A SURVEY OF REINFORCEMENT LEARNING FROM HUMAN FEEDBACK

A PREPRINT

Timo Kaufmann
LMU Munich, MCML Munich
Munich, Germany
timo.kaufmann@ifi.lmu.de

Paul Weng
Duke Kunshan University
Kunshan, China
paul.weng@dukekunshan.edu.cn

Viktor Bengs
LMU Munich, MCML Munich
Munich, Germany
viktor.bengs@ifi.lmu.de

Eyke Hüllermeier
LMU Munich, MCML Munich
Munich, Germany
eyke@ifi.lmu.de

December 2023

ABSTRACT

Reinforcement learning from human feedback (RLHF) is a variant of reinforcement learning (RL) that learns from human feedback instead of relying on an engineered reward function. Building on prior work on the related setting of preference-based reinforcement learning (PbRL), it stands at the intersection of artificial intelligence and human-computer interaction. This positioning offers a promising avenue to enhance the performance and adaptability of intelligent systems while also improving the alignment of their objectives with human values. The training of large language models (LLMs) has impressively demonstrated this potential in recent years, where RLHF played a decisive role in targeting the model's capabilities toward human objectives. This article provides a comprehensive overview of the fundamentals of RLHF, exploring the intricate dynamics between machine agents and human input. While recent focus has been on RLHF for LLMs, our survey adopts a broader perspective, examining the diverse applications and wide-ranging impact of the technique. We delve into the core principles that underpin RLHF, shedding light on the symbiotic relationship between algorithms and human feedback, and discuss the main research trends in the field. By synthesizing the current landscape of RLHF research, this article aims to provide researchers as well as practitioners with a comprehensive understanding of this rapidly growing field of research.

Keywords Preference-Based Reinforcement Learning · Human Feedback · Preference Learning · Interactive Learning · Reward Learning · Reward Modeling · Agent Alignment · Human-in-the-Loop · Human-Computer Interaction · Human-Compatible AI

Contents

1 Introduction

In reinforcement learning (RL), an agent traditionally navigates through an environment and attempts to make optimal actions or decisions through a process of trial and error. Whether a decision is optimal or not is determined solely by reward signals. These signals have to be defined manually based on measurements of the agent's performance, ensuring that the learning agent receives the necessary signals to learn the correct behavior. Designing a reward function by hand, however, is challenging. Success is hard to formally define and measure in many applications. Beyond that, a sparse signal of success may not be well suited for agent learning – resulting in the need for *reward shaping* (Ng et al. 1999), where the reward signal is transformed into one that is more suitable for learning. This often makes the reward signal more susceptible to spurious correlations, however – behaviors that are rewarded because they are usually correlated with the true objective but are not valuable in themselves. This ultimately cumulates in the issue of *reward hacking* (Skalse et al. 2022b), where learning agents exploit reward-specific loopholes to achieve undesired outcomes while still generating high rewards.

In response to these challenges, reinforcement learning from human feedback (RLHF) has emerged as a practically meaningful alternative that introduces a critical human-in-the-loop component to the standard RL learning paradigm. In a nutshell, RLHF differs from RL in that the objective is defined and iteratively refined by the human in the loop instead of being specified ahead of time. This approach not only has the potential to overcome the limitations and issues of classical RL methods but also has potential benefits for agent alignment, where the agent's learning goals are more closely aligned with human values, promoting ethically sound and socially responsible AI systems.

RLHF has seen a number of successful applications, advances in methodology, and theoretical insights since the last comparable survey (Wirth et al. 2017). The applications range from the domain of large language models (LLMs) (OpenAI 2022) over image generation (Lee et al. 2023), continuous control (Christiano et al. 2017) and games (Ibarz et al. 2018) to robotics (Hejna et al. 2023a). At the same time, there have been a lot of developments of the methodology since the last comparable survey (Wirth et al. 2017). Examples of methodological developments include using data augmentation and semi-supervised learning approaches to improve sample complexity (Park et al. 2022), using meta-learning to quickly adapt the learned preferences to new tasks (Ren et al. 2022), fusing multiple feedback types (Palan et al. 2019), using self-supervised representation learning to increase feedback efficiency (Metcalf et al. 2022), actively synthesizing hypothetical behaviors for queries (Reddy et al. 2020), and optimizing queries for ease of answer (Bıyık et al. 2020b). Finally, there have been some achievements with regard to theoretical results for the field of RLHF, providing new insights but also new questions for the fundamental mathematical problems underlying the modeling of the learning scenario in RLHF.

In this survey, we, therefore, discuss the current state of affairs with regard to the ongoing research in RLHF, classify the current approaches as well as concisely describe their main characteristics, and give a brief overview of the application areas.

1.1 Why Human Feedback

In conventional RL, the agent's objective is defined by a reward function which it aims to maximize (Sutton et al. 2018). Specifying this reward function can be challenging, particularly in complex domains: What would be a suitable reward function for a robot assisting humans in a household environment or for autonomous vehicles navigating through a busy urban environment? Moreover, even well-defined reward functions can lead to unexpected behaviors due to distributional shifts or over-optimization, raising practical and safety concerns. Learning the agent's objective from human feedback circumvents reward engineering challenges and fosters robust training, with the reward function dynamically refined and adjusted to distributional shifts as the agent learns.

Feedback vs. Demonstrations The field of inverse RL aims to infer reward functions from human demonstrations (Arora et al. 2021). While this can partially resolve reward engineering challenges, it faces inherent difficulties: (i) It is generally not possible to robustly identify rewards from demonstrations (Cao et al. 2021a), (ii) it is only applicable in scenarios where good demonstrations can be obtained, (iii) it struggles to outperform the demonstrator, and (iv) humans often do not demonstrate the behavior they would prefer a machine to adopt (Basu et al. 2017). Interactive feedback, by contrast, can use active queries to differentiate between human preferences and irrelevant noise, is much easier to provide than demonstrations, does not require near-optimal performance from the human evaluators, and elicits preferences on the behavior that a human would prefer from the machine. Interactive feedback can also be used to complement demonstrations, in which case it can be used to shape and refine capabilities learned through initial training, like behavioral cloning, thereby preventing overfitting to demonstrated behavior (Abramson et al. 2022).

Table 1: Feedback types classified as belonging to PbRL, SSRL, and RLHF as defined in this survey.

Feedback Type	PbRL	SSRL	RLHF
Binary trajectory comparisons	✓	✗	✓
Trajectory rankings	✓	✗	✓
State preferences	✓	✗	✓
Action preferences	✓	✗	✓
Binary critique	✗	✓	✓
Scalar feedback	✗	✓	✓
Corrections	✗	✗	✓
Action advice	✗	✗	✓
Implicit feedback	✗	✗	✓
Natural language	✗	✗	✓

Avoiding Reward Engineering Reward engineering in RL presents significant challenges, as accurately specifying reward functions is notoriously difficult (Amodei et al. 2016; Knox et al. 2023). These challenges can be mitigated by utilizing human feedback, which enables training agents for tasks that are hard to define manually and helps avoid safety issues arising from misaligned rewards (Skalse et al. 2022b). Safety issues related to a misalignment between the agent's and the human's objectives are studied as the AI alignment problem (Gabriel 2020), in particular agent alignment and value alignment (Kirchner et al. 2022). Although the effectiveness of RLHF in resolving these alignment issues is debated (Christiano 2023), it presents a promising approach to enhance alignment (Leike et al. 2018).

Excessive optimization for poorly specified rewards often leads to unintended behaviors. Agents may exploit simulation flaws for higher rewards (Lehman et al. 2020; Baker et al. 2020) or engage in *reward hacking* (Skalse et al. 2022b), where the behavior maximizes the specified reward but deviates from the intended objective. This is evident in cases where agents focus on intermediate rewards without achieving the actual goal (Clark et al. 2016) or prematurely exit a game to avoid negative rewards (Saunders et al. 2018). The root of these issues is that the reward function does not properly reflect the actual learning task. While these issues may seem trivial in game-like environments, their implications are far more serious in safety-critical contexts such as healthcare and autonomous driving. In these settings, it is crucial to prevent misaligned reward functions from leading to harmful outcomes, like a care robot causing injury or a self-driving car jeopardizing road safety.

1.2 The Origins of Reinforcement Learning from Human Feedback

Learning behavior from human feedback has long been studied as a subfield of RL, but methods and terminology have evolved over time. Early methods, as discussed in more detail by Knox (2012), focused on learning directly from human rewards (Isbell et al. 2001; Knox et al. 2008). This survey, however, focuses on more indirect approaches of inferring the objective from human feedback.

Reinforcement learning from human feedback (RLHF) in its modern guise has its origin in the setting of preference-based reinforcement learning (PbRL) as introduced independently by Akrour et al. (2011) and Cheng et al. (2011). The original idea of preference-based reinforcement learning (PbRL) is to infer the objective from qualitative feedback, such as pairwise preferences between behaviors or between actions given states, instead of quantitative feedback in the form of numerical rewards. The term RLHF was coined as an alternative later on (Askell et al. 2021; Ouyang et al. 2022; OpenAI 2022), though initially referring to the same concept of learning behavior from relative feedback.

Disentangling PbRL and RLHF is challenging due to their overlapping use in the literature. For instance, Christiano et al. (2017) themselves are using the term PbRL, yet are often cited as a seminal reference for RLHF (Daniels-Koch et al. 2022; Ouyang et al. 2022). This indicates the interchangeability of these terms. Practically, RLHF is often associated with reward modeling and deep RL, while PbRL is often linked to direct policy optimization in traditional RL settings. This is underlined by Jeon et al. (2020), who characterize PbRL as limited to direct policy learning from preferences. This is in contrast with other sources, however, who include reward learning within the scope of RLHF (Christiano et al. 2017; Wirth et al. 2017).

Despite the overlapping and sometimes conflicting usage, RLHF is increasingly viewed as a generalization of PbRL. While both involve human feedback to define RL objectives, PbRL primarily focuses on relative feedback, such as binary comparisons and rankings. RLHF not only includes these aspects but also extends to a wider range of feedback types (Metz et al. 2023). Table 1 gives an exemplary overview of our interpretation of these terms.

Another concept, semi-supervised reinforcement learning (SSRL), introduced by Christiano (2016) and discussed by Amodei et al. (2016), refers to an RL setting where an agent receives feedback on a subset of its experiences. The initial discussions of SSRL focused on absolute feedback on subsets of the agent's experiences, making the concept complementary to PbRL. In contrast to PbRL and RLHF, the term SSRL seems to be used less in the recent literature.

In our work, we adopt the viewpoint that RLHF is a broader category that encompasses various approaches where human feedback is used to define the objective of an RL agent. In this definition, RLHF encompasses both PbRL and SSRL. As the definitions and distinctions between these terms are not universally agreed upon, these distinctions are based on our interpretation of the current predominant usage of these terms in the literature.

1.3 Scope of the Survey

This section outlines the criteria guiding our selection of approaches in the realm of RLHF. We focus on works that rely on a reward model as the sole source of information about the objective. This reward model should then be learned in an interactive, online, scalable, and asynchronous manner. The following will describe each of these criteria in more detail.

Reward Modeling We focus on approaches that learn a reward model from human feedback and then use this model to train a policy. Although it is possible to directly optimize a policy from human feedback (Wirth et al. 2017), thereby performing RLHF without reward learning, this approach has been practiced only rarely so far. The decomposition into reward learning and policy training offers many conceptual and practical benefits. Among those benefits are the direct applicability of supervised learning techniques for the reward model and the possibility of evaluating the reward model in isolation. In addition to that, the decomposition naturally leads to a form of semi-supervised learning, enabling the agent to use labeled episodes for reward model training while leveraging unlabelled episodes to refine its behavior and explore the environment. Note, however, that there is a recent trend in the direction of direct policy learning in the domain of language model fine-tuning (see Section 5.2.3).

Human Defined While there are many approaches that include humans in the RL loop, in this survey, we focus on approaches where human feedback is the only source of truth about the objective. This excludes approaches to reward shaping, feature engineering, and other forms of human guidance that are supplementary to a given objective.

Interactive and Online We also put an emphasis on providing feedback in an interactive, online manner. This excludes imitation learning, learning from demonstration, and pure inverse RL. While we do not directly cover inverse RL in this survey, combinations of inverse RL methods with interactive improvements of the reward function are in scope and employed by some of the surveyed methods. See Sections 3.3 and 5.5.1 for a discussion of those approaches.

Scalable and Asynchronous We focus on works in which the human is included in the loop, but the agent is not blocked by the human's feedback, and the human does not need to be present continuously. This distinguishes RLHF from more direct methods of incorporating a human into the RL loop, and we believe that this is key for practicality and efficiency.

In addition to these criteria, we mainly focus on works published after 2017 since earlier works are surveyed by Wirth et al. (2017). Nevertheless, some works from this period are revisited from time to time in order to elaborate on certain concepts that are still state of the art or have significantly shaped it. Exceptions are made if the used methods are of interest for RLHF approaches.

1.4 Prior Surveys

Based on the criteria mentioned in the previous section, we will first differentiate our survey from other surveys in marginally related subject areas sharing the common theme of human-in-the-loop RL. Then, we will describe the differences between our survey and previous surveys or survey-like articles that exist within the RLHF field.

1.4.1 Human-in-the-Loop RL

Human participation in machine learning, particularly in guiding machine learners, is a much-studied scenario. This field, commonly referred to as human-in-the-loop (HITL) machine learning, can be further divided into subfields based on various criteria, e.g., the ones detailed in Section 1.3. Prior surveys of these subfields are compiled in Table 2 and briefly summarized in the following.

Table 2: An overview of prior surveys of human-in-the-loop RL. ✓ indicates that the criterion is a main focus of the survey, (✓) indicates that the criterion is partially addressed, while ✗ indicates that the criterion is not covered.

Reference	**Topic**	**Reward Modelling**	**Human Defined**	**Interactive and Online**	**Scalable and Async.**
Wu et al. (2022)	HITL machine learning.	✗	✗	✗	✗
Najar et al. (2021)	RL with human advice.	(✓)	(✓)	✓	(✓)
Lin et al. (2020a)	Social feedback.	✗	(✓)	✓	✗
Poole et al. (2022)	Human brain signals.	✗	✓	✓	✗
Arzate Cruz et al. (2020)	Interactive RL for HCI.	✗	✗	✓	✗
Osa et al. (2018)	Imitation learning.	✗	✓	✗	✗
Arora et al. (2021)	Inverse RL.	✓	✓	✗	✓
Bignold et al. (2021)	Assisted RL.	✗	✗	(✓)	✗
Luketina et al. (2019)	Language-informed RL.	✗	✗	✗	✗
Zhang et al. (2021)	Human guidance.	✗	✓	✓	✗
Ji et al. (2023a)	AI Alignment.	✓	✗	✓	✓
Liu et al. (2023b)	RLHF for LLMs.	✗	✗	✗	✗
Ours	RLHF.	✓	✓	✓	✓

Human-in-the-Loop Wu et al. (2022) survey HITL machine learning in general. They also cover some applications of RLHF (for LLMs in particular) but do not give a detailed overview. Similarly broad in scope, Najar et al. (2021) study the setting of RL with human advice, which they define as 'teaching signals that can be communicated by the teacher to the learning system without executing the task.' While this setting subsumes RLHF, the broad generality limits the depth in which their survey can cover RLHF approaches.

Interactive RL RLHF can be considered a sub-field of interactive RL, which studies RL algorithms that learn in interaction with humans. This interaction can take the form of feedback defining an objective, resulting in the RLHF setting, but can also, e.g., be used to drive exploration or speed up the agent's learning process.

Arzate Cruz et al. (2020) survey interactive RL from an human-computer interaction (HCI) viewpoint, exploring various ways humans can influence RL agents, with a particular focus on reward definition based on human feedback, without a predefined environmental reward function. Due to the breadth of their survey, they do not cover many works in this area. The survey by Lin et al. (2020a) centers on interactive RL using human social cues, like gestures and spoken language, but does not cover the reward modeling aspect. Similarly, the study by Poole et al. (2022) examines RL with direct feedback from human brain signals, such as through brain-computer interfaces, also not focusing on reward modeling.

Demonstrations Learning from demonstrations, in the form of behavior cloning (Osa et al. 2018) and inverse RL (Arora et al. 2021), shares the goal of RLHF to learn behavior from human input. In contrast to RLHF, however, it requires demonstrations of the desired behavior instead of feedback, and these demonstrations are usually not provided interactively and online. This limits their applications and also their final performance due to the availability of near-optimal demonstrations. Nonetheless, imitation and demonstration can be a useful component of an RLHF system but are not the main focus of this survey. However, we will discuss the intersection between these fields in some parts whenever necessary.

Assisted RL Bignold et al. (2021) review the field of assisted RL, where an agent may receive external information (for example, from a human) that aids it in action selection. While updates to the reward function are one of the possible effects of advice in this setting (in addition to action selection or modifications of the agent's internal state), it is usually assumed that an initial reward function is given and the extent of the updates is limited to reward shaping or supplementary reward signals. In contrast to RLHF, the external information does not define the task but only helps the agent in achieving it. Closely related to this, Luketina et al. (2019) survey RL assisted by natural language. In addition to this assistance setting, they also discuss approaches that infer a language-conditioned reward function. However, they discuss this setting rather briefly and use techniques from inverse RL and not RLHF.

Guidance In their survey on human guidance, Zhang et al. (2021) delve into various aspects related to RLHF. Although they touch on aspects such as reward learning, it is not the primary emphasis of their work. Instead, their main focus lies on exploring more immediate approaches that do not involve the learning of a reward model.

Table 3: An overview of prior RLHF-specific surveys and articles with substantial review components. ✓ indicates that the aspect is addressed, (✓) indicates that the aspect is partially addressed, while ✗ indicates that the aspect is not covered.

Reference (Focus)	Beyond Comparisons	Label Collection	RM Training	Theory	App. and Benchmarks
Wirth et al. (2017) (preference-based RL).	✗	(✓)	(✓)	✗	(✓)
Abdelkareem et al. (2022) (recent advances of PbRL).	✗	✗	(✓)	(✓)	(✓)
Jeon et al. (2020) (reward-rational implicit choice).	✓	✗	(✓)	✗	✗
Metz et al. (2023) (feedback types).	✓	✓	✗	✗	✗
Casper et al. (2023) (open issues in RLHF).	✓	✗	✓	✗	(✓)
Ours (fundamentals, recent advances, and trends).	✓	✓	✓	✓	✓

AI Alignment Ji et al. (2023a) provide a general overview of AI alignment, i.e., the challenge of aligning the objectives of an intelligent system with those of its human operators. This survey covers RLHF in some detail. As AI alignment is a very broad field, however, the article nevertheless does not go into as much depth on the topic of RLHF as we do here.

Applications Liu et al. (2023b) give an overview of current applications of RLHF methods for LLMs such as ChatGPT and GPT-4. Even though it currently enjoys a lot of attention, it is only a specific application area for RLHF. Our survey adopts a broader perspective, examining the diverse applications and impact of RLHF encompassing application areas beyond LLMs.

1.4.2 PbRL and RLHF

There have been previous surveys or survey-like articles that are closely related to RLHF. Table 3 gives a brief overview of how these articles differ from ours, which we will explain in more detail below.

Preference-Based RL Previous surveys in the domain of RLHF often focus on PbRL, where feedback is limited to binary preferences (see Section 1.2). An illustrative example of this is the survey by Wirth et al. (2017), which is a direct precursor to our work. In contrast to our work, they concentrate on binary preferences for trajectories and primarily survey methods that learn policies without deriving a reward model. Since then, the reward-modeling approach has become dominant in the field, and other approaches have extended RLHF to new feedback types. Abdelkareem et al. (2022) give another more recent literature review of PbRL. While this review focuses on reward modeling and includes some recent work, it is far less comprehensive than our review, as many aspects are only touched upon and partly overlap with those of Wirth et al. (2017).

Feedback Types Although not a survey per se, Jeon et al. (2020) propose reward-rational implicit choice as a unifying framework to comprehend many previous studies in PbRL and RLHF. To illustrate the generality of their approach, they overview different feedback types used in previous work and explain how they fit into their framework. Similarly, it is not strictly a survey, Metz et al. (2023) propose a common framework for studying user interaction and interface design for multiple feedback types. As part of their work, they provide a classification of feedback types and a brief overview of RLHF approaches. Nevertheless, many facets of RLHF are not dealt with at all in those studies, as they are not primarily survey articles. In contrast to Jeon et al. (2020) and Metz et al. (2023), our survey has a more extensive coverage going beyond their study of feedback types and also discussing more recent work.

Open Problems Casper et al. (2023) provide a detailed overview of the open questions and limitations of RLHF with a particular focus on aspects of security, governance, and transparency. In their article, reward modeling is also covered, as is human feedback, which goes beyond preference comparisons, but other aspects, such as theoretical approaches or an overview of existing benchmarks, are not included. Thus, it can be seen as a supplementary article that is ideal for further reading once being familiarized with the topic through our survey.

All in all, our survey can be seen as the canonical continuation of Wirth et al. (2017), which examines the evolution of the field of PbRL to the more modern and general field of RLHF. This includes a thorough description of the basics as well as an in-depth discussion of current advances and trends in the field.

1.5 Outline

In the next section, we begin with an introduction to the basics by revisiting the most important concepts from the standard RL setting, which are also naturally important in RLHF (Section 2). We then dive into the RLHF topic by outlining the most studied scenario of reward model learning from pairwise preferences. Using this introductory and illustrative example scenario, we explain the basic framework of the RLHF alongside its three main components of (human) feedback, label collection (feedback acquisition), and reward model learning. These three main components will essentially form the structure of our survey. In Section 3, we turn our attention to the human feedback component and provide an overview of the different types of feedback as well as their key attributes. The important concepts in terms of label collection are then explained in Section 4, followed by learning the reward model in Section 5. Section 6 is devoted to an overview of recent progress on the theoretical side of RLHF, including approaches involving a theoretical guarantee, as well as theoretical insights into the relationship between standard RL and RLHF. Finally, Section 7 highlights some interesting practical applications of RLHF and the existing benchmarks before Section 8 concludes the survey by pointing out some possible avenues for future work.

2 Preliminaries

In this section, we recall the basic setting and the most important concepts of RL and RLHF. In the course of this review, we will fix the notation that will be used throughout the survey. We first introduce what is probably the most studied RLHF scenario, i.e., learning a reward model from binary trajectory comparisons. On the basis of this introductory and illustrative example scenario, we explain the basic framework of RLHF with its main components and briefly discuss the respective roles of these components in the learning process. We will also briefly touch on active learning, which has a strong connection to the feedback collection component.

Notations For any integer $n \in \mathbb{N}$, we denote by $[n]$ the set $\{1, 2, \ldots, n\}$. For any set S, $\Delta(S)$ denotes the set of probability distributions over S. We use $\mathbb{P}(E)$ for denoting the probability of some event E, while $\mathbb{E}[X]$ is used to denote the expected value of a random variable X. In some cases, we will write $\mathbb{E}_P[\cdot]$ or similar variants to emphasize that the distribution for the expected value is governed by the probability distribution $P \in \Delta(S)$. Moreover, we will write $X \sim P$ if a random variable X is distributed according to a probability distribution P.

2.1 Reinforcement Learning

Reinforcement learning (RL) (Sutton et al. 2018) is the setting of learning behavior from rewarded interaction with an environment. Such a learning environment is formalized as an Markov decision process (MDP), which is a model for sequential decision-making. In an MDP, an agent iteratively observes its current state, takes an action that causes the transition to a new state, and finally receives a reward that depends on the action's effectiveness. Formally, an MDP is defined as a tuple $(\mathcal{S}, \mathcal{A}, P, R, d_0, \gamma)$ where

- $\mathcal{S}$ is a set of states (the *state space*),
- $\mathcal{A}$ is a set of actions (the *action space*),
- $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a transition function (the *transition dynamics*),
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function,
- $d_0 \in \Delta(\mathcal{S})$ is a distribution over initial states,
- and $\gamma \in [0, 1]$ is a discount factor.

The transition function P defines the dynamics of the environment: For any state s and action a, the value $P(s, a)(s')$, also sometimes denoted $P(s' \mid s, a)$, is the probability of reaching the state s' after executing a in s. In light of this, we will also refer to the transition function sometimes simply as the *transition dynamics.* For a given state and action, the transition is conditionally independent of all previous states and actions, which is known as the *Markov property* and the reason for the naming as an MDP. The value $R(s, a) \in \mathbb{R}$ provides an immediate evaluation after performing action a in state s, which is also called the (instantaneous) reward. It is also quite possible that the instantaneous reward is 0 for some states, and one only receives a reward in specific states, for example, in so-called *terminal* states

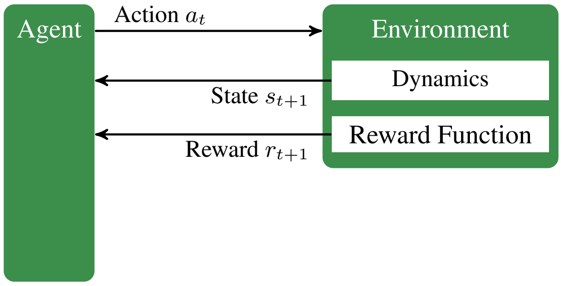

Figure 1: The standard RL setting.

for which the transition function is zero. When both the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are finite, we call the MDP a *tabular* MDP.

