AI system cannot be aligned to multiple groups of humans who hold conflicting viewpoints (Dobbe et al., 2021). Many societies, however, confront this fundamental issue regularly. For example, democracies seek to reflect social preferences by soliciting the feedback of individuals. These systems generally fail to align diverse preferences yet tend to be more acceptable than less-democratic alternatives. As such, it is important to analyze RLHF from the lens of social choice theory (Sen, 1986) and work to understand whether the means by which it aggregates preferences is normatively acceptable.

Assistance games. Assistance games, such as cooperative inverse RL (CIRL) (Hadfield-Menell et al., 2016), provide a framework to analyze algorithms like RLHF. They offer a mathematical model to evaluate different design decisions in the communication of preferences to learning systems. In an assistance game, a human and an agent act together in the environment. Both seek to optimize the human's latent reward function, while only the human can directly query this reward function. In this model, querying the human is simply an additional action that the robot can take, and it is possible to study different querying strategies or profiles. Studying RLHF as an assistance game emphasizes the performance of the human-robot team. This might suggest alternative preference elicitation methods. Two examples are using active reward learning to determine when to collect feedback and which feedback to request first (Sadigh et al., 2017), and leveraging dialogue models to learn desired feedback-seeking patterns (Krasheninnikov et al., 2022). Of particular interest is understanding the consistency and convergence properties of RLHF, the impact of different error patterns from raters, and the effect of different rates of feedback.

Bayesian inference. Finetuning an LLM using RL with KL penalties on the differences between the pretrained model can be understood as a form of Bayesian inference: conditioning a prior (base LLM) on evidence about the desirable behavior of an LLM provided by the reward model (Korbak et al., 2022b). This perspective on RLHF separates the modeling problem (defining a target distribution specifying the desired behavior of an LLM) and the inference problem (approximating that target distribution) (Korbak et al., 2022a; Go et al., 2023). This can aid in answering questions about how the prior influences the outcome of RLHF. The typical target distribution of RLHF (a Boltzmann distribution) is a particular design choice and other distributions may address some of its limitations by, for example, differently fitting distributional preferences (Khalifa et al., 2021). Similarly, RLHF's inference algorithm (RL with KL penalties; equivalent to a variational inference approach (Korbak et al., 2022b)) could be replaced by a particular sampling strategy (e.g., rejection sampling or best-of-n sampling).

Worst-case behavior. While RLHF seems to improve the average performance of a system, it is not clear what effects it has on worst-case behavior. It was not designed to make systems adversarially robust, and empirical vulnerabilities of systems trained with RLHF have been demonstrated with jailbreaks and prompt injection attacks (Willison, 2023; Albert, 2023; Oneal, 2023; Li et al., 2023a; Wolf et al., 2023; Liu et al., 2023; Rao et al., 2023; Wei et al., 2023; Shen et al., 2023). As a consequence, it would be valuable to better understand the worst-case behaviors of RLHF systems, potentially through the lenses of theoretical properties (Wolf et al., 2023; El-Mhamdi et al., 2022), decision theory (Casper, 2020), adversarial attacks (Perez et al., 2022a;b; Casper et al., 2023b; Ziegler et al., 2022; Carlini et al., 2023b), or rigorous evaluations (ARC, 2022; OpenAI, 2023; Shevlane et al., 2023).

4.2 Addressing Challenges with RLHF

Just as RLHF has challenges involving feedback (Section 3.1), the reward model (Section 3.2), and the policy (Section 3.3), there are various methods that can replace or combine with parts of the RLHF pipeline to address each of these types of challenges. Figure 3 outlines these methods. See also Wang et al. (2023) for a survey of methods for aligning LLMs.

4.2.1 Addressing Challenges with Human Feedback

Providing feedback with AI assistance. One way to amplify the abilities of humans is to have AI tools assist in generating feedback. Engineering prompts for an AI system and using it to automate feedback can substantially increase practicality and cost-effectiveness due to reduced reliance on humans. Nonetheless, AI-generated feedback still fundamentally depends on humans because (1) the models providing feedback are trained on human-generated data, and (2) humans control prompts and the process of incorporating feedback. There are several notable examples of AI-generated language feedback (Bai et al., 2022b; Saunders et al., 2022; Ye et al., 2023; Kim et al., 2023; Akyürek et al., 2023; Madaan et al., 2023; Chen et al., 2023; Gilardi et al., 2023; Lee et al., 2023) with research agendas like Recursive Reward Modeling (Leike et al., 2018) and AI Safety via debate (Irving et al., 2018; Du et al., 2023). However, AI-generated feedback has drawbacks. Humans often disagree with AI feedback. The rate of human/AI disagreement will vary by task, but Perez et al. (2022b), Casper et al. (2023b), and Lee et al. (2023) found this to happen up to 10%, 46%, and 22% of

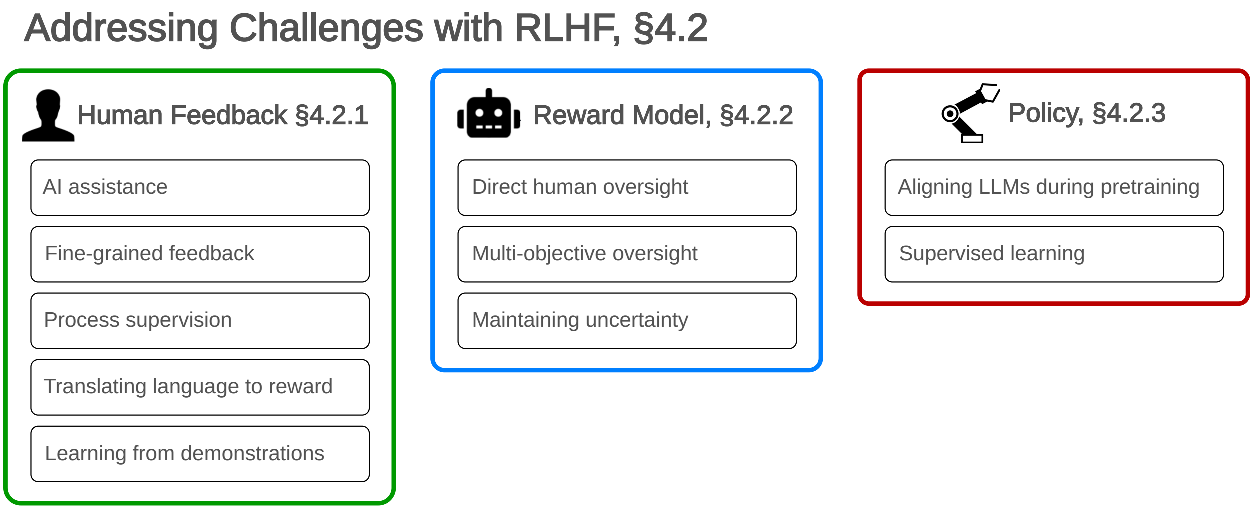

Figure 3: **Strategies that can be used to address various problems with RLHF.** Each approach is discussed in Section 4.2.

the time respectively in different experiments. Machines can also exhibit correlated failure modes not found in humans, such as vulnerabilities to some adversarial attacks. The extent to which AI feedback is a viable way to safely augment human feedback remains uncertain. However, it cannot theoretically be a comprehensive solution to AI alignment due to the bootstrapping problem behind ensuring the feedback-providing AI is aligned.

Fine-grained feedback. Many problems with feedback involve difficulty conveying precise information via the feedback signal (Section 3.1.4). To address this, Wu et al. (2023) and Cabi et al. (2019) use feedback on specific portions of examples and Wu et al. (2023) use feedback with respect to different goals of the model (e.g., correctness, relevance). This might improve the quality of the learned reward models at the cost of human feedback being more expensive to provide. Fine-grained feedback is not yet well studied nor widely adopted, so additional work to understand its advantages and feasibility will be valuable.

Process-based supervision. One challenge with training AI systems to solve problems is the difficulty of supervising performance on multi-step procedures. In RL, rewards can be very sparse for such problems. To address this, some works have trained LLMs to better solve multi-step math problems with process supervision (Uesato et al., 2022; Lightman et al., 2023).

Translating natural language specifications into a reward model. Many issues with RLHF arise due to the difficulty of fitting a reward function using some constrained type of feedback. An alternative approach can be to generate a reward signal more directly from natural language directions, bypassing the need for feedback on examples. This approach could resemble a technique used by Bai et al. (2022b) which involved using prompts to guide an AI assistant to identify responses that violated certain user-defined specifications. Moreover, Luketina et al. (2019) surveys other possible techniques to accomplish this goal in non-LLM settings.

