On the Challenges and Practices of Reinforcement Learning from Real Human Feedback

Timo Kaufmann*✉, Sarah Ball*, Jacob Beck,
Eyke Hüllermeier, and Frauke Kreuter

LMU Munich, Geschwister-Scholl-Platz 1, 80539 Munich, Germany
{timo.kaufmann,eyke}@ifi.lmu.de,
{sarah.ball,jacob.beck,frauke.kreuter}@stat.uni-muenchen.de

Abstract. Reinforcement learning from human feedback (RLHF) is a variant of reinforcement learning (RL) that does not require an engineered reward function but instead learns from human feedback. Due to its increasing popularity, various authors have studied how to learn an accurate reward model from only few samples, making optimal use of this feedback. Because of the cost and complexity of user studies, however, this research is often conducted with synthetic human feedback. Such feedback can be generated by evaluating behavior based on ground-truth rewards which are available for some benchmark tasks. While this setting can help evaluate some aspects of RLHF, it differs from practical settings in which synthetic feedback is not available. Working with real human feedback brings additional challenges that cannot be observed with synthetic feedback, including fatigue, inter-rater inconsistencies, delay, misunderstandings, and modality-dependent difficulties. We describe and discuss some of these challenges together with current practices and opportunities for further research in this paper.

Keywords: Reinforcement learning · RLHF · Human feedback.

1 Introduction

Reinforcement learning (RL) is a general framework of solving tasks by rewarded interaction with an environment. In contrast to supervised learning, which learns from a set of examples labeled with their solutions, RL can learn from experience without the need for such labels. It only requires a reward signal that can evaluate possible behaviors. As such, RL could be broadly useful in many domains. Yet, until recently, RL has barely been deployed in practice. This discrepancy can largely be attributed to two reasons: The data-inefficiency of RL, i.e., the amount of (potentially unsafe or expensive) interactions with the environment that is necessary to learn a useful behavior, and the difficulty of correctly specifying the desired behavior.

The second issue, the difficulty of specification, is the main concern of this paper. We focus on reinforcement learning from human feedback (RLHF) in

* Equal contribution

particular, which is a class of methods that employs human feedback for task specification. In the regular RL setting without human feedback, tasks are specified by a numerical reward signal occasionally provided to the agent after taking an action. This reward is usually determined by a reward function that can be evaluated automatically. Such a function is easy to specify in games, wherefore so many of the prominent successes of RL have been in these domains, such as Atari [43], Go [54] and StarCraft [57]. Now contrast these examples with tasks such as robotic manipulation, self-driving cars, household robots and care robots. Tasks in all of these domains are much harder to specify with a reward function.

Learning from Human Feedback. RLHF modifies the RL setting to learn from human feedback, commonly in the form of pairwise comparisons, instead of pre-specified reward functions. Humans are asked to provide feedback on a small subset of the agent's experiences and the agent is trained to behave in accordance with that feedback. Section 2 describes this setting in more detail.

As previously noted, data-inefficiency and difficulty of task specification can be considered the two main limitations of RL. Since RLHF at least partially solves the task specification problem, it has recently seen a number of successful applications in domains in which interaction with the environment is cheap and without great risk, e.g., language [48], simulated continuous control [15] and games [27]. There has also been some success applying RLHF to robotics [24], which requires greater care with data efficiency.

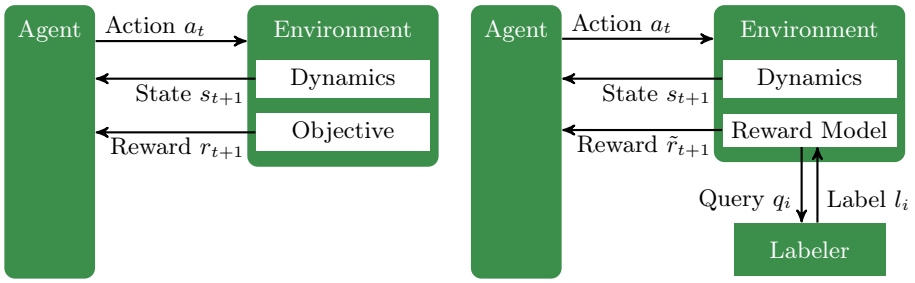
Insufficiency of Synthetic Feedback. While RLHF has generally been recognized to be a useful tool to specify objectives for RL agents, we argue that there is a lack of research into the feedback collection itself. Many recent works have attempted to elaborate on the technical aspects of RLHF: Develop new methods and algorithms to make more efficient use of human feedback. While the technical aspects are well-researched, the question of how to optimally design user studies, i.e., methods of gathering human feedback, received less attention. Since user studies can be difficult and expensive, many recent advances have only been evaluated with synthetic feedback. This is possible by choosing benchmark tasks for which ground-truth rewards are available, such as games with a pre-defined score function, and then using these rewards to generate synthetic preferences, e.g., preferring behaviors with higher reward.

Such an evaluation with synthetic feedback is useful to assess some properties of RLHF methods such as sample efficiency or final performance. In fact, synthetic feedback should be a core component of a RLHF evaluation suite since it allows for many cheap experiments and enables consistent, fair, and systematic comparisons of different methods. It is not sufficient on its own however: Synthetic feedback can miss many realities that arise when interacting with real humans, such as fatigue, distraction, inter-labeler inconsistencies, labeling delay, prior exposure or the relative difficulty of different feedback modalities. These realities are of crucial importance for practical applications of RLHF, in which ground-truth rewards and therefore synthetic feedback is generally not available, and should not be neglected in research either.

As we have identified in this section, there is a gap between the common practice of RLHF evaluation with synthetic feedback and the settings for which the practical deployment of RLHF is best-suited. In this work, we attempt to close this gap by giving an overview of common practices in user study design, discussing challenges that arise with human feedback, and identifying some of the research opportunities that real human feedback enables. The remainder of this paper will first introduce some preliminaries necessary to understand the context of this work (Section 2), then examine challenges (Section 3) and opportunities (Section 4) posed by real human feedback, as well as relevant decisions in user study design (Section 5) and finally discuss our findings and propose avenues for future work (Section 6).