In an MDP, an H*-step trajectory* τ is a sequence of $H \in \mathbb{N}\setminus\{0\}$ pairs of state-action ending in a terminal state. Formally, it is given by $\tau = (s_0, a_0, s_1, a_1, \dots, s_H)$. Given $t_0 \geq 0$ and $H' \leq H$, we can define a *segment* $\sigma = (s_{t_0}, a_{t_0}, s_{t_0+1}, a_{t_0+1}, \dots, s_{H'})$, which refers to a continuous sequence of steps within a larger trajectory. A trajectory τ's *return* $R(\tau)$ is the accumulated (discounted) rewards collected along this trajectory:

$$R(\tau) = \sum_{h=0}^{H-1} \gamma^h R(s_h, a_h)\,. \tag{1}$$

Note that we here use the same notation for the return and the reward function, but both have different signatures (trajectory vs. state-action pair). We can also define return $R(\sigma)$ of a segment σ in a similar manner. The return is well defined even if the horizon H is infinite as long as $\gamma < 1$. If the MDP is a tabular MDP and any trajectory has finite length, i.e., H is necessarily finite, we call the MDP *finite,* and otherwise *infinite.*

A *policy* specifies how to select actions in a state, either deterministically or stochastically. In the former case, a policy is simply a mapping $\pi : \mathcal{S} \to \mathcal{A}$ from states to actions, while in the latter, it is a mapping $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ from states to probability distributions over actions. Since the deterministic case is a special case of the stochastic one, we assume the latter case in the following.

The basic interaction loop is depicted in Fig. 1: The agent chooses an action $a_t \sim \pi(s_t)$ based on its policy and the current state. As a consequence, the environment transitions into the new state $s_{t+1} \sim P(s_t, a_t)$, governed by the transition dynamics. The agent observes this new state and the reward $r_{t+1} \sim R(s, a)$, after which this interaction cycle is restarted.

In this setting, the RL agent aims at learning a policy that maximizes the expected return

$$J(\pi) = \mathbb{E}_{d_0, P, \pi}[R(\tau)]\,,$$

where the expectation is with respect to policy π, transition function P, and initial distribution d_0. To solve this problem, two families of RL approaches have been considered: *model-based* RL and *model-free* RL. The methods in the first family learn a model (i.e., P, R) of the underlying MDP to help solve the RL problem, while the methods in the second directly try to obtain a good policy without learning an MDP model. The second family can be further decomposed into two main categories: *value-based* methods and *policy search* methods. In deep RL, both value functions and policies are approximated with neural networks.

Value-based methods (e.g., DQN and its variants (Mnih et al. 2015; Hessel et al. 2018)) aim at learning the Q-function Q^* of an optimal policy. The Q-function of a policy π is defined by:

$$Q_\pi(s, a) = R(s, a) + \gamma \mathbb{E}_{P,\pi}\left[\sum_{h=0}^{H-1} \gamma^h R(s_h, a_h)\right],$$

where in the expectation, $s_0 \sim d_0$, $a_0 \sim \pi(\cdot \mid s_0)$, and $a_h \sim \pi(\cdot \mid s_h)$ as well as $s_h \sim P(\cdot \mid s_{h-1}, a_{h-1})$ for $h \in [H]$. A policy can be naturally designed from a Q-function by choosing an action in a greedy manner in each state: $\pi(s) = \arg\max_a Q(s, a)$. Note that for a deterministic optimal policy π^* it holds that $J(\pi^*) = \mathbb{E}_{d_0}[Q^*(s, \pi^*(s))]$.

Similar to the action-value function Q, we can also define the state-value function

$$V_\pi(s) = \gamma \mathbb{E}_{P,\pi}\left[\sum_{h=0}^{H-1} \gamma^h R(s_h, a_h) \,|\, s_0 = s\right].$$

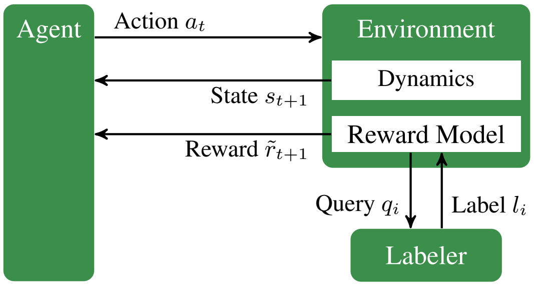

Figure 2: RLHF with reward modeling.

Its value for some state s is the expected return when starting in that state and then always using the policy π. It is related to the Q-function by means of

$$V_\pi(s) = \mathbb{E}_{a \sim \pi(s)} \left[Q_\pi(s, a) \right]$$

for any state $s \in \mathcal{S}$.

In contrast, policy search methods directly aim at finding a good policy in some parametrized policy space. The most data-efficient algorithms in this class of methods follow an actor-critic scheme where both an actor (i.e., a policy) and a critic (i.e., usually its Q-value function) are learned at the same time. Typical representative methods here are PPO (Schulman et al. 2017), TD3 (Fujimoto et al. 2018), or SAC (Haarnoja et al. 2018).

RL algorithms can further be classified as either *on-policy* or *off-policy*. In an on-policy algorithm, such as PPO, only the recently generated transitions are used for training. In contrast, in an off-policy algorithm, such as DQN (or its variants), TD3, or SAC, the agent can be updated with transitions not necessarily generated by its current policy. While on-policy training is usually more stable, off-policy training enables more data-efficient learning by reusing samples from a replay buffer that stores past transitions.

2.2 Preference-Based MDPs

In contrast to standard RL as described in the previous section, RLHF does not assume that a reward signal is available. It instead assumes the existence of an *oracle* (e.g., *human labeler*) that can provide information about the reward in a specific indirect manner. More precisely, in the RLHF, the agent can make queries q_i to the oracle, which in practice means asking for human feedback, and in response, the agent receives a label l_i, which in general gives a hint about the reward. In principle, the query can be made asynchronously to the actual conventional RL cycle. See Fig. 2 for an illustration.

In the most common setting, the oracle can compare two (segments of) trajectories, but various other cases have been considered, as we shall see later on. For the former case, RLHF is based on the setting of preference-based MDPs (Gilbert et al. 2017; Wirth et al. 2017), which can be defined as an MDP model without reward function, but where comparisons of trajectories are available.

2.3 Reward Learning

RLHF approaches can be divided into two categories, depending on whether a utility-based approach is used for reward modeling or an alternative criterion that is detached from a utility concept is used (Gilbert et al. 2016b; Gilbert et al. 2016a; Wirth et al. 2017). Most works fall into the first category, on which this overview focuses. Such approaches assume a human-dependent utility function that can be used as a reward function in order to apply standard RL methods. Next, we will describe the commonly used approach for reward learning for the common setting of binary trajectory comparisons.

The prevalent approach to learning a utility function from observations of pairwise comparisons is based on the Bradley-Terry model (Bradley et al. 1952), which stipulates a probabilistic model for the oracle (human labeler):

$$\mathbb{P}(\tau_1 \succ \tau_2) = \frac{1}{1 + \exp(R(\tau_2) - R(\tau_1))},$$

where $\succ$ means "preferred to" and $R(\tau)$ corresponds to the utility (i.e., return in the context of RL) of τ. Note that this utility function is a kind of surrogate function for the true reward function, which is (tacitly) assumed to induce the same optimal policy as the true reward function. For a given data set $\mathcal{D} = \{\tau_1^i \succ \tau_2^i \mid i \in [N]\}$, a utility function R_ψ

parameterized by ψ can then be learned by the maximum likelihood principle (or equivalently using a cross-entropy loss):

$$\max_{\psi} \prod_{i=1}^{N} \frac{1}{1 + \exp(R_\psi(\tau_2^i) - R_\psi(\tau_1^i))} \,. \tag{2}$$

In the context of RL, since $R_\psi(\tau) = \sum_{h=0}^{H-1} \gamma^h R_\psi(s_h, a_h)$, (2) can then directly be used to train a function approximator (e.g., single or ensemble of neural network) to approximate R.

This entire modeling approach accommodates the case of a noisy or unreliable oracle, in which case the Bradley-Terry model can be understood as the generative model of the answers from the oracle (or labels provided by the human labeler). When the oracle is reliable, more direct methods based on preference elicitation to recover the reward function have been studied (Regan et al. 2009; Regan et al. 2011; Weng et al. 2013; Gilbert et al. 2015; Sadigh et al. 2017). In this survey, we will focus on the general case where the oracle may be noisy.

Note that in contrast to the typical way of preference learning, the learned reward function is used to train an RL agent and not directly to compare trajectories. This discrepancy in the objective function in the reward learning part and how the learned rewards are used may lead to suboptimal policies (Lindner et al. 2021).

2.4 Reinforcement Learning from Human Feedback

In the RLHF setting as illustrated in Fig. 2, the learning agent needs to solve an RL task without having access to a reward function. To this end, the agent usually simultaneously learns an approximation of the reward function (via the assumed utility function) and an RL policy. Therefore, a generic RLHF algorithm consists of repeating two phases: (1) reward learning and (2) RL training. The first phase can itself be decomposed into two main steps: (i) generate queries to ask the oracle, (ii) train a reward function approximator with the answers provided by the oracle. The RL training part is more conventional and is usually directly based on running a deep RL algorithm using the currently trained reward function approximator.

Algorithm 1 Generic RLHF Algorithm in an Actor-Critic Scheme

```
 1: Initialize parameters $\theta$ (policy), $\phi$ (critic), and $\psi$ (reward)
 2: Initialize replay buffer $\mathcal{B}$ with randomly-generated trajectories
 3: for $i = 1, \ldots, N$ do
 4:     // Reward learning
 5:     Generate queries from $\mathcal{B}$
 6:     Update $\mathcal{D}$ with answers to queries from the oracle
 7:     Update $\psi$ using $\mathcal{D}$ (e.g., to maximize Eq. (2))
 8:     // RL training
 9:     Update $\mathcal{B}$ with new trajectories generated with $\pi_\theta$
10:     Update $\theta$ (actor) using $\mathcal{B}$ and $R_\psi$
11:     Update $\phi$ (critic) using $\mathcal{B}$ and $R_\psi$
12: end for
```

This basic generic algorithm is summarized in Algorithm 1, where an off-policy actor-critic scheme is assumed to be used for the RL training part, but other RL policy learning approaches can, of course, also be used here. For an on-policy algorithm, such as PPO (Schulman et al. 2017), only the recently generated transitions are used for training. For a DQN-like algorithm, lines 9 to 11 would be replaced by a loop where transitions are generated by a behavior policy based on the current estimate of the Q-function (e.g., ε-greedy algorithm) and the Q network is updated using mini-batches of transitions sampled from the replay buffer $\mathcal{D}$.

An efficient RLHF algorithm needs to overcome several difficulties which are specific to this setting:

- The oracle may provide various types of feedback (see Section 3). The questions of what information some given feedback provides and how observed feedback can be exploited need to be answered (see Section 5).
- Informative queries need to be generated to minimize the efforts of the oracle, which is crucial when it is a human (see Section 4). Active learning techniques (see next subsection) can be adapted to face this challenge.
- The RL agent is actually trained in a non-stationary environment since the reward approximator is concurrently updated. The RL training part needs, therefore, to account for this factor (e.g., using non-vanishing learning rates).

- There is also the question of how the agent's performance can be meaningfully evaluated, especially if the reward function is not known (see Section 7).
- Collecting feedback directly from humans introduces its own challenges, such as the question of a suitable user interface and the associated issues of delay between query and feedback observation, or the feedback variability and reliability (see Section 4). This may explain why many studies evaluate novel RLHF algorithms with simulated feedback.

A standard RL algorithm can be run in the RL training part, as done in most previous work in RLHF (although this may not be the best approach). This suggests that any improvements in a standard deep RL method (e.g., auxiliary losses (Gelada et al. 2019), planning in learned model (Hafner et al. 2020), curriculum learning (Narvekar et al. 2020), or data augmentation (Laskin et al. 2020; Lee et al. 2020; Lin et al. 2020b)) may potentially be transferred to the RLHF setting. In addition, most previous work in RLHF directly uses trajectories stored in replay buffer $\mathcal{D}$ to synthesize queries. An interesting research direction to explore in RLHF would be to specifically generate trajectories in order to be able to synthesize more informative queries (instead of only generating trajectories that are beneficial for RL training). This would lead to tackle a novel exploration-exploitation dilemma: Shall we visit state-action pairs that may be bad but may help better learn the reward function, or shall we visit state-action pairs that we currently think are good? This is further discussed in Section 5.5.3.

In RLHF, since the oracle is a human or a group of humans, reducing the number of queries is crucial to limit the labeling cost. Therefore, the reward learning part requires techniques similar to those proposed in active learning, which we recall next.

2.5 Active Learning

In active learning (Settles 2012), the task is to strategically select data points for labeling to minimize the amount of labeled data required to achieve a desired level of performance, particularly valuable in scenarios like RLHF where labeling is costly. Unlike batch learning, where labeled data is predetermined, active learning empowers the learner to actively select the most informative unlabeled instances for labeling, maximizing the learning process with limited labeled data. We will only briefly introduce the active learning task here and then discuss the strategies for creating informative queries considered thus far in Section 4.

For RLHF with pairwise comparisons, this setting can be formally described as follows. Suppose there is a set of N pairs of trajectories $\{(\tau_1^i, \tau_2^i) \mid i = 1, \ldots, N\}$, where each pair (τ_1^i, τ_2^i) can be interpreted as an unlabeled instance. To efficiently learn a reward function to explain observed pairwise comparisons, an agent can select a set of unlabeled pairs (possibly a singleton) to query an oracle to obtain their labels.

At a high level, the main idea in active learning is to query data points to quickly reduce the epistemic (i.e., reducible) uncertainty about the predictions of the learned model, although other aspects can be important, such as the representativeness of the queried data points (see Wang et al. (2023a) for a survey). Two main representations are considered to describe this epistemic uncertainty: either using an ensemble of models or using a Bayesian representation. In both cases, a first basic approach selects queries using uncertainty-based criteria in order to focus on instances with high prediction uncertainty as measured by, e.g., variance or entropy computed over predictions. In contrast to the first approach, where the selection criteria are instance-based, a second approach considers criteria that may depend on all instances. Possible options are, for instance, expected model change, expected error reduction, or density-weighted uncertainty-based criteria. Here, the weights in the expectation or density allow us to take into account the distributional information about the instances and, therefore, to focus on the higher-density regions.

3 Feedback

Feedback mechanisms are fundamental to the success of any RL system. In the standard setting as described in Section 2.1, the RL agents expect feedback in the form of scalar immediate rewards. These rewards are most commonly determined by a hand-engineered reward function, which can be used to evaluate any state-action combination. As discussed in Section 1.1, it is desirable to allow humans to refine behavior interactively through feedback instead of requiring them to pre-specify a reward function.

While a human could, in principle, assign rewards to each of the agent's actions directly, thereby taking the role of the reward function, this is usually impractical for multiple reasons. The main challenge is the human effort required to provide rewards on a sufficiently regular basis, i.e., at least once per episode. In addition to that, directly integrating human rewards into the RL loop would require these rewards immediately, which impedes the learning pace while waiting for human feedback. Finally, the standard RL setting expects numeric state-action rewards, which is challenging to provide in a consistent manner.

In contrast to directly rewarding each of the agent's actions, RLHF as discussed in this survey (see Section 1.3) harnesses indirect and asynchronous feedback methods. Such methods avoid the challenges of immediate numeric rewards and are also better aligned with human interaction patterns, resulting in improved learning progress and human user experience.

A feedback type is a kind of interaction in which a human conveys some information about their preferences. Examples include pairwise comparisons and direct critiques. This section is concerned with the many ways in which human feedback can be expressed and used. Several previous works have already studied and sorted feedback types by listing the most common ones (Jeon et al. 2020) and discussing their attributes and dimensions (Metz et al. 2023; Lindner et al. 2022). The attributes and classes described in this section build upon this prior work and can be considered a synthesis and extension of it.

The remainder of this section will start by discussing relevant attributes of feedback types that can be used to classify them (Section 3.1). We will then discuss common classes and examples of interactive feedback types (Section 3.2) as well as some non-interactive types that can serve as initializations (Section 3.3).

3.1 Attributes of Feedback Types

Feedback types may differ on many dimensions, some of which relate to the way feedback is given (arity, involvement), others to the form of the query instance it is given on (granularity, abstraction), and yet others to features of the human interaction (intent, explicitness). The attributes we discuss in this section are based on the framework proposed by Metz et al. (2023). We have adjusted and added terminology where it aids clarity, generalized the distinction between relative and absolute feedback to 'arity', and added the categories of co-generative involvement and literal intent. We have also systematically analyzed a set of exemplary feedback types with respect to these dimensions (see the next section), extending the initial work in that direction by Metz et al. (2023) who consider a smaller set of abstract classes. Furthermore, we systematically analyze a set of exemplary feedback types in the next section, expanding on the initial examination of a smaller range of abstract classes by Metz et al. (2023). In the following, we will introduce each of the six attributes in more detail.

Arity This attribute describes whether a single instance is evaluated in isolation (*unary*) or relative to other instances (*binary*, n*-ary*). Unary feedback is often convenient for detailed and descriptive feedback but lacks any grounding and therefore puts a great burden on the human to provide consistent feedback. Non-unary feedback always has an implicit grounding but relates to the instances being comparable. While n-ary feedback, such as rankings, can provide more information than binary feedback, it also puts a higher cognitive burden on the labeler.

Involvement The labeler may either passively *observe* an instance, actively *generate* it, or coactively participate in its generation (*co-generation*). Passive involvement poses the smallest challenge to the labelers since it does not require the ability to demonstrate the task. It can also easily be directed at the most informative examples with active learning techniques. Unfortunately, passive feedback often cannot match the information density of active feedback. It is, therefore, common to combine both types to first initialize the reward model from (possibly very suboptimal) active feedback and then refine it from passive feedback. Between these two extremes is co-generative feedback, in which a human can share control with the agent. This can be less demanding than active feedback and makes it possible to direct the human's attention to the most informative samples, but it is still more taxing than purely passive involvement.

Granularity Feedback may also differ on the granularity of the instances being evaluated. This ranges from whole *episode* recordings over partial *segments* to feedback on individual *steps* (i.e., states, actions, or state-action pairs). A more coarse-grained granularity has the advantage of giving humans more context and getting feedback for larger sections of behavior but also poses credit assignment problems. Finer-grained feedback is much easier to learn from and conveys more information, but is often impractical or tedious for humans to provide. Note that we only classify a type of feedback as "episode" granularity if it *requires* entire episodes. If it is compatible with partial segments as well, we classify it as "segment" even if the discussed source paper uses entire episodes.

Abstraction This describes whether feedback is given directly on raw *instances*, e.g., behavior recordings (see granularity) or on abstract *features* of the instances. While feature-level information can be easier to learn from, extracting useful features is challenging. In some contexts, it may also be harder for a human to make abstract judgments rather than more intuitive instance-level judgments. Note that this always refers to the level of abstraction that the user sees, which may differ from the features used as inputs for the reward model. Types

Table 4: An overview of the common classes and their defining attributes. When an attribute is not specified, this means that it is not a defining feature of the class and may vary in different instantiations.

Class	Granularity	Involvement	Arity	Abstraction	Intent	Explicitness
Critique	–	Observed	Unary	–	Evaluative	Explicit
Comparisons	–	Observed	2+	–	Evaluative	Explicit
Inter-Temporal	Segment	Observed	Unary	–	Evaluative	Explicit
Task Desc.	Episode	Observed	–	Feature	Descriptive	Explicit
Social Behavior	Segment	Observed	Unary	Instance	Literal	Implicit
Improvements	Episode	Co-generative	Unary	Instance	–	–
E-Stops	Episode	Observed	Unary	Instance	Literal	Implicit
Importance	–	Observed	–	–	Descriptive	Explicit
Feature Traces	Segment	Active	Unary	Instance	Descriptive	Explicit
Similarity Queries	–	Observed	Ternary	–	Descriptive	Explicit

of feedback that depend on active generation (see involvement), such as improvements, generally work on a raw instance level.

Intent The assumed human intent can be important for feedback processing. A human may be *evaluative*, *instructive*, or *descriptive* in their explicit feedback, while they are generally *literal* in their implicit feedback. Evaluative, instructive, and descriptive feedback is pedagogical in nature, aiming to teach a reward function, whereas literal feedback is a byproduct of a human actor's efforts to optimize the reward function directly. While evaluative, instructive, and literal feedback is generally given within the context of a particular query instance, descriptive feedback can describe the task more generally, e.g., through a partial reward function.

The distinction between literal and pedagogical feedback was introduced by Milli et al. (2020). They argue that humans with pedagogical intent (i.e., evaluative, instructive, or descriptive) may act differently compared to humans with literal intent. Even though they find that assuming the (wrong) literal intent can still lead to better reward inference, it still indicates that it can be important to know this intent to choose the right human model (see Section 5.1.5).

Explicitness Humans may communicate *explicitly* for the purposes of feedback or *implicitly* as a side-effect of actions directed at other purposes. While explicit information is easiest to learn from, implicit information is often much more readily available and can possibly communicate more detailed preferences.

3.2 Common Classes

Even though the concrete types of feedback used in the literature are rarely exactly the same, they can generally be sorted into a set of common classes. We will describe a selection of those classes, their defining attributes, and examples of concrete instances in the literature in the following. Table 4 gives an overview of the classes and their attributes as described in Section 3.1.

3.2.1 Critique

Critique is arguably the most direct type of feedback. In this setting, the human expresses their preference by directly critiquing an instance of agent behavior, often in the form of binary feedback. Note that critique, as considered in this survey, is distinct from directly supplying a reward signal since it is given in an asynchronous and indirect manner (see Section 1.3). The defining features of critique are that the human passively observes the behavior (**involvement**), gives feedback on a single instance (**arity**), and does so explicitly (**explicitness**) with an evaluative **intent**. The feedback may be given for any **granularity** and on any level of **abstraction**.

There are many examples of critique feedback in the literature. Xiao et al. (2020) employ binary feedback on individual state and action pairs. Although they learn a shaping reward that complements an environment reward signal, the same technique could be used without environment reward. Huang et al. (2023) extend this to multi-label feedback, allowing the user to distinguish between a regular good or bad action and a terminal action that achieves the goal or fails to do so. They map these classes to scalar values and then learn a reward model by regression. Wang et al. (2020) present an approach to learning a reward model from noisy critiques in the form of human physiological signals (brain signals) using active querying. In contrast to this action-level feedback, Fu et al. (2018b) and Singh et al. (2019) rely on binary outcome success labels. Fu et al. (2018b) introduce the basic approach, which Singh et al. (2019) extend by moving to an off-policy setting and including online queries, thereby reducing the reliance on many positive examples by interactively correcting false positives.

In addition to learning the main reward function, critique can also be used for safety evaluation. Cosner et al. (2022) train a secondary reward model focused on safety from binary action critiques. This is in addition to the main reward model, which is trained from comparisons in their approach. Note that this secondary safety model could, in principle, be trained with any of the feedback types discussed here, using methods identical to the ones used for reward learning.

3.2.2 Comparisons

Binary comparisons and rankings are among the most common types of feedback. The defining features of comparisons are that the human passively observes the behavior (**involvement**), gives relative feedback on multiple instances (**arity**), and does so explicitly (**explicitness**) with an evaluative **intent**. It is most commonly given on a segment (**granularity**), but other granularities are possible in principle (Cosner et al. 2022). Similarly, comparisons are commonly requested on an instance level (**abstraction**), but this is not a requirement.

Comparisons were first used for direct policy learning (Akrour et al. 2011; Cheng et al. 2011), but were later extended to the reward-learning setting (Wirth et al. 2016; Christiano et al. 2017). The most common setting (Christiano et al. 2017) relies on pairwise comparisons of trajectory segments, but comparisons of individual states or actions were also considered in early PbRL works (Fürnkranz et al. 2012).

This basic setting has been extended and modified in various ways. To reduce noise in the labels, it is common to extend the binary choice by giving the labelers the option to avoid the hard choice and instead indicate incomparability, uncertainty, or perceived similarity (Holladay et al. 2016). The most common interpretation of this intermediate option is "equally preferable" (Liu et al. 2023a), e.g., as if each trajectory had an equal probability of being preferred in the preference predictor. Ibarz et al. (2018) instead provide two intermediate options: The previously discussed "equally preferable" and an "incomparable" option which results in simply omitting the query from the data set. In contrast to this, Verma et al. (2023a) explicitly state that they do not allow for the equally preferred option, arguing that these cases are rare enough not to matter very much. Another line of research suggests that more precision in the expression of pairwise preferences, such as softening the hard binary choice to scalar feedback indicating the strength of a preference (Wilde et al. 2022), can be beneficial for preference learning. Other extensions change the pairwise setting to choices among larger choice sets (Ziegler et al. 2020) or even full rankings (Myers et al. 2022; Brown et al. 2020; Ouyang et al. 2022). Ziegler et al. (2020) note that for language tasks, a larger choice set can amortize the time needed for a labeler to get acquainted with the context necessary to understand a query. Basu et al. (2019) propose to use hierarchical queries, i.e., a sequence of pairwise comparisons that build up on each other.

Pairwise comparisons provide many benefits over absolute ratings, such as reduced bias, inconsistencies, and subjectivity (Yannakakis et al. 2015), but also convey relatively little information per label. Tien et al. (2023) study the weaknesses of pairwise-comparison-based reward learning. Since the amount of information provided for each label is small, these models are prone to causal confusion, i.e., misattributing reward to noise and misidentifying the reward function.

3.2.3 Inter-Temporal Feedback

One limitation of trajectory comparisons is that they require a set of roughly comparable trajectories. In many real-world environments, starting conditions or even the agent's current task may vary between episodes. In these cases, it is hard for human labelers to compare trajectories from these episodes, limiting the usefulness of comparison feedback. One way to remedy this limitation is to provide feedback within a single trajectory. Instead of comparing a set of instances with each other as in regular comparative feedback, inter-temporal feedback conveys relative judgments over different states in time within a single instance. The defining features of inter-temporal feedback are that it is given explicitly (**explicitness**) on segment (**granularity**) while passively observing (**involvement**) a single instance (**arity**) of the agent's behavior with evaluative **intent**. It is most commonly given on raw instances, but any level of **abstraction** is possible in principle. There are two main ways to convey this feedback: *Reward sketching* and *inter-temporal preferences*.

Reward sketching, as introduced by Cabi et al. (2020), involves users sketching a visual representation of the reward function over time. This type of feedback, which can be given by sketching a graph with the mouse while watching a behavior recording, provides intuitive, per-timestep reward annotations. Rahtz et al. (2022) also adopted this approach, referring to it as "one of the highest-bandwidth feedback mechanisms currently available".