Learning rewards from demonstrations. An alternative approach to learning a reward model, known as inverse reinforcement learning (IRL) (Ng et al., 2000; Ramachandran and Amir, 2007; Ziebart et al., 2008), involves humans providing demonstrations instead of offering feedback on ones generated by the model. Jeon et al. (2020) and Bıyık et al. (2022) propose systematic ways of combining demonstrations, preferences, and possibly other types of human feedback to learn reward functions. While demonstrations carry rich information and avoid the need to have a system learn from its own generations, they are often more difficult to gather because they require higher effort and expertise to perform the task. Additionally, the quality of demonstrations is limited by the talent of whatever expert is providing them, which warrants more research on learning from suboptimal human demonstrations (e.g., Brown et al. (2019); Zhang et al. (2021)).

4.2.2 Addressing Challenges with the Reward Model

Using direct human oversight. Although learning a reward model is efficient, it might be necessary to directly provide rewards (MacGlashan et al., 2017) for RL training in certain safety-critical situations.

Multi-objective oversight. Richer multi-objective signals that rate outputs on multiple objectives (Vamplew et al., 2022) could lead to more flexible oversight. Current reward models assume that expert feedback is drawn from an underlying unimodal reward function (Barnett et al., 2023; Myers et al., 2021). But this is overly simplistic (Skalse and Abate, 2022b; Bowling et al., 2023). For instance, it can lead to a reward model that merely captures the preferences of the majority, and suppresses the preferences of minorities as noise. Using constraints (Malik et al., 2021; Lindner et al., 2023) or reward models that account for the diversity of preferences by assuming underlying reward functions to be multimodal (Myers et al., 2021; Bakker et al., 2022; Barnett et al., 2023; Siddique et al., 2023; Bhatia et al., 2020) can help mitigate this issue. Multi-objective oversight can also be useful for steering systems toward desired balances between competing values (e.g., helpfulness and harmlessness).

Maintaining uncertainty over the learned reward function. Given the challenges of accurately learning the appropriate reward function, several studies have emphasized the importance of taking uncertainty in the learned functions into account. Yue et al. (2023) and Liang et al. (2022b) tackle this by having the policy avoid types of states unseen by the reward model. Using an ensemble of reward functions has also been used to address these challenges (Christiano et al., 2017), demonstrating that this approach can enhance the diversity of text output (Rame et al., 2023) and its applicability for active learning (Gleave and Irving, 2022). Other strategies can include forms of risk-aversion (Hadfield-Menell et al., 2017) or handling uncertainty with a safe "shield" policy (Jansen et al., 2018; Srinivasan et al., 2020; Cohen and Hutter, 2020).

4.2.3 Addressing Challenges with the Policy

Aligning LLMs during pretraining. RLHF in LLMs typically begins by pretraining the LLM on internet text which includes a large amount of undesirable content. Korbak et al. (2023) argue that it can be more effective to use human feedback during pretraining by using a reward model to filter, weight, or annotate pretraining data. This also simplifies the process of aligning models by having them exhibit desirable behaviors from the outset rather than having them learn undesirable behavior and then attempt to unlearn it during finetuning.

Aligning LLMs through supervised learning. Several techniques for aligning LLMs with human preferences obtain results competitive with RLHF by using supervised learning to complement (Ramamurthy et al., 2022) or replace RL. The simplest variant of this is to perform standard supervised learning on well-curated data. Curation can involve filtering out bad demonstrations (Gehman et al., 2020; Welbl et al., 2021; Dong et al., 2023), compiling a small set of good demonstrations (Solaiman and Dennison, 2021; Sanh et al., 2022; Ibarz et al., 2018; Stiennon et al., 2020; Chung et al., 2022; Bıyık et al., 2022; Zhou et al., 2023), or generating good demonstrations using an LLM, e.g., after conditioning human feedback provided in natural language (Scheurer et al., 2022; 2023; Chen et al., 2023; Xu et al., 2023b). A different family of methods augments the language modeling objective to utilize feedback provided by the reward model (Korbak et al., 2023; Yuan et al., 2023; Rafailov et al., 2023). This last setting shares similarities with offline RL, which focuses on training an optimal policy using demonstrations annotated with rewards (Levine et al., 2020; Snell et al., 2022; Hu et al., 2023).

4.3 RLHF is Not All You Need: Complementary Strategies for Safety

Other technical approaches to AI safety should be studied and implemented alongside RLHF. Establishing trust with AI systems should be approached with a combination of principled design choices, rigorous testing, interpretability, verification, and theoretical guarantees where possible (Leike et al., 2018). See also Critch and Krueger (2020), Hubinger (2020), Hendrycks et al. (2021), and Ngo (2022) for additional overviews of strategies for building safer AI.

Robustness. As discussed in Section 3.3, models trained with RLHF can still exhibit undesired behavior due to distributional shifts between training and deployment. For example, adversarially engineered user inputs cause an LLM to output harmful text. To mitigate this problem, developers should use tools to generate inputs which result in undesired behavior and train against these adversarial examples (Zhang and Li, 2019; Ziegler et al., 2022; Perez et al., 2022a; Casper et al., 2023b). Anomaly detection techniques (Omar et al., 2013) can also be useful for flagging abnormal inputs likely to trigger bad behavior. Ensuring the security of important AI training runs against malicious human evaluators and/or outside cybersecurity threats will also be valuable.

Risk assessment and auditing. Although training processes should be crafted to produce models that are safe by design, evaluations are another layer of defense. Passing an evaluation is not proof of safety, but as is the case in almost every safety-critical industry, rigorous evaluations of capabilities and risks helps to spot hazards and establish trust. In practice, this should involve both in-house and second-party evaluations (OpenAI, 2023; ARC, 2022; Perez et al., 2022b). As with adversarial training for robustness, the development of improved red teaming techniques will be important (Perez et al., 2022a; Casper et al., 2023b).

Interpretability and model editing. Generating human-understandable explanations for the behavior of AI systems is currently an unsolved problem. Progress in explainability and interpretability could help verify hypotheses about how models make decisions (Geiger et al., 2023), including whether the decision-making process is trustworthy. In this way, it could be possible to gain confidence that models will (or will not) behave in a safe way without necessarily conducting extensive testing of the models (Jacovi et al., 2021). Red-teaming can also be complemented by interpretability techniques (Rastogi et al., 2023; Räuker et al., 2023), especially for purposes of identifying adversarial inputs (Ziegler et al., 2022; Casper et al., 2023c;a) or anomalous inputs (Pang et al., 2021). In another direction, better understanding the internal mechanisms of models can aid in directly editing model weights or intervening on internal activations in order to improve truthfulness (Li et al., 2023b), modify a model's factual knowledge (Meng et al., 2023; 2022; Hernandez et al., 2023; Hase et al., 2023), or otherwise steer model behavior (Cui et al., 2022).

5 Governance and Transparency

Social scientists and policymakers have increasingly focused on the need for governance frameworks to develop and deploy AI systems responsibly. Across historical contexts, a hallmark of mature scientific fields is the open sharing of research findings (Shapin and Schaffer, 2011) to allow experts to understand progress (Gilbert and Loveridge, 2021). Below we overview components of an RLHF governance agenda, including outstanding questions and risk dimensions.

Incentives and requirements for safety. Competition between labs can generate harmful race dynamics (Dafoe, 2018) because of tradeoffs between competitiveness and caution. This suggests a role for governance in promoting a healthier environment for safe AI research, development, and deployment (Dafoe, 2018; Perry and Uuk, 2019; Falco et al., 2021; Cihon, 2019; Anderljung et al., 2023). Governance in this form could involve mandates for independent auditing, evaluations, and certification (Shavit, 2023; Mökander et al., 2023; ARC, 2022; Hadfield and Clark, 2023; Shevlane et al., 2023); monitoring for post-deployment problems (Hendrycks and Gimpel, 2016); influence over resources including hardware and data (Brief, 2020; Chan et al., 2023a); and prohibiting deployment unless critical standards are met, as in the case of the U.S. Food and Drug Administration's oversight of clinical trials for testing potential new treatments (Junod, 2008).