2 Preliminaries

RLHF lies in the intersection of classical RL, active learning, and preference elicitation. In this section, we first give a short introduction to RL, then highlight how RLHF differs from the standard setting, and finally introduce the field of preference elicitation.



(a) RL (b) RLHF

Fig. 1: Contrasting the standard RL setting with RLHF in its most common formulation, using a reward model.

Reinforcement Learning The goal of RL [56] is to learn behavior from rewarded interaction with an environment. The environment is commonly formalized as a Markov decision process (MDP) defined by a tuple $(\mathcal{S}, \mathcal{A}, P, R)$. Here, $\mathcal{S}$ is the set of possible states, $\mathcal{A}$ the set of possible actions, $P : \mathcal{S} \times \mathcal{A} \to \mathbb{P}(\mathcal{S})$ the probabilistic state transition function[1], and $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function.

In the standard RL setting depicted in Figure 1a, the agent's objective is defined by the reward function. In each time step t, the agent chooses an action

[1] $\mathbb{P}(S)$ denotes the set of probability distributions over S.

$a_t \in \mathcal{A}$ from the set of available actions. The environment updates its state in accordance with this action and determines a reward. Both the new state $s_{t+1} \sim P(s_t, a_t)$ and the reward $r_{t+1} = R(s_t, a_t)$ are observed by the agent before picking its next action. This interaction continues until some termination criterion is met at the time horizon T, marking the end of an episode. Learning usually occurs over many of these episodes. The goal of RL then is to learn a policy $\pi : S \to \mathbb{P}(\mathcal{A})$ that maps states to actions or, in the case of a probabilistic policy, to a distribution over actions. The policy should be optimized to maximize the expected discounted cumulative sum of future rewards within an episode, formalized as $J(\pi, s_0) = \mathbb{E}_{\pi,s_0}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$, where $\gamma \in [0, 1)$ is a discount factor. Note that this may require the agent to trade off immediate rewards for larger rewards in the future.

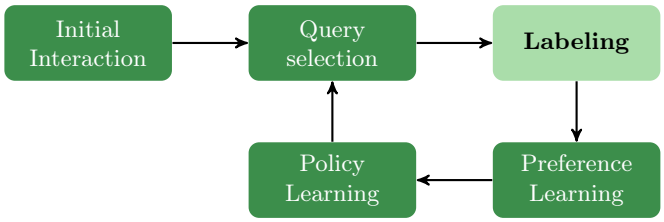


Fig. 2: The RLHF training cycle. While many recent works have studied query selection as well as preference- and policy learning, the labeling aspect is understudied and the main focus of this paper.

Reinforcement Learning from Human Feedback The simplest way to learn from human feedback would be to let a human directly specify the rewards for each of an agent's actions, i.e., r_t in Figure 1a. Unfortunately, this poses its own challenges: Consistent scalar rewards are not easy for humans to provide, human feedback is costly, and it does not scale to the amount of training an RL agent needs. For this reason, human feedback is usually employed in an indirect manner in RLHF approaches: Human labelers are asked to give feedback on behavior, and this feedback is used to train a reward model that can give rewards on behalf of the human. The feedback is decoupled from the agent's training process, and because of this, can be given at a slower rate and does not need to cover every interaction. It may also be given in a form that is convenient to the human. Figure 1b depicts this interaction: The agent's actions are rewarded by a reward model, which is in turn trained on a dataset of experiences labeled by human feedback. These labels l_i are provided asynchronously by a labeler in response to queries q_i posed by the query selection mechanism.

In the most common setting [15], this is done in a cycle of gathering experiences and querying preferences over these experiences. This common cycle is depicted in Figure 2. In the following, we will describe each of the steps.

Initial Interaction In the first step, we gather initial interaction between the agent and the environment with some prior policy (e.g., random behavior). The agent's experiences are saved. This results in a pool of stored experiences, which we can use in the next step.

Query Selection After a certain number of episodes, we select a set of queries to ask the human labelers about the stored experiences. In the most common case, these queries consist of pairwise comparisons of alternative behaviors.

Labeling The queries are then presented to a human labeler, who can give a response, e.g., indicate which behavior they prefer. Instead of asking real humans, it is also possible to synthesize this feedback based on ground-truth rewards when available. Since these rewards are generally only available for RL benchmark tasks, this is only viable for evaluation of research and not for practical applications. As highlighted in this work, synthetic feedback has its limitations and it is important to research the impact of real human feedback as well.

Preference Learning The queries together with the labels are then used to train a model of the human's preferences, the reward model (see Figure 1b).

Policy Learning After the reward model is trained, the agent is deployed in the environment again. The reward model is used as the reward function and the policy is optimized, e.g., with Trust Region Policy Optimization [51] or similar RL techniques, to maximize these rewards.

During policy learning, the agent gathers a new set of experiences, and the latter is used for a new iteration of query selection, labeling, and preference- and policy learning. The whole cycle is then repeated until a termination condition is met.

Preference Elicitation Learning a reward model can be seen as a special case of preference elicitation. The goal of preference elicitation is to gather information about an individual's preferences in a systematic and structured manner. This can be accomplished by interacting with the individual, e.g., as part of a user study, and attempting to understand the mechanisms around their feedback. Possible techniques include direct questioning, surveys, interactive interfaces and passive observation of behavior. The challenge lies in designing effective methods that can accurately capture individual preferences while minimizing biases and cognitive limitations.

Keeping human behavior in mind when designing labeling tasks can have multiple benefits. Unveiling and tackling unwanted biases within the feedback mechanism can be a lever to improve data quality. As a consequence, the number of required labeled instances might decrease. User studies can be enriched by the use of additional data sources such as surveys (of the feedbacking individuals) or paradata such as mouse movement, eye tracking or response time.