Inter-temporal preferences were introduced by Abramson et al. (2022). In this setting, humans give feedback on multiple points of a trajectory, indicating whether an agent makes progress towards or regresses from a goal. This is then interpreted as preferences relative to the other labeled and unlabelled points. The authors note that one potential downside of this feedback type is that labelers may tend to give preferences on short-term actions that are easy to judge, failing to communicate long-horizon preferences. Cui et al. (2018) propose a similar type of feedback, in which humans segment a trajectory into good and bad parts. This makes it possible to derive many state-action labels from few segmentation points.

3.2.4 Task Descriptions

Task descriptions consist of direct, concept-level descriptions of the task the agent is supposed to complete. This is most commonly conveyed in the form of information about the reward function, i.e., by proxy rewards or reward queries. Descriptive feedback does not generally refer to any particular behavior instance but instead gives global direction for the entire task. However, in line with our selection criteria (Section 1.3), we only consider task descriptions that are interactive and online within the context of one or multiple observations to fill holes in the initial description. The defining features of task descriptions are that the labeler passively observes the agent's behavior (**involvement**) and gives feedback explicitly (**explicitness**) on a feature-level (**abstraction**) with descriptive intent (**intent**). Descriptive feedback may be given with respect to a single or multiple instances (**arity**), although it generally refers to multiple instances. It is most commonly given on an episode (**granularity**), but other granularities are possible in principle.

Task descriptions are most commonly provided in the form of incomplete reward functions. He et al. (2022) demonstrate this using proxy reward functions, which are preliminary reward functions that might not cover all edge cases, to guide the agent toward learning the actual reward function. Alternatively, Mindermann et al. (2018) and Hadfield-Menell et al. (2017b) suggest querying about the reward function. They allow users to choose from a set of understandable, linear proxy rewards or to specify which features are more critical in the linear reward structure.

3.2.5 Social Behavior

Humans give rich implicit social feedback in the form of facial reactions and gestures when interacting with agents. The defining attributes of this type of feedback are that it is given implicitly (**explicitness**) on passively observed (**involvement**) segments (**granularity**) with respect to a single instance (**arity**) and literal **intent**.

Cui et al. (2021) propose a framework to learn reward functions from such social behavior. They suggest a two-phase training setup. In the first phase, they ground the implicit feedback by use of incentives, i.e., they incentivize humans to have a known objective. After learning a mapping from feedback to reward, they use regular RL techniques to learn a policy. Note that the learned reward function can be seen as conditional on human implicit feedback and, therefore, they require a human in the loop throughout training.

3.2.6 Improvements

Improvements are a form of feedback in which the human improves on the agent's behavior, either by intervening as the agent acts or by providing a corrected behavior after the agent acts. To improve an episode, it is usually necessary to observe the entire episode (**granularity**) at the instance level (**abstraction**). In this type of feedback, the human both observes and demonstrates behavior, resulting in co-generative **involvement**. Improvements generally relate to a single reference trajectory being improved (unary **arity**), although an improvement could also be interpreted as a binary comparison between the improved and the non-improved trajectory. Improvements are most commonly provided explicitly with instructive **intent**.

We distinguish between post-facto improvements, calling them *corrections*, and improvements made while the agent is acting, calling them *interventions*. The key difference is that the uncorrected trajectory is available in the case of corrections, while it can only be estimated in the case of interventions.

Interventions can be considered to be an instance of the shared autonomy setting since the agent and the user share autonomy to reach a common goal. This setting is studied by Abramson et al. (2022), who ask humans to intercede on agent failure. They use this to collect targeted demonstrations where the agent is the weakest and to identify challenging situations for their evaluations. The gathered data is then used for behavior cloning and reward model training. The fact that a correction occurred is not directly used as feedback. In contrast to this Losey et al. (2022) observe that the interventions themselves are intentional and therefore carry information about the true objective. Losey et al. (2018) propose to incorporate uncertainty for active learning and risk-sensitive deployment in this setting. Li et al. (2021) further extend the intervention setting to learn from a sequence of correlated physical corrections without needing to wait until the trajectory is completed.

The correction case is closely related to the setting of coactive learning (Shivaswamy et al. 2015), in which a learning system repeatedly proposes a solution which a user may correct to reach a common goal. Jain et al. (2015) treat corrections as a demonstration while Jeon et al. (2020) propose (but do not evaluate) an alternative interpretation of inferring implicit preferences from comparisons by assuming the corrected trajectory is preferred over the original one.

3.2.7 E-Stops

Emergency stops (e-stops) (Ghosal et al. 2023) are an active type of feedback. In this type of feedback, the human may intervene with the agent's behavior by stopping it, i.e., they may choose to stop the agent's current trajectory at any point. This is closely related to interventions but, in contrast to those, does not suggest an alternative action. The defining features of e-stops are that the human both passively observes the agent's behavior (**involvement**), gives absolute feedback on a single instance (**arity**) on the instance level (**abstraction**) and does so implicitly (**explicitness**) as a side-effect of regular interaction. The **intent** is literal due to the implicit nature. For the purposes of intervention, the human usually observes the full episode (**granularity**). Due to the small amount of infrequent information they provide, e-stops should only be considered as a supplementary feedback type.

E-stops are primarily intended to prevent bad behavior and only implicitly convey information about the correct behavior. This interaction and the arising incentives have been formalized in the form of the "off-switch game" by Hadfield-Menell et al. (2017a). Jeon et al. (2020) propose to interpret this as a form of reward-rational feedback, where the 'off' choice maps to the trajectory with the robot remaining still after the off switch has been triggered. Kahn et al. (2021) demonstrate that a robot can learn to navigate using such feedback.

3.2.8 Importance

Another form of supplementary feedback may come in the form of importance labels, communicating which parts of the observation are important for the objective. Its defining features are that the importance information itself does not contribute towards generating behavior samples (observed **involvement**), is of descriptive **intent**, and is given explicitly (**explicitness**). **Granularity**, **arity**, and **abstraction** may vary depending on the primary feedback type. Since importance feedback needs a base task with respect to which the importance is defined, it cannot be used on its own but is rather a supplementary type of feedback.

One way to convey this information is by labeling salient parts of a visual input. This is explored by Guan et al. (2021), who augment pairwise comparisons with manually annotated visual saliency maps, informing the algorithm which parts of the visual input contributed to the decision. They leverage these annotations for data augmentation by assuming that random perturbations to irrelevant (non-salient) regions do not impact the human preferences. Basu et al. (2018) take an even more direct approach by combining comparative feedback with direct feature queries, i.e., asking the user which feature is important for inferring the reward.

3.2.9 Feature Traces

While many approaches either rely on hard-coded features or learn a model entirely end-to-end, it is also possible to actively elicit new features from human feedback. Feature traces were proposed by Bobu et al. (2022) as an approach to actively learn new relevant features. This type of feedback relies on a human operator to demonstrate a behavior in which a certain feature of interest, such as the distance to a sensitive object, monotonically increases or decreases. They make it possible to extend the set of features once the current set can no longer adequately explain the human feedback supplied through another type of feedback. The defining characteristics of feature traces are that they are of descriptive (**intent**) and explicitly (**explicitness**) given in an active manner (**involvement**) for a single (**arity**) segment (**granularity**) on an instance-level **abstraction**.

Feature traces are strongly related to inter-temporal preferences (Section 3.2.3) since both types rely on changes in feature- or reward values in the course of a single trajectory. Bobu et al. (2022) propose to learn from feature traces by leveraging a Bradley-Terry model to learn the feature values, similar to other approaches that use such a model to learn reward values. Similar to importance feedback, feature traces rely on another type of feedback to actually make use of the learned features and is, therefore, a purely supplementary form of feedback. For instance, Bobu et al. (2022) use intervention feedback to train a reward model on the set of features derived using feature traces.

3.2.10 Similarity Queries

Similarity queries are a feedback type aimed at learning a representation conforming to a notion of similarity and difference in the trajectory space. That aim is closely aligned with that of feature queries, though the actual queries are more similar to comparisons. The queries consist of triples of trajectories, with one anchor and two alternatives, for which the human has to decide which pair is more similar. Responses to similarity queries are given on observed behavior (**involvement**) with ternary **arity**, descriptive **intent**, and explicit **feedback**, while the **granularity** and **abstraction** may vary. This type of feedback was first introduced by Bobu et al. (2023), who used it to learn representations for reward learning. Similar to feature traces and importance, this is a supplementary type of feedback.

3.3 Initializations

Some modes of communicating reward functions are not interactive nor online and, therefore, do not directly fit within the scope of this survey (Section 1.3). Since these are often used to initialize a reward function for later refinement with some of the previously discussed interactive feedback types, they are still worth mentioning.

Initializations are most commonly given by examples of successful task completions, either in the form of terminal or goal states (Xie et al. 2018), expert demonstrations (Fu et al. 2018a; Abramson et al. 2022), labeled demonstrations (Du et al. 2023), or ranked demonstrations (Brown et al. 2019). Since this is not the main focus of our survey, we refer to the literature on inverse RL (Arora et al. 2021) for further details.

4 Label Collection

In this section, we explain how preference data can be collected for training a reward model. We first overview how queries can be generated for a given query type, then how the query type itself can be selected, and finally discuss issues arising in such human-computer interaction.

4.1 Active Learning

Following the terminology used in Bayesian active learning, we call the criterion to measure the quality of a query *acquisition function*, which is used to select which queries are presented to a human labeler. We first review the various acquisition functions used in RLHF for query selection and then discuss the extension to the choice of feedback type.

4.1.1 Acquisition Function

One core problem that needs to be tackled in RLHF is that of learning about the human's preferences. This problem shares some similarities with the active learning setting since the agent can actively query a human teacher about those preferences. However, in contrast to standard active learning, which usually assumes a supervised learning setting, in RLHF, the agent needs to solve this problem in the context of RL. This means that it can both influence the distribution of the data (i.e., the transitions) and decide which data should be labeled.

As the RL agent is trained and its learned policy changes, the trajectories it generates will naturally evolve. Most work directly uses the trajectories obtained during RL training for preferences learning. Alternatively, the agent can also generate trajectories specifically to be used for querying (not necessarily for RL training), possibly with a learned transition model to control the sampling cost (e.g., Reddy et al. (2020)). This kind of active generation of queries can possibly lead to more informative ones.

In order to efficiently learn a suitable reward model, the agent must generate and select queries (Line 5 from Algorithm 1) so that it can quickly learn a good strategy using those queries. This selection is performed via a criterion, usually called *acquisition function*, which allows the queries to be compared. An efficient acquisition function may need to include various factors, such as: uncertainty, on-policy data, query simplicity, trajectory quality, query diversity, and query cost, which will be discussed one by one in the following. As a side remark, as noted by Habibian et al. (2022), interestingly, the queries asked by an RL agent also reveal its current reward learning stage.

Uncertainty This corresponds to how uncertain the agent is about the ground-truth reward function. We are particularly interested in epistemic uncertainty in this setting, i.e., uncertainty that can be reduced by additional queries. This is usually one of the most important aspects to consider when deciding which query to ask. This uncertainty is usually either represented as a probability distribution (i.e., belief) in Bayesian approaches or using an ensemble of reward networks to approximate this belief.

With a belief representation, various classic Bayesian acquisition functions have been considered, such as the *probability of improvement* or the *expected improvement* (e.g., Daniel et al. (2014)). However, they require maintaining a distribution over reward functions (e.g., using Gaussian processes (Daniel et al. 2014)) and, therefore, may not be suitable for more complex domains due to the computational complexity. Thus, simpler alternative criteria have been considered, such as *information gain* (Mindermann et al. 2018; Bıyık et al. 2020a) or *variance*. Although computationally heavy, *volume removal* (i.e., the minimum volume of the hypothesis set removed by an answer to a query) has also been studied (Sadigh et al. 2017; Basu et al. 2018; Basu et al. 2019).

When using an ensemble instead of a direct belief representation, these criteria for epistemic uncertainty reduction correspond to measures of disagreement within the ensemble. Previous criteria could possibly be applied, but one popular candidate is the *variance* of the ensemble outputs (Lee et al. 2021b; Metcalf et al. 2022; Gleave et al. 2022b).

Recently, the *average entropy* of the ensemble outputs (assuming a Bradley-Terry model for the human answers) has also been used (Lee et al. 2021b; Lee et al. 2021a; Park et al. 2022). However, note that it does not quantify epistemic uncertainty but rather the uncertainty in the human's answers, as provided by the response model of the human. Therefore, this criterion may not be suitable in the RLHF setting since it amounts to focusing on the queries for which an answer is expected to be the most random (according to the Bradley-Terry model). By definition of this model, the segments in the pairwise comparisons are the most similar in terms of returns and are, therefore, the hardest to answer for the human.

In addition to epistemic uncertainty, one may also take utilities (i.e., returns or expected returns in RL) into account to select queries. In a Bayesian setting, this leads to acquisition functions such as *expected value of information* (Myers et al. 2023) or *information gain over return differences* (Lindner et al. 2021), while in a non-Bayesian setting, the notion of *regret*, which measures the difference of performance between a policy optimal for the ground-truth reward function and a policy optimal for a learned reward function, can be used (Wilde et al. 2020).

On-Policy Data Only focusing on uncertainty is likely insufficient or inefficient in RL. In particular, it may be important to favor more on-policy trajectories to guarantee the relevance of the generated queries. Indeed, improving reward learning in state-action regions that may never be visited with the current policy would lead to wasteful queries (Lindner et al. 2021). One simple approach to ensure that the data is more on-policy is by favoring more recently-generated trajectories (Eberhard et al. 2022).

Query Simplicity Selecting queries only based on their informativeness may lead to queries that are hard to answer for a human, which is, for example, the case for the average entropy. The ease of answering a query is important to alleviate the cognitive load of the human oracle. Some work specifically takes this aspect into account, for instance, by considering the similarity of consecutive queries (Racca et al. 2019) or the information gain. For this latter criterion, Bıyık et al. (2020b) show that in contrast to volume removal, it naturally leads to queries that are easier to answer for a human because information gain can be increased when the uncertainty in the human answer is lower.

Trajectory Quality Most approaches directly use the trajectories generated during RL training. Especially early in training, these can be very bad with respect to the ground-truth reward function. In addition to that, they can be irrelevant or even contradictory for a given task (Katz et al. 2021). Building queries on such trajectories may lead to unnatural queries for a human to respond to, such as comparing a very bad trajectory with an irrelevant one. Katz et al. (2021) measure trajectory quality by optimizing over sampled reward functions.

Query Diversity When asking many queries (in sequence or in batch), the diversity of the queries becomes especially crucial to avoid asking redundant queries. Most work follows a very myopic approach: Queries are often selected from a randomly generated set of potential queries, and sequences of queries are not really coordinated. Few work (Bıyık et al. 2018; Wilde et al. 2018) specifically tackles the former point, but the latter point is rarely considered, which may be due to its computational intractability. Indeed, planning ahead a sequence of queries would amount to solving a sequential decision-making problem under uncertainty over a combinatorial action space (i.e., the set of possible queries).

Query Cost The cost of generating queries may also be an important factor if the interaction of the human is live since it may not be practical to let the human wait before showing any queries (Bıyık et al. 2018). In that case, it may be more important to quickly show some relatively good queries instead of computing the most informative ones.

Although this factor may not translate directly into an acquisition function, it may influence the choice of the acquisition function and its implementation in a given problem.

Since various different acquisition functions have been considered, some effort (Lee et al. 2021b; Lee et al. 2021a) has been made to compare them. Generally speaking, uncertainty-based criteria (e.g., variance or average entropy) seem to often perform better empirically compared to random selection, a query diversity-based criterion alone or combined with an uncertainty-based criterion. In addition, Eberhard et al. (2022) empirically observed that a variance-based criterion performs better than a selection method only based on trajectory recency. Surprisingly, random selection has been shown to perform competitively in some cases (Christiano et al. 2017; Ibarz et al. 2018). Thus, a better understanding of which acquisition function should be preferred in which situation or domain is still an open question.

In addition, combinations of different criteria have naturally also been evaluated. For instance, Reddy et al. (2020) use four acquisition functions (high uncertainty, high novelty, high reward, low reward) in parallel. This approach has also been validated in a 3D environment (Rahtz et al. 2022). A more sophisticated approach consists of considering a portfolio of acquisition functions and learning to select them using a multi-armed bandit approach (Hoffman et al. 2011).

Various extensions to the basic setting have also been investigated. In the context of multiple human labelers, the issue of selecting reliable teachers to query arises (Daniels-Koch et al. 2022). Assuming all teachers have the same preferences, this can be modeled by incorporating a rationality coefficient β into a Bradley-Terry model and estimating this factor:

$$\max_{\theta} \prod_{i=1}^{N} \frac{1}{1+\exp(\beta(R_\psi(\tau_2^i) - R_\psi(\tau_1^i)))}\,, \tag{3}$$

where a higher β corresponds to a more reliable human (see Section 5.1.2). The setting in which this assumption does not hold, i.e., the labelers' reward functions differ (a setting already considered in inverse RL (Choi et al. 2012)), has also been studied recently (Xue et al. 2023a; Dong et al. 2023; Myers et al. 2022; Bakker et al. 2022). Interestingly, a noisy oracle may provide more information than a completely reliable oracle because the frequency of erroneous answers given by the former is related to how much a segment is preferred to the other one (Xu et al. 2020; Chan et al. 2021). In contrast, only a binary preorder over segments can be inferred from the answers of a reliable and deterministic oracle, which may not be enough to recover the true reward function.

4.1.2 Adaptive Choice of Feedback Type

In addition to selecting queries within a given feedback type, it is also possible to actively select the feedback type itself (Fitzgerald et al. 2023). The best choice of feedback type can depend on many factors, such as human rationality as well as task-dependent factors, some of which may change during the labeling process. Ghosal et al. (2023) formalize this setting as one in which we try to select a feedback design (or feedback type) x out of the space of possible designs $\mathcal{X}$ such that the expected information gain over the distribution of reward functions is maximized for the next human response. Concretely, the goal is to choose a feedback design by means of

$$x = \operatorname*{argmax}_{x \in \mathcal{X}} \mathbb{E}_{c_h \sim P(c_h \mid x)} \big[D_{KL}\big(\mathbb{P}(\theta \mid c_h, x) \mid \mathbb{P}(\theta)\big) \big]\,,$$

where c_h is the human response to a query defined by x and $\mathbb{P}(\theta)$ is the prior distribution over reward functions.

Ghosal et al. (2023) find that the most informative feedback type depends on the (type-dependent) rationality of the human labeler (see Section 4.2.1). More precisely, it is shown that the most informative feedback depends on the rationality factor, e.g., while demonstrations are more informative than comparisons when the human is highly rational, comparisons should be preferred in less-rational settings. Given that this rationality might change due to factors such as fatigue or an individual labeler's capabilities, this suggests that adaptively adjusting the feedback type during the labeling process may be worthwhile. Further study of this relationship is a promising area for future work.

4.2 Challenges of Human Labeling

This section explores the label collection process, which follows after query selection. This task intersects with several related disciplines, especially within the social sciences, as it encompasses the design of interactions to facilitate informative query responses. A prominent field in this area is psychometrics (Furr 2021), which focuses on measuring psychological attributes, including preferences. Similarly, survey research (Fowler 2013) is dedicated to developing techniques for gathering information from individuals via surveys. Human-computer interaction plays a significant role as well, investigating the design of user interfaces tailored for preference elicitation (Pommeranz et al. 2012). Moreover, preference label collection is also necessary for discrete choice experiments within health economics (Ryan et al. 2008), where it is used for the assessment of service values.

4.2.1 Psychology-Aware Preference Elicitation

Understanding human psychology is essential for effective preference elicitation in RLHF systems. Human decision-making is complex, often diverging from traditional rational choice models due to cognitive, social, and emotional factors. This complexity is exemplified by phenomena like fatigue, which can affect the reliability of choices based on the order of queries. This section investigates these phenomena, exploring how constructive preferences, biases, framing effects, and social interactions shape the observed choices. Recognizing and addressing these psychological underpinnings is key to developing more accurate and reliable systems. In this section, we will discuss various psychological phenomena, such as cognitive biases and response biases, and related effects (fallacies, biases, heuristics, psychological phenomena impacting decision-making processes), which may falsify labels by adding systematic bias or noise.

Preference learning methods typically assume the existence of inherent, stable preferences that can be elicited through querying. Contrary to this assumption, psychological research, such as the work by Lichtenstein et al. (2006), indicates that preferences are often constructed during the elicitation process and may vary with the method of elicitation or over time. This suggests that the feedback type not only affects elicitation's effectiveness but also shapes preferences. Systematic biases, noise, and other psychological factors may influence observed choices, challenging the traditional models of human choice used to infer latent utilities (see Section 5.1). The elicitation method, query presentation, and context thus play a critical role in shaping measured preferences, compounded by cognitive biases and irrationalities.

The influence of psychological phenomena on preference learning has been well-documented in the literature, especially within the context of explicit preference elicitation for recommender systems. For instance, Tran et al. (2021) provide a thorough discussion of the relationship between psychology and recommender systems. Similarly, Atas et al. (2021) review how preference construction is influenced by cognitive biases, personality traits, and emotional states in recommender systems, discussing effects like serial position, framing, anchoring, choice overload, and preference visibility. In a more specialized discussion, Mandl et al. (2011) focus on cognitive biases in the context of consumer decision-making and its interaction with recommender systems. Mandl et al. (2011) specifically address cognitive biases in consumer decision-making in interaction with recommender systems. Finally, Kaufmann et al. (2023) link these psychological aspects to RLHF, discussing the common practice of using synthetic instead of real human feedback for algorithm evaluation and highlighting the limitations of that approach. They further discuss challenges posed by real human feedback, many of which are related to the concepts discussed in the following paragraphs, as well as the opportunities provided by integrating psychological insights into RLHF systems.

Constructive preferences are closely related to *framing effects*, which refer to changes in elicited preferences based on how tasks or alternatives are described, even when these descriptions are essentially equivalent. For example, presenting a choice as a loss versus a gain can lead to different decisions despite identical outcomes. Moreover, *serial position effects*, commonly known as primacy and recency effects, also play a significant role. These effects describe the tendency for the beginning and end of an experience to influence subjective experience disproportionately. This phenomenon becomes particularly relevant in scenarios like video choices, where the initial or concluding segments might disproportionately affect preferences. Atas et al. (2021) discuss both of these effects in the context of recommender systems.

Ordering effects pose another challenge in preference elicitation, where the sequence of queries can affect responses. Day et al. (2012) outline several factors contributing to these effects: institutional learning, changing preferences, and varying levels of cognitive effort. Institutional learning involves gaining familiarity with the task and feedback type, which can enhance labelers' expertise and, consequently, the accuracy of their responses. However, due to the constructive nature of preferences, their preferences may evolve during the elicitation process, leading to changing preferences. This evolution might also be influenced by *anchoring effects*, where previously seen instances bias responses. Furthermore, cognitive effort levels can fluctuate due to factors like fatigue or boredom. This is closely related to *choice overload*, a form of fatigue from excessive choices, as discussed by Atas et al. (2021) and bounded rationality, as explored by Chen et al. (2013). In such scenarios, labelers might opt out of making a choice when overwhelmed by options. Bounded rationality refers to the limitations in human decision-making capabilities, particularly when processing large amounts of information, which aligns with the concept of choice overload. To address these challenges, studies like Bıyık et al. (2020b) and Zhang et al. (2022) propose methods to reduce cognitive effort in responding to queries. Bıyık et al. (2020b) focus on posing queries that are straightforward for humans to answer, while Zhang et al. (2022) enhance the human evaluation process by presenting queries in a user-friendly format.

When multiple labelers collaborate on the same task in preference elicitation, as is studied, e.g., by Barnett et al. (2023) and Daniels-Koch et al. (2022), this may lead to another set of biases if they have the opportunity to exchange information. This exchange may either be direct or indirect through observing the system's predictions, which are based on the other labeler's feedback. Such interactions can affect their preferences through several mechanisms, as identified by Atas et al. (2021): anchoring effects, transfer of emotional states, and conflict avoidance. *Anchoring effects*, for instance, occur when a labeler's choices are influenced by the knowledge of others' preferences or system predictions, a phenomenon also discussed under the term *preference visibility*. This bias can lead labelers to align their preferences with the anchors they are exposed to, which is a significant consideration in recommender systems. Understanding these biases is crucial for designing RLHF systems that mitigate the influence of labeler interactions on preference construction.

The effects previously discussed stem from systemic biases in preference expression. In addition to these biases, choices may also be affected by noise. This is commonly discussed under the term stochastic rationality, where an agent's behavior is rational with respect to an unobserved random state. The reward-rational implicit choice framework, as introduced by Jeon et al. (2020), addresses this by integrating a rationality factor β into the human choice model (see Eq. (3)). This factor's impact has been further examined by Ghosal et al. (2023) through synthetic experiments and user studies, demonstrating that accurately estimating this type-dependent rationality coefficient can enhance learning performance and guide feedback type selection (see Section 4.1.2). However, a practical method for estimating this factor remains a challenge. While Ghosal et al. (2023) use calibration feedback with a known latent utility function for estimation, such an approach is not feasible for most tasks. In a related study, Daniels-Koch et al. (2022) investigate a scenario with multiple teachers, focusing on the agent's ability to select the most knowledgeable or rational teacher. Therefore, developing more advanced methods to estimate this factor, along with understanding its variability due to factors like fatigue or other ordering effects, presents a vital area for future research in preference elicitation

Incorporating psychological insights into the preference-learning components of RLHF systems is essential for optimizing their efficacy. A key area of focus should be research aimed at mitigating biases and harnessing cognitive aspects of preference formation. For instance, designing user interfaces that minimize framing effects and developing algorithms that account for ordering and serial positioning are crucial steps. In this realm, Metz et al. (2023) propose a configurable user interface, called RLHF-Blender, for studying various feedback modalities and their combinations. Additionally, the study by Krening et al. (2018) on the impact of feedback type, such as binary critiques versus action advice, on task performance and labeler satisfaction highlights the significant role of feedback type in preference elicitation. Furthermore, the work of Pommeranz et al. (2012) in user-interaction design underlines the importance of having an expressive feedback type to increase user engagement.