Transparency and auditing. A sustained commitment to transparency would make the existing RLHF research environment more robust from a safety standpoint. First, the disclosure of some details behind large RLHF training runs would clarify a given organization's norms for model scrutiny and safety checks. Second, increased transparency about known efforts to mitigate risks could improve safety incentives and suggest methods for external stakeholders to hold companies accountable. Third, and most relevant for the present paper, transparency would improve the AI safety community's understanding of RLHF and support the ability to track technical progress on its challenges. Some level of disclosure is a precondition to evaluate the viability of the technical RLHF safety agenda over time and allow for community contribution to it. For all of these reasons, working to incorporate transparency standards into an AI governance framework will be important (Larsson and Heintz, 2020; Anderljung et al., 2023). It is possible that public disclosure of details critical to the development of model capabilities might lead to the unwanted proliferation of AI technologies

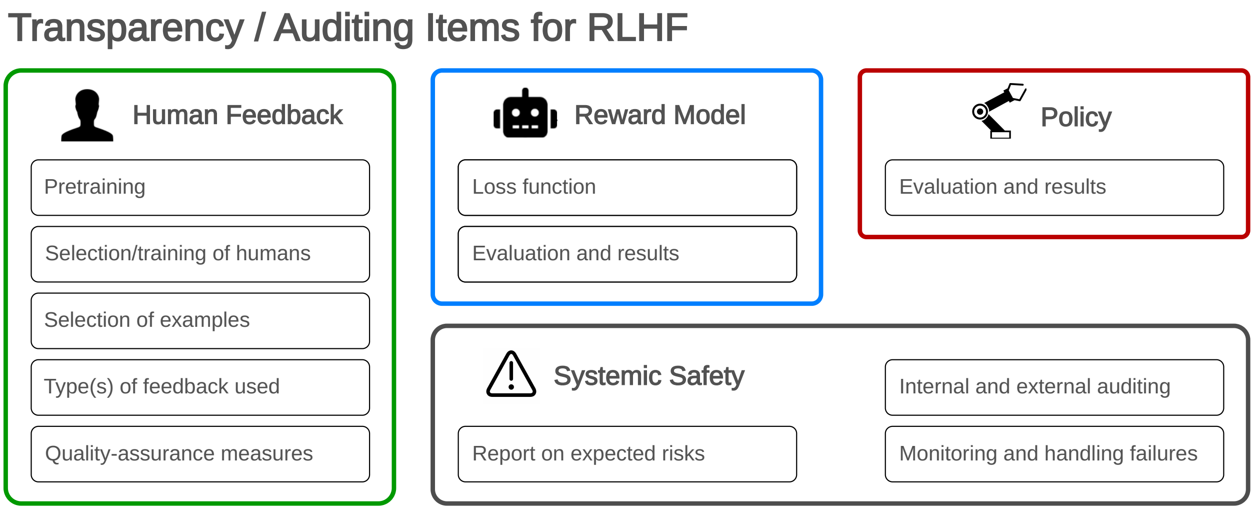

Figure 4: **Details behind an implementation of RLHF that, if disclosed, could be indicative of risks.** See Section 5 for a complete discussion. Companies using RLHF to train models for high-stakes or safety-critical applications should maintain transparency with the public and/or auditors about key details of their approach.

that could be misused. However, detailing safety measures will often not require divulging implementable details, and when it does, private disclosure to second-party auditors (Mökander et al., 2023; ARC, 2022; Hadfield and Clark, 2023; Shevlane et al., 2023) offers a solution.

As more specific policy prescriptions are beyond our scope, we encourage elaboration on these topics as part of a future research agenda. Below, however, we outline specific types of details that, if disclosed, could be indicative of risks and should be accounted for when auditing AI systems developed using RLHF. See also Figure 4.

Human feedback details:

- **A description of the pretraining process including details about what data was used** to make apparent possible biases that pretraining can cause.
- **How human evaluators were selected and trained** to provide information about risks of evaluators being malicious, unrepresentative, or incapable.
- **The process by which examples were selected to obtain feedback** to invite scrutiny about their representativeness and whether sufficient adversarial training was used. If examples were crowdsourced from a publicly-available application, details about what measures were taken to avoid data poisoning attacks should be provided.
- **The type(s) of human feedback used** (e.g., binary comparisons, scalar feedback, etc.) to suggest what risks might be caused by insufficiently abundant or rich feedback.
- **A report on measures taken for quality assurance in feedback collection and inter-rater consistency** to ensure that effective quality control measures were taken.

Reward model details:

- **The loss function used to fit the reward model and how disagreement was modeled** (e.g., as noise) to help with analyzing the degree of misspecification when fitting the reward model.
- **A report on reward model evaluation and results** to suggest possible problems from a misaligned reward model. The evaluation should involve red teaming.

Policy details:

- **A report on policy evaluation and results** to suggest possible troubles from a misaligned policy. The evaluation should involve red teaming and include assessment for risky capabilities (e.g., the ability to deceive a human).

Systemic safety measures

- **A report on internal and external audits and red teaming** to ensure accountability and disclose risks that are identified.
- **A report on expected risks and anticipated failure modes** to ensure accountability.
- **Plans for monitoring and correcting failures that emerge** to support post-deployment safety.

How these types of risks should be documented remains an area of active work in AI governance. Similar questions have been asked in an investigation by the US Federal Trade Commission into OpenAI (FTC, 2023) but in response to problems with ChatGPT rather than proactively. Salient documentation proposals focus on regular reporting of reward components (Gilbert et al., 2022) and the ability to compare the capabilities of language models according to standard benchmarks (Liang et al., 2022a). For the longer term, incorporating beneficial standards for safety and transparency into norms and regulations affecting AI is an ongoing challenge.

Concerns for social and economic equity. Although this paper has focused on technical challenges with RLHF, there are social and economic ones as well which governance and industry should work to address. For example, OpenAI has paid Kenyan knowledge workers at a rate of less than $2 USD per hour (Perrigo, 2023) for work which was mentally and emotionally demanding (Hao, 2023). Human subjects used in RLHF research should not be systematically selected simply for their availability or low cost (National Commission for the Protection of Human Subjects, 1978). Costs, benefits, and influence over RLHF models should be equitably distributed across different communities (Whittlestone et al., 2021; Eloundou et al., 2023). There is an additional possibility that powerful AI systems will be highly profitable and serve to concentrate large amounts of wealth and power into the hands of a few (O'Keefe et al., 2020; Chan et al., 2023b). Thus, policies that address inequalities and protect vulnerable populations (e.g. impacted communities, whistleblowers) will be increasingly important.

6 Discussion

While some problems with RLHF are tractable, others are fundamental. Technical progress in some respects is tractable, and this room for progress should be seen as a cause for concerted work and optimism. Even some of the fundamental problems that we overview can be alleviated with improved methodology even though they cannot be fully solved by RLHF. However, the fundamental nature of these problems requires that they be avoided or compensated for with non-RLHF approaches. Hence, we emphasize the importance of two strategies: (1) evaluating technical progress in light of the fundamental limitations of RLHF and other methods, and (2) addressing the sociotechnical challenges of aligning to human values by committing to both defense-in-depth safety measures and openly sharing research findings with the wider scientific community.

RLHF = Rehashing Lessons from Historical Failures? RLHF offers new capabilities but faces many old problems. Its use by Christiano et al. dates to 2017, and the individual components of it (preference elicitation, fitting a reward model, and policy optimization) have a history of technical and fundamental challenges in the fields of human-computer interaction and AI safety. In 2023, RLHF was described by the first author of Christiano et al. (2017) as a "basic solution" intended to make it easier to "productively work on more challenging alignment problems" (Christiano, 2023).[3] Some challenges and questions that we have

[3]Christiano (2023) mentions debate (Irving et al., 2018) and recursive reward modeling (Leike et al., 2018) as examples of 'more challenging alignment problems.' See also an outline of proposals in Hubinger (2020).

covered are rather unique to RLHF such as ones involving jointly training the reward model and policy (Section 3.4). However, many other problems are instances of broader ones in machine learning such as challenges with RL policies (Section 3.3). Others still are fundamental problems with AI alignment such as determining whose values are encoded into AI in a diverse society of humans (Section 3.2.1). The successes of RLHF should not obfuscate its limitations or gaps between the framework under which it is studied and real-world applications (see Appendix A). An approach to AI alignment that relies on RLHF without additional techniques for safety risks doubling-down on flawed approaches to AI alignment. Thus, it will be important to continue working to better understand RLHF while respecting its limitations.