Giving feedback for the purposes of RL training differs from most common labeling and preference elicitation tasks in that it is usually an online active learning setting. It is online, because the queries for RLHF training are usually generated on-the-fly by an agent to optimize the current estimate of the human preferences (see Figure 2). In contrast to most labeling tasks, RLHF queries

therefore come from a changing distribution instead of a fixed dataset. The setting is also active, because the agent normally generates more experiences than we can label, wherefore a subset of these experiences must be selected for labeling. We can even direct the agent's behavior to generate more informative experiences, further actively shaping the data stream. These properties make feedback for RLHF distinct from most other labeling tasks.

3 Challenges of Real Human Feedback

Synthetic feedback is always reliable, consistent and fast. These characteristics cannot be attributed to real human feedback, which instead poses many challenges. We will discuss some of these challenges in this section, starting with the underlying behavioral patterns of individual labelers followed by a discussion of the disagreements that can arise as a consequence of these challenges.

3.1 Labeler Behavior

When responding in any kind of interaction such as a labeling task, a web survey or an interview, human response generally follows a multitude of biasing patterns. Some of them, as discussed in the following, are particularly important for user study design:

Response Bias Individuals may display several response biases, such as *acquiescence bias* (tendency to agree) [21], *primacy/recency effects* (tendency to select the first/last piece of information in an array) [16, 45], or *satisficing* (opting for quick and easy responses instead of thoughtful consideration) [21, 34]. An example of satisficing is a response strategy named "*straightlining*" that describes the repeated selection of the same response option irrespective of the information at hand [25]. Task design choices are likely to facilitate or weaken the resulting biases and label noise. For example, asking for a respondent's approval may foster acquiescence bias, or ways of displaying the information may facilitate primacy/recency bias. Some of these choices and interactions are discussed in more detail in Section 5.

Even though the existing literature lacks studies that empirically assess and quantify response bias in the context of RLHF, a multitude of previous research mentions and acknowledges the presence and importance of this source of cognitive bias. Examples are given by Koyama et al. [33], who experiment with adjusting multiple parameters of an image labeling task such as brightness or contrast in order to reduce the cognitive load and Early et al. [18], who mention the potential presence of recency bias in RL.

Fatigue As with any task, participants are likely to fatigue throughout the duration of the labeling task, potentially decreasing the quality of future responses and leading to inconsistencies. RL labelers fatiguing critically throughout their task might lead to more total labels needed due to diminished label quality. In addition, non-random ordering of the queries could

systematically disadvantage the label quality of later units. See Section 5 for further discussion of the importance of query order.

While, to our knowledge, no study has previously attempted to quantify fatigue bias in RL settings, multiple studies indicate awareness of the potential source of bias [5, 38]. The quantitative assessment of fatigue has also received some attention in survey methodological research: Jeong et al. [31] assess fatigue bias and find evidence for severely altered response behavior in later stages the labeling process while Hart et al. [23] examine how fatigue leads to biased responses in order to reduce the respondent's anticipated burden.

Experience In contrast to fatiguing throughout the duration of a task, individuals might also add to their task-specific experience. In general, experience and perceived task difficulty are likely to impact the quality of an individual's labels.

This is particularly relevant for RLHF labeling tasks since the agent's objective and environment may be unfamiliar to the labeler at first and the queries are quite repetitive afterwards, with environment and objective usually staying unchanged.

Expertise In addition to gaining experience during the feedback task, labelers may also have different amounts of expertise with respect to RL, the target task or labeling in general to start with. On the one end of this spectrum are the designers of the RL task, on the other end are inexperienced crowd workers unfamiliar with any labeling task.

The papers that we surveyed take different approaches in this respect. For instance, the authors of Christiano et al. [15] provide feedback for one of the tasks themselves, thereby relying on labelers with a high level of expertise. For other tasks, however, they rely on contractors with a lower expertise level. Instead of authors or contractors, Bignold et al. [9] employ university students without prior ML knowledge.

Motivation Another important driver of human behavior is the underlying motivation. There is a vast literature on studying how motivation influences performance in areas such as employment [14, 36, 50] and education [12, 42, 44].

In the context of RLHF, there might be differences in the motivation of crowdworkers and researchers. While crowdworkers' motivations are mainly financial [40] and therefore extrinsic, a researcher labeling data for their own model might instead have intrinsic motivation to provide optimal feedback. Studies show that intrinsic motivation leads to higher performance than extrinsic motivation, with extrinsic motivation even having the potential to lead to adverse outcomes [35]. Motivational patterns can therefore aggravate or weaken response biases, such as straightlining, where labelers always respond in the same way.

We are not aware of any RLHF-specific literature that directly addresses the influence of motivation on labeling quality. However, Bignold et al. [9] indirectly address the question of how to increase labeler motivation and find that labelers prefer informative over evaluative feedback, resulting in better

and more feedback per episode and longer participation of labelers providing instructive feedback compared to those providing evaluative feedback.

Misunderstandings The labelers may misunderstand the task the agent is trying to solve, rendering their feedback misleading. To mitigate this, existing work has often either relied on in-person studies [7, 10], where such misunderstandings are easier to correct, provided extensive guidelines to the labelers [15, 49, 55], or gave labelers real-time feedback and the opportunity to ask clarifying questions through chat [49, 55].

Distractions Study participants may be distracted during a study, leading to inconsistent behavior. This is particularly challenging in online studies, since researchers cannot control or observe the participant's environment or actions.

Drawing from the survey literature, studies find that participants might be distracted from background noise in the environment where they participate but also from their own actions when browsing multiple websites during the study [3, 52, 64]. The evidence of whether these distractions in online surveys lead to reduced data quality is mixed, however (see, e.g., Aizpurua et al. [1], Ansolabehere and Schaffner [3], Sendelbah et al. [52], Wenz [58]). Concerning the RLHF-specific literature, we are not aware of any study investigating the impact of distractions on feedback quality or comparing the controlled environment of in-person studies with the uncontrolled environment of remote studies.

Uneven Labeling Rate Humans may take anywhere between seconds to minutes to respond to a query. Learning algorithms should therefore be able to accommodate possible delays. An example of challenges arising due to delay is described by Christiano et al. [15], where the training process on one task was disrupted because of one labeler deviating from the usual feedback schedule.