The integration of these research findings into RLHF systems points to a clear need for a more multidisciplinary approach. Drawing insights from related fields like behavioral economics and psychology can provide valuable methodologies and perspectives. Addressing irrational choice patterns and enhancing the quality of human feedback remain critical challenges. As we continue to develop and refine these systems, the focus should be on creating robust frameworks that align learning processes with human behavior, effectively managing the inherent complexity and variability of human feedback.

4.2.2 Importance of Researcher-Labeler Agreement

High-quality labels are important for the final policy in an RLHF process. Early work on fine-tuning language models using RLHF noticed a mismatch between the researcher's goals and the (paid) labeler's actual labels (researcher-labeler disagreement). Ziegler et al. (2020) note that researchers agreed with each other about 60% of the time (on 4-way comparisons, where random choice would result in 25% agreement), while agreeing with labelers only 38% or 46% of the time (depending on the task). Stiennon et al. (2020) attempt to reduce these disagreements by maintaining a hands-on relationship with the labelers and thereby ensuring high researcher-labeler agreement. Concretely, they provide on-boarding with detailed instructions, keep an open channel of communication between researchers and labelers, and give feedback to the labelers. They evaluate the researcher-labeler agreement and reach an agreement rate of $77\% \pm 2\%$.

Due to the inherently subjective nature of the task (researchers agreed with each other in $73\% \pm 4\%$ of the cases), a much higher agreement rate cannot be expected. Ouyang et al. (2022) also report the agreement rates on a different task (instruction fine-tuning instead of summarization) and find that labelers agree with each other in $72.6 \pm 1.5\%$ of the time, after a screening procedure that, amongst others, selects labelers that agree with researcher labels.

The importance of quality does not trump the importance of quantity, however. Indeed, Stiennon et al. (2020) note that excluding low-confidence samples from the data set generally did not help with reward model training. This indicates that even though quality is important, a larger quantity is still generally better.

The scale of the labeled data set required for effective training and refinement varies widely, impacting the quality of the resulting models. Studies have shown a broad range in data set sizes, from tens of labels in smaller studies (Jain et al. 2015) to hundreds in more complex scenarios (Christiano et al. 2017). Larger-scale applications may require thousands (Guan et al. 2021) or even millions of labels (Abramson et al. 2022), each bringing its own challenges in ensuring label accuracy and consistency. This variability in data set size underscores the need for rigorous label quality control measures across different scales. In smaller data sets, each label carries more weight, making accuracy and precision critical. Conversely, in larger data sets, the challenge lies in maintaining consistency and mitigating systematic biases that might emerge from the sheer volume of data.

Similarly, the labeling setting varies in the surveyed works, from author-provided feedback (Kim et al. 2023), over small in-person studies (Katz et al. 2021), to larger remote studies (Kim et al. 2023). Each setting provides unique challenges to ensure high-quality labels.

Various works have suggested measures to improve label quality. Hagendorff et al. (2022) discuss the possible failure modes of the labeling task in more detail, for example, discussing systematic biases and conflicting motivation, and propose concrete changes to the training and evaluation methodology to alleviate these. Glaese et al. (2022) suggest providing labelers with multiple natural language rules and collecting preference labels for each rule individually to improve label quality. This is related to Bai et al. (2022b), who propose to generate feedback automatically based on such a set of rules and a language model.

5 Reward Model Training

In this section, we delve deeper into the aspect of reward model learning, which we briefly touched on in Section 2.3. We will discuss various aspects associated with the topic, such as the different reward model inputs or model architectures considered hitherto or the various possibilities regarding the training process.

5.1 Human Feedback Model

The basic premise underlying the majority of approaches in RLHF is that human feedback is directly related to the reward function to be learned. To this end, the human feedback must first be captured in a sound mathematical framework that establishes the connection to the reward function. On a high level, one can break down (almost) all feedback types in Section 3.2 to a choice scenario: The human chooses one specific feedback option (label) from an available (possibly infinite) pool of possible feedback options (choice sets)[1]. Here, the query that is made specifies the explicit contents of the choice set, e.g., if the query is to compare two trajectories, then the choice set consists of all possible outcomes for these two trajectories.

Assuming that human choices are not always optimal, one obtains a fruitful mathematical framework when focusing on the probability

$$\mathbb{P}\left(c \text{ is chosen} \,|\, \mathcal{C}\right) , \tag{4}$$

where $\mathcal{C}$ is the set of possible choices and $c \in \mathcal{C}$ one explicit choice. For the RLHF scenario, where the agent asks queries q_i and the human gives labels l_i as feedback (see Section 2.4), the choice set is specified by a function of the query. Formally, $\mathcal{C} = m(q)$ for some mapping m that maps a query q to the set of all possible candidate labels extractable from q for the specific feedback type. For example, if the query is to rank a finite number of trajectories, then the choice set is the set of all possible rankings that can occur for the trajectories involved.

With this view, we can therefore place (4) in the RLHF context and write

$$\mathbb{P}\left(\text{label } l \text{ is provided} \,|\, m(q)\right) \tag{5}$$

for the probability that a human labeler returns a label l from all possible candidate labels that can be extracted from a given query q One could also recover the noiseless scenario if the latter probability distribution is degenerated for all possible candidate label sets.

5.1.1 Boltzmann Distribution

Human choice models as in (4) have been studied for a long time in various scientific fields such as psychology (Thurstone 1927), economics (Train 2009) or behavioral science (Cattelan 2012). Accordingly, there are many different choice models to resort to for (5), which, in some cases, are the same models, just under different names. A popular class of such human choice models assumes every choice option c to be equipped with a (latent) utility u_c, which the human perceives in a perturbed way. This perturbation is modeled by means of perturbation random variables ϵ_c that perturb the utility in an additive way, so that (4) becomes

$$\mathbb{P}\left(c \text{ is chosen} \,|\, \mathcal{C}\right) = \mathbb{P}\left(c = \underset{c \in \mathcal{C}}{\operatorname{argmax}}\ u_c + \epsilon_c\right) . \tag{6}$$

The translation for the RLHF setting for (5) is then accordingly

$$\mathbb{P}\left(\text{label } l \text{ is provided} \,|\, m(q)\right) = \mathbb{P}\left(l = \underset{l \in m(q)}{\operatorname{argmax}}\ u_l + \epsilon_l\right) , \tag{7}$$

[1] This point of view goes back to the work of Jeon et al. (2020).

and we shall now stick to the RLHF translation from now on. These probabilities depend on the specific distributional assumptions that are made on the perturbation variables that only for specific cases lead to a closed-form of the right-hand sides of the latter displays. When stipulating a standard Gumbel distribution for the perturbations, one always obtains a closed form that is proportional to the exponential utility of the provided label:

$$\mathbb{P}\left(\text{label } l \text{ is provided} \,|\, m(q)\right) \propto \exp(u_l)\,. \tag{8}$$

This is known as the *Boltzmann distribution* that also appears in a perhaps slightly modified version in various different subfields of machine learning and statistics. When restricting to discrete (choice) sets for $m(q)$, this distribution is also known as the multinomial logit model (Train 2009) or Gibbs measure (Georgii 2011), and as the Bradley-Terry model (Bradley et al. 1952) when the choice sets consist of pairs. All of these also have a link to the Plackett-Luce model (Luce 1959; Plackett 1975), which is a probability distribution on the space of total orders or rankings (see Alvo et al. (2014) for details).

This model is often used for various reasons. A particularly compelling reason is the closed analytic form, which in turn makes it possible to obtain a closed form for the gradient with respect to the utilities. Another reason is that this model satisfies Luce's axiom of choice (Plackett 1975), which requires the probability of choosing an option from a pool of choice options not being affected by the presence or absence of other options in the pool. In this way, coherent decision-making is ensured, which, however, might be challenged as humans are likely not making fully rational decisions (see Section 4.2.1). Jeon et al. (2020) show that the usage of the Boltzmann distribution is justified by the principle of maximum entropy. More precisely, they show that it is the maximum entropy distribution over choices for a so-called satisficing human decision maker, i.e., one who is making in expectation a choice with an optimal reward up to some slack $\epsilon > 0$.

To build the bridge between reward learning and the modeling of human feedback, the Boltzmann distribution can be used by assuming that the utilities can be represented as a function of the reward function, usually featuring the return of a trajectory. More specifically, one assumes a *grounding function* ψ that maps choice options (or labels) to the set of distributions over trajectories and sets the utility of a label l as

$$u_l := \mathbb{E}_{\tau \sim \psi(l)}[R(\tau)]\,. \tag{9}$$

Table 1 in Jeon et al. (2020) provides an overview of the different grounding functions that lead to a specific feedback type. It is worth noting that one can also easily find a grounding function for the feedback type of a (partial) order over trajectories as considered, for instance, by Myers et al. (2022). Moreover, one can generalize this modeling approach by using (partial) segments instead of trajectories.

Although this general human feedback model has been much in use and shown to be useful for the sake of human alignment, it is not without its critics (see Lindner et al. (2022) or Section 3.2.1 in Casper et al. (2023)). This has led to different adaptions of the general model based on the Boltzmann distribution that will be discussed in the following. Moreover, we will also concisely review other human feedback models that have been in use besides the Boltzmann distribution, discuss relevant work on the consequences or robustness of human feedback model misspecification, and highlight contributions on varying the standard assumptions on the nature of the human feedback.

5.1.2 Human-Specific Rationality Coefficient

The Boltzmann distribution in (8) can be extended by a rationality coefficient $\beta \in [0, \infty)$ that reflects the precision of the human labeler[2]:

$$\mathbb{P}\left(\text{label } l \text{ is provided} \,|\, m(q)\right) = \mathbb{P}\left(l = \underset{l \in m(q)}{\operatorname{argmax}} \; \beta\, u_l + \epsilon_l\right) \propto \exp(\beta \cdot u_l)\,. \tag{10}$$

The higher β, the more (10) resembles a pointmass distribution modeling a highly rational human labeler (decision-maker) that is always able to identify the option with highest utility, while the lower β, the more (10) resembles a uniform distribution modeling a highly irrational human labeler (decision-maker) acting purely at random. Without this extension, the commonly considered Boltzmann distribution (or Bradley-Terry model in the common case of pairwise comparisons) in (8) assumes a rationality coefficient of 1. Ghosal et al. (2023) show in their experiments that the estimation of this coefficient can indeed positively influence reward learning. For the estimation, however, a calibration reward function is needed, as the rationality coefficient is otherwise not identifiable (Bengs et al. 2020). Similar findings are shown by Daniels-Koch et al. (2022), who model the rationality coefficient as a query-dependent function that might differ for the human labelers (see Section 5.1.6).

Another alternative to the rationality coefficient for representing irrational humans is achieved by introducing a query-independent error probability (Christiano et al. 2017). To be more precise, it is assumed that the human labeler only

[2] This can be achieved by multiplying the utilities u_l with β.

adheres to the Boltzmann distribution in (8) in 90% of cases and otherwise makes completely random decisions. This formulation is similar to Huber's contaminated model (Mu et al. 2023).

5.1.3 Alternative Utility Notions

Knox et al. (2022) show that the Boltzmann model does not generally lead to an identifiable reward function using (9) by presenting three concrete scenarios for which identification is not possible. The root cause of the non-identifiability is the usage of a trajectory's return as the utility in (9). They, therefore, suggest using a trajectory's regret as an alternative, which provably leads to identifiable rewards.

A trajectory's regret is the negated sum of the optimal policy's advantage over each state-action pair in the trajectory. Empirically, it has been shown that this modification improves the alignment of the learned strategy with human preferences. The downside of this alternative is that regret depends on the unknown optimal policy. Recently, it has also been suggested to consider Q-values of a human policy as the utilities (Myers et al. 2023), while Holladay et al. (2016) used differences of cost functions that depend on the available choice set and the human's uncertainty.

5.1.4 Human Feedback Models Beyond Boltzmann

While the human feedback model based on the Boltzmann distribution is the most widely used model nowadays, other models have also been considered in the literature. In particular, for the probability in (4) other models such as the Thurstone model (Wilson et al. 2012; Kupcsik et al. 2018; Bıyık et al. 2020a), the ridge-noise model (Schoenauer et al. 2014), the binary model (Sugiyama et al. 2012) or mixed forms thereof (Wirth et al. 2016) have been considered. Of these models, only the Thurstone model (Thurstone 1927) has a similar interpretation as the Boltzmann distribution based on perturbed utilities, only differing in the distribution of the perturbance random variables.

Link functions Another possibility, which is particularly popular in theoretical work on RLHF (see Section 6), is the use of other functions on the right-hand sides of Eq. (8) than the exponential function. The concept is primarily used for pairwise comparisons of trajectories. It essentially states that the probability of the result of a pairwise comparison between two trajectories is the difference of their utility values under a so-called *link function*. More specifically, let $q = \{\tau_1, \tau_2\}$ be the query to compare the trajectories τ_1 and τ_2, then, assuming a link function $\Phi : \mathbb{R} \to [0, 1]$, one models the probability in (5) for l representing a preference for τ_1 as

$$\mathbb{P}\left(\text{label } l \text{ is provided} \,|\, m(q)\right) = \mathbb{P}\left(\tau_1 \succ \tau_2 \,|\, m(\{\tau_1, \tau_2\})\right) = \Phi(u_{\tau_1} - u_{\tau_2})\,. \tag{11}$$

For l representing a preference for τ_2, one proceeds similarly. The minimal assumptions on the link functions are that

(i) it is (strictly) monotonically increasing to take into account that trajectories with higher utilities will also have a higher chance to be picked;

(ii) $\Phi(x) = 1 - \Phi(-x)$ to ensure that $\mathbb{P}\left(\tau_1 \succ \tau_2 \,|\, m(\{\tau_1, \tau_2\})\right) = 1 - \mathbb{P}\left(\tau_1 \prec \tau_2 \,|\, m(\{\tau_1, \tau_2\})\right)$.

Note that the second property implies $\Phi(x) = 1/2$ so that trajectories with the same utility also have the same chance of being selected. Any cumulative distribution function of a symmetric continuous random variable fulfills these two conditions. The two most common link functions that both fulfill the conditions are the linear link function given by

$$\Phi(x) = \max\{0, \min\{1, 1/2 \cdot (1 + x)\}\}$$

and the logistic link function given by

$$\Phi(x) = \frac{1}{1 + \exp(-x)}\,.$$

Both are cumulative distribution functions: The linear link function is the cumulative distribution function of a continuous uniform distribution on $[0, 1]$. In contrast, the logistic link function is the cumulative distribution function of a logistic distribution with location parameter 0 and scale parameter 1. Moreover, both are intensively studied in theoretical approaches (see Section 6.1), and the latter leads to (8) (when restricted on pairwise comparisons) and is a special case of the softmax function.

Two-Staged Choice Model Bobu et al. (2020b) propose the Limiting Errors due to Similar Selections (LESS) model that is inspired by the attribute rule model suggested by Gul et al. (2014). It assumes a feature map for trajectories and a (similarity) function mapping trajectory features and trajectories to integers and uses a two-stage process for modeling the human feedback (or choice): First, choosing a trajectory feature according to the Boltzmann distribution and then a trajectory with the (logarithmic) similarity functions as the utilities within the Boltzmann distribution. Their experiments show that this model can capture human feedback more appropriately than the standard Boltzmann distribution.

Generative Model Abramson et al. (2022) evaluate the usage of a generative model for learning from human preferences. More specifically, instead of assuming some underlying utility as in the Bradley-Terry model, they attempt to train a model to generate the human feedback (inter-temporal preferences in this case, see Section 3.2.3) and directly interpret this feedback as reward. However, they found that this empirically does not perform as well as the inter-temporal Bradley-Terry model.

5.1.5 Misspecification

The human feedback model may be misspecified in various ways. Milli et al. (2020) investigate the problem of misspecifying the nature of human feedback that can be either literal or pedagogical. The former means that the human gives targeted feedback for solving the actual RL problem, while the latter means that the human gives targeted feedback that is deemed helpful for the learner. They show theoretically and empirically that the case of a learner assuming a pedagogical with an actual literal human always performs worse than the reversed case, i.e., a learner assuming a literal with an actual pedagogical human.

A related question is studied by Freedman et al. (2021), namely, what if the learner makes incorrect assumptions about the choices from which the human selects its feedback? They consider different types of such choice set misspecification and show that depending on the type of misspecification, the performances might vary drastically, even leading to no losses at all in specific cases.

In the field of inverse RL, the general question of the robustness of reward learning in terms of a misspecified human feedback model is theoretically investigated by Skalse et al. (2023a). It turns out that the optimality model is not robust with respect to any misspecification, the Boltzmann model is robust for quite a range of specific misspecifications, and the degree of robustness of the maximal causal entropy model lies between the latter two. Even though these results are primarily derived for inverse RL, they also have similar immediate implications for RLHF.

5.1.6 Diverse Preferences

One potential issue with the RLHF framework is that it does not specify whose preferences to align to. In practice, it is common to aggregate preference feedback from multiple users and learn a single reward model to try to predict the mean preference. In contrast, Bakker et al. (2022) investigate this question more explicitly and proposes to learn multiple reward functions, which can then be aggregated in arbitrary manners and even be utilized to find consensus among people with different preferences. Myers et al. (2022) consider the case of multiple humans that provide the feedback and use a mixture model of Plackett-Luce models to represent the feedback more accurately. With a stronger focus on the active retrieval of human feedback, Freedman et al. (2023) model the problem of selecting a suitable human labeler as a variant of the multi-armed bandit problem (Lattimore et al. 2020) that they call hidden-utility bandit. In this variant, the agent has in each decision round the choice between two things: (i) drawing a bandit arm and receiving a hidden arm-dependent utility and observing an item, or (ii) asking a human for a preference for two pairs of items but incurring a human-specific query cost. The feedback mechanism of an individual human is modeled via a Boltzmann distribution with a known individual rationality coefficient. The same modeling of human feedback is also considered by Barnett et al. (2023), who, however, use a Bayesian approach to determine which person should be queried in order to obtain the most informative feedback in expectation. Daniels-Koch et al. (2022) model the rationality coefficient as a query-dependent function that might differ for the human labelers.

5.1.7 Relaxation of the Markov Assumption

Most works assume that the human feedback is given based on a latent Markovian reward model, i.e., the return of a trajectory τ decomposes into a sum of independent rewards over state-action pairs (see (1)). Early et al. (2022) relax this assumption by dropping the need for the Markov property, such that the instantaneous reward might depend on hidden states. Similarly, Kim et al. (2023) avoid the Markov assumption by utilizing a transformer as the preference model. A similar effect may be achieved by learning a state representation with a recurrent network in which the rewards are Markov, similar to the approach taken by Hafner et al. (2023), but we are not aware of any work exploring this. Abramson et al. (2022) work in a non-Markovian setting as well by utilizing memory-augmented networks for both the policy and the reward model.

5.2 Utility Learning

After choosing a human model to relate feedback to utilities, we can use the observed feedback to recover the latent utilities. This utility learning can be reduced to a standard supervised learning problem and, therefore, is commonly solved with the techniques of empirical risk minimization or Bayesian approaches, both of which will be discussed in the following.

5.2.1 Risk Minimization

The most prevalent variant for learning the reward function, already been presented in Section 2.3, is a special case of empirical risk minimization. The general approach of empirical risk minimization for reward function learning, assuming an underlying human feedback model with utilities u_{l_i} as in (9), is to find the minimizer of

$$\mathcal{L}(R;\mathcal{D}) = \sum_{i=1}^{N} \ell \left(u_{l_i}(R), m(q_i) \right) , \tag{12}$$

where $\mathcal{D} = \{(l_i, q_i)\}_{i=1}^N$ is the given data set of observed label and query pairs and $\ell : \mathbb{R} \times \mathcal{Q}$ is a suitable loss function, where $\mathcal{Q}$ denotes the set of all possible label sets. To provide an example, let us consider the common case of pairwise trajectory comparisons. Here the grounding function ψ is the identity, the queries are pairs of trajectories $q_i = \{\tau_1^i, \tau_2^i\}$ and the labels $l_i \in \{\tau_1^i \succ \tau_2^i, \tau_1^i \prec \tau_2^i\} = m(q_i)$ are the human's preference over the two trajectories. For a given query q_i, we then obtain (2) as a special case of (12) by using the loss function:

$$\begin{aligned} \ell(u_{l_i}(R), m(q_i)) &= -\log \left(\frac{1}{1 + \exp(u_{l_i}(R) - u_{m(q_i)\setminus l_i}(R))} \right) \\ &= -\log \left(\frac{1}{1 + \exp(\mathbb{E}_{\tau\sim\psi(l_i)}[R(\tau)] - \mathbb{E}_{\tau\sim\psi(m(q_i)\setminus l_i)}[R(\tau)])} \right) . \end{aligned}$$

This is the negative log-likelihood for the Boltzmann distribution for the observational pair $(l_i, m(q_i))$.

For the entire learning process, a model class $\mathcal{R}$ is assumed for the reward function R. This model class is usually a parameterized class of functions, such as, for example, the class of linear reward functions (Katz et al. 2021)

$$\mathcal{R} = \{R_{\boldsymbol{\theta}}(s,a) = \boldsymbol{\theta}^\top \boldsymbol{\phi}(s,a) \,|\, (s,a) \in \mathcal{S} \times \mathcal{A}, \boldsymbol{\theta} \in \mathbb{R}^d\},$$

where $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is some state-action feature mapping. This entails that good features are known in advance such that rewards can be expressed as a linear combination of those features. Using such linear models may lead to reward model misspecification. Studying this setting, Bobu et al. (2020a) propose to adapt the hyperparameter β in Eq. (3) to account for this issue.

Since the assumption of a linear reward model may be too strong in practice, most recent work is based on non-linear models, especially using differentiable models, but other cases have been investigated as well. In the former case, for instance, decision trees have been considered to learn an interpretable reward model (see Section 5.2.5). In the latter case, simple multilayer perceptron (MLP) has naturally been considered, but more recent deep learning architectures are more commonly used in the recent literature. Thus, especially with partially observable domains, the reward network may be composed of a state-action encoder followed by fully connected layers. For instance, Abramson et al. (2022) combine ResNet blocks for image processing, a learnable embedding table, a multi-modal transformer, LSTMs, and MLPs. Besides, Kim et al. (2023) utilize a Transformer-based architecture (Vaswani et al. 2017), motivated by the observation that rewards are often non-Markovian.

In addition to the usual empirical risk in (12), it is also typical, as in supervised machine learning, to add a regularization function to prevent overfitting:

$$\mathcal{L}(R_{\boldsymbol{\theta}};\mathcal{D}) = \sum_{i=1}^{N} \ell(u_{c_i}(R_{\boldsymbol{\theta}}), \mathcal{C}_i) + \lambda_r(\boldsymbol{\theta}) , \tag{13}$$

where $\lambda_r : \Theta \to \mathbb{R}_+$ is a regularization function defined on the parameter space Θ of the underlying reward model class. For instance, Christiano et al. (2017) simply use L2 regularization and also considers dropout in some domains.

The main supervised loss to train the reward model can also be augmented with additional losses corresponding to auxiliary tasks to avoid overfitting and improve generalizability. For instance, Abramson et al. (2022) use a behavior cloning loss and add a policy head to the reward network, thereby preventing that the reward model drifts too far from its initialization from the policy.

5.2.2 Bayesian Approach

As is generally the case in supervised machine learning, there is also the variant of using Bayesian modeling for learning a target object instead of the (empirical) minimization of a loss function. To this end, one starts with a prior distribution ρ over the parameter space of the reward function that is updated in light of the data set $\mathcal{D}$ by means of Bayes theorem:

$$\mathbb{P}(\theta \,|\, \mathcal{D}) \propto L_\theta(\mathcal{D}) \cdot \rho(\theta) ,$$

where $L_\theta(\mathcal{D}) = \prod_{i=1}^{N} \mathbb{P}\left(l_i \text{ is provided} \,|\, m(q_i), \theta\right)$ is the likelihood of the data under the assumed human feedback model with reward function R_θ. Such an approach is used for pairwise trajectory comparisons, for instance, by Schoenauer et al. (2014) for the noisy-ridge model or by Sadigh et al. (2017) for the Boltzmann distribution as the human feedback model. In inverse RL, such Bayesian approaches have been considered as well (see Section 4.3 in Arora et al. (2021)).

Instead of assuming that the reward functions are parameterized, one can use the reward functions directly as a parameter class and use a prior distribution over them. This could, for example, be a Gaussian process as initially considered by Kupcsik et al. (2018) for pairwise trajectory comparisons and adapted in later works (Bıyık et al. 2020a; Cosner et al. 2022). Here, again, it is worth mentioning that such considerations have also been made in inverse RL before (see Section 4.3 in Arora et al. (2021)).

5.2.3 Direct Methods

Note that the two-step approach of utility learning and policy optimization is not the only viable path to learning a policy from human feedback. While the previous two subsections have discussed the case in which we learn a reward function from observed preferences by assuming a human feedback model, an emerging branch of the literature is concerned with circumventing the reward-learning step and using preferences directly to address the actual RL problem. Concrete approaches are DPO (Rafailov et al. 2023), SLiC-HF (Zhao et al. 2023), OPPO (Kang et al. 2023), or DPPO (An et al. 2023). Azar et al. (2023) introduce an objective called Ψ-preference optimization (ΨPO) that unifies the objective functions in DPO and RLHF. More specifically, for a specific instantiation of Ψ, the objective in ΨPO recovers the latter two.

It is worth noting that approaches for directly learning the policy from preferences have been considered in the past as well (Wilson et al. 2012; Fürnkranz et al. 2012; Wirth et al. 2013a; Wirth et al. 2013b; Busa-Fekete et al. 2014). In Sections 3.2.1 and 3.2.2 in the survey by Wirth et al. (2017), these methods are explained in more detail.