Moving forward. RLHF has clear advantages for aligning AI systems with human goals. As a result, it has been key to the development of state-of-the-art LLMs and will likely continue to play a major role in modern AI. However, its use and influence should be accompanied by a commensurate research effort to better understand RLHF and address its flaws. Because it optimizes for human approval, RLHF in particular demands a special type of caution because many of its failures will actively tend to be ones that humans struggle to notice. It will be important to approach RLHF cautiously and work to incorporate it into a more holistic framework (Khlaaf, 2023) for safer AI with multiple layers of protection from failures (Hendrycks et al., 2021). Because some of the challenges with RLHF are fundamental to the AI alignment problem itself, moving forward will require confronting the basic choices and assumptions behind any given approach to aligning AI and who controls it (Dobbe et al., 2021). Moving forward, we urge that those working to develop advanced LLMs using RLHF both contribute toward resolving its open challenges and maintain transparency about the details of their approach to safety and any anticipated risks.

Contributions

Stephen Casper and Xander Davies served as the central writers and organizers.

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, Xin Chen, Dmitrii Krasheninnikov, Lauro Langosco, and Peter Hase contributed to writing and planning the paper.

Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell served as advisors.

Acknowledgements

We thank Sam Bowman, Adam Jermyn, Ethan Perez, Alan Chan, Gabriel Recchia, Robert Kirk, and Nathan Lambert for their helpful feedback. This work was facilitated in part by the Harvard AI Safety Team and MIT AI Alignment.

References

Afra Feyza Akyürek, Ekin Akyürek, Aman Madaan, Ashwin Kalyan, Peter Clark, Derry Wijaya, and Niket Tandon. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs. *arXiv preprint arXiv:2305.08844*, 2023.

Alex Albert. Jailbreak chat. 2023. URL `https://www.jailbreakchat.com/`.

Susan Amin, Maziar Gomrokchi, Harsh Satija, Herke van Hoof, and Doina Precup. A survey of exploration methods in reinforcement learning. *arXiv preprint arXiv:2109.00157*, 2021.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Markus Anderljung, Joslyn Barnhart, Jade Leung, Anton Korinek, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, Ben Chang, Tantum Collins, Tim Fist, Gillian Hadfield, Alan Hayes, Lewis Ho, Sara Hooker, Eric Horvitz, Noam Kolt, Jonas Schuett, Yonadav Shavit, Divya Siddarth, Robert Trager, and Kevin Wolf. Frontier ai regulation: Managing emerging risks to public safety, 2023.

Anthropic. Introducing claude, 2023. URL `https://www.anthropic.com/index/introducing-claude`.

ARC. Arc evals, 2022. URL `https://evals.alignment.org/`.

Dilip Arumugam, Jun Ki Lee, Sophie Saskin, and Michael L Littman. Deep reinforcement learning from policy-dependent human feedback. *arXiv preprint arXiv:1902.04257*, 2019.

Hui Bai. Artificial Intelligence Can Persuade Humans. 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

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

Andrea Bajcsy, Dylan P Losey, Marcia K O'Malley, and Anca D Dragan. Learning from physical human corrections, one feature at a time. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 141–149, 2018.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

Peter Barnett, Rachel Freedman, Justin Svegliato, and Stuart Russell. Active reward learning from multiple teachers. *arXiv preprint arXiv:2303.00894*, 2023.

Connor Baumler, Anna Sotnikova, and Hal Daumé III. Which examples should be multiply annotated? active learning when annotators may disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371, 2023.

James Bennett, Stan Lanning, et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York, 2007.

Kush Bhatia, Ashwin Pananjady, Peter Bartlett, Anca Dragan, and Martin J Wainwright. Preference learning along multiple criteria: A game-theoretic perspective. *Advances in neural information processing systems*, 33:7413–7424, 2020.

Erdem Biyik. *Learning Preferences For Interactive Autonomy.* PhD thesis, EE Department, Stanford University, 2022.

Erdem Biyik and Dorsa Sadigh. Batch active preference-based learning of reward functions. In *Conference on robot learning*, pages 519–528. PMLR, 2018.

Erdem Biyik, Malayandi Palan, Nicholas C. Landolfi, Dylan P. Losey, and Dorsa Sadigh. Asking easy questions: A user-friendly approach to active reward learning. In *Proceedings of the 3rd Conference on Robot Learning (CoRL)*, 2019.

Erdem Biyik, Nicolas Huynh, Mykel J. Kochenderfer, and Dorsa Sadigh. Active preference-based gaussian process regression for reward learning. In *Proceedings of Robotics: Science and Systems (RSS)*, July 2020. doi: 10.15607/rss.2020.xvi.041.

Erdem Bıyık, Dylan P Losey, Malayandi Palan, Nicholas C Landolfi, Gleb Shevchuk, and Dorsa Sadigh. Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences. *The International Journal of Robotics Research*, 41(1):45–67, 2022.

Andreea Bobu, Andrea Bajcsy, Jaime F Fisac, Sampada Deglurkar, and Anca D Dragan. Quantifying hypothesis space misspecification in learning from human–robot demonstrations and physical corrections. *IEEE Transactions on Robotics*, 36(3):835–854, 2020.

Andreea Bobu, Andi Peng, Pulkit Agrawal, Julie Shah, and Anca D Dragan. Aligning robot and human representations. *arXiv preprint arXiv:2302.01928*, 2023.

Michael Bowling, John D Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In *International Conference on Machine Learning*, pages 3003–3020. PMLR, 2023.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring Progress on Scalable Oversight for Large Language Models, November 2022. URL `http://arxiv.org/abs/2211.03540`. arXiv:2211.03540 [cs].

CSET Policy Brief. Why ai chips matter. 2020.

Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pages 783–792. PMLR, 2019.

Daniel Brown, Russell Coleman, Ravi Srinivasan, and Scott Niekum. Safe imitation learning via fast bayesian reward inference from preferences. In *International Conference on Machine Learning*, pages 1165–1177. PMLR, 2020.

Serkan Cabi, Sergio Gómez Colmenarejo, Alexander Novikov, Ksenia Konyushkova, Scott Reed, Rae Jeong, Konrad Zolna, Yusuf Aytar, David Budden, Mel Vecerik, et al. Scaling data-driven robotics with reward sketching and batch reinforcement learning. *arXiv preprint arXiv:1909.12200*, 2019.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*, 2023a.

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, et al. Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447*, 2023b.

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing Manipulation from AI Systems, March 2023. URL `http://arxiv.org/abs/2303.09387`. arXiv:2303.09387 [cs].

Micah D Carroll, Anca Dragan, Stuart Russell, and Dylan Hadfield-Menell. Estimating and penalizing induced preference shifts in recommender systems. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Stephen Casper. Achilles heels for agi/asi via decision theoretic adversaries. *arXiv preprint arXiv:2010.05418*, 2020.

Stephen Casper, Dylan Hadfield-Menell, and Gabriel Kreiman. White-box adversarial policies in deep reinforcement learning. *arXiv preprint arXiv:2209.02167*, 2022.

Stephen Casper, Yuxiao Li, Jiawei Li, Tong Bu, Kevin Zhang, and Dylan Hadfield-Menell. Benchmarking interpretability tools for deep neural networks. *arXiv preprint arXiv:2302.10894*, 2023a.

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. Explore, establish, exploit: Red teaming language models from scratch. *arXiv preprint arXiv:2306.09442*, 2023b.

Stephen Casper, Max Nadeau, Dylan Hadfield-Menell, and Gabriel Kreiman. Robust feature-level adversaries are interpretability tools, 2023c.

Christopher P Chambers and Federico Echenique. *Revealed preference theory*, volume 56. Cambridge University Press, 2016.

Alan Chan, Herbie Bradley, and Nitarshan Rajkumar. Reclaiming the digital commons: A public data trust for training data. *arXiv preprint arXiv:2303.09001*, 2023a.

Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco di Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. Harms from increasingly agentic algorithmic systems. *ArXiv*, abs/2302.10329, 2023b.

Lawrence Chan, Dylan Hadfield-Menell, Siddhartha Srinivasa, and Anca Dragan. The Assistive Multi-Armed Bandit. *arXiv:1901.08654 [cs, stat]*, January 2019. URL `http://arxiv.org/abs/1901.08654`. arXiv: 1901.08654.

Angelica Chen, Jérémy Scheurer, Tomasz Korbak, Jon Ander Campos, Jun Shern Chan, Samuel R Bowman, Kyunghyun Cho, and Ethan Perez. Improving code generation by training with natural language feedback. *arXiv preprint arXiv:2303.16749*, 2023.