3.2 Labeler Disagreements

As a result of the challenges discussed in Section 3.1, disagreements within the labels are nearly unavoidable. These can occur in multiple ways:

Intra-Labeler Disagreement Inconsistencies within the responses of a single labeler are a natural result of the limited and biased cognitive capacities of a human labeler, which can potentially lead to the same labeler giving different responses to the same query, depending on context.

Inter-Labeler Disagreement Multiple labelers may have different opinions, levels of expertise, or perspectives on specific queries. They may also interpret the queries differently. Disagreement between multiple labelers will, therefore, always occur.

While employing multiple labelers may lead to disagreements, it is often necessary to create a dataset of the scale needed for practical applications. The studies we surveyed generally use less than 20 labelers [7, 10, 15], however it is likely that recent industrial applications of RLHF, such as ChatGPT [48],

used many more labelers to produce datasets of sufficient size. Unfortunately the precise numbers are not public in this case.
In addition to scale, a diverse set of labelers may also help to reduce biases carried by a single labeler. Barnett et al. [6] even observe improved reward learning when valuing the variance within and between labelers instead of the established practice of not distinguishing between labelers.

Researcher-Labeler Disagreement Relatedly, the preferences of the (paid) labelers may differ from the concepts the researchers wish to convey. For instance, Ziegler et al. [63] show that there is a significant difference between evaluations of the paper authors and Scale AI freelance workers in four natural-language-related tasks (38% agreement in a sentiment assessment task and 46% agreement on a summarization task). However, they note that there is also only 60% agreement between the paper's authors for a small subset of the labeling tasks. The authors argue that this is because, in such language tasks, it is often difficult to agree due to a lack of a clear ground truth. This result shows that different levels of expertise might influence the outcome or quality of the training data, but it also highlights that some disagreement is likely unavoidable. Despite this potential influence, we observed that many papers do not comment on whom they asked for feedback.

4 Opportunities of Real Human Feedback

As discussed in Section 3, real human feedback poses many challenges when compared to synthetic feedback. Each challenge also presents an opportunity for further research which can yield improvements to the RLHF method however. These opportunities, by their nature, can only be researched with real human feedback and are of key importance to practical application of RLHF. In this section, we will make some of these research directions implied by the challenges discussed in Section 3 explicit. We will also discuss how real human feedback may provide additional benefits through the use of implicit information.

4.1 Optimizing the Labeling Task

One of the opportunities afforded by real human feedback is to study potential improvements to the labeling task, i.e., the way queries are asked and the form in which feedback is given. By framing this task in a human-friendly way, it is possible to get more information for the same amount of human time.

When using synthetic feedback, it is common to compare alternative approaches by the number of labels required to reach a specific performance. This is not a fair comparison, however, as some choices may increase the difficulty for the human while others may decrease it. Working with real human feedback enables us to measure labeling time in addition to the number of labels provided.

One opportunity is to optimize the feedback modality for ease of answer, thereby getting more feedback for the same amount of human time and effort.

The ideal framing of the labeling task also depends on the context. For example, pairwise comparisons can prove challenging in goal-conditioned settings, where two compared behaviors may aim to solve different goals and be difficult to compare directly.

While extending or replacing the comparison setting poses new challenges in encoding, interpreting and learning from these possibly richer forms of feedback, there exists initial work proposing unified frameworks for this purpose. Jeon et al. [30] propose the framework of reward-rational implicit choice, which interprets human feedback (regardless of its form) as a choice from a set of possibly infinitely many alternatives. The framework assumes that this choice is Boltzmann-rational with respect to a reward function. By utilizing an appropriate grounding function that maps feedback to behavior, this enables learning a reward function consistent with the observed feedback. Notably, the framework even extends to language feedback, where the grounding function maps an utterance to all compatible behaviors (e.g., following an instruction or avoiding concerns expressed in natural language), and the choice set comprises all possible utterances. It is then possible to infer a reward function such that the observed feedback is Boltzmann-rational with respect to that function. Additionally, Metz et al. [41] contribute a common encoding for multiple types of feedback, thereby further simplifying the use of rich and diverse sources of feedback as proposed in this section.

Extensions to Comparison Queries Many studies in the RLHF space use pairwise comparisons, meaning they show multiple alternative behaviors side-by-side and ask the human to pick a favored one [15]. This common setting can be extended in multiple ways, including augmenting the labels by explanations, providing additional response options and asking for labels for longer interactions.

Explanations. One extension consist of allowing participants to state which feature or which visual region influenced their decision the most [7, 22]. Guan et al. [22] study the amount of human time needed to provide visual saliency information in addition to binary feedback and find that it incurs little additional effort while resulting in improved sample efficiency.

Response Options. Another extension to the comparison setting is the addition of more response options beyond binary preference. One way to do this is to allow labelers to reject comparisons when they cannot give a certain answer ("soft choice setting"). This option can reduce labeling noise [26]. Wilde et al. [59] go one step further by allowing the labeler to give quasi-continuous feedback with a slider bar. They compare this to the soft choice setting and find that while scale feedback is slightly less easy to use than soft choice, the former significantly improved learning in their experiments. The authors state that the gains in learning outweigh the drop in easiness, although the participants rated both scale and soft choice feedback equally expressive.

Long Interactions. Another example is asking labelers to review more extended interactions, giving multiple bits of feedback per interaction. An early exploration of this idea is presented by Interactive Agents Team et al. [28].

Optimizing Query Presentation Another way to improve the labeling task is by improving how queries are presented to the labelers. Zhang et al. [62] explore how clustering different comparisons instead of presenting them one after the other can increase feedback efficiency. Based on visualization and dimensionality-reduction techniques, they design an interactive user interface allowing the human to label a subset of the state space. Even though their user studies with real human feedback are somewhat limited, they find that training efficiency increases for the same amount of human time in some simple MuJoCo tasks. This is closely related to user-interface-related design decisions discussed in Section 3.

Optimizing Query Selection In addition to the presentation, the selection of queries also significantly impacts perceived difficulty. For example, Bıyık et al. [10] find that information-gain-based query selection leads to easier queries when compared to volume-removal-based selection.