5.2.4 Partial Identifiability

A crucial question when it comes to learning the reward function is whether the reward function can be identified at all. If two reward functions induce exactly the same human feedback model, the reward function is called partially identifiable or ambiguous. Skalse et al. (2023c) study this topic for the Boltzmann distribution as the underlying human feedback model when demonstrations (inverse RL) or pairwise trajectory preferences are given as feedback. For demonstrations, this question has also been examined in other works (Ng et al. 2000; Dvijotham et al. 2010; Kim et al. 2021; Cao et al. 2021a).

5.2.5 Interpretability

The field of explainable artificial intelligence (XAI) has emerged in recent years to improve the transparency and explainability of models or even to enable them in the first place. Roughly speaking, the aim is to resort to more interpretable methods or provide explanations for both experts and non-experts, shedding light on why a certain input in a (black box) model leads to a certain result. Explanations can take different forms, as can the ways to ensure the transparency of models, and for a detailed overview, we refer to Barredo Arrieta et al. (2020). It is worth noting that the field has grown so extensively over the years that even dedicated overviews for the field of interpretable and explainable RL are by now available (Puiutta et al. 2020; Glanois et al. 2022; Qing et al. 2023; Milani et al. 2023).

For the branch of RLHF, the existing works are quite sparse and mostly limited to using tree models as transparent and explainable models for learning the reward function (Bewley et al. 2022b; Bewley et al. 2022a; Kalra et al. 2022; Bewley et al. 2023; Kalra et al. 2023). Another way to realize explainability within RLHF suggested by Zhang et al. (2023a) is to learn simultaneously the reward function and the importance of states using a weight network. Assuming that for (long) trajectories, only a few states are important for the preference outcome, their framework can be used to select samples for explainability purposes. Moreover, a perturbation analysis is suggested to evaluate explanations in a quantitative manner using the learned state importance weights.

5.2.6 Online Improvements

Christiano et al. (2017) demonstrate that it is important to improve the reward model online, which is also confirmed by Gao et al. (2023) as otherwise, the issue of overoptimization of an imperfect reward model may occur. Abramson et al. (2022) give an example of this: They attempt to fine-tune an agent initialized with behavioral cloning with an engineered reward function and find that it fails to generalize and actually worsens the performance. They also compare RLHF with a reward model trained offline with iterative improvement and find that this leads to better performance, even sometimes exceeding human performance. However, this is not generally the case, as McKinney et al. (2022) show that reward models trained online together with a policy may not be effective when a completely new policy is trained.

5.2.7 Combining Different Feedback Types

An additional extension that is of interest is how the learning process can incorporate different sources of feedback. So far, three options have been considered in the literature: First, the combination of demonstrations and preferences in a model-free approach (Ibarz et al. 2018) or a passive-active preference-based reward learning approach (Palan et al. 2019; Bıyık et al. 2022a). Second, the combination of pairwise preferences and ordinal labels for actions in order to ensure safety-aware learning (Cosner et al. 2022). And third, the combination of pairwise preferences, demonstrations, and ratings for reducing the cognitive load for the human labeler (Koppol et al. 2020).

5.3 Evaluating Learned Reward Functions

A central question when it comes to learning the reward function is how to evaluate the learned reward function and how reward functions can be compared with each other. For this purpose, a couple of different approaches are available.

Rollout Method In inverse RL, a common method for evaluation is the rollout method (Ng et al. 1999; Fu et al. 2018a; Ibarz et al. 2018; Brown et al. 2019). In this approach, one first learns an optimal policy for the learned reward function and then estimates the value of this policy for online trajectory rollouts using the known ground-truth reward function. This approach can be transferred to RLHF as well. In many cases, however, the true reward function is unknown, and in safety-critical areas such as medicine or autonomous driving, such online rollouts cannot be executed.

Off-policy Evaluations As a possible alternative, so-called off-policy evaluations, which estimate the value of the optimal policy on the basis of an available data set, have been established. For coping with biases or large variances due to policy mismatch, approaches using importance sampling (Precup et al. 2000), regression- or classification-based methods (Paduraru 2013; Le et al. 2019; Irpan et al. 2019), or combinations of these (Jiang et al. 2016) have been proposed. The problem with these approaches, however, is that the traces of the explicit sources of error through policy learning or reward learning are blurred.

Distance Functions Yet another alternative, which has been advanced in the seminal paper by Gleave et al. (2022a), is using a suitable distance function for reward functions. Suitable here means that two reward functions, which differ only by certain transformations such as potential shaping (Ng et al. 2000) or positive scaling, should have zero distance if these transformations do not change the policy ranking with regard to the expected return. For this purpose, Gleave et al. (2022a) present a pseudometric, called Equivalent-Policy Invariant Comparison (EPIC) distance, that is determined in three steps: First, mapping two reward functions to a so-called *canonicalization form* that is invariant to transformations of the latter kind. Second, normalizing these canonicalization forms by means of a specific weighted L_2-norm whose weights are determined by a distribution over the transitions. Finally, the EPIC distance is the weighted L_2-norm distance of the normalized canonicalization forms.

Even if some attractive properties, above all a Lipschitz continuity in terms of the EPIC distance of two reward functions for the difference of the value functions of the induced optimal policies is shown, this distance has its shortcomings. One of these is that the canonicalization form used by EPIC distance does not encode sufficient knowledge about the underlying transition dynamics, which might lead to unreliable distances when evaluating reward functions on physically non-realizable transitions. To this end, Wulfe et al. (2022) propose the Dynamics-Aware Reward Distance (DARD), which uses a slightly different form of canonicalization but restricts the evaluation of the reward functions to transitions that are approximately physically feasible.

Recently, EPIC-like distances (Jenner et al. 2022) and STAndardised Reward Comparison (STARC) metrics (Skalse et al. 2023b), which are entire classes of pseudometrics on the space of all reward functions were proposed that generalize the three-step approach underlying the EPIC distance (and DARD) by parameterizing each of the steps. Specifically, the canonicalization function in the first step, the normalization in the second, and the metric in the third step are kept variable. If these three functional parameters fulfill certain requirements, then the resulting distance has some appealing properties, e.g., being a pseudometric that is zero if and only if the two reward functions induce the same ordering of policies or imply upper and lower bounds on value function differences.

Visual and Human Inspection For an evaluation by visual inspection, Jenner et al. (2021) propose a method for preprocessing reward functions by transforming them into simpler but equivalent reward functions for better interpretability. Related to this and the rollout method, the quality of the reward function learned can also be evaluated by a human (or expert) by examining the behavior of the agent on the target task.

5.4 Reward Model Inputs

Besides the feedback type, another factor is the modality of the reward model input data. This usually consists of the agent's observations and actions. Observations can range from true state to high dimensional inputs (e.g., images), while actions can range from discrete finite actions to continuous actions.

For instance, many typical RL benchmarks are in the continuous control domain (e.g., MuJoCo simulated robotics tasks) with true state representations and simple discrete actions. In such problems, Christiano et al. (2017) train reward models from these inputs.

When no compact state representation is available, raw images are often used in control tasks, which makes the learning of rewards more challenging since the setting becomes partially observable and the reward function is generally not Markov with respect to the observations. In such cases, the conventional trick of approximating a true state with a sequence of frames is often employed. This approach is used, for instance, by Christiano et al. (2017) to train reward models on the Atari benchmark suite.

More recently, many applications of RLHF are in the natural language processing (NLP) domain. In these settings, the policy takes natural language as both input and output while the reward model takes it as input (e.g., Ouyang et al. (2022)). Naturally, more complex scenarios (e.g., with both language and vision inputs (Abramson et al. 2022)) have also been studied.

5.5 Increasing Feedback Efficiency

Maximizing feedback efficiency is vital in RLHF due to the high cost of human feedback. This section delves into methods that enhance learning from limited human feedback. We discuss methods that leverage prior offline data, methods that use (partially unlabelled) data more efficiently, and methods that aim to gather more informative data.

5.5.1 Using Prior Data

There are often large amounts of prior data available at little or no additional cost. While this data generally was generated for other tasks, many basic human preferences are the same for various tasks and can often even be extracted from completely unsupervised data such as text corpora. By leveraging this prior data, we can greatly reduce the amount of feedback necessary to learn the current task's objective. We explore various methods, including foundation models, meta- and transfer learning, pertaining, and reward model initialization.

Meta- and Transfer Learning Meta- and transfer learning techniques in reward model training exploit the commonalities across different objectives, facilitating quick adaptation to new tasks. Ren et al. (2022) develop a broadly applicable meta-reward model, pre-trained on a diverse set of tasks to capture a wide range of preference patterns, enabling efficient adaptation to new tasks with fewer examples. Xie et al. (2018) use a similar meta-learning approach to build a goal classifier across multiple visuomotor tasks. Closely related to these meta-learning approaches, Hejna et al. (2023a) integrate few-shot learning principles, optimizing their approach for scenarios where only a few examples are available for adapting to new tasks. In the domain of transfer learning, Liu et al. (2023a) explore zero-shot transfer of preferences, a method that enables adapting preferences without additional data from the new task. In a different vein, but closely related to meta- and transfer learning, Mendez et al. (2018) tackle the lifelong inverse RL problem, focusing on inferring reward functions for multiple tasks over time, which involves knowledge transfer between tasks. Collectively, these studies underscore the potential of meta- and transfer learning in enhancing the efficiency and applicability of reward models in RLHF.

Leveraging Foundation Models Foundation models, i.e., models trained on large amounts of often unlabeled data, can acquire significant knowledge about basic human preferences. A language model trained to predict the next token in a text corpus, for example, may learn to complete the sentence 'Frank was mad that his vacuum robot broke the vase,' thereby learning that humans prefer non-destructive behavior. These learned preferences can then be leveraged in RLHF approaches. Du et al. (2023) exemplify such an approach by training a success detector using a pre-trained vision-language model (Flamingo). Their approach utilizes a data set of trajectories with binary success labels, employing a non-interactive training method.

Reward Model Initialization It is often beneficial to initialize the reward model with parameters from a model trained on a related task. This strategy is particularly common in language model fine-tuning, where self-supervised pretraining is a common practice. In such scenarios, it becomes logical to use these pre-trained models for initializing not just the policy but also the reward model. This methodology is adopted by Askell et al. (2021) and Ouyang et al. (2022). Specifically, Ouyang et al. (2022) use a pretrained language model for the reward model, opting for a smaller model relative to the policy to mitigate unstable learning. Notably, while they apply supervised fine-tuning to the policy before the RLHF phase, the reward model is initialized directly from the language model without any preliminary fine-tuning. This approach's applicability extends beyond language models to other areas. A notable example is Abramson et al. (2022), who, in the control domain, begin by training a policy through contrastive self-supervised learning and behavioral cloning. They then add an MLP head to the policy for the prediction of cumulative rewards.

Preference Model Pretraining Preference-model pretraining (Askell et al. 2021; Bai et al. 2022a) leverages prior offline data to pretrain the preference model before training it on policy samples. Askell et al. (2021) note that in the case of language models, noisy preference data can be readily obtained from sources such as rated Reddit comments, preferred Stack Overflow answers, and reverted Wikipedia edits. They leverage this as a pretraining step to increase data efficiency. This is in addition to regular language model pretraining, as discussed in the previous paragraph. A similar approach could be applied to control in case prior data and some means of inferring preferences, such as human corrections, are available. Even if no inferred preferences are available, Verma et al. (2023c) show that it can be beneficial to pre-train the preference model to predict close to constant reward on an initial set of trajectories. This avoids excessive fitting of the policy to random initialization differences in the reward function.

Supervised Representation Learning A compact representation that captures all relevant information for human preferences while minimizing noise can greatly enhance preference learning efficiency. It may also generalize better than a representation learned end-to-end as part of the preference learning task, which may contain spurious correlations. Bobu et al. (2022) address this by proposing the learning of features through explicit human feedback using feature traces. Feature traces (see Section 3.2.9) involve human labelers explicitly teaching relevant features one by one by demonstrating behavior in which the feature monotonically increases or decreases. This method directly aligns the learned representation with human-identified features, enhancing preference learning efficiency but requiring detailed human input. However, feature traces require labelers to be able to identify and articulate relevant features, which can be challenging. Bobu et al. (2023) offer an alternative approach with their Similarity-based Implicit Representation Learning (SIRL) method. SIRL learns representations from similarity queries (see Section 3.2.10), where human labelers provide feedback on whether behaviors are similar or different concerning the features that matter to them. This method captures a broader range of human notions of similarity without needing explicit feature knowledge, thus reducing the cognitive load on human labelers. In summary, while both approaches emphasize human feedback's centrality in representation learning, they differ in their methods of gathering this feedback. The feature traces used by Bobu et al. (2022) require specific feature knowledge, whereas SIRL used by Bobu et al. (2023) utilizes more intuitive similarity assessments, potentially offering a more user-friendly way to capture human preferences.

These diverse methods of utilizing prior data demonstrate the potential for enhancing data efficiency in RLHF, enabling more effective learning from limited human feedback.

5.5.2 Using Data More Efficiently

Beyond the application of prior data, several techniques can enhance the efficiency of data utilization in training processes. This section will discuss a range of such methods, including self-supervised and semi-supervised training, as well as the integration of inductive biases and data augmentation strategies. These approaches are designed to make the most of the available human interactions and improve the final performance of RLHF models.

Self-Supervised Auxiliary Tasks Self-supervised training enhances data efficiency in reward model training by using unannotated data to capture information about the task. This technique extends beyond the scope of pretraining methods, as discussed in the prior section, to incorporating concurrent auxiliary tasks to maximize the utility of available data. A prevalent technique, as applied by Abramson et al. (2022), Brown et al. (2020), and Metcalf et al. (2022), involves adding self-supervised losses to enhance representation learning for rewards. Abramson et al. (2022) implement a contrastive task where the reward network differentiates between observations that are consistent between multiple modalities and those that are not, blending this with preference learning loss and behavioral cloning. Brown et al. (2020) add multiple auxiliary tasks such as inverse and forward dynamics modeling, temporal distance prediction, and variational autoencoder training. Similarly, Metcalf et al. (2022) use the self-predictive representations technique (Schwarzer et al. 2022) to learn state representations that encode environmental dynamics, enabling a linear model to anticipate successor states, thereby forming an efficient basis for preference learning and significantly boosting sample efficiency. However, auxiliary losses for better representation learning are not the only approach to leverage self-supervised training. An alternate approach by Verma et al. (2022) involves identifying important states using attention weights from a world model transformer and state importance estimates based on a preference predicting transformer. These estimates can aid credit assignment for observed preferences, further optimizing the training process.

Semi-Supervised Training Semi-supervised training, blending labeled and unlabeled data, can leverage the unlabeled data to glean information about the task and the environment. This is most commonly done by generating pseudo-labels for the unlabeled data, either by leveraging model predictions or by making assumptions. The first approach is utilized by Cao et al. (2021b) and Zhan et al. (2021), which use generative models and GAN-based methods to mimic human preferences. Similarly, Park et al. (2022) expand their data set with high-confidence unlabeled samples based on the preference predictor's evaluations. The second strategy, making assumptions to augment data, is showcased by Zhou et al. (2020). They generate preference data by assuming that (i) human-written examples are better than model-written examples, (ii) human-written and model-written examples are indistinguishable amongst themselves, and (iii) generations of later model iterations are better than those of earlier ones.

Data Augmentation Data augmentation focuses on creating additional examples from existing labeled data. Temporal augmentation is particularly effective in RLHF, involving trajectory data. This is exemplified by Brown et al. (2019) and Park et al. (2022) who base their augmentation on the premise that preferences for complete trajectories can be extrapolated to cropped segments, allowing the generation of multiple derivative pairs from a single labeled trajectory pair. Park et al. (2022) additionally explore state modifications, such as random re-scaling and Gaussian noise addition, finding temporal cropping to be the most effective, with noise sometimes negatively impacting performance. In a similar vein, Verma et al. (2023b) focus on augmenting trajectories by concentrating on changing elements in observations and perturbing the other parts, based on the premise that movement indicates importance in image-based observations. Complementing these methods, Abramson et al. (2022) employ augmentation by randomly altering instructions and language responses, thus creating artificial examples of non-preferred behavior. These diverse data augmentation methods collectively enhance the training data set, contributing to the increased robustness and efficacy of RLHF models.

5.5.3 Gathering Better Data

In addition to leveraging unlabeled data and using labels more efficiently, sample efficiency can be further increased by collecting more informative samples in the first place. This can either be achieved by selecting more informative samples from the experience buffer or by generating more informative experiences. While selecting informative samples from the experience buffer is addressed under active learning (see Section 4.1.1), this section focuses on generating more informative experiences.

While we are not aware of many works in this area, one possible approach involves steering the agent's exploration towards regions of the state space where human feedback would be most beneficial. Liang et al. (2022) implement this by employing intrinsic motivation, driven by the estimated uncertainty of the reward model, to guide the agent's exploration. This highlights the potential of not just using data more efficiently but also generating data in a more targeted manner.

6 Theory

The field of RLHF has recently made some progress in terms of theoretical results, which we will discuss in this section. First, we consider the contributions where the goal is to learn a provable (near) optimal policy both in an online and offline fashion or even in a way that falls in between. Then, we discuss and highlight recent contributions related to different theoretical aspects of RLHF, such as its relation to the standard reward-based RL. Tables 5 and 6 provide a concise overview of the results for the online or offline policy learning setting. Here, $\mathcal{N}_{\mathcal{F}}(\epsilon, d)$ denotes the ϵ-covering number of a set $\mathcal{F}$ under some metric d[3]. It is worth mentioning that all works have two standard assumptions, namely that the reward function is bounded and that the ground-truth reward, human feedback model, or transition dynamic is an element of the considered model space, respectively.

[3] The ϵ-covering number is the minimum integer N such that there exists a subset $\mathcal{F}' \subset \mathcal{F}$ with $|\mathcal{F}'| = N$, and for any $f \in \mathcal{F}$, there exists some $f' \in \mathcal{F}'$ satisfying $d(f, f') \leq \epsilon$.

6.1 Policy Learning

In the literature focusing on theoretical results, there is a distinction (similar to the distinction made in standard RL) between an offline and online setting. In the former, learning is based on a given fixed data set, usually previously collected through an interaction with the environment. In contrast, in the online environment, one interacts directly with the environment to learn from real-time feedback and continuously updates one's strategies based on the feedback received, allowing the agent to learn and adapt as it engages with the environment.

Online Learning The first work dealing with the question of theoretical guarantees for learning an optimal policy from trajectory comparison feedback (see Section 3.2) in an online manner is by Novoseller et al. (2020). It laid the foundation for a paradigm subsequently embraced by many subsequent research endeavors: Adapting learning algorithms from the dueling or preference-based bandit literature (Yue et al. 2009; Sui et al. 2018; Bengs et al. 2021) to the underlying situation with additional states. The preference-based bandit problem can be viewed as a preference-based RL problem with one state, so state transition dynamics must be considered accordingly for a fruitful adaptation. It is worth mentioning that Jain et al. (2015) used a quite similar idea before for feedback in the form of corrections (see Section 3.2) by resorting to the coactive learning setting (Shivaswamy et al. 2012). Assuming the existence of a ground-truth context-trajectory scoring function and that the user's feedback is informative, the Preference Perceptron algorithm by Shivaswamy et al. (2012) is used and analyzed in terms of its cumulative regret.

Novoseller et al. (2020) suggest the Dueling Posterior Sampling (DPS), which is an adaptation of the self-sparring algorithm (Sui et al. 2017). It takes a Bayesian perspective on the problem and defines a Dirichlet prior on the dynamics and a Gaussian prior on the rewards that are subsequently updated, while the trajectories to be compared by the human labeler are chosen based on their (posterior) probability of being optimal[4]. Assuming a linear link function (see Section 5.1.4) as well as a tabular MDP, it is shown that DPS is (i) consistent, i.e., converges in distribution to the optimal policy, and (ii) achieves an asymptotic expected regret bound (see Table 5).

Xu et al. (2020) combine dynamic programming and policy search with a black-box preference-based bandit algorithm for each state to design routines that return an almost optimal (a.k.a. ε-optimal) policy with high probability[5] The first routine, called Preference-based Policy Search (PPS), requires access to a simulator, while the second routine, called Preference-based Exploration and Policy Search (PEPS), gets rid of this requirement by exploring the state space by means of an auxiliary synthetic reward function. By assuming that the probability of one policy dominating another policy is bounded uniformly over all states from below by a multiplicative of their value function, they show generic upper bounds for both routines on the number of pairwise trajectory comparisons (see Table 5). If these dominance probabilities have even more structural properties, such as fulfilling stochastic transitivity or stochastic triangle inequality (see Haddenhorst et al. (2020) and Bengs et al. (2021)), then these upper bounds can be further refined.

A follow-up work by Saha et al. (2023) assumes a feature embedding of trajectories that gives rise to a feature embedding of policies and adapts the MaxInP algorithm (Saha 2021) for contextual dueling bandits by essentially viewing the policy embeddings as the contexts. More precisely, assuming a logistic link function (see Section 5.1.4), confidence sets for the expected scores of the policies are constructed based on the maximum likelihood estimate (MLE), and the two policies with the highest uncertainty in terms of maximal variance are used to sample a trajectory, respectively, to be compared. In this way, the logistic preference-based reinforcement learning (LPbRL) is derived and also extended to the case of unknown dynamics by taking the uncertainty regarding the dynamics into account when constructing the confidence sets. For both cases, i.e., known or unknown dynamics, upper bounds on the regret of LPbRL are shown (see Table 5).

In contrast to previous work that all considers tabular MDPs, Chen et al. (2022) consider the case of a general unknown human feedback model and unknown dynamics each from function classes with a finite Eluder dimension[6] (Russo et al. 2013). They propose and analyze the Preference-based Optimistic Planning (PbOP) algorithm, which essentially follows a similar design as LPbRL but uses least-square estimates for the human feedback model and transitions dynamics along with confidence sets based on them. Moreover, they derive lower bounds for the regret of any learning algorithm by reducing the once-per-episode-feedback RL problem (Chatterji et al. 2021) to the PbRL problem. Finally, they extend their analysis to the case of K-wise comparisons, where one obtains all $\binom{K}{2}$ pairwise comparisons for K

[4] The latter probability is assessed by posterior sampling; a commonly used technique in the bandit literature used by so-called Thompson Sampling strategies, see Lattimore et al. (2020) for more details.

[5] This is a so-called PAC learning setting (Valiant 1984) in which the goal of finding the/an optimal object is relaxed to finding a "good enough" object, usually specified by some distance measure on the object domain.

[6] Roughly speaking, the Eluder dimension of a function class refers to the number of worst-case errors one must make to identify an unknown function from that class.

Table 5: An overview of approaches, their assumptions, goals, and properties for online policy learning. $\widetilde{\mathcal{O}}$ is used to hide $\log$-factors. T is the number of iterations of the respective algorithm.