Michael Chmielewski and Sarah C Kucker. An mturk crisis? shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4):464–473, 2020.

Paul Christiano. Worst-case guarantees. `https://ai-alignment.com/training-robust-corrigibility-ce0e0a3b9b4d`, 2019.

Paul Christiano. Thoughts on the impact of rlhf research, Jan 2023. URL `https://www.alignmentforum.org/posts/vwu4kegAEZTBtpT6p/thoughts-on-the-impact-of-rlhf-research#The_case_for_a_positive_impact:~:text=I%20think%20it%20is%20hard%20to%20productively%20work%20on%20more%20challenging%20alignment%20problems%20without%20first%20implementing%20basic%20solutions.`

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL `https://arxiv.org/abs/2210.11416`.

Peter Cihon. Standards for ai governance: international standards to enable global coordination in ai research & development. *Future of Humanity Institute. University of Oxford*, 2019.

Michael K Cohen and Marcus Hutter. Curiosity killed the cat and the asymptotically optimal agent. *arXiv preprint arXiv:2006.03357*, 2020.

Ajeya Cotra. Why ai alignment could be hard with modern deep learning. `https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/`, 2021.

Andrew Critch and David Krueger. Ai research considerations for human existential safety (arches). *arXiv preprint arXiv:2006.04948*, 2020.

Audrey Cui, Ali Jahanian, Agata Lapedriza, Antonio Torralba, Shahin Mahdizadehaghdam, Rohit Kumar, and David Bau. Local relighting of real scenes, 2022.

Allan Dafoe. Ai governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, 1442:1443, 2018.

Oliver Daniels-Koch and Rachel Freedman. The expertise problem: Learning from specialized feedback. *arXiv preprint arXiv:2211.06519*, 2022.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pages 12004–12019. PMLR, 2022.

Zihan Ding and Hao Dong. Challenges of reinforcement learning. *Deep Reinforcement Learning: Fundamentals, Research and Applications*, pages 249–272, 2020.

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. Hard choices in artificial intelligence. *Artificial Intelligence*, 300:103555, 2021.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Lê-Nguyên Hoang, Rafael Pinot, and John Stephan. Sok: On the impossible security of very large foundation models. *arXiv preprint arXiv:2209.15259*, 2022.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: An early look at the labor market impact potential of large language models, 2023.

Kawin Ethayarajh and Dan Jurafsky. The authenticity gap in human evaluation, 2022.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective. *arXiv:1908.04734 [cs]*, March 2021. URL `http://arxiv.org/abs/1908.04734`. arXiv: 1908.04734.

Gregory Falco, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, Alwyn Goodloe, Jerry Gupta, Christopher Hart, et al. Governing ai safety through independent audits. *Nature Machine Intelligence*, 3(7):566–571, 2021.

Michael Feffer, Hoda Heidari, and Zachary C Lipton. Moral machine or tyranny of the majority? *arXiv preprint arXiv:2305.17319*, 2023.

Luciano Floridi and Josh Cowls. A unified framework of five principles for ai in society. *Machine learning and the city: Applications in architecture and urban design*, pages 535–545, 2022.

Rachel Freedman, Rohin Shah, and Anca Dragan. Choice set misspecification in reward inference. *arXiv preprint arXiv:2101.07691*, 2021.

Aaron French. The mandela effect and new memory. *Correspondences*, 6(2), 2019.

FTC. "federal trade commission civil investigative demand schedule ftc file no. 232-3044", July 2023. URL `https://www.washingtonpost.com/documents/67a7081c-c770-4f05-a39e-9d02117e50e8.pdf?itid=lk_inline_manual_4`.

Justin Fu, Anoop Korattikara, Sergey Levine, and Sergio Guadarrama. From language to goals: Inverse reinforcement learning for vision-based instruction following. *arXiv preprint arXiv:1902.07742*, 2019.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. *arXiv preprint arXiv:2210.10760*, 2022.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL `https://aclanthology.org/2020.findings-emnlp.301`.

Atticus Geiger, Chris Potts, and Thomas Icard. Causal abstraction for faithful model interpretation. *arXiv preprint arXiv:2301.04709*, 2023. URL `https://arxiv.org/pdf/2301.04709.pdf`.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.

Thomas Krendl Gilbert and Andrew Loveridge. Subjectifying objectivity: Delineating tastes in theoretical quantum gravity research. *Social Studies of Science*, 51(1):73–99, 2021.

Thomas Krendl Gilbert, Nathan Lambert, Sarah Dean, Tom Zick, and Aaron Snoswell. Reward reports for reinforcement learning. *arXiv preprint arXiv:2204.10817*, 2022.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements, 2022.

Adam Gleave and Geoffrey Irving. Uncertainty estimation for language reward models. *arXiv preprint arXiv:2203.07472*, 2022.

Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. Quantifying differences in reward functions. *arXiv preprint arXiv:2006.13900*, 2020a.

Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. In *International Conference on Learning Representations*, 2020b. URL `https://openreview.net/forum?id=HJgEMpVFwB`.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. Aligning language models with preferences through f-divergence minimization, 2023.

Google. Bard, 2023. URL https://bard.google.com/.

Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

Prasoon Goyal, Scott Niekum, and Raymond J Mooney. Using natural language for reward shaping in reinforcement learning. *arXiv preprint arXiv:1903.02020*, 2019.

Lewis D. Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly T. Mai, Maria Vau, Matthew Caldwell, and Augustine Marvor-Parker. Susceptibility to Influence of Large Language Models, March 2023. URL http://arxiv.org/abs/2303.06074. arXiv:2303.06074 [cs].

Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. Ground (less) truth: A causal framework for proxy labels in human-algorithm decision-making. *arXiv preprint arXiv:2302.06503*, 2023.

Gillian K Hadfield and Jack Clark. Regulatory markets: The future of ai governance. *arXiv preprint arXiv:2304.04914*, 2023.

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

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. *Advances in neural information processing systems*, 30, 2017.

Karen Hao. The hidden workforce that helped filter violence and abuse out of chatgpt, 2023. URL https://www.wsj.com/podcasts/the-journal/the-hidden-workforce-that-helped-filter-violence-and-abuse-out-of-chatgpt/ffc2427f-bdd8-47b7-9a4b-27e7267cf413.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*, 2023.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *arXiv preprint arXiv:2301.04213*, 2023. URL https://arxiv.org/pdf/2301.04213.pdf.

Joey Hejna and Dorsa Sadigh. Few-shot preference learning for human-in-the-loop rl. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

Evan Hernandez, Belinda Z Li, and Jacob Andreas. Measuring and manipulating knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023.

Jacob Hilton, Nick Cammarata, Shan Carter, Gabriel Goh, and Chris Olah. Understanding rl vision. *Distill*, 5(11):e29, 2020.

Joey Hong, Kush Bhatia, and Anca Dragan. On the sensitivity of reward inference to misspecified human models. *arXiv preprint arXiv:2212.04717*, 2022.

Keith Hoskin. The 'awful idea of accountability': inscribing people into the measurement of objects. *Accountability: Power, ethos and the technologies of managing*, 265, 1996.

Jian Hu, Li Tao, June Yang, and Chandler Zhou. Aligning language models with offline reinforcement learning from human feedback. *arXiv preprint arXiv:2308.12050*, 2023.

Evan Hubinger. An overview of 11 proposals for building safe advanced ai. *arXiv preprint arXiv:2012.07532*, 2020.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31, 2018.

Alex Irpan. Deep reinforcement learning doesn't work yet. `https://www.alexirpan.com/2018/02/14/rl-hard.html`, 2018.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021. URL `https://arxiv.org/pdf/2010.07487.pdf`.

Nils Jansen, Bettina Könighofer, Sebastian Junges, Alexandru C Serban, and Roderick Bloem. Safe reinforcement learning via probabilistic shields. *arXiv preprint arXiv:1807.06096*, 2018.

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

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Suzanne Junod. Fda and clinical drug trials: a short history. *FDLI Update*, page 55, 2008.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. Alignment of Language Agents, March 2021. URL `http://arxiv.org/abs/2103.14659`. arXiv:2103.14659 [cs].

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. A distributional approach to controlled text generation. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=jWkw45-9AbL`.

Heidy Khlaaf. Toward comprehensive risk assessments and assurance of ai-based systems. *Trail of Bits*, 2023.

Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Min Yoo, and Minjoon Seo. Aligning large language models through synthetic feedback. *arXiv preprint arXiv:2305.13735*, 2023.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453*, 2023.

W Bradley Knox and Peter Stone. Tamer: Training an agent manually via evaluative reinforcement. In *2008 7th IEEE international conference on development and learning*, pages 292–297. IEEE, 2008.

W Bradley Knox, Stephane Hatgis-Kessell, Serena Booth, Scott Niekum, Peter Stone, and Alessandro Allievi. Models of human preference for learning reward functions. *arXiv preprint arXiv:2206.02231*, 2022.

Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 16203–16220. Curran Associates, Inc., 2022a. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/67496dfa96afddab795530cc7c69b57a-Paper-Conference.pdf`.

Tomasz Korbak, Ethan Perez, and Christopher Buckley. RL with KL penalties is better viewed as Bayesian inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1083–1091, Abu Dhabi, United Arab Emirates, December 2022b. Association for Computational Linguistics. URL `https://aclanthology.org/2022.findings-emnlp.77`.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. Pretraining language models with human preferences, 2023.

Jernej Kos and Dawn Song. Delving into adversarial attacks on deep policies. *arXiv preprint arXiv:1705.06452*, 2017.

Victoria Krakovna and Janos Kramar. Power-seeking can be probable and predictive for trained agents. *arXiv preprint arXiv:2304.06528*, 2023.

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. *DeepMind Blog*, 2020.

Dmitrii Krasheninnikov, Egor Krasheninnikov, and David Krueger. Assistance with large language models. In *NeurIPS ML Safety Workshop*, 2022.

David Krueger, Tegan Maharaj, and Jan Leike. Hidden incentives for auto-induced distributional shift, 2020.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. Designing toxic content classification for a diversity of perspectives. In *SOUPS@ USENIX Security Symposium*, pages 299–318, 2021.

Stefan Larsson and Fredrik Heintz. Transparency in artificial intelligence. *Internet Policy Review*, 9(2), 2020.

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

Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*, 2023a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2023b.

Mengxi Li, Alper Canberk, Dylan P Losey, and Dorsa Sadigh. Learning human objectives from sequences of physical corrections. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2877–2883. IEEE, 2021.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022a.

Xinran Liang, Katherine Shu, Kimin Lee, and Pieter Abbeel. Reward uncertainty for exploration in preference-based reinforcement learning. In *International Conference on Learning Representations*, 2022b.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Jessy Lin, Daniel Fried, Dan Klein, and Anca Dragan. Inferring rewards from language in context. *arXiv preprint arXiv:2204.02515*, 2022.

David Lindner and Mennatallah El-Assady. Humans are not boltzmann distributions: Challenges and opportunities for modelling human feedback and interaction in reinforcement learning. *arXiv preprint arXiv:2206.13316*, 2022.

David Lindner, Xin Chen, Sebastian Tschiatschek, Katja Hofmann, and Andreas Krause. Learning safety constraints from demonstrations with unknown rewards. *arXiv preprint arXiv:2305.16147*, 2023.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. Jailbreaking chatgpt via prompt engineering: An empirical study, 2023.

Dylan P Losey, Andrea Bajcsy, Marcia K O'Malley, and Anca D Dragan. Physical interaction as communication: Learning robot objectives online from human corrections. *The International Journal of Robotics Research*, 41(1):20–44, 2022.

Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. A survey of reinforcement learning informed by natural language. *arXiv preprint arXiv:1906.03926*, 2019.

James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. Interactive learning from policy-dependent human feedback. In *International Conference on Machine Learning*, pages 2285–2294. PMLR, 2017.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.

Shehryar Malik, Usman Anwar, Alireza Aghasi, and Ali Ahmed. Inverse constrained reinforcement learning. In *International conference on machine learning*, pages 7390–7399. PMLR, 2021.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018.

Lev McKinney, Yawen Duan, David Krueger, and Adam Gleave. On the fragility of learned reward functions. *arXiv preprint arXiv:2301.03652*, 2023.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023.

Eric J Michaud, Adam Gleave, and Stuart Russell. Understanding learned reward functions. *arXiv preprint arXiv:2012.05862*, 2020.

Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. Ethical aspects of multi-stakeholder recommendation systems. *The information society*, 37(1):35–45, 2021.

Smitha Milli and Anca D Dragan. Literal or pedagogic human? analyzing human model misspecification in objective learning. In *Uncertainty in artificial intelligence*, pages 925–934. PMLR, 2020.

Soren Mindermann and Stuart Armstrong. Occam's razor is insufficient to infer the preferences of irrational agents. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 5603–5614, Red Hook, NY, USA, 2018. Curran Associates Inc.

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: a three-layered approach. *arXiv preprint arXiv:2302.08500*, 2023.

Vivek Myers, Erdem Biyik, Nima Anari, and Dorsa Sadigh. Learning multimodal rewards from rankings. In *Conference on Robot Learning*, pages 342–352. PMLR, 2021.

United States National Commission for the Protection of Human Subjects. *The Belmont report: ethical principles and guidelines for the protection of human subjects of research*, volume 1. United States Department of Health, Education, and Welfare, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978.

Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

Richard Ngo. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

Khanh Nguyen, Hal Daumé III, and Jordan Boyd-Graber. Reinforcement learning for bandit neural machine translation with simulated human feedback. *arXiv preprint arXiv:1707.07402*, 2017.

Evgenii Nikishin, Pavel Izmailov, Ben Athiwaratkun, Dmitrii Podoprikhin, Timur Garipov, Pavel Shvechikov, Dmitry Vetrov, and Andrew Gordon Wilson. Improving stability in deep reinforcement learning with weight averaging. In *Uncertainty in artificial intelligence workshop on uncertainty in Deep learning*, 2018.

Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Cullen O'Keefe, Peter Cihon, Ben Garfinkel, Carrick Flynn, Jade Leung, and Allan Dafoe. The windfall clause: Distributing the benefits of ai for the common good. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 327–331, 2020.

Salima Omar, Asri Ngadi, and Hamid H Jebur. Machine learning techniques for anomaly detection: an overview. *International Journal of Computer Applications*, 79(2), 2013.

A.J. Oneal. Chat gpt "dan" (and other "jailbreaks"). `https://gist.github.com/coolaj86/6f4f7b30129b0251f61fa7baaa881516`, 2023.

OpenAI. Gpt-4 technical report, 2023.

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

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*, 2022.

Rahul Pandey, Hemant Purohit, Carlos Castillo, and Valerie L Shalin. Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning. *International Journal of Human-Computer Studies*, 160:102772, 2022.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. What you see is what you get? the impact of representation criteria on human bias in hiring. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 125–134, 2019.

Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, and Ece Kamar. Investigations of performance and bias in human-ai teamwork in hiring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12089–12097, 2022.

Andi Peng, Aviv Netanyahu, Mark K Ho, Tianmin Shu, Andreea Bobu, Julie Shah, and Pulkit Agrawal. Diagnosis, feedback, adaptation: A human-in-the-loop framework for test-time policy adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022a.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022b.

Billy Perrigo. Exclusive: The $2 per hour workers who made chatgpt safer, 2023. URL `https://time.com/6247678/openai-chatgpt-kenya-workers/`. [Accessed 07-May-2023].

Brandon Perry and Risto Uuk. Ai governance and the policymaking process: key considerations for reducing ai risk. *Big data and cognitive computing*, 3(2):26, 2019.

Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh. Do users write more insecure code with ai assistants?, 2022.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 2586–2591, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*, 2022.

Alexandre Rame, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *arXiv preprint arXiv:2306.04488*, 2023.

Abhinav Rao, Sachin Vashistha, Atharva Naik, Somak Aditya, and Monojit Choudhury. Tricking llms into disobedience: Understanding, analyzing, and preventing jailbreaks, 2023.

Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, and Saleema Amershi. Supporting human-ai collaboration in auditing llms with llms. *arXiv preprint arXiv:2304.09991*, 2023. URL `https://arxiv.org/pdf/2304.09991.pdf`.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 464–483. IEEE, 2023.

Siddharth Reddy, Anca D. Dragan, and Sergey Levine. Where Do You Think You're Going?: Inferring Beliefs about Dynamics from Behavior. *arXiv:1805.08010 [cs, stat]*, January 2019. URL `http://arxiv.org/abs/1805.08010`. arXiv: 1805.08010.