Evaluating Alternative Feedback Modalities An alternative to extending the commonly used comparison queries is to explore the use of alternative feedback modalities. Possible choices include *instructive* forms of feedback such as demonstrations [27] and corrections [29], *evaluative* forms of feedback such as critiques [4, 32] and ordinal feedback [37, 61], and *comparative* forms of feedback such as pairwise comparisons [15] and rankings [20, 46].

Developing Techniques for Aided Evaluation Another way to enhance human time efficiency while collecting feedback is to aid the human in evaluation. For example, Guan et al. [22] use object tracking and object detection to simplify the annotation of visually salient objects. The effectiveness of this aided evaluation can only be studied with real human feedback.

4.2 Utilizing Available Information

Another opportunity for RLHF research enabled by real human feedback is to make use of information that human labelers give implicitly.

Implicit Feedback Humans leak a lot of information by implicit behaviors [17], such as gestures, facial expressions, vocalizations, tone of voice, non-verbal cues, and response delay when providing labels. This implicit feedback can be challenging to detect and interpret since it varies from human to human. Credit assignment to past or even anticipated events poses a further challenge. Nonetheless, provided they can be incorporated successfully, these cues can be an additional source of feedback with no additional human effort.

One example of such implicit feedback is visual saliency information, which can be used to augment preference labels. While Guan et al. [22] studied this as an implicit form of feedback, this information is always given implicitly and only needs to be collected. If we can use such implicit information, it may be possible to learn more accurate models from fewer samples. As a step in the direction of learning from implicit feedback, Jeon et al. [30] propose the framework of reward-rational implicit choice to learn from explicit and implicit forms of feedback.

Implicit Reward Shaping Human labelers often give feedback not only based on the task performance but also on the agent's progress and whether or not its current behavior may eventually lead to the correct behavior. While this often violates assumptions on how feedback is given and can therefore be challenging, it can also be an opportunity if used correctly since such shaped rewards can simplify the policy learning problem. For example, Christiano et al. [15] find that training from human feedback sometimes outperforms synthetic 'oracle' feedback, likely due to this implicit reward shaping.

5 Design Decisions

With human behavior being all but error-free, task designers must carefully consider a variety of decisions to avoid task design-driven bias (see Section 3) and make use of the opportunities afforded by human feedback (see Section 4). Those design choices are, of course, interrelated and context-dependent. For instance, if it is necessary to have longer trajectories, the mental load of processing two more complex and long trajectories might be too high if they are shown simultaneously.

5.1 Study Setup

Many design decisions that one has to make in the context of a user study are related to the general setup of the study. In the following, we list a selection of these decisions that we consider to be of particular importance concerning RL feedback:

Order The resulting set of labels will be impacted by how the queries or different stages of a labeling task are ordered. In addition to increasing fatigue or expertise, social psychologists have observed patterns of "contrast" and "assimilation" [11]. While a contrast effect describes an individual perceiving a piece of information more dissimilar from a previous piece of information (e.g., the judgment of the height of strangers), an assimilation effect is present when a piece of information seems to be more similar to the previous piece of information. As a classic example, a crooked politician makes other politicians appear less trustworthy. Beck et al. [8] observe some indication for a contrast effect in a hate speech-related labeling task.

Guidelines The initial guidelines, tutorials, and examples might have a significant anchoring effect on the subsequent labeling behavior. This introductory material should be carefully selected or crafted. Generally, adding initial guidelines to a labeling task seems to benefit the resulting data quality [19, 47].

Incentives The incentive structure of a labeling task is a crucial determinant of the motivation and response behavior of the labelers. In the design process, a decision must be taken for either a fixed wage per task or a fixed wage per time. Past research on incentives in labeling tasks has reported mixed results. Multiple studies could not find increased label quality through performance-based bonus payments [53, 60]. In addition, some observations have been made that higher wages increase the quantity but not the quality of labeling work [13, 39].

Quality Control Crowdsourcing, in particular, poses the challenge of quality control since study participants may be incentivized to complete as many queries as possible instead of focusing on accuracy. However, crowdsourcing platforms like Prolific or Scale AI have developed quality controls that might reduce this problem. Scale AI, for instance, uses benchmark labeling, which serves to screen out labelers that do not meet a certain standard (see 63).

Participant Selection The participant selection has a significant influence on response quality and labeler expertise. It can be influenced by prior screening, manual selection, or choice of crowdsourcing or contracting platform.

5.2 User Interface

Another design choice is presented by the user interface through which the study participants can communicate their preferences. An example of the importance of this choice is described by Amodei et al. [2], who observe that the learned reward function does not capture the desired behavior due to the labelers' inability to judge depth in a two-dimensional video. The agent may actively learn to exploit such interface-driven limitations of the labelers since it is rewarded for any behavior that *appears* correct to humans. The user interface is closely related to the feedback modality (see Section 4.1), i.e., the form in which the human labelers are expected to give feedback.

6 Discussion and Future Work

Reinforcement learning from human feedback inherently depends on human feedback. In Section 3, we identified several challenges posed by this feedback in the context of RL training. Throughout this work, we discussed existing approaches that address some of these challenges. However, we also noticed in Section 4 that every challenge also represents an opportunity for further research for improving the RLHF method. Finally, we introduced important design decisions that can be leveraged to overcome the challenges and capitalize on the opportunities in Section 5. Given these open areas for improvement, we believe that the human aspect of RLHF has been understudied thus far.

We suspect that one reason for this lack of research is the RLHF research community's lack of experience in conducting user studies. We provide an introduction to the challenges faced in user study design in this work and believe future work focused on reducing the friction introduced by this unfamiliarity would be worthwhile. One possible avenue to accomplish this would be to develop frameworks for crowdsourcing labels in the online active learning setting posed by RLHF. Such a framework could make feedback collection more attainable for academic researchers by simplifying interaction with crowdsourcing platforms.