Algorithm (Reference)	**Algorithmic approach**	**Assumptions**	**Target(s) and goal(s) of learner**	**Theoretical guarantee(s)**
Dueling Posterior Sampling (DPS) (Novoseller et al. 2020)	Leveraging Posterior Sampling from dueling bandits	Linear link function, tabular MDP	Bayes regret minimization w.r.t. optimal policy based on trajectory comparison feedback	Asymptotic regret rate: $\mathcal{O}\left(\lvert\mathcal{S}\rvert\sqrt{\lvert\mathcal{A}\rvert T\log(\lvert\mathcal{A}\rvert)}\right)$
Logistic Preference Reinforcement Learning (LPbRL) (Saha et al. 2023)	Leveraging MaxInP from contextual dueling bandits	Logistic link function, tabular MDP, linear rewards & d-dimensional feature embedding of trajectories	Expected regret minimization w.r.t. optimal policy based on trajectory comparison feedback	Transition dynamics: 1. Known $\widetilde{\mathcal{O}}\left(\lvert\mathcal{S}\rvert Hd\sqrt{T\log(T)}\right)$ 2. Unknown $\widetilde{\mathcal{O}}((\sqrt{d}+H^2+\lvert\mathcal{S}\rvert)\sqrt{dT}$ $+\sqrt{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert TH})$
Preference-based Optimistic Planning (PbOP) (Chen et al. 2022)	Leveraging MaxInP; general function approximation	General human feedback model class $\mathcal{F}_{\mathbb{T}}$ and general transition dynamic class $\mathcal{F}_{\mathbb{P}}$ with finite l_2-norm ρ-Eluder dimension $d_{\mathbb{T}}^{(2)}(\rho)$ and $d_{\mathbb{P}}^{(2)}(\rho)$, respectively	High probability regret minimization w.r.t. optimal policy based on trajectory comparison feedback	$\widetilde{\mathcal{O}}\left(\sqrt{d_{\mathbb{P}}(\frac{1}{T})HT\log\left(\mathcal{N}_{\mathcal{F}_{\mathbb{P}}}\left(\frac{1}{T},d\right)\right)}\right)$ $+\widetilde{\mathcal{O}}\left(\sqrt{d_{\mathbb{T}}(\frac{1}{T})T\log\left(\mathcal{N}_{\mathcal{F}_{\mathbb{T}}}\left(\frac{1}{T},d\right)\right)}\right)$ d being the ℓ-infinity norm $\lvert\lvert\cdot\rvert\rvert_\infty$
Preference-based Policy Search (PPS) (Xu et al. 2020)	Dynamic programming, policy search, (ϵ,δ)-PAC black-box dueling bandit algorithm and simulator	Uniform dependence of policy preference probabilities on value function differences, tabular MDP, (ϵ,δ)-PAC dueling bandit algorithm with $\Psi(K,\varepsilon,\delta)\varepsilon^{-\alpha}$ sample complexity for K arms	(ϵ,δ)-PAC for optimal policy based on trajectory comparison feedback	Simulator step bound $\mathcal{O}\left(\frac{H^{\alpha+1}\lvert\mathcal{S}\rvert\Psi(\lvert\mathcal{A}\rvert,\varepsilon/H,\delta/\lvert\mathcal{S}\rvert)}{\varepsilon^{\alpha}}\right)$ Sample complexity bound $\mathcal{O}\left(\frac{H^{\alpha}\lvert\mathcal{S}\rvert\Psi(\lvert\mathcal{A}\rvert,\varepsilon/H,\delta/\lvert\mathcal{S}\rvert)}{\varepsilon^{\alpha}}\right)$
Preference-based Exploration & Policy Search (PEPS) (Xu et al. 2020)	Similar to PPS, instead of simulator using an auxiliary synthetic reward function	Same as PPS and stochastic triangle inequality of trajectory comparisons preferences	(ϵ,δ)-PAC for optimal policy based on trajectory comparison feedback	Step complexity bound $\widetilde{\mathcal{O}}\left(\frac{H^{\alpha+1}\lvert\mathcal{S}\rvert^2\Psi(\lvert\mathcal{A}\rvert,\varepsilon/H,\delta/\lvert\mathcal{S}\rvert)}{\varepsilon^{\alpha+1}}\right)$ Comparison complexity bound $\mathcal{O}\left(\frac{H^{\alpha}\lvert\mathcal{S}\rvert\Psi(\lvert\mathcal{A}\rvert,\varepsilon/H,\delta/\lvert\mathcal{S}\rvert)}{\varepsilon^{\alpha}}\right)$
UCBVI-Planning (Kong et al. 2022)	Optimistic least-squares value iteration, maximum information gain, value iteration based on pessimistic expected value function estimation	Binary rewards for state-action pairs based on human response model $f\in\mathcal{F}_H$ with bounded noise $\Delta>0$, compliant and tabular/linear MDP with dimension d	(ϵ,δ)-PAC for optimal policy based on binary state-action reward feedback	Tabular MDP: $\mathcal{O}\Bigg(\frac{H^4\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert\log\left(\frac{H\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\epsilon\delta}\right)}{\epsilon^2}$ $+\frac{H^3\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert\log\left(\frac{H\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\epsilon\delta}\right)}{\epsilon}\Bigg)$ Linear MDP: $\mathcal{O}\left(\frac{\lvert\mathcal{A}\rvert^2d^5d_{\mathcal{F}_H}H^4\log\left(\frac{H\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\epsilon\delta\Delta}\right)}{\epsilon^2}\right)$
Preference-based & Randomized Least-Squares Value Iteration (PR-LSVI) (Wu et al. 2023)	Least-squares value iteration with perturbed state-action-wise reward model	General differentiable link function Φ, linear MDP, linear rewards with d-dimensional feature embedding of trajectories	Expected regret minimization w.r.t. optimal policy and/or low trajectory comparison feedback complexity steered by $\epsilon\in[0,1]$	Expected regret bound: $\widetilde{\mathcal{O}}\Big(\epsilon Td^{1/2}+\sqrt{T}\cdot d^3H^{5/2}\gamma$ $+d^{17/2}H^{11/2}\gamma^3\Big)$ Comparison complexity bound: $\widetilde{\mathcal{O}}\left(d^4(\kappa+R_{\max})^2/\epsilon^2\right)$ $\kappa=\inf_{x\in[-R_{\max},R_{\max}]}\Phi'(x)$

many queried trajectories. In essence, the regret term coming from the human feedback model class improves by a factor of $\sqrt{K}$.

Wu et al. (2023) consider a similar learning scenario as Saha et al. (2023) but with the additional objective to keep the number of queries of trajectory comparisons low, which is a combination of two competing objectives also studied in the bandit literature (Degenne et al. 2019). For this purpose, they suggest the Preference-based and Randomized Least-Squares Value Iteration (PR-LSVI) algorithm, which combines least-squares value iteration with a perturbed state-action-based reward model with Gaussian noise for regret minimization; a similar idea to CoLSTIM suggested for contextual dueling bandits (Bengs et al. 2022). More specifically, in each time step, the policy maximizing the value function of the perturbed state-action-based reward model and the policy maximizing the later in the previous time steps are "played". By sampling trajectories of these two policies and computing their expected absolute reward difference (based on the perturbed state-action-based reward model) as a measure of uncertainty, preference feedback for these two trajectories is queried if the uncertainty exceeds a certain threshold. Moreover, they also suggest a posterior sampling counterpart of this algorithm, the Preference-based Thompson Sampling (PbTS) algorithm, and analyze it in terms of Bayesian quantities.

Offline Learning Zhu et al. (2023) study the performance of a greedy policy trained from a data set consisting of trajectory pairs along with the observed preference that is assumed to be generated by means of a Bradley-Terry model with linear rewards. For this purpose, different results with respect to the MLE of the Bradley-Terry model for different feedback scenarios are derived that are quite of independent interest. In particular, they show concentration inequalities of the MLE for trajectory-based comparison feedback and additionally its asymptotic normality for action-based comparison feedback that also holds for K-wise comparisons. Based on these, it is shown that the greedy policy using the MLE in the case of action-based feedback might fail while using a pessimistic MLE leads to minimax-rates with respect to the performance gap[7]. The latter is also shown to be true in the case of trajectory-based feedback. Technically, the pessimistic MLE is realized by taking the policy that has the largest pessimistic expected value function, i.e., the lowest realization of the value function within a hyperparameter-dependent confidence region around the MLE. Further results of independent interest are the inferred theoretical guarantees for maximum entropy inverse RL (Ziebart et al. 2008) and action-based inverse RL algorithms (Neu et al. 2009).

The simple model assumptions underlying (Zhu et al. 2023) were then replaced by more sophisticated assumptions in some subsequent work. The linear reward assumption has been replaced by more general reward function classes by Zhan et al. (2023a) and Li et al. (2023). In addition, Zhan et al. (2023a) also consider more general unknown human feedback models and construct the confidence regions for the pessimistic approach directly from the log-likelihood function. The resulting approach, called FREEHAND, is analyzed in terms of its performance gap, for which some problem-dependent coefficients, the per-step, per-trajectory, and transition concentrability coefficient, are introduced. On the basis of a lower bound, it is shown that the per-trajectory concentrability coefficient should naturally appear in the bound on the performance gap. Moreover, the concentrability coefficient is shown to be upper bounded by the constant appearing in the special case of linear rewards considered by Zhu et al. (2023). Finally, it is worth mentioning that both trajectory-based and action-based comparison feedback are considered.

Assuming a dynamic discrete choice model (Rust 1987) underlying the given data set of observed trajectories (without explicitly observed preferences), Li et al. (2023) suggest the Dynamic-Choice-Pessimistic-Policy-Optimization (DCPPO) algorithm. It first estimates the reward model using this assumption and then learns a policy in a (pessimistic) value iteration manner from the estimated reward model. In the case of a linear MDP and a known model class that entails both the value and the reward function of the dynamic discrete choice model, DCPPO is analyzed with respect to its performance gap. This is done for the case of a linear function model class as well as a subset of a reproducing kernel Hilbert space (RKHS) as the model class.

Focusing on the estimation of the weight parameter in the Bradley-Terry model for the action-based feedback under label differential privacy conditions (Dwork 2008), Chowdhury et al. (2023) analyze two estimation procedures, MLE and stochastic gradient descent (SGD), under similar assumptions as in Zhu et al. (2023). In both cases, the cost of ensuring label differential privacy is a multiplicative factor.

Reward collapse, a term introduced by Song et al. (2023), describes the issue when rank-based training methods for LLMs lead to the same reward distribution regardless of the prompts used in the final training steps. The authors show that this occurs because the rank-based approach does not adequately account for prompt-related information. To address this problem, the authors propose a family of utility functions as well as an optimization method that successfully creates prompt-dependent reward distributions, effectively mitigating the collapse of rewards during training.

[7] The expected difference between the optimal value function and the value function of the used policy.

Table 6: Overview of approaches, their assumptions, goals, and properties for offline policy learning with a data set of size n. $\widetilde{\mathcal{O}}$ is used to hide $\log$-factors. $R_{\max}$ is a bound on the reward.

Algorithm (Reference)	**Algorithmic approach**	**Assumptions**	**Target(s) and goal(s) of learner**	**Theoretical guarantee(s)**
Pessimistic MLE (Zhu et al. 2023)	Greedy policy for pessimistic expected value function estimation	Logistic link function, linear reward function for a state-pair feature embedding with some regularity assumptions on weights, known transition dynamics	High probability bound for the performance gap based on trajectory-based (and action-based) feedback	$\mathcal{O}\left(e^{2HR_{\max}}\sqrt{\frac{d+\log(1/\delta)}{n}}\right)$
oFfline Reinforce-mEnt lEarning with HumAN feeDback (FREEHAND) (Zhan et al. 2023a)	Greedy policy for pessimistic expected value function estimation	General differentiable link function Φ, general bounded reward function class $\mathcal{F}_r$ and general transition dynamic class $\mathcal{F}_{\mathbb{P}}$	High probability bound for the performance gap based on trajectory-based (and action-based) feedback	Transition dynamics: 1. Known $\mathcal{O}\left(\sqrt{\frac{C_r^2\kappa^2\log(\mathcal{N}_{\mathcal{F}_r}(1/N,\vert\cdot\vert)/\delta)}{n}}\right)$ 2. Unknown $\mathcal{O}\left(\sqrt{\frac{C_r^2\kappa^2\log(\mathcal{N}_{\mathcal{F}_r}(1/N,\vert\cdot\vert)/\delta)}{n}}\right) +$ $\mathcal{O}\left(R_{\max}\sqrt{\frac{C_P^2\kappa^2\log(\mathcal{N}_{\mathcal{F}_{\mathbb{P}}}(1/N,\vert\cdot\vert)/\delta)}{n}}\right)$ $\kappa = \inf_{x\in[-R_{\max},R_{\max}]}\Phi'(x)$ C_r, C_P reward and transition concentrability coefficient
Dynamic-Choice-Pessimistic-Policy-Optimization (DCPPO) (Li et al. 2023)	Value iteration based on pessimistic expected value function estimation	Dynamic discrete choice model, linear MDP, linear reward function for a state-pair feature embedding with some regularity assumptions on weights, known model class entailing the value and reward function of the dynamic discrete choice model	High probability bound for the performance gap based on action-based feedback	Linear model class: $\mathcal{O}\left(\vert\mathcal{A}\vert d^{3/2}H^2e^H\sqrt{\frac{\log(dHn/\delta)}{n}}\right)$ RKHS model class with different eigenvalue decay: $\mathcal{O}\left(\tilde{d}He^H\vert\mathcal{A}\vert\sqrt{\mu\log(nR_{\max}H/\delta)}\right)$ μ-finite spectrum, $\mathcal{O}\left(\tilde{d}He^H\vert\mathcal{A}\vert\sqrt{(\log(nR_{\max}H)/\delta)^{1+1/\mu}}\right)$ μ-exponential decay, $\mathcal{O}\left(\tilde{d}He^H\vert\mathcal{A}\vert(nR_{\max})^{\kappa^*}\sqrt{\log(nR_{\max}H/\delta)}\right)$ μ-polynomial decay, $\kappa^* = \frac{d+1}{2(\mu+d)} + \frac{1}{\mu(1-2\tau)-1}$, $\tilde{d}$ =population effective · sampling effective dimension
LCBVI-Tabular-Offline (Kong et al. 2022)	Maximum information gain for reward querying, value iteration based on pessimistic expected value function estimation for policy learning	Binary rewards for state-action pairs based on human response model with bounded noise, compliant and tabular MDP	High probability bound for the performance gap based on binary state-action reward feedback	Linear model class: $\mathcal{O}\Big(H\sqrt{\vert\mathcal{S}\vert\log(\vert\mathcal{S}\vert\vert\mathcal{A}\vert Hn/\delta)}$ $\cdot\,\mathbb{E}_{\pi^*}\left[\sum_{h=1}^{H}(N_h(s_h,a_h)+1)^{-1/2}\right]\Big)$ N_h are numbers of visit time

Blending Online and Offline Learning Kong et al. (2022) study the problem of optimal policy learning from critique feedback (see Section 3.2), i.e., binary rewards for state-action pairs, with as few queries to the human as possible. They assume an underlying ground-truth human feedback model that leads to a positive evaluation for a state-action pair if it exceeds a specific threshold evaluated at that pair. In addition, the learning process consists of two phases: First, exploring the environment in an unsupervised manner, and then querying user feedback in an active reward learning phase to learn the human feedback model. This learning process is again analyzed in two variants: Either the exploration phase was performed externally, and a data set consisting of trajectories is provided (offline), or this data set is actively collected itself (online). For both variants, an active learning algorithm is proposed that essentially selects query points (state-action pairs) that provide the most information gain given the points already designated to be queried. For the online variant, an exploration strategy based on optimistic least-squares value iteration (Jin et al. 2020) is also introduced for tabular or linear MDPs. In both variants, policy learning is carried out by a pessimistic value iteration with the empirical transitions and the estimated reward function, resulting in UCBVI-Planning (online) and LCBVI-Tabular-Offline (offline). Under the assumption of bounded noise (Massart et al. 2006) or low-noise assumption (Korba et al. 2017; Haddenhorst et al. 2021), bounds on the performance gap of both algorithms are derived.

The question of the ideal experimental design for RLHF is addressed by Zhan et al. (2023b), in particular, how to separate the process of data acquisition (e.g., trajectories to be evaluated) from the process of retrieving human feedback to avoid constantly involving humans in the training loop. Assuming linear rewards, the Bradley-Terry model and either a transition oracle (e.g., available for tabular or low-rank MDPs) or a linear MDP they suggest the expeRimental dEsiGn for queryIng huMan prEference (REGIME) algorithm that first samples exploratory trajectories indented to be as informative as possible for learning the reward via MLE and then applies a greedy policy based on the reward learned by the latter. They explicitly show that REGIME requires less human feedback to be queried in order to output an ϵ-optimal policy at the end than the approach by Saha et al. (2023).

6.2 Preference-Based vs. Reward-Based Learning

There have been some theoretical analyses regarding the question in how far, or if at all, preference-based feedback in the form of trajectory comparisons is more s more suitable compared to numerical feedback. Ji et al. (2023b) suggest a human rating model for this purpose in the numerical feedback case and analyze the LCB algorithm (Jin et al. 2021) in order to compare it with the pessimistic MLE (Zhu et al. 2023). It is shown that under specific assumptions, LCB has a constant performance gap, while the preference-based pessimistic MLE under similar assumptions has a similar bound as in Table 6.

Wang et al. (2023b) provide reduction-based algorithms that can directly utilize state-of-the-art results in reward-based RL for RLHF with utility-based and general state-action and trajectory-based comparison feedback. They show, in general, how theoretical results of the underlying standard RL algorithm can be translated to theoretical results for the resulting preference-based RL algorithm. For some special cases, such as MDPs with finite Eluder dimension and utility-based preference feedback, the theoretical guarantees are explicitly derived using state-of-the-art RL algorithms that are qualitatively similar to explicit preference-based RL algorithms.

7 Applications and Benchmarks

The field of RLHF has advanced significantly in the last few years, with increasing interest driven by prominent applications. First and foremost are applications to large language models, exemplified by ChatGPT (OpenAI 2022). This section starts by providing a sample of such applications, showcasing how this technology is being utilized in fields as varied as robotics, language processing, image generation, and more. We will also delve into libraries that provide foundational support for RLHF research, enabling researchers and practitioners to experiment with and refine a range of approaches. We then explore a spectrum of benchmarks that have been developed to standardize and simplify the evaluation of new approaches, offering insights into their performance in different settings. Finally, and closely related to those benchmarks, we will discuss common evaluation practices.

7.1 Applications

RLHF finds applications across various domains, showcasing its versatility in addressing complex and nuanced tasks. The most prominent application is ChatGPT (OpenAI 2022), which is an example of an application in the domain of language models. Beyond that, however, applications extend across diverse domains such as control tasks, generative models, and recommender systems. This section provides an overview of notable works applying RLHF in different areas.

Control and Interactive Environments There is a long history of using control environments as benchmark tasks for RL. In addition to the breadth of available environments, control applications are of particular interest because tasks are often hard to specify. Christiano et al. (2017) demonstrated the effectiveness RLHF in games as well as simulated continuous control tasks, matching the performance of RL agents trained on ground-truth rewards with a fraction of the feedback. Extending to robotics, Ding et al. (2023) trained a reward model for diverse tasks with a single robot, achieving human-like behavior. Kupcsik et al. (2018) applied RLHF for precise robot-to-human handovers. Similarly, Abramson et al. (2022) used RLHF in the Playhouse simulator, a platform for sensorimotor task training, and Milani et al. (2022) showcase an application in the context of the MineRL Basalt competition for Minecraft tasks.

Generative Models in Language and Imaging Generative models, i.e., models that generate new data instead of just predicting labels, can be framed as an RL setting in which a policy assembles the output through its actions. In the context of language models, this means that the language model is interpreted as a policy with tokens as actions. Using this reframing, we can use RLHF approaches to fine-tune generative models to produce preferred outputs. ChatGPT (OpenAI 2022) and GPT-4 (OpenAI 2023) are prime examples of language models fine-tuned using RLHF. These applications build on earlier work, such as by Ouyang et al. (2022), Ziegler et al. (2020) and Glaese et al. (2022). This method extends to text summarization (Gao et al. 2018; Gao et al. 2020; Stiennon et al. 2020), dialogue summarization (Chen et al. 2023), and question answering (Nakano et al. 2022). In image generation, Lee et al. (2023) and Xu et al. (2023) demonstrate the use of reward modeling for text-to-image tasks, while Pinto et al. (2023) and Kazemi et al. (2020) explore RLHF applications in broader computer vision tasks.

Recommender Systems In the context of recommender systems, Xue et al. (2023b) have shown the potential of RLHF in optimizing for long-term engagement. Although it is, in principle, possible to algorithmically evaluate policies in this domain, these rewards are sparse. To combat this, Xue et al. (2023b) use RLHF to distill sparse, global feedback into a dense reward model.

These diverse applications underscore RLHF's adaptability and its growing importance in various technological domains, paving the way for innovative solutions and enhanced human-computer interactions.

7.2 Supporting Libraries

Several libraries have emerged that aim to provide a toolset for implementing and experimenting with RLHF and reward learning algorithms, contributing to the ease and efficiency of research and development. One notable example is the `imitation` library (Gleave et al. 2022c). It encompasses a collection of imitation and reward learning algorithms, including those introduced in the seminal work by Christiano et al. (2017). Two other libraries, `APReL` (Bıyık et al. 2022b) and `POLAR` (Tucker et al. 2022), focus on the Bayesian setting. Bıyık et al. (2022b) provide a specialized framework for preference-based reward learning with a focus on Bayesian methods. Meanwhile, Tucker et al. (2022) introduce a framework designed for Bayesian reward learning from multiple modalities, including pairwise preferences, in MATLAB. Finally, the domain of language model fine-tuning, the `trlX` library (Castricato et al. 2023) offers a toolkit specifically designed for language model training. It specializes in the fine-tuning of transformer-based language models, treating the language model as the policy in an RLHF setup.

Due to the many interacting components and the human element in RLHF research, implementing new ideas and running experiments can be quite challenging. The discussed libraries reduce this challenge and make RLHF research more approachable to many researchers.

7.3 Benchmarks

Due to the difficulty of reproducible evaluations without a ground-truth objective and with humans in the loop, benchmarks play an important role in the advancement and evaluation of RLHF approaches. Several benchmarks have been proposed, each focusing on different applications and challenges.

One such benchmark is B-Pref (Lee et al. 2021a), which focuses on control tasks with synthetic feedback. B-Pref aims to provide simulated human feedback that captures some irrationalities, thereby coming closer to evaluation with real human feedback than other approaches. At the same time, by relying entirely on synthetic feedback, the results are reproducible and cost-effective to generate. In a similar vein, Freire et al. (2020) propose a set of environments designed to diagnose common problems in reward learning. These environments help in identifying and addressing the typical challenges that arise in RLHF scenarios.

A more application-driven benchmark with a complex environment is given by the MineRL BASALT competition (Shah et al. 2021; Milani et al. 2022). The competition proposes the challenge of solving tasks defined by natural language descriptions in Minecraft based on human feedback. Writing hand-engineered reward functions is very challenging in that setting, which makes it a good benchmark for methods based on human feedback. The competition is method-agnostic in principle, also considering non-RL approaches such as behavioral cloning. It is also agnostic for the feedback type, which may include demonstrations, comparisons, and others. The initial data set consists of demonstrations, however. The final evaluation is performed by humans through pairwise comparisons.

In the domain of language modeling, Truthful QA (Lin et al. 2022) serves as a benchmark that measures the truthfulness of models. Also, in the context of language models, Ramamurthy et al. (2023) introduce a set of pre-trained reward models, learned from human feedback, as benchmarks. These models serve as reference points for evaluating new RLHF techniques against established standards.

Together, these benchmarks provide a diverse and comprehensive suite of tests that drive the development and refinement of RLHF methods, ensuring they are robust, effective, and capable of handling a wide range of real-world scenarios.

7.4 Evaluation

Evaluating RLHF poses unique challenges, particularly in scenarios without clear ground-truth task specifications. Evaluations generally focus on either the learned policy or the reward model, each shedding light on different aspects of system performance.

Policy Evaluation Assessing learned behavior is crucial for the evaluation of an RLHF system. In domains with ground-truth rewards, these can be used for policy evaluation (Christiano et al. 2017). However, many RLHF applications lack this clarity. Ouyang et al. (2022), for instance, evaluate the quality of language model responses by having labelers rate the output quality on a test set of prompts, highlighting the significance of human judgment in assessing model outputs. Jain et al. (2015) use direct Likert-scale scores for evaluations, including self-assessments by trainers and cross-evaluations by others. Losey et al. (2022) extend this with a Likert-scale survey and free-form participant comments, comparing evaluations based on known true rewards with subjective experiences. Moreover, Abramson et al. (2022) employ a multi-stage evaluation scheme that includes scripted probe tasks, a standardized test suite evaluated by humans, and full interactive assessments, demonstrating the need for diverse and thorough evaluation methodologies in RLHF.

Reward Model Evaluation Direct reward model evaluation complements policy assessment. While reward model accuracy is a more direct measure of preference-learning success, the ultimate goal is inducing effective policies. A perfectly accurate reward model is often not necessary to induce a good policy, which is the actual goal of RLHF. Therefore, both evaluation methods are ideally used in combination. Jain et al. (2015) also use a ranking loss method for test sets of trajectories, compared against expert evaluations with known Likert-scores. This approach provides quantitative measures of the reward model's fidelity. In addition, Wilde et al. (2023) compare parameter-based and reward-based evaluation measures for learned reward functions, identifying strengths and weaknesses in both methods and contributing to a more nuanced understanding of reward model assessment in RLHF. These approaches provide a quantitative measure of the reward model's accuracy in reflecting human preferences and expert judgments. For a detailed discussion of reward model evaluation, also refer to Section 5.3.

Policy- and reward model evaluation both offer insights into the performance of an RLHF approach. Ideally both measures should be combined to enable quick iteration and give insights into both the preference learning performance as well as the quality of the learned behavior.

8 Discussion and Conclusion

In this survey, we have provided an overview of the current state of RLHF, highlighting its evolution from PbRL and examining its broad applications across various domains like control, natural language processing, and computer vision. While our survey captures the current state and many significant trends and advancements in RLHF, we acknowledge the rapid expansion of this field and the inevitable limitations in covering every extension and application in depth. We will discuss some of these extensions, open questions, and conclusions in this section.

We have specifically focused on RLHF methods where a reward function is learned online from human feedback. There have been some recent works that are outside of this scope and yet propose promising new methods to learn human-aligned objectives. One of them is learning reward functions offline, as recently investigated by Shin et al. (2023). An alternative approach is to learn objectives from a pre-trained AI system instead of human feedback. This has been termed RL from AI feedback (RLAIF) and leverages foundation models as a source of preference (Bai et al.

2022b). Alternatively, instead of a reward function, it is possible to train a policy directly (An et al. 2023; Busa-Fekete et al. 2014) or to learn a Q-function from preference feedback (Hejna et al. 2023b; Cheng et al. 2011).

Most work on RLHF implicitly assumes that tasks can be specified by maximization of expected accumulated scalar rewards. This assumption, called the *reward hypothesis* (Silver et al. 2021), is under active debate (Lambert 2021; Vamplew et al. 2022; Bowling et al. 2022; Skalse et al. 2022a) in the RL community. Recent approaches in RLHF are, for instance, considering more complex objective functions, such as multi-objective frameworks involving non-linear aggregation of expected accumulated vector rewards (Qian et al. 2023).

Many more extensions of RLHF are inspired by revisiting classic RL topics under the RLHF lens. This is exemplified by studies on exploration (Liang et al. 2022), reward feature learning (Katz et al. 2021), reward design (Ma et al. 2023), reward shaping (Xiao et al. 2020), multi-task RL (Ouyang et al. 2022; Abramson et al. 2022; Myers et al. 2023), hindsight experience replay (Zhang et al. 2023b), interpretability (Bewley et al. 2022b), safe RL (Dai et al. 2023; Cosner et al. 2022), and fair RL (Siddique et al. 2023). As discussed in Section 4.2, the intersection of RLHF with HCI also offers a fertile ground for future research, especially for refining feedback mechanisms. It is crucial to keep human psychology in mind when designing these systems and to learn from other related fields that already studied such issues extensively.

In addition to those extensions, current RLHF methods also have challenges and limitations to be aware of. Casper et al. (2023) offer a thorough analysis of these issues and limitations, highlighting the practical constraints of current approaches. Adding to that, from a theoretical perspective, a primary challenge lies in further relaxing underlying assumptions. This requires striking a delicate balance: On the one hand, ensuring the assumptions are not overly restrictive to encompass a broad range of practical use cases, and on the other, maintaining the feasibility of theoretical guarantees for computationally efficient algorithms. Key questions in this context are whether it is possible to design algorithms that do not need to actively maintain a policy space and eliminate sub-optimal policies nor rely on a computation oracle. Recent work such as Wang et al. (2023b) or Wu et al. (2023) give hope that this may be possible.