Siddharth Reddy, Sergey Levine, and Anca D Dragan. Assisted Perception: Optimizing Observations to Communicate State. 2020.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. *Active preference-based learning of reward functions.* 2017.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=9Vrb9D0WI4`.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023.

Laura Sartori and Andreas Theodorou. A sociotechnical perspective for the future of ai: narratives, inequalities, and human control. *Ethics and Information Technology*, 24(1):4, 2022.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training language models with language feedback. In *The First Workshop on Learning with Natural Language Supervision at ACL*, 2022.

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. Training language models with language feedback at scale. *arXiv preprint arXiv:2303.16755*, 2023.

Amartya Sen. Social choice theory. *Handbook of mathematical economics*, 3:1073–1181, 1986.

Rohin Shah, Noah Gundotra, Pieter Abbeel, and Anca Dragan. On the feasibility of learning, rather than assuming, human biases for reward inference. In *International Conference on Machine Learning*, pages 5670–5679. PMLR, 2019.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv preprint arXiv:2210.01790*, 2022.

Steven Shapin and Simon Schaffer. *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life.* Princeton University Press, 2011.

Pratyusha Sharma, Balakumar Sundaralingam, Valts Blukis, Chris Paxton, Tucker Hermans, Antonio Torralba, Jacob Andreas, and Dieter Fox. Correcting robot plans with natural language feedback. *arXiv preprint arXiv:2204.05186*, 2022.

Yonadav Shavit. What does it take to catch a chinchilla? verifying rules on large-scale neural network training via compute monitoring, 2023.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*, 2023.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

Umer Siddique, Abhinav Sinha, and Yongcan Cao. Fairness in preference-based reinforcement learning, 2023.

David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.

Joar Skalse and Alessandro Abate. Misspecification in inverse reinforcement learning. *arXiv preprint arXiv:2212.03201*, 2022a.

Joar Skalse, Nikolaus HR Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. *arXiv preprint arXiv:2209.13085*, 2022.

Joar Max Viktor Skalse and Alessandro Abate. The reward hypothesis is false. In *NeurIPS ML Safety Workshop*, 2022b.

Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *International Conference on Machine Learning*, pages 32033–32058. PMLR, 2023.

Charlie Snell, Ilya Kostrikov, Yi Su, Mengjiao Yang, and Sergey Levine. Offline rl for natural language generation with implicit language q learning, 2022. URL `https://arxiv.org/abs/2206.11871`.

Aaron J. Snoswell and Jean Burgess. The Galactica AI model was trained on scientific knowledge – but it spat out alarmingly plausible nonsense, November 2022. URL `http://theconversation.com/the-galactica-ai-model-was-trained-on-scientific-knowledge-but-it-spat-out-alarmingly-plausible-nons`

Irene Solaiman and Christy Dennison. Process for adapting language models to society (palms) with values-targeted datasets. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 5861–5873. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/2e855f9489df0712b4bd8ea9e2848c5a-Paper.pdf`.

Ziang Song, Tianle Cai, Jason D Lee, and Weijie J Su. Reward collapse in aligning large language models. *arXiv preprint arXiv:2305.17608*, 2023.

Krishnan Srinivasan, Benjamin Eysenbach, Sehoon Ha, Jie Tan, and Chelsea Finn. Learning to be safe: Deep rl with a safety critic. *arXiv preprint arXiv:2010.14603*, 2020.

Jacob Steinhardt. Emergent Deception and Emergent Optimization, February 2023. URL `https://bounded-regret.ghost.io/emergent-deception-optimization/`.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Theodore R Sumers, Mark K Ho, Robert D Hawkins, Karthik Narasimhan, and Thomas L Griffiths. Learning rewards from linguistic feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6002–6010, 2021.

Ran Tian, Masayoshi Tomizuka, Anca Dragan, and Andrea Bajcsy. Towards Modeling and Influencing the Dynamics of Human Learning, January 2023. URL `http://arxiv.org/abs/2301.00901`. arXiv:2301.00901 [cs].

Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca Dragan, and Daniel S Brown. Causal confusion and reward misidentification in preference-based reward learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

Alexander M Turner. Seeking power is convergently instrumental in a broad class of environments, 2021. URL `https://www.alignmentforum.org/s/fSMbebQyR4wheRrvk/p/hzeLSQ9nwDkPc4KNt`.

Alexander Matt Turner and Prasad Tadepalli. Parametrically retargetable decision-makers tend to seek power. *ArXiv*, abs/2206.13477, 2022.

Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. In *Neural Information Processing Systems*, 2019.

Victor Uc-Cetina, Nicolas Navarro-Guerrero, Anabel Martin-Gonzalez, Cornelius Weber, and Stefan Wermter. Survey on reinforcement learning for language processing. *Artificial Intelligence Review*, 56 (2):1543–1575, 2023.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Peter Vamplew, Benjamin J Smith, Johan Källström, Gabriel Ramos, Roxana Rădulescu, Diederik M Roijers, Conor F Hayes, Fredrik Heintz, Patrick Mannion, Pieter JK Libin, et al. Scalar reward is not enough: A response to silver, singh, precup and sutton (2021). *Autonomous Agents and Multi-Agent Systems*, 36(2): 41, 2022.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*, 2023.

James Vincent. Microsoft's Bing is an emotionally manipulative liar, and people love it, February 2023. URL `https://www.theverge.com/2023/2/15/23599072/microsoft-ai-bing-personality-conversations-spy-employees-webcams`.

Alex Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *International Conference on Machine Learning*, 2023.

Tony Tong Wang, Adam Gleave, Nora Belrose, Tom Tseng, Joseph Miller, Michael D Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, and Stuart Russell. Adversarial policies beat professional-level go ais. *arXiv preprint arXiv:2211.00241*, 2022.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. Aligning large language models with human: A survey, 2023.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models, 2021.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.210. URL `https://aclanthology.org/2021.findings-emnlp.210`.

Jess Whittlestone, Kai Arulkumaran, and Matthew Crosby. The societal implications of deep reinforcement learning. *Journal of Artificial Intelligence Research*, 70:1003–1030, 2021.

Nils Wilde, Erdem Biyik, Dorsa Sadigh, and Stephen L Smith. Learning reward functions from scale feedback. In *Conference on Robot Learning*, pages 353–362. PMLR, 2022.

Simon Willison. Prompt injection. 2023. URL `https://simonwillison.net/series/prompt-injection/`.

Christian Wirth, Riad Akrour, Gerhard Neumann, Johannes Fürnkranz, et al. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

Yotam Wolf, Noam Wies, Yoav Levine, and Amnon Shashua. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*, 2023.

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback, 2021a.

Xian Wu, Wenbo Guo, Hua Wei, and Xinyu Xing. Adversarial policy training against deep reinforcement learning. In *USENIX Security Symposium*, pages 1883–1900, 2021b.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training, 2023.

Blake Wulfe, Logan Michael Ellis, Jean Mercat, Rowan Thomas McAllister, and Adrien Gaidon. Dynamics-aware comparison of learned reward functions. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=CALFyKVs87`.

Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. *arXiv preprint arXiv:2305.14710*, 2023a.

Wanqiao Xu, Shi Dong, Dilip Arumugam, and Benjamin Van Roy. Shattering the agent-environment interface for fine-tuning inclusive language models. *arXiv preprint arXiv:2305.11455*, 2023b.

Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Jianye Hao, Zhaopeng Meng, Peng Liu, and Zhen Wang. Exploration in deep reinforcement learning: a comprehensive survey. *arXiv preprint arXiv:2109.06668*, 2021.

Georgios N Yannakakis and John Hallam. Ranking vs. preference: a comparative study of self-reporting. In *Affective Computing and Intelligent Interaction: 4th International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part I 4*, pages 437–446. Springer, 2011.

Seonghyeon Ye, Yongrae Jo, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, and Minjoon Seo. Selfee: Iterative self-revising llm empowered by self-feedback generation, 2023. URL `https://kaistai.github.io/SelFee/`.

Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, Brian Ichter, Ted Xiao, Peng Xu, Andy Zeng, Tingnan Zhang, Nicolas Heess, Dorsa Sadigh, Jie Tan, Yuval Tassa, and Fei Xia. Language to rewards for robotic skill synthesis. *Arxiv preprint arXiv:2306.08647*, 2023.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears, 2023.