In parallel to our work, Metz et al. [41] took a first step into the direction of reducing friction through standardized tooling. They propose a configurable interface for giving feedback on behavior as well as a common encoding for many different feedback modalities. In accordance with the claims made in this paper, they acknowledge the importance of the human factors of RLHF, advocate for the need for systematic empirical studies with real humans and discuss the importance of reducing the friction and collaborating across disciplines. Future extensions could further aid researchers in conducting these studies by providing examples of usage, documentation and integration with crowd-sourcing services. While there are many similarities between our work and the paper by Metz et al. [41], the two differ in focus and can be seen as complementary. While they focus on learning from diverse sources of feedback and therefore discuss attributes of different feedback modalities, we rather focus on human aspects and describe individual challenges such as response biases in more detail.

In addition to this meta-work on reducing research friction, studying the challenges and opportunities discussed in Sections 3 and 4 would be another important direction for future work. This would provide a greater understanding of optimal user study design for RLHF, hopefully enabling us to apply it to more settings with less human effort and, as a side-effect, provide examples that would further reduce the previously discussed unfamiliarity. We hope that future research will pay increased attention to the challenges and opportunities posed by real human feedback.

Acknowledgements This publication was supported by the Munich Center for Machine Learning (MCML) and LMUexcellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder as well as by the Hightech Agenda Bavaria. This paper was also supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

Bibliography

[1] Aizpurua, E., Heiden, E.O., Park, K.H., Wittrock, J., Losch, M.E.: Investigating Respondent Multitasking and Distraction Using Self-reports and Interviewers' Observations in a Dual-frame Telephone Survey. Survey Methods: Insights from the Field (SMIF) (Nov 2018), https://doi.org/10.13094/SMIF-2018-00006

[2] Amodei, D., Christiano, P., Ray, A.: Learning from human preferences (Jun 2017), URL https://openai.com/research/learning-from-human-preferences, (accessed 2023-05-25)

[3] Ansolabehere, S., Schaffner, B.F.: Distractions: The Incidence and Consequences of Interruptions for Survey Respondents. Journal of Survey Statistics and Methodology **3**(2), 216–239 (Jun 2015), https://doi.org/10.1093/jssam/smv003

[4] Argall, B., Browning, B., Veloso, M.: Learning by demonstration with critique from a human teacher. In: Proceedings of the ACM/IEEE International Conference on Human-robot Interaction, pp. 57–64, Association for Computing Machinery (Mar 2007), https://doi.org/10.1145/1228716.1228725

[5] Arzate Cruz, C., Igarashi, T.: A Survey on Interactive Reinforcement Learning: Design Principles and Open Challenges. In: Proceedings of the 2020 ACM Designing Interactive Systems Conference, pp. 1195–1209, Association for Computing Machinery (Jul 2020), https://doi.org/10.1145/3357236.3395525

[6] Barnett, P., Freedman, R., Svegliato, J., Russell, S.: Active Reward Learning from Multiple Teachers. In: The AAAI Workshop on Artificial Intelligence Safety (Feb 2023)

[7] Basu, C., Singhal, M., Dragan, A.D.: Learning from Richer Human Guidance: Augmenting Comparison-Based Learning with Feature Queries. In: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, pp. 132–140, Association for Computing Machinery (Feb 2018), https://doi.org/10.1145/3171221.3171284

[8] Beck, J., Eckman, S., Chew, R., Kreuter, F.: Improving Labeling Through Social Science Insights: Results and Research Agenda. In: Chen, J.Y.C., Fragomeni, G., Degen, H., Ntoa, S. (eds.) HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence, pp. 245–261, Springer Nature Switzerland (2022), https://doi.org/10.1007/978-3-031-21707-4_19

[9] Bignold, A., Cruz, F., Dazeley, R., Vamplew, P., Foale, C.: Human Engagement Providing Evaluative and Informative Advice for Interactive Reinforcement Learning. Neural Computing and Applications (Jan 2022), https://doi.org/10.1007/s00521-021-06850-6

[10] Bıyık, E., Palan, M., Landolfi, N.C., Losey, D.P., Sadigh, D.: Asking Easy Questions: A User-Friendly Approach to Active Reward Learning. In: Pro-

ceedings of the Conference on Robot Learning, pp. 1177–1190, PMLR (May 2020), URL https://proceedings.mlr.press/v100/b-iy-ik20a.html

[11] Bless, H., Schwarz, N.: Chapter 6 - Mental Construal and the Emergence of Assimilation and Contrast Effects: The Inclusion/Exclusion Model. In: Advances in Experimental Social Psychology, vol. 42, pp. 319–373, Academic Press (Jan 2010), https://doi.org/10.1016/S0065-2601(10)42006-7

[12] Broussard, S.C., Garrison, M.E.B.: The Relationship Between Classroom Motivation and Academic Achievement in Elementary-School-Aged Children. Family and Consumer Sciences Research Journal **33**(2), 106–120 (2004), https://doi.org/10.1177/1077727X04269573

[13] Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? Perspectives on Psychological Science **6**(1), 3–5 (Jan 2011), https://doi.org/10.1177/1745691610393980

[14] Cerasoli, C.P., Nicklin, J.M., Ford, M.T.: Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. Psychological Bulletin **140**(4), 980–1008 (Jul 2014), https://doi.org/10.1037/a0035661

[15] Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep Reinforcement Learning from Human Preferences. In: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc. (2017), URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

[16] Crano, W.D.: Primacy versus recency in retention of information and opinion change. The Journal of Social Psychology **101**, 87–96 (1977), https://doi.org/10.1080/00224545.1977.9923987

[17] Cui, Y., Zhang, Q., Knox, B., Allievi, A., Stone, P., Niekum, S.: The EMPATHIC Framework for Task Learning from Implicit Human Feedback. In: Proceedings of the 2020 Conference on Robot Learning, pp. 604–626, PMLR (Oct 2021), URL https://proceedings.mlr.press/v155/cui21a.html

[18] Early, J., Bewley, T., Evers, C., Ramchurn, S.: Non-Markovian Reward Modelling from Trajectory Labels via Interpretable Multiple Instance Learning. Advances in Neural Information Processing Systems **35**, 27652–27663 (Dec 2022), URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b157cfde6794e93b2353b9712bbd45a5-Abstract-Conference.html