Although RLHF has significantly contributed to the advancements in LLMs and other areas of machine learning, it is a domain still in its infancy with many unanswered questions and inherent limitations. Despite and because of these challenges, it is ripe for further advancements in theory and practice, hopefully resulting in even more robust algorithms making more efficient use of human feedback. It remains intriguing to what extent RLHF will continue to shape the fields of natural language processing, RL, robotics, AI alignment, and beyond in the future.

Acknowledgements This publication was supported by 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 work has also been supported in part by the program of National Natural Science Foundation of China (No. 62176154) and a collaborative project funded by NetEase.

References

Abdelkareem, Youssef, Shady Shehata, and Fakhri Karray (2022). “Advances in Preference-based Reinforcement Learning: A Review”. In: *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2527–2532. DOI: 10.1109/SMC53654.2022.9945333.

Abramson, Josh, Arun Ahuja, Federico Carnevale, Petko Georgiev, Alex Goldin, Alden Hung, Jessica Landon, Jirka Lhotka, Timothy Lillicrap, Alistair Muldal, George Powell, Adam Santoro, Guy Scully, Sanjana Srivastava, Tamara von Glehn, Greg Wayne, Nathaniel Wong, Chen Yan, and Rui Zhu (2022). *Improving Multimodal Interactive Agents with Reinforcement Learning from Human Feedback*. arXiv: 2211.11602. preprint.

Akrour, Riad, Marc Schoenauer, and Michele Sebag (2011). “Preference-Based Policy Learning”. In: *Proceedings of Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. Springer, pp. 12–27. DOI: 10.1007/978-3-642-23780-5_11.

Alvo, Mayer and Philip L.H. Yu (2014). *Statistical Methods for Ranking Data*. Springer. DOI: 10.1007/978-1-4939-1471-5.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané (2016). *Concrete Problems in AI Safety*. arXiv: 1606.06565. preprint.

An, Gaon, Junhyeok Lee, Xingdong Zuo, Norio Kosaka, Kyung-Min Kim, and Hyun Oh Song (2023). “Direct Preference-based Policy Optimization without Reward Modeling”. In: Conference on Neural Information Processing Systems (NeurIPS). URL: https://openreview.net/forum?id=FkAwlqBuyO.

Arora, Saurabh and Prashant Doshi (2021). “A Survey of Inverse Reinforcement Learning: Challenges, Methods and Progress”. In: *Artificial Intelligence* 297, p. 103500. DOI: 10.1016/j.artint.2021.103500.

Arzate Cruz, Christian and Takeo Igarashi (2020). “A Survey on Interactive Reinforcement Learning: Design Principles and Open Challenges”. In: *Proceedings of the ACM Designing Interactive Systems Conference (DIS)*. Association for Computing Machinery, pp. 1195–1209. DOI: 10.1145/3357236.3395525.

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

Atas, Müslüm, Alexander Felfernig, Seda Polat-Erdeniz, Andrei Popescu, Thi Ngoc Trang Tran, and Mathias Uta (2021). “Towards Psychology-Aware Preference Construction in Recommender Systems: Overview and Research Issues”. In: *Journal of Intelligent Information Systems* 57.3, pp. 467–489. DOI: 10.1007/s10844-021-00674-5.

Azar, Mohammad Gheshlaghi, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos (2023). *A General Theoretical Paradigm to Understand Learning from Human Preferences*. arXiv: 2310.12036. preprint.

Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan (2022a). *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*. arXiv: 2204.05862. preprint.

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan (2022b). *Constitutional AI: Harmlessness from AI Feedback*. arXiv: 2212.08073. preprint.

Baker, Bowen, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch (2020). “Emergent Tool Use From Multi-Agent Autocurricula”. In: *Proceedings of International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=SkxpxJBKwS.

Bakker, Michiel, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, and Christopher Summerfield (2022). “Fine-Tuning Language Models to Find Agreement among Humans with Diverse Preferences”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 35, pp. 38176–38189. URL: https://proceedings.neurips.cc/paper/2022/hash/f978c8f3b5f399cae464e85f72e28503-Abstract-Conference.html.

Barnett, Peter, Rachel Freedman, Justin Svegliato, and Stuart Russell (2023). “Active Reward Learning from Multiple Teachers”. In: AAAI 2023 Workshop on Artificial Intelligence Safety.

Barredo Arrieta, Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera (2020). “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI”. In: *Information Fusion* 58, pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

Basu, Chandrayee, Erdem Bıyık, Zhixun He, Mukesh Singhal, and Dorsa Sadigh (2019). “Active Learning of Reward Dynamics from Hierarchical Queries”. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 120–127. DOI: 10.1109/IROS40897.2019.8968522.

Basu, Chandrayee, Mukesh Singhal, and Anca D. Dragan (2018). “Learning from Richer Human Guidance: Augmenting Comparison-Based Learning with Feature Queries”. In: *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Association for Computing Machinery, pp. 132–140. DOI: 10.1145/3171221.3171284.

Basu, Chandrayee, Qian Yang, David Hungerman, Mukesh Singhal, and Anca D. Dragan (2017). “Do You Want Your Autonomous Car To Drive Like You?” In: *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Association for Computing Machinery, pp. 417–425. DOI: 10.1145/2909824.3020250.

Bengs, Viktor, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier (2021). “Preference-Based Online Learning with Dueling Bandits: A Survey”. In: *Journal of Machine Learning Research* 22.7, pp. 1–108. ISSN: 1533-7928. URL: http://jmlr.org/papers/v22/18-546.html.

Bengs, Viktor and Eyke Hüllermeier (2020). “Preselection Bandits”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 778–787. URL: https://proceedings.mlr.press/v119/bengs20a.html.

Bengs, Viktor, Aadirupa Saha, and Eyke Hüllermeier (2022). "Stochastic Contextual Dueling Bandits under Linear Stochastic Transitivity Models". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 1764–1786. URL: https://proceedings.mlr.press/v162/bengs22a.html.

Bewley, Tom, Jonathan Lawry, and Arthur Richards (2023). *Learning Interpretable Models of Aircraft Handling Behaviour by Reinforcement Learning from Human Feedback*. arXiv: 2305.16924. preprint.

Bewley, Tom, Jonathan Lawry, Arthur Richards, Rachel Craddock, and Ian Henderson (2022a). *Reward Learning with Trees: Methods and Evaluation*. arXiv: 2210.01007. preprint.

Bewley, Tom and Freddy Lécué (2022b). "Interpretable Preference-based Reinforcement Learning with Tree-Structured Reward Functions". In: *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. IFAAMAS, pp. 118–126. URL: https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p118.pdf.

Bignold, Adam, Francisco Cruz, Matthew E. Taylor, Tim Brys, Richard Dazeley, Peter Vamplew, and Cameron Foale (2021). "A Conceptual Framework for Externally-Influenced Agents: An Assisted Reinforcement Learning Review". In: *Journal of Ambient Intelligence and Humanized Computing*, pp. 3621–3644. DOI: 10.1007/s12652-021-03489-y.

Bıyık, Erdem, Nicolas Huynh, Mykel Kochenderfer, and Dorsa Sadigh (2020a). "Active Preference-Based Gaussian Process Regression for Reward Learning". In: *Proceedings of Robotics: Science and Systems (RSS)*. Vol. 16. ISBN: 978-0-9923747-6-1. URL: http://www.roboticsproceedings.org/rss16/p041.html.

Bıyık, Erdem, Dylan P. Losey, Malayandi Palan, Nicholas C. Landolfi, Gleb Shevchuk, and Dorsa Sadigh (2022a). "Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences". In: *The International Journal of Robotics Research* 41.1, pp. 45–67. DOI: 10.1177/02783649211041652.

Bıyık, Erdem, Malayandi Palan, Nicholas C. Landolfi, Dylan P. Losey, and Dorsa Sadigh (2020b). "Asking Easy Questions: A User-Friendly Approach to Active Reward Learning". In: *Proceedings of the Conference on Robot Learnin (CoRL)*. PMLR, pp. 1177–1190. URL: https://proceedings.mlr.press/v100/b-iy-ik20a.html.

Bıyık, Erdem and Dorsa Sadigh (2018). "Batch Active Preference-Based Learning of Reward Functions". In: *Proceedings of the Conference on Robot Learning (CoRL)*. PMLR, pp. 519–528. URL: https://proceedings.mlr.press/v87/biyik18a.html.

Bıyık, Erdem, Aditi Talati, and Dorsa Sadigh (2022b). "APReL: A Library for Active Preference-based Reward Learning Algorithms". In: *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 613–617. DOI: 10.1109/HRI53351.2022.9889650.

Bobu, Andreea, Andrea Bajcsy, Jaime F. Fisac, Sampada Deglurkar, and Anca D. Dragan (2020a). "Quantifying Hypothesis Space Misspecification in Learning From Human–Robot Demonstrations and Physical Corrections". In: *IEEE Transactions on Robotics* 36.3, pp. 835–854. DOI: 10.1109/TRO.2020.2971415.

Bobu, Andreea, Yi Liu, Rohin Shah, Daniel S. Brown, and Anca D. Dragan (2023). "SIRL: Similarity-based Implicit Representation Learning". In: *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Association for Computing Machinery, pp. 565–574. DOI: 10.1145/3568162.3576989.

Bobu, Andreea, Dexter R. R. Scobee, Jaime F. Fisac, S. Shankar Sastry, and Anca D. Dragan (2020b). "LESS Is More: Rethinking Probabilistic Models of Human Behavior". In: *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Association for Computing Machinery, pp. 429–437. DOI: 10.1145/3319502.3374811.

Bobu, Andreea, Marius Wiggert, Claire Tomlin, and Anca D Dragan (2022). "Inducing Structure in Reward Learning by Learning Features". In: *The International Journal of Robotics Research* 41.5, pp. 497–518. DOI: 10.1177/02783649221078031.

Bowling, Michael, John D. Martin, David Abel, and Will Dabney (2022). *Settling the Reward Hypothesis*. arXiv: 2212.10420. preprint.

Bradley, Ralph Allan and Milton E. Terry (1952). "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons". In: *Biometrika* 39.3/4, pp. 324–345. DOI: 10.2307/2334029. JSTOR: 2334029.

Brown, Daniel S., Russell Coleman, Ravi Srinivasan, and Scott Niekum (2020). "Safe Imitation Learning via Fast Bayesian Reward Inference from Preferences". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 1165–1177. URL: https://proceedings.mlr.press/v119/brown20a.html.

Brown, Daniel S., Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum (2019). "Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 783–792. URL: https://proceedings.mlr.press/v97/brown19a.html.

Busa-Fekete, Róbert, Balázs Szörényi, Paul Weng, Weiwei Cheng, and Eyke Hüllermeier (2014). "Preference-Based Reinforcement Learning: Evolutionary Direct Policy Search Using a Preference-Based Racing Algorithm". In: *Machine Learning* 97.3, pp. 327–351. DOI: 10.1007/s10994-014-5458-8.

Cabi, Serkan, Sergio Gómez Colmenarejo, Alexander Novikov, Ksenia Konyushova, Scott Reed, Rae Jeong, Konrad Zolna, Yusuf Aytar, David Budden, Mel Vecerik, Oleg Sushkov, David Barker, Jonathan Scholz, Misha Denil, Nando de Freitas, and Ziyu Wang (2020). "Scaling Data-Driven Robotics with Reward Sketching and Batch Reinforcement Learning". In: *Proceedings of Robotics: Science and Systems (RSS)*. Vol. 16. DOI: 10.15607/RSS.2020.XVI.076.

Cao, Haoyang, Samuel Cohen, and Lukasz Szpruch (2021a). "Identifiability in Inverse Reinforcement Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34. Curran Associates, Inc., pp. 12362–12373. URL: https://proceedings.neurips.cc/paper/2021/hash/671f0311e2754fcdd37f70a8550379bc-Abstract.html.

Cao, Zehong, KaiChiu Wong, and Chin-Teng Lin (2021b). "Weak Human Preference Supervision for Deep Reinforcement Learning". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.12, pp. 5369–5378. DOI: 10.1109/TNNLS.2021.3084198.

Casper, Stephen, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell (2023). *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*. arXiv: 2307.15217. preprint.

Castricato, Louis, Alex Havrilla, Shahbuland Matiana, Duy V. Phung, Aman Tiwari, Jonathan Tow, and Maksym Zhuravinsky (2023). *trlX: A Scalable Framework for RLHF*. Zenodo. DOI: 10.5281/zenodo.8076391.

Cattelan, Manuela (2012). "Models for Paired Comparison Data: A Review with Emphasis on Dependent Data". In: *Statistical Science* 27.3, pp. 412–433. DOI: 10.1214/12-STS396.

Chan, Lawrence, Andrew Critch, and Anca Dragan (2021). *Human Irrationality: Both Bad and Good for Reward Inference*. arXiv: 2111.06956. preprint.

Chatterji, Niladri, Aldo Pacchiano, Peter Bartlett, and Michael Jordan (2021). "On the Theory of Reinforcement Learning with Once-per-Episode Feedback". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34. Curran Associates, Inc., pp. 3401–3412. URL: https://proceedings.neurips.cc/paper/2021/hash/1bf2efbbe0c49b9f567c2e40f645279a-Abstract.html.

Chen, Jiaao, Mohan Dodda, and Diyi Yang (2023). "Human-in-the-Loop Abstractive Dialogue Summarization". In: *Findings of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, pp. 9176–9190. DOI: 10.18653/v1/2023.findings-acl.584.

Chen, Li, Marco de Gemmis, Alexander Felfernig, Pasquale Lops, Francesco Ricci, and Giovanni Semeraro (2013). "Human Decision Making and Recommender Systems". In: *ACM Transactions on Interactive Intelligent Systems* 3.3, 17:1–17:7. DOI: 10.1145/2533670.2533675.

Chen, Xiaoyu, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang (2022). "Human-in-the-Loop: Provably Efficient Preference-based Reinforcement Learning with General Function Approximation". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 3773–3793. URL: https://proceedings.mlr.press/v162/chen22ag.html.

Cheng, Weiwei, Johannes Fürnkranz, Eyke Hüllermeier, and Sang-Hyeun Park (2011). "Preference-Based Policy Iteration: Leveraging Preference Learning for Reinforcement Learning". In: *Proceedings of Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. Springer, pp. 312–327. DOI: 10.1007/978-3-642-23780-5_30.

Choi, Jaedeug and Kee-eung Kim (2012). "Nonparametric Bayesian Inverse Reinforcement Learning for Multiple Reward Functions". In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 25. Curran Associates, Inc., pp. 314–322. URL: https://proceedings.neurips.cc/paper/2012/hash/140f6969d5213fd0ece03148e62e461e-Abstract.html.

Chowdhury, Sayak Ray and Xingyu Zhou (2023). "Differentially Private Reward Estimation from Preference Based Feedback". In: ICML 2023 Workshop on The Many Facets of Preference-Based Learning. URL: https://openreview.net/forum?id=TqzYmBPSGC.

Christiano, Paul (2016). *Semi-Supervised Reinforcement Learning*. Medium. URL: https://ai-alignment.com/semi-supervised-reinforcement-learning-cf7d5375197f (visited on 12/14/2023).

– (2023). *Thoughts on the Impact of RLHF Research*. URL: https://www.alignmentforum.org/posts/vwu4kegAEZTBtpT6p/thoughts-on-the-impact-of-rlhf-research.

Christiano, Paul, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei (2017). "Deep Reinforcement Learning from Human Preferences". In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 30. Curran Associates, Inc., pp. 4299–4307. URL: https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Clark, Jack and Dario Amodei (2016). *Faulty Reward Functions in the Wild*. OpenAI Blog. URL: https://openai.com/blog/faulty-reward-functions/ (visited on 02/17/2023).

Cosner, Ryan, Maegan Tucker, Andrew Taylor, Kejun Li, Tamas Molnar, Wyatt Ubelacker, Anil Alan, Gabor Orosz, Yisong Yue, and Aaron Ames (2022). "Safety-Aware Preference-Based Learning for Safety-Critical Control". In: *Proceedings of the Annual Learning for Dynamics and Control Conference (L4DC)*. PMLR, pp. 1020–1033. URL: https://proceedings.mlr.press/v168/cosner22a.html.

Cui, Yuchen and Scott Niekum (2018). "Active Reward Learning from Critiques". In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6907–6914. DOI: 10.1109/ICRA.2018.8460854.

Cui, Yuchen, Qiping Zhang, Brad Knox, Alessandro Allievi, Peter Stone, and Scott Niekum (2021). "The EMPATHIC Framework for Task Learning from Implicit Human Feedback". In: *Proceedings of the Conference on Robot Learning (CoRL)*. PMLR, pp. 604–626. URL: https://proceedings.mlr.press/v155/cui21a.html.

Dai, Josef, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang (2023). *Safe RLHF: Safe Reinforcement Learning from Human Feedback*. arXiv: 2310.12773. preprint.

Daniel, Christian, Malte Viering, Jan Metz, Oliver Kroemer, and Jan Peters (2014). "Active Reward Learning". In: *Proceedings of Robotics: Science and Systems (RSS)*. Vol. 10. ISBN: 978-0-9923747-0-9. URL: http://www.roboticsproceedings.org/rss10/p31.html.

Daniels-Koch, Oliver and Rachel Freedman (2022). "The Expertise Problem: Learning from Specialized Feedback". In: NeurIPS 2022 Workshop on ML Safety. URL: https://openreview.net/forum?id=I7K975-H1Mg.

Day, Brett, Ian J. Bateman, Richard T. Carson, Diane Dupont, Jordan J. Louviere, Sanae Morimoto, Riccardo Scarpa, and Paul Wang (2012). "Ordering Effects and Choice Set Awareness in Repeat-Response Stated Preference Studies". In: *Journal of Environmental Economics and Management* 63.1, pp. 73–91. DOI: 10.1016/j.jeem.2011.09.001.

Degenne, Rémy, Thomas Nedelec, Clement Calauzenes, and Vianney Perchet (2019). "Bridging the Gap between Regret Minimization and Best Arm Identification, with Application to A/B Tests". In: *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, pp. 1988–1996. URL: https://proceedings.mlr.press/v89/degenne19a.html.

Ding, Zihan, Yuanpei Chen, Allen Z. Ren, Shixiang Shane Gu, Hao Dong, and Chi Jin (2023). *Learning a Universal Human Prior for Dexterous Manipulation from Human Preference*. arXiv: 2304.04602. preprint.

Dong, Zibin, Yifu Yuan, Jianye Hao, Fei Ni, Yao Mu, Yan Zheng, Yujing Hu, Tangjie Lv, Changjie Fan, and Zhipeng Hu (2023). *AlignDiff: Aligning Diverse Human Preferences via Behavior-Customisable Diffusion Model*. arXiv: 2310.02054. preprint.

Du, Yuqing, Ksenia Konyushkova, Misha Denil, Akhil Raju, Jessica Landon, Felix Hill, Nando de Freitas, and Serkan Cabi (2023). *Vision-Language Models as Success Detectors*. arXiv: 2303.07280. preprint.

Dvijotham, Krishnamurthy and Emanuel Todorov (2010). "Inverse Optimal Control with Linearly-Solvable MDPs". In: *Proceedings of the International Conference on Machine Learning (ICML)*. Omnipress, pp. 335–342. URL: https://icml.cc/Conferences/2010/papers/571.pdf.

Dwork, Cynthia (2008). "Differential Privacy: A Survey of Results". In: *Proceedings of Theory and Applications of Models of Computation (TAMC)*. Springer, pp. 1–19. DOI: 10.1007/978-3-540-79228-4_1.

Early, Joseph, Tom Bewley, Christine Evers, and Sarvapali Ramchurn (2022). "Non-Markovian Reward Modelling from Trajectory Labels via Interpretable Multiple Instance Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 35, pp. 27652–27663. URL: https://proceedings.neurips.cc/paper/2022/hash/b157cfde6794e93b2353b9712bbd45a5-Abstract-Conference.html.

Eberhard, André, Houssam Metni, Georg Fahland, Alexander Stroh, and Pascal Friederich (2022). "Actively Learning Costly Reward Functions for Reinforcement Learning". In: NeurIPS 2022 Workshop on AI for Accelerated Materials Design. URL: https://openreview.net/forum?id=eFHNEv6G9fF.

Fitzgerald, Tesca, Pallavi Koppol, Patrick Callaghan, Russell Quinlan Jun Hei Wong, Reid Simmons, Oliver Kroemer, and Henny Admoni (2023). "INQUIRE: INteractive Querying for User-aware Informative REasoning". In: *Proceedings of the Conference on Robot Learning (CoRL)*. PMLR, pp. 2241–2250. URL: https://proceedings.mlr.press/v205/fitzgerald23a.html.

Fowler Jr., Floyd J. (2013). *Survey Research Methods*. SAGE Publications. 185 pp. ISBN: 978-1-4833-1240-8.

Freedman, Rachel, Rohin Shah, and Anca Dragan (2021). "Choice Set Misspecification in Reward Inference". In: *Proceedings of the Workshop on Artificial Intelligence Safety 2020 Co-Located with the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence*. Vol. 2640. CEUR. URL: https://ceur-ws.org/Vol-2640/#paper_14.

Freedman, Rachel, Justin Svegliato, Kyle Wray, and Stuart Russell (2023). *Active Teacher Selection for Reinforcement Learning from Human Feedback*. arXiv: 2310.15288. preprint.

Freire, Pedro, Adam Gleave, Sam Toyer, and Stuart Russell (2020). "DERAIL: Diagnostic Environments for Reward And Imitation Learning". In: NeurIPS 2020 Workshop on Deep Reinforcement Learning.

Fu, Justin, Katie Luo, and Sergey Levine (2018a). "Learning Robust Rewards with Adverserial Inverse Reinforcement Learning". In: *Proceedings of International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=rkHywl-A-.

Fu, Justin, Avi Singh, Dibya Ghosh, Larry Yang, and Sergey Levine (2018b). “Variational Inverse Control with Events: A General Framework for Data-Driven Reward Definition”. In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 31. Curran Associates, Inc., pp. 8547–8556. URL: https://proceedings.neurips.cc/paper/2018/hash/c9319967c038f9b923068dabdf60cfe3-Abstract.html.

Fujimoto, Scott, Herke Hoof, and David Meger (2018). “Addressing Function Approximation Error in Actor-Critic Methods”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 1587–1596. URL: https://proceedings.mlr.press/v80/fujimoto18a.html.

Fürnkranz, Johannes, Eyke Hüllermeier, Weiwei Cheng, and Sang-Hyeun Park (2012). “Preference-Based Reinforcement Learning: A Formal Framework and a Policy Iteration Algorithm”. In: *Machine Learning* 89.1, pp. 123–156. DOI: 10.1007/s10994-012-5313-8.

Furr, R. Michael (2021). *Psychometrics: An Introduction*. SAGE Publications. 505 pp. ISBN: 978-1-07-182409-2.

Gabriel, Iason (2020). “Artificial Intelligence, Values, and Alignment”. In: *Minds and Machines* 30.3, pp. 411–437. DOI: 10.1007/s11023-020-09539-2.

Gao, Leo, John Schulman, and Jacob Hilton (2023). “Scaling Laws for Reward Model Overoptimization”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 10835–10866. URL: https://proceedings.mlr.press/v202/gao23h.html.

Gao, Yang, Christian M. Meyer, and Iryna Gurevych (2018). “APRIL: Interactively Learning to Summarise by Combining Active Preference Learning and Reinforcement Learning”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 4120–4130. DOI: 10.18653/v1/D18-1445.

– (2020). “Preference-Based Interactive Multi-Document Summarisation”. In: *Information Retrieval Journal* 23.6, pp. 555–585. DOI: 10.1007/s10791-019-09367-8.

Gelada, Carles, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare (2019). “DeepMDP: Learning Continuous Latent Space Models for Representation Learning”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 2170–2179. URL: https://proceedings.mlr.press/v97/gelada19a.html.

Georgii, Hans-Otto (2011). *Gibbs Measures and Phase Transitions*. Walter de Gruyter. 561 pp. ISBN: 978-3-11-025029-9.

Ghosal, Gaurav R., Matthew Zurek, Daniel S. Brown, and Anca D. Dragan (2023). “The Effect of Modeling Human Rationality Level on Learning Rewards from Multiple Feedback Types”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 5, pp. 5983–5992. DOI: 10.1609/aaai.v37i5.25740.

Gilbert, Hugo, Olivier Spanjaard, Paolo Viappiani, and Paul Weng (2015). “Reducing the Number of Queries in Interactive Value Iteration”. In: *Proceedings of Algorithmic Decision Theory (ADT)*. Springer International Publishing, pp. 139–152. DOI: 10.1007/978-3-319-23114-3_9.

Gilbert, Hugo and Paul Weng (2016a). “Quantile Reinforcement Learning”. In: Asian Workshop on Reinforcement Learning.

Gilbert, Hugo, Paul Weng, and Yan Xu (2017). “Optimizing Quantiles in Preference-Based Markov Decision Processes”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1, pp. 3569–3575. DOI: 10.1609/aaai.v31i1.11026.

Gilbert, Hugo, Bruno Zanuttini, Paolo Viappiani, Paul Weng, and Esther Nicart (2016b). “Model-Free Reinforcement Learning with Skew-Symmetric Bilinear Utilities”. In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 252–261. URL: http://auai.org/uai2016/proceedings/papers/91.pdf.

Glaese, Amelia, 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 (2022). *Improving Alignment of Dialogue Agents via Targeted Human Judgements*. arXiv: 2209.14375. preprint.

Glanois, Claire, Paul Weng, Matthieu Zimmer, Dong Li, Tianpei Yang, Jianye Hao, and Wulong Liu (2022). *A Survey on Interpretable Reinforcement Learning*. arXiv: 2112.13112. preprint.

Gleave, Adam, Michael D. Dennis, Shane Legg, Stuart Russell, and Jan Leike (2022a). “Quantifying Differences in Reward Functions”. In: *Proceedings of International Conference on Learning Representations (ICLR)*. URL: https://iclr.cc/virtual/2021/poster/3348.