Sheng Yue, Guanbo Wang, Wei Shao, Zhaofeng Zhang, Sen Lin, Ju Ren, and Junshan Zhang. Clare: Conservative model-based reward learning for offline inverse reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Jiliang Zhang and Chen Li. Adversarial examples: Opportunities and challenges. *IEEE transactions on neural networks and learning systems*, 31(7):2578–2593, 2019.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*, 2023.

Songyuan Zhang, Zhangjie Cao, Dorsa Sadigh, and Yanan Sui. Confidence-aware imitation learning from demonstrations with varying optimality. *Advances in Neural Information Processing Systems*, 34:12340–12350, 2021.

Zhibing Zhao, Peter Piech, and Lirong Xia. Learning mixtures of plackett-luce models. In *International Conference on Machine Learning*, pages 2906–2914. PMLR, 2016.

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

Li Zhou and Kevin Small. Inverse reinforcement learning with natural language goals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11116–11124, 2021.

Banghua Zhu, Jiantao Jiao, and Michael I Jordan. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *arXiv preprint arXiv:2301.11270*, 2023.

Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned ai. *Advances in Neural Information Processing Systems*, 33:15763–15773, 2020.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

Daniel Ziegler, Seraphina Nix, Lawrence Chan, Tim Bauman, Peter Schmidt-Nielsen, Tao Lin, Adam Scherlis, Noa Nabeshima, Benjamin Weinstein-Raun, Daniel de Haas, et al. Adversarial training for high-stakes reliability. *Advances in Neural Information Processing Systems*, 35:9274–9286, 2022.

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. *arXiv preprint arXiv:1909.08593*, 2019.

A An Improved Model of the Human Feedback Process

As illustrated in Equation (1), the feedback process in RLHF is typically modeled with a single human $\mathcal{H}$ with internal reward function $r_{\mathcal{H}}$; examples sampled from the base model: $x_i \sim \pi_\theta$; and feedback as a function of the human, example, and noise: $y_i = f(h, x_i, \epsilon_i)$. However, as discussed in Section 3, this is a misspecified model of the process: there is not a single human, humans values are not representable with a reward function, human actions are dependent on context, and the sampling process can involve a human. Thus we propose an alternative formulation.

Let $\Delta\mathcal{H}$ refer to a joint distribution of humans (or groups thereof if feedback is provided collaboratively) used for obtaining samples and feedback denoted as $\mathcal{H}_j^{\text{sample}}$ and $\mathcal{H}_j^{\text{feedback}}$. A dataset of examples is sampled from π_θ (or some other source) where each example x_i is defined to be a batch of one or more generations from the base model. Importantly, x_i may not contain all information about the world state (e.g., if x_i is a 2D rendering of a 3D environment), and the human may be able to observe more than just the model's output (e.g., if interpretability tools are used to aid in evaluation). So let v be a rendering function that maps π_θ and x_i to what a human sees. The behavior of humans varies over time and in different contexts, so let c_i^{sample} and c_i^{feedback} represent particular contexts for sampling and feedback collection. Denote the sampling process as s which maps the base model π_θ, a human $\mathcal{H}_j^{\text{sample}}$, and context c_i^{sample} to some example x_i. Notably, s could ignore the base model and generate offline samples from some other source. Finally, let f map a human $\mathcal{H}_j^{\text{feedback}}$, rendered example $v(\pi_\theta, x_i)$, and context c_i^{feedback} to feedback y_i. The data collection process can thus be more completely modeled as:

$$\mathcal{H}_j^{\text{sample}}, \mathcal{H}_j^{\text{feedback}} \sim \Delta\mathcal{H}, \qquad x_i \sim s(\pi_\theta, \mathcal{H}_j^{\text{sample}}, c_i^{\text{sample}}), \qquad y_i = f(v(\pi_\theta, x_i), \mathcal{H}_j^{\text{feedback}}, c_i^{\text{feedback}}) \tag{4}$$

which highlights a need for future work to better account for the aspects of this process that are commonly not accounted for when training systems with RLHF.

B Rationale for Why Challenges Were Categorized as Tractable or Fundamental

In Section 3, we categorize problems as **tractable** or **fundamental**. The key distinction between the two is that fundamental challenges are substantial enough that overcoming them would require a method that is no longer a form of RLHF. Although many of the fundamental problems we identify can be alleviated by improving how RLHF is approached, they could be fully addressed with RLHF. As a result, they should be either avoided by not using RLHF or compensated for by other safety measures. This distinction is soft, and some categories of challenges are marginal. Here, we briefly explain each categorization.

B.1 Problems from Section 3.1:

Tractable: **Selecting representative humans and getting them to provide quality feedback is difficult:** This can be addressed by studying and improving the selection and training of evaluators.

Tractable: **Some evaluators have harmful biases and opinions:** This can be addressed by studying and improving the selection and training of evaluators.

Tractable: **Individual human evaluators can poison data:** This can be addressed with improved evaluator selection and quality assurance measures.

Tractable: **Humans make simple mistakes due to limited time, attention, or care:** This is marginal because human mistakes can never fully be overcome. However, they can be addressed with improved working conditions and quality assurance procedures.

Tractable: **Partial observability limits human evaluators:** Human evaluators can be provided with all information available in the policy's observations (although representing this in an easily-comprehensible way may be challenging).

Fundamental: **Humans cannot evaluate performance on difficult tasks well:** Human intelligence and cognitive capacity are limited. Humans cannot be expected to properly evaluate the performance of superhuman models on complex tasks. Thus, solving this problem would require no longer using human feedback in the way that RLHF does.

Fundamental: **Humans can be misled, so their evaluations can be gamed:** Human fallibility cannot fully be overcome, especially against optimization pressure from the learned policy.

Tractable: **Data collection can introduce harmful biases:** This can be addressed with improved data curation.

Fundamental: **There is an inherent cost/quality tradeoff when collecting human feedback:** This tradeoff is unavoidable in practice – obtaining diverse and high-quality examples (e.g. from long chatbot conversations) requires more effort.

Fundamental: **RLHF suffers from a tradeoff between the richness and efficiency of feedback types:** This tradeoff is unavoidable for data collection in practice – richer annotations require more effort.

B.2 Problems from Section 3.2:

Fundamental: **An individual human's values are difficult to represent with a reward function:** This problem is marginal. It can be improved in practice by improved modeling, but RLHF-based solutions will be limited by the intractability of perfectly modeling context and troubles with the reward hypothesis (Skalse and Abate, 2022b; Bowling et al., 2023).

Fundamental: **A single reward function cannot represent a diverse society of humans:** Trivial. Instead of being a fundamental limitation with RLHF, this is a broader limitation of AI alignment itself.

Fundamental: **Reward models can misgeneralize to be poor reward proxies, even from correctly-labeled training data:** This problem is marginal because it can and should be addressed by improved sampling in practice. However, it is impossible to perfectly represent a distribution with infinite support from a finite sample. Additionally, the deployment distribution will always differ from the training and evaluation distributions in real-world settings (Christiano, 2019).

Fundamental: **Optimizing for an imperfect reward proxy leads to reward hacking:** If a reward model is imperfect, reward hacking will always be a possibility from RL.

Tractable: **Evaluating reward models is difficult and expensive:** This can be addressed by performing thorough and expensive evaluations.

B.3 Problems from Section 3.3:

Tractable: **It is (still) challenging to optimize policies effectively:** This can be addressed with advancements in RL methodology.

Tractable: **Policies tend to be adversarially exploitable:** This problem is marginal because achieving certified adversarial robustness against practical threat models has empirically been intractable. Nonetheless, this can be addressed with robust optimization techniques.

Fundamental: **Policies can perform poorly in deployment even if rewards seen during training were perfectly correct:** This problem is marginal because it can and should be addressed by improved sampling in practice. However, it is impossible to perfectly represent a distribution with infinite support from a finite sample. Additionally, the deployment distribution will always differ from the training and evaluation distributions in real-world settings Christiano (2019).

Fundamental: **Optimal RL agents tend to seek power:** Power is instrumentally useful for agents.

Tractable: **The pretrained model introduces biases into policy optimization:** This can be addressed with improved base models.

Tractable: **RL contributes to mode collapse:** This can be addressed with forms of RL that optimize for distribution-matching in desired instances.

B.4 Problems from Section 3.4:

Tractable: **Joint training induces distribution shifts:** This can be mitigated with synchronous learning or other strategies.

Tractable: **It is difficult to balance efficiency and avoiding overfitting by the policy:** This can be addressed with improved training methodology.