[19] Fort, K., Ehrmann, M., Nazarenko, A.: Towards a methodology for named entities annotation. In: Proceedings of the Third Linguistic Annotation Workshop, pp. 142–145, Association for Computational Linguistics (Aug 2009), ISBN 978-1-932432-52-7

[20] Fürnkranz, J., Hüllermeier, E.: Preference Learning and Ranking by Pairwise Comparison. In: Fürnkranz, J., Hüllermeier, E. (eds.) Preference Learning, pp. 65–82, Springer (2010), https://doi.org/10.1007/978-3-642-14125-6_4

[21] Groves, R.M., Fowler Jr, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R.: Survey Methodology. John Wiley & Sons, 2 edn. (2009), ISBN 978-0-470-46546-2

[22] Guan, L., Verma, M., Guo, S., Zhang, R., Kambhampati, S.: Widening the Pipeline in Human-Guided Reinforcement Learning with Explanation and Context-Aware Data Augmentation. In: Advances in Neural Information Processing Systems (Oct 2021), URL https://proceedings.neurips.cc/paper/2021/hash/b6f8dc086b2d60c5856e4ff517060392-Abstract.html

[23] Hart, T.C., Rennison, C.M., Gibson, C.: Revisiting Respondent "Fatigue Bias" in the National Crime Victimization Survey. Journal of Quantitative Criminology **21**(3), 345–363 (Sep 2005), https://doi.org/10.1007/s10940-005-4275-4

[24] Hejna, D.J., Sadigh, D.: Few-Shot Preference Learning for Human-in-the-Loop RL. In: Proceedings of The 6th Conference on Robot Learning, pp. 2014–2025, PMLR (Mar 2023), ISSN 2640-3498, URL https://proceedings.mlr.press/v205/iii23a.html

[25] Herzog, A.R., Bachman, J.G.: Effects of Questionnaire Length on Response Quality. The Public Opinion Quarterly **45**(4), 549–559 (1981), https://doi.org/10.1086/268687

[26] Holladay, R., Javdani, S., Dragan, A., Srinivasa, S.: Active comparison based learning incorporating user uncertainty and noise. In: RSS Workshop on Model Learning for Human-Robot Communication (2016)

[27] Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., Amodei, D.: Reward learning from human preferences and demonstrations in Atari. In: Advances in Neural Information Processing Systems, vol. 31, Curran Associates, Inc. (2018), URL https://proceedings.neurips.cc/paper/2018/hash/8cbe9ce23f42628c98f80fa0fac8b19a-Abstract.html

[28] Interactive Agents Team, D., Abramson, J., Ahuja, A., Carnevale, F., Georgiev, P., Goldin, A., Hung, A., Landon, J., Lhotka, J., Lillicrap, T., Muldal, A., Powell, G., Santoro, A., Scully, G., Srivastava, S., von Glehn, T., Wayne, G., Wong, N., Yan, C., Zhu, R.: Improving Multimodal Interactive Agents with Reinforcement Learning from Human Feedback (Nov 2022), URL http://arxiv.org/abs/2211.11602

[29] Jain, A., Wojcik, B., Joachims, T., Saxena, A.: Learning Trajectory Preferences for Manipulators via Iterative Improvement. In: Advances in Neural Information Processing Systems, vol. 26, Curran Associates, Inc. (2013), URL https://proceedings.neurips.cc/paper/2013/hash/c058f544c737782deacefa532d9add4c-Abstract.html

[30] Jeon, H.J., Milli, S., Dragan, A.: Reward-rational (implicit) choice: A unifying formalism for reward learning. In: Advances in Neural Information Processing Systems, vol. 33, pp. 4415–4426, Curran Associates, Inc. (2020), URL https://proceedings.neurips.cc/paper/2020/hash/2f10c1578a0706e06b6d7db6f0b4a6af-Abstract.html

[31] Jeong, D., Aggarwal, S., Robinson, J., Kumar, N., Spearot, A., Park, D.S.: Exhaustive or exhausting? Evidence on respondent fatigue in long surveys. Journal of Development Economics **161**, 102992 (Mar 2023), https://doi.org/10.1016/j.jdeveco.2022.102992

[32] Judah, K., Roy, S., Fern, A., Dietterich, T.: Reinforcement Learning Via Practice and Critique Advice. In: Proceedings of the AAAI Conference on

Artificial Intelligence, vol. 24, pp. 481–486 (Jul 2010), https://doi.org/10.1609/aaai.v24i1.7690
[33] Koyama, Y., Sato, I., Sakamoto, D., Igarashi, T.: Sequential line search for efficient visual design optimization by crowds. ACM Transactions on Graphics **36**(4), 48:1–48:11 (Jul 2017), https://doi.org/10.1145/3072959.3073598
[34] Krosnick, J.A., Alwin, D.F.: An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. The Public Opinion Quarterly **51**(2), 201–219 (1987), https://doi.org/10.1086/269029
[35] Kuvaas, B., Buch, R., Weibel, A., Dysvik, A., Nerstad, C.G.L.: Do intrinsic and extrinsic motivation relate differently to employee outcomes? Journal of Economic Psychology **61**, 244–258 (Aug 2017), https://doi.org/10.1016/j.joep.2017.05.004
[36] Lawler, E.E.: Motivation in Work Organizations. Brooks/Cole Publishing Co (1973), ISBN 0-8185-0088-3
[37] Li, K., Tucker, M., Bıyık, E., Novoseller, E., Burdick, J.W., Sui, Y., Sadigh, D., Yue, Y., Ames, A.D.: ROIAL: Region of Interest Active Learning for Characterizing Exoskeleton Gait Preference Landscapes. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 3212–3218 (May 2021), https://doi.org/10.1109/ICRA48506.2021.9560840
[38] Li, Z., Shi, L., Cristea, A.I., Zhou, Y.: A Survey of Collaborative Reinforcement Learning: Interactive Methods and Design Patterns. In: Designing Interactive Systems Conference 2021, pp. 1579–1590, Association for Computing Machinery (Jun 2021), https://doi.org/10.1145/3461778.3462135
[39] Litman, L., Robinson, J., Rosenzweig, C.: The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. Behavior Research Methods **47**(2), 519–528 (Jun 2015), https://doi.org/10.3758/s13428-014-0483-x
[40] Martin, D., Hanrahan, B.V., O'Neill, J., Gupta, N.: Being a turker. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 224–235, Association for Computing Machinery (Feb 2014), https://doi.org/10.1145/2531602.2531663
[41] Metz, Y., Lindner, D., Baur, R., Keim, D.A., El-Assady, M.: RLHF-Blender: A Configurable Interactive Interface for Learning from Diverse Human Feedback. In: ICML 2023 Workshop Interactive Learning with Impplicit Human Feedback (Jun 2023), URL https://openreview.net/forum?id=JvkZtzJBFQ
[42] Mitchell, J.V.: Interrelationships and predictive efficacy for indices of intrinsic, extrinsic, and self-assessed motivation for learning. Journal of Research & Development in Education **25**, 149–155 (1992), ISSN 0022-426X
[43] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (Feb 2015), https://doi.org/10.1038/nature14236

[44] Muogbo, U.S.: The Influence of Motivation on Employees' Performance: A Study of Some Selected Firms in Anambra State. AFRREV IJAH: An International Journal of Arts and Humanities **2**(3), 134–151 (2013), https://doi.org/10.4314/ijah.v2i3

[45] Murphy, J., Hofacker, C., Mizerski, R.: Primacy and Recency Effects on Clicking Behavior. Journal of Computer-Mediated Communication **11**(2), 522–535 (2006), https://doi.org/10.1111/j.1083-6101.2006.00025.x

[46] Myers, V., Bıyık, E., Anari, N., Sadigh, D.: Learning Multimodal Rewards from Rankings. In: Proceedings of the 5th Conference on Robot Learning, pp. 342–352, PMLR (Jan 2022), URL https://proceedings.mlr.press/v164/myers22a.html

[47] Nédellec, C., Bessieres, P., Bossy, R.R., Kotoujansky, A., Manine, A.P.: Annotation guidelines for machine learning-based named entity recognition in microbiology. In: Proceeding of Data and Text Mining for Integrative Biology Workshop 17. European Conference on Machine Learning 10. European Conference on Principles and Practice of Knowledge Discovery in Databases, Springer-Verlag (2006)

[48] OpenAI: ChatGPT: Optimizing Language Models for Dialogue (2022), URL https://openai.com/blog/chatgpt/, (accessed 2023-02-02)

[49] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems, vol. 35, pp. 27730–27744 (Dec 2022), URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[50] Porter, L.W., Lawler, E.E.: Managerial Attitudes and Performance. R.D. Irwin (1968)

[51] Schulman, J., Levine, S., Abbeel, P., Jordan, M., Moritz, P.: Trust Region Policy Optimization. In: Proceedings of the 32nd International Conference on Machine Learning, pp. 1889–1897, PMLR (Jun 2015), ISSN 1938-7228, URL https://proceedings.mlr.press/v37/schulman15.html

[52] Sendelbah, A., Vehovar, V., Slavec, A., Petrovčič, A.: Investigating respondent multitasking in web surveys using paradata. Computers in Human Behavior **55**, 777–787 (Feb 2016), https://doi.org/10.1016/j.chb.2015.10.028

[53] Shaw, A.D., Horton, J.J., Chen, D.L.: Designing incentives for inexpert human raters. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, pp. 275–284, Association for Computing Machinery (Mar 2011), https://doi.org/10.1145/1958824.1958865

[54] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of Go with deep neural networks and tree search. Nature **529**(7587), 484–489 (Jan 2016), https://doi.org/10.1038/nature16961

[55] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D.M., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.: Learning to summarize from human feedback (Feb 2022), URL http://arxiv.org/abs/2009.01325
[56] Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. The MIT Press, second edition edn. (2018), ISBN 978-0-262-03924-6
[57] Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J.P., Jaderberg, M., Vezhnevets, A.S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T.L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C., Silver, D.: Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature **575**(7782), 350–354 (Nov 2019), https://doi.org/10.1038/s41586-019-1724-z
[58] Wenz, A.: Do Distractions During Web Survey Completion Affect Data Quality? Findings From a Laboratory Experiment. Social Science Computer Review **39**(1), 148–161 (Feb 2021), https://doi.org/10.1177/0894439319851503
[59] Wilde, N., Bıyık, E., Sadigh, D., Smith, S.L.: Learning Reward Functions from Scale Feedback. In: Proceedings of the 5th Conference on Robot Learning, pp. 353–362, PMLR (Jan 2022), URL https://proceedings.mlr.press/v164/wilde22a.html
[60] Yin, M., Chen, Y., Sun, Y.A.: The Effects of Performance-Contingent Financial Incentives in Online Labor Markets. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 27, pp. 1191–1197 (Jun 2013), https://doi.org/10.1609/aaai.v27i1.8461
[61] Zap, A., Joppen, T., Fürnkranz, J.: Deep Ordinal Reinforcement Learning. In: Brefeld, U., Fromont, E., Hotho, A., Knobbe, A., Maathuis, M., Robardet, C. (eds.) Machine Learning and Knowledge Discovery in Databases, pp. 3–18, Springer International Publishing (2020), https://doi.org/10.1007/978-3-030-46133-1_1
[62] Zhang, D., Carroll, M., Bobu, A., Dragan, A.: Time-Efficient Reward Learning via Visually Assisted Cluster Ranking. In: NeurIPS Workshop on Human in the Loop Learning (Dec 2022)
[63] Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-Tuning Language Models from Human Preferences (Jan 2020), URL http://arxiv.org/abs/1909.08593
[64] Zwarun, L., Hall, A.: What's going on? Age, distraction, and multitasking during online survey taking. Computers in Human Behavior **41**, 236–244 (Dec 2014), https://doi.org/10.1016/j.chb.2014.09.041