Gleave, Adam and Geoffrey Irving (2022b). *Uncertainty Estimation for Language Reward Models*. arXiv: 2203.07472. preprint.

Gleave, Adam, Mohammad Taufeeque, Juan Rocamonde, Erik Jenner, Steven H. Wang, Sam Toyer, Maximilian Ernestus, Nora Belrose, Scott Emmons, and Stuart Russell (2022c). *Imitation: Clean Imitation Learning Implementations*. arXiv: 2211.11972. preprint.

Guan, Lin, Mudit Verma, Sihang Guo, Ruohan Zhang, and Subbarao Kambhampati (2021). "Widening the Pipeline in Human-Guided Reinforcement Learning with Explanation and Context-Aware Data Augmentation". In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 21885–21897. URL: https://proceedings.neurips.cc/paper/2021/hash/b6f8dc086b2d60c5856e4ff517060392-Abstract.html.

Gul, Faruk, Paulo Natenzon, and Wolfgang Pesendorfer (2014). "Random Choice as Behavioral Optimization". In: *Econometrica* 82.5, pp. 1873–1912. DOI: 10.3982/ECTA10621.

Haarnoja, Tuomas, Aurick Zhou, Pieter Abbeel, and Sergey Levine (2018). "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 1861–1870. URL: https://proceedings.mlr.press/v80/haarnoja18b.html.

Habibian, Soheil, Ananth Jonnavittula, and Dylan P. Losey (2022). "Here's What I've Learned: Asking Questions That Reveal Reward Learning". In: *ACM Transactions on Human-Robot Interaction* 11.4, 40:1–40:28. DOI: 10.1145/3526107.

Haddenhorst, Björn, Viktor Bengs, Jasmin Brandt, and Eyke Hüllermeier (2021). "Testification of Condorcet Winners in Dueling Bandits". In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. PMLR, pp. 1195–1205. URL: https://proceedings.mlr.press/v161/haddenhorst21a.html.

Haddenhorst, Björn, Eyke Hüllermeier, and Martin Kolb (2020). "Generalized Transitivity: A Systematic Comparison of Concepts with an Application to Preferences in the Babington Smith Model". In: *International Journal of Approximate Reasoning* 119, pp. 373–407. DOI: 10.1016/j.ijar.2020.01.007.

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell (2017a). "The Off-Switch Game". In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, pp. 220–227. DOI: 10.24963/ijcai.2017/32.

Hadfield-Menell, Dylan, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan (2017b). "Inverse Reward Design". In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 30. Curran Associates, Inc., pp. 6765–6774. URL: https://proceedings.neurips.cc/paper/2017/hash/32fdab6559cdfa4f167f8c31b9199643-Abstract.html.

Hafner, Danijar, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi (2020). "Dream to Control: Learning Behaviors by Latent Imagination". In: *Proceedings of International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=S1lOTC4tDS.

Hafner, Danijar, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap (2023). *Mastering Diverse Domains through World Models*. arXiv: 2301.04104. preprint.

Hagendorff, Thilo and Sarah Fabi (2022). *Methodological Reflections for AI Alignment Research Using Human Feedback*. arXiv: 2301.06859. preprint.

He, Jerry Zhi-Yang and Anca D. Dragan (2022). "Assisted Robust Reward Design". In: *Proceedings of the Conference on Robot Learning (CoRL)*. PMLR, pp. 1234–1246. URL: https://proceedings.mlr.press/v164/he22a.html.

Hejna, Donald Joseph and Dorsa Sadigh (2023a). "Few-Shot Preference Learning for Human-in-the-Loop RL". In: *Proceedings of the Conference on Robot Learning (CoRL)*. PMLR, pp. 2014–2025. URL: https://proceedings.mlr.press/v205/iii23a.html.

Hejna, Joey and Dorsa Sadigh (2023b). "Inverse Preference Learning: Preference-based RL without a Reward Function". In: ICML 2023 Workshop on Interactive Learning with Implicit Human Feedback. URL: https://openreview.net/forum?id=ut9y3udeAo.

Hessel, Matteo, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver (2018). "Rainbow: Combining Improvements in Deep Reinforcement Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1, pp. 3215–3222. DOI: 10.1609/aaai.v32i1.11796.

Hoffman, Matthew, Eric Brochu, and Nando de Freitas (2011). "Portfolio Allocation for Bayesian Optimization". In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, pp. 327–336. DOI: 10.5555/3020548.3020587.

Holladay, Rachel, Shervin Javdani, Anca Dragan, and Siddhartha Srinivasa (2016). "Active Comparison Based Learning Incorporating User Uncertainty and Noise". In: RSS 2016 Workshop on Model Learning for Human-Robot Communication.

Huang, Jie, Jiangshan Hao, Rongshun Juan, Randy Gomez, Keisuke Nakamura, and Guangliang Li (2023). "GAN-Based Interactive Reinforcement Learning from Demonstration and Human Evaluative Feedback". In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4991–4998. DOI: 10.1109/ICRA48891.2023.10160939.

Ibarz, Borja, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei (2018). "Reward Learning from Human Preferences and Demonstrations in Atari". In: *Advances in Neural Information Processing Systems*

(NIPS). Vol. 31. Curran Associates, Inc., pp. 8022–8034. URL: https://proceedings.neurips.cc/paper/2018/hash/8cbe9ce23f42628c98f80fa0fac8b19a-Abstract.html.

Irpan, Alexander, Kanishka Rao, Konstantinos Bousmalis, Chris Harris, Julian Ibarz, and Sergey Levine (2019). "Off-Policy Evaluation via Off-Policy Classification". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32. Curran Associates, Inc., pp. 5438–5449. URL: https://proceedings.neurips.cc/paper/2019/hash/b5b03f06271f8917685d14cea7c6c50a-Abstract.html.

Isbell, Charles, Christian R. Shelton, Michael Kearns, Satinder Singh, and Peter Stone (2001). "A Social Reinforcement Learning Agent". In: *Proceedings of the International Conference on Autonomous Agents (AGENTS)*. Association for Computing Machinery, pp. 377–384. DOI: 10.1145/375735.376334.

Jain, Ashesh, Shikhar Sharma, Thorsten Joachims, and Ashutosh Saxena (2015). "Learning Preferences for Manipulation Tasks from Online Coactive Feedback". In: *The International Journal of Robotics Research* 34.10, pp. 1296–1313. DOI: 10.1177/0278364915581193.

Jenner, Erik and Adam Gleave (2021). "Preprocessing Reward Functions for Interpretability". In: NeurIPS 2021 Workshop on Cooperative AI.

Jenner, Erik, Joar Max Viktor Skalse, and Adam Gleave (2022). "A General Framework for Reward Function Distances". In: NeurIPS 2022 Workshop on ML Safety. URL: https://openreview.net/forum?id=Hn21kZHiCK.

Jeon, Hong Jun, Smitha Milli, and Anca Dragan (2020). "Reward-Rational (Implicit) Choice: A Unifying Formalism for Reward Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. Curran Associates, Inc., pp. 4415–4426. URL: https://proceedings.neurips.cc/paper/2020/hash/2f10c1578a0706e06b6d7db6f0b4a6af-Abstract.html.

Ji, Jiaming, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao (2023a). *AI Alignment: A Comprehensive Survey*. arXiv: 2310.19852. preprint.

Ji, Xiang, Huazheng Wang, Minshuo Chen, Tuo Zhao, and Mengdi Wang (2023b). *Provable Benefits of Policy Learning from Human Preferences in Contextual Bandit Problems*. arXiv: 2307.12975. preprint.

Jiang, Nan and Lihong Li (2016). "Doubly Robust Off-policy Value Evaluation for Reinforcement Learning". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 652–661. URL: https://proceedings.mlr.press/v48/jiang16.html.

Jin, Chi, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan (2020). "Provably Efficient Reinforcement Learning with Linear Function Approximation". In: *Proceedings of the Conference on Learning Theory (COLT)*. PMLR, pp. 2137–2143. URL: https://proceedings.mlr.press/v125/jin20a.html.

Jin, Ying, Zhuoran Yang, and Zhaoran Wang (2021). "Is Pessimism Provably Efficient for Offline RL?" In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 5084–5096. URL: https://proceedings.mlr.press/v139/jin21e.html.

Kahn, Gregory, Pieter Abbeel, and Sergey Levine (2021). "LaND: Learning to Navigate From Disengagements". In: *IEEE Robotics and Automation Letters* 6.2, pp. 1872–1879. DOI: 10.1109/LRA.2021.3060404.

Kalra, Akansha and Daniel S. Brown (2022). "Interpretable Reward Learning via Differentiable Decision Trees". In: NeurIPS 2022 Workshop on ML Safety. URL: https://openreview.net/forum?id=3bk40MsYjet.

– (2023). *Can Differentiable Decision Trees Learn Interpretable Reward Functions?* arXiv: 2306.13004. preprint.

Kang, Yachen, Diyuan Shi, Jinxin Liu, Li He, and Donglin Wang (2023). "Beyond Reward: Offline Preference-guided Policy Optimization". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 15753–15768. URL: https://proceedings.mlr.press/v202/kang23b.html.

Katz, Sydney M., Amir Maleki, Erdem Bıyık, and Mykel J. Kochenderfer (2021). *Preference-Based Learning of Reward Function Features*. arXiv: 2103.02727. preprint.

Kaufmann, Timo, Sarah Ball, Jacob Beck, Frauke Kreuter, and Eyke Hüllermeier (2023). "On the Challenges and Practices of Reinforcement Learning from Real Human Feedback". In: ECML PKDD 2023 Workshop Towards Hybrid Human-Machine Learning and Decision Making.

Kazemi, Hadi, Fariborz Taherkhani, and Nasser M. Nasrabadi (2020). "Preference-Based Image Generation". In: *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3393–3402. DOI: 10.1109/WACV45572.2020.9093406.

Kim, Changyeon, Jongjin Park, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee (2023). "Preference Transformer: Modeling Human Preferences Using Transformers for RL". In: *Proceedings of International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=Peot1SFDX0.

Kim, Kuno, Shivam Garg, Kirankumar Shiragur, and Stefano Ermon (2021). "Reward Identification in Inverse Reinforcement Learning". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 5496–5505. URL: https://proceedings.mlr.press/v139/kim21c.html.

Kirchner, Jan H., Logan Smith, Jacques Thibodeau, Kyle McDonell, and Laria Reynolds (2022). *Researching Alignment Research: Unsupervised Analysis*. arXiv: 2206.02841. preprint.

Knox, W. Bradley (2012). “Learning from Human-generated Reward”. The University of Texas at Austin. URL: https://repositories.lib.utexas.edu/items/20b9e8a1-a78d-4844-816f-3c0b0a4c848a.

Knox, W. Bradley, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone (2023). “Reward (Mis)Design for Autonomous Driving”. In: *Artificial Intelligence* 316, p. 103829. DOI: 10.1016/j.artint.2022.103829.

Knox, W. Bradley, Stephane Hatgis-Kessell, Serena Booth, Scott Niekum, Peter Stone, and Alessandro Allievi (2022). *Models of Human Preference for Learning Reward Functions*. arXiv: 2206.02231. preprint.

Knox, W. Bradley and Peter Stone (2008). “TAMER: Training an Agent Manually via Evaluative Reinforcement”. In: *Proceedings of the IEEE International Conference on Development and Learning (ICDL)*, pp. 292–297. DOI: 10.1109/DEVLRN.2008.4640845.

Kong, Dingwen and Lin Yang (2022). “Provably Feedback-Efficient Reinforcement Learning via Active Reward Learning”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 35, pp. 11063–11078. URL: https://proceedings.neurips.cc/paper/2022/hash/476c289f685e27936aa089e9d53a4213-Abstract-Conference.html.

Koppol, Pallavi, Henny Admoni, and Reid Simmons (2020). “Iterative Interactive Reward Learning”. In: ICML 2020 Workshop on Participatory Approaches to Machine Learning.

Korba, Anna, Stéphan Clemencon, and Eric Sibony (2017). “A Learning Theory of Ranking Aggregation”. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, pp. 1001–1010. URL: https://proceedings.mlr.press/v54/korba17a.html.

Krening, Samantha and Karen M. Feigh (2018). “Interaction Algorithm Effect on Human Experience with Reinforcement Learning”. In: *ACM Transactions on Human-Robot Interaction* 7.2, 16:1–16:22. DOI: 10.1145/3277904.

Kupcsik, Andras, David Hsu, and Wee Sun Lee (2018). “Learning Dynamic Robot-to-Human Object Handover from Human Feedback”. In: *Robotics Research: Volume 1*. Springer Proceedings in Advanced Robotics. Springer International Publishing, pp. 161–176. DOI: 10.1007/978-3-319-51532-8_10.

Lambert, Nathan (2021). *Reward Is Not Enough*. Democratizing Automation. URL: https://robotic.substack.com/p/reward-is-not-enough (visited on 02/17/2023).

Laskin, Misha, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas (2020). “Reinforcement Learning with Augmented Data”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. Curran Associates, Inc., pp. 19884–19895. URL: https://proceedings.neurips.cc/paper/2020/hash/e615c82aba461681ade82da2da38004a-Abstract.html.

Lattimore, Tor and Csaba Szepesvári (2020). *Bandit Algorithms*. Cambridge University Press. DOI: 10.1017/9781108571401.

Le, Hoang, Cameron Voloshin, and Yisong Yue (2019). “Batch Policy Learning under Constraints”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 3703–3712. URL: https://proceedings.mlr.press/v97/le19a.html.

Lee, Kimin, Kibok Lee, Jinwoo Shin, and Honglak Lee (2020). “Network Randomization: A Simple Technique for Generalization in Deep Reinforcement Learning”. In: *Proceedings of International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=HJgcvJBFvB.

Lee, Kimin, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu (2023). *Aligning Text-to-Image Models Using Human Feedback*. arXiv: 2302.12192. preprint.

Lee, Kimin, Laura Smith, Anca Dragan, and Pieter Abbeel (2021a). “B-Pref: Benchmarking Preference-Based Reinforcement Learning”. In: Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (NeurIPS Datasets and Benchmarks). URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/d82c8d1619ad8176d665453cfb2e55f0-Abstract-round1.html.

Lee, Kimin, Laura M. Smith, and Pieter Abbeel (2021b). “PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 6152–6163. URL: https://proceedings.mlr.press/v139/lee21i.html.

Lehman, Joel, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J. Bentley, Samuel Bernard, Guillaume Beslon, David M. Bryson, Nick Cheney, Patryk Chrabaszcz, Antoine Cully, Stephane Doncieux, Fred C. Dyer, Kai Olav Ellefsen, Robert Feldt, Stephan Fischer, Stephanie Forrest, Antoine Fŕenoy, Christian Gagńe, Leni Le Goff, Laura M. Grabowski, Babak Hodjat, Frank Hutter, Laurent Keller, Carole Knibbe, Peter Krcah, Richard E. Lenski, Hod Lipson, Robert MacCurdy, Carlos Maestre, Risto Miikkulainen, Sara Mitri, David E. Moriarty, Jean-Baptiste Mouret, Anh Nguyen, Charles Ofria, Marc Parizeau, David Parsons, Robert T. Pennock, William F. Punch, Thomas S. Ray, Marc Schoenauer, Eric Schulte, Karl Sims, Kenneth O. Stanley, François Taddei, Danesh Tarapore, Simon Thibault, Richard Watson, Westley Weimer, and Jason Yosinski (2020). “The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities”. In: *Artificial Life* 26.2, pp. 274–306. DOI: 10.1162/artl_a_00319.

Leike, Jan, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg (2018). *Scalable Agent Alignment via Reward Modeling: A Research Direction*. arXiv: 1811.07871. preprint.

Li, Mengxi, Alper Canberk, Dylan P. Losey, and Dorsa Sadigh (2021). "Learning Human Objectives from Sequences of Physical Corrections". In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2877–2883. DOI: 10.1109/ICRA48506.2021.9560829.

Li, Zihao, Zhuoran Yang, and Mengdi Wang (2023). "Reinforcement Learning with Human Feedback: Learning Dynamic Choices via Pessimism". In: ICML 2023 Workshop on Interactive Learning with Implicit Human Feedback. URL: https://openreview.net/forum?id=gxM2AUFMsK.

Liang, Xinran, Katherine Shu, Kimin Lee, and Pieter Abbeel (2022). "Reward Uncertainty for Exploration in Preference-based Reinforcement Learning". In: *Proceedings of International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=OWZVD-l-ZrC.

Lichtenstein, Sarah and Paul Slovic, eds. (2006). *The Construction of Preference*. Cambridge University Press. DOI: 10.1017/CBO9780511618031.

Lin, Jinying, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li (2020a). "A Review on Interactive Reinforcement Learning From Human Social Feedback". In: *IEEE Access* 8, pp. 120757–120765. DOI: 10.1109/ACCESS.2020.3006254.

Lin, Stephanie, Jacob Hilton, and Owain Evans (2022). "TruthfulQA: Measuring How Models Mimic Human Falsehoods". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*. Association for Computational Linguistics, pp. 3214–3252. DOI: 10.18653/v1/2022.acl-long.229.

Lin, Yijiong, Jiancong Huang, Matthieu Zimmer, Yisheng Guan, Juan Rojas, and Paul Weng (2020b). "Invariant Transform Experience Replay: Data Augmentation for Deep Reinforcement Learning". In: *IEEE Robotics and Automation Letters* 5.4, pp. 6615–6622. DOI: 10.1109/LRA.2020.3013937.

Lindner, David and Mennatallah El-Assady (2022). "Humans Are Not Boltzmann Distributions: Challenges and Opportunities for Modelling Human Feedback and Interaction in Reinforcement Learning". In: IJCAI-ECAI 2022 Workshop on Communication in Human-AI Interaction.

Lindner, David, Matteo Turchetta, Sebastian Tschiatschek, Kamil Ciosek, and Andreas Krause (2021). "Information Directed Reward Learning for Reinforcement Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34. Curran Associates, Inc., pp. 3850–3862. URL: https://proceedings.neurips.cc/paper/2021/hash/1fa6269f58898f0e809575c9a48747ef-Abstract.html.

Liu, Runze, Yali Du, Fengshuo Bai, Jiafei Lyu, and Xiu Li (2023a). *Zero-Shot Preference Learning for Offline RL via Optimal Transport*. arXiv: 2306.03615. preprint.

Liu, Yiheng, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge (2023b). "Summary of ChatGPT-Related Research and Perspective towards the Future of Large Language Models". In: *Meta-Radiology* 1.2, p. 100017. DOI: 10.1016/j.metrad.2023.100017.

Losey, Dylan P., Andrea Bajcsy, Marcia K. O'Malley, and Anca D. Dragan (2022). "Physical Interaction as Communication: Learning Robot Objectives Online from Human Corrections". In: *The International Journal of Robotics Research* 41.1, pp. 20–44. DOI: 10.1177/02783649211050958.

Losey, Dylan P. and Marcia K. O'Malley (2018). "Including Uncertainty When Learning from Human Corrections". In: *Proceedings of the Conference on Robot Learning (CoRL)*. PMLR, pp. 123–132. URL: http://proceedings.mlr.press/v87/losey18a.html.

Luce, R. Duncan (1959). *Individual Choice Behavior*. Individual Choice Behavior. John Wiley, pp. xii, 153. xii, 153.

Luketina, Jelena, Nantas Nardelli, Gregory Farquhar, Jakob Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel (2019). "A Survey of Reinforcement Learning Informed by Natural Language". In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, pp. 6309–6317. DOI: 10.24963/ijcai.2019/880.

Ma, Yecheng Jason, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar (2023). *Eureka: Human-Level Reward Design via Coding Large Language Models*. arXiv: 2310.12931. preprint.

Mandl, Monika, Alexander Felfernig, Erich Teppan, and Monika Schubert (2011). "Consumer Decision Making in Knowledge-Based Recommendation". In: *Journal of Intelligent Information Systems* 37.1, pp. 1–22. DOI: 10.1007/s10844-010-0134-3.

Massart, Pascal and Élodie Nédélec (2006). "Risk Bounds for Statistical Learning". In: *The Annals of Statistics* 34.5, pp. 2326–2366. DOI: 10.1214/009053606000000786.

McKinney, Lev E., Yawen Duan, David Krueger, and Adam Gleave (2022). "On The Fragility of Learned Reward Functions". In: NeurIPS 2022 Workshop on Deep Reinforcement Learning. URL: https://openreview.net/forum?id=9gj9vXfeS-y.

Mendez, Jorge, Shashank Shivkumar, and Eric Eaton (2018). “Lifelong Inverse Reinforcement Learning”. In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 31. Curran Associates, Inc., pp. 4507–4518. URL: https://papers.nips.cc/paper/2018/hash/2d969e2cee8cfa07ce7ca0bb13c7a36d-Abstract.html.

Metcalf, Katherine, Miguel Sarabia, and Barry-John Theobald (2022). *Rewards Encoding Environment Dynamics Improves Preference-based Reinforcement Learning*. arXiv: 2211.06527. preprint.

Metz, Yannick, David Lindner, Raphaël Baur, Daniel A. Keim, and Mennatallah El-Assady (2023). “RLHF-Blender: A Configurable Interactive Interface for Learning from Diverse Human Feedback”. In: ICML 2023 Workshop on Interactive Learning with Implicit Human Feedback. URL: https://openreview.net/forum?id=JvkZtzJBFQ.

Milani, Stephanie, Anssi Kanervisto, Karolis Ramanauskas, Sander Schulhoff, Brandon Houghton, Sharada Mohanty, Byron Galbraith, Ke Chen, Yan Song, Tianze Zhou, Bingquan Yu, He Liu, Kai Guan, Yujing Hu, Tangjie Lv, Federico Malato, Florian Leopold, Amogh Raut, Ville Hautamäki, Andrew Melnik, Shu Ishida, João Henriques, Robert Klassert, Walter Laurito, Lucas Cazzonelli, Cedric Kulbach, Nicholas Popovic, Marvin Schweizer, Ellen Novoseller, Vinicius Goecks, Nicholas Waytowich, David Watkins, Josh Miller, and Rohin Shah (2022). “Towards Solving Fuzzy Tasks with Human Feedback: A Retrospective of the MineRL BASALT 2022 Competition”. In: *Proceedings of the NeurIPS 2022 Competitions Track*. PMLR, pp. 171–188. URL: https://proceedings.mlr.press/v220/milani22a.html.

Milani, Stephanie, Nicholay Topin, Manuela Veloso, and Fei Fang (2023). “Explainable Reinforcement Learning: A Survey and Comparative Review”. In: *ACM Computing Surveys*. DOI: 10.1145/3616864.

Milli, Smitha and Anca D. Dragan (2020). “Literal or Pedagogic Human? Analyzing Human Model Misspecification in Objective Learning”. In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. PMLR, pp. 925–934. URL: https://proceedings.mlr.press/v115/milli20a.html.

Mindermann, Sören, Rohin Shah, Adam Gleave, and Dylan Hadfield-Menell (2018). “Active Inverse Reward Design”. In: ICML 2018 Workshop on Goals in Reinforcement Learning.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis (2015). “Human-Level Control through Deep Reinforcement Learning”. In: *Nature* 518.7540 (7540), pp. 529–533. DOI: 10.1038/nature14236.

Mu, Weiyan and Shifeng Xiong (2023). “On Huber’s Contaminated Model”. In: *Journal of Complexity* 77, p. 101745. DOI: 10.1016/j.jco.2023.101745.

Myers, Vivek, Erdem Bıyık, Nima Anari, and Dorsa Sadigh (2022). “Learning Multimodal Rewards from Rankings”. In: *Proceedings of the Conference on Robot Learning (CoRL)*. PMLR, pp. 342–352. URL: https://proceedings.mlr.press/v164/myers22a.html.

Myers, Vivek, Erdem Bıyık, and Dorsa Sadigh (2023). “Active Reward Learning from Online Preferences”. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7511–7518. DOI: 10.1109/ICRA48891.2023.10160439.

Najar, Anis and Mohamed Chetouani (2021). “Reinforcement Learning With Human Advice: A Survey”. In: *Frontiers in Robotics and AI* 8, p. 584075. DOI: 10.3389/frobt.2021.584075.

Nakano, Reiichiro, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman (2022). *WebGPT: Browser-assisted Question-Answering with Human Feedback*. arXiv: 2112.09332. preprint.

Narvekar, Sanmit, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone (2020). “Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey”. In: *Journal of Machine Learning Research* 21.181, pp. 1–50. ISSN: 1533-7928. URL: http://jmlr.org/papers/v21/20-212.html.

Neu, Gergely and Csaba Szepesvári (2009). “Training Parsers by Inverse Reinforcement Learning”. In: *Machine Learning* 77.2, pp. 303–337. DOI: 10.1007/s10994-009-5110-1.

Ng, Andrew Y., Daishi Harada, and Stuart J. Russell (1999). “Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. Morgan Kaufmann Publishers Inc., pp. 278–287. ISBN: 978-1-55860-612-8.

Ng, Andrew Y. and Stuart J. Russell (2000). “Algorithms for Inverse Reinforcement Learning”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. Morgan Kaufmann Publishers Inc., pp. 663–670. ISBN: 978-1-55860-707-1.

Novoseller, Ellen, Yibing Wei, Yanan Sui, Yisong Yue, and Joel Burdick (2020). “Dueling Posterior Sampling for Preference-Based Reinforcement Learning”. In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. PMLR, pp. 1029–1038. URL: https://proceedings.mlr.press/v124/novoseller20a.html.

OpenAI (2022). *ChatGPT: Optimizing Language Models for Dialogue*. URL: https://openai.com/blog/chatgpt (visited on 02/02/2023).

– (2023). *GPT-4 Technical Report*. OpenAI. URL: https://cdn.openai.com/papers/gpt-4.pdf.

Osa, Takayuki, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters (2018). "An Algorithmic Perspective on Imitation Learning". In: *Foundations and Trends® in Robotics* 7.1-2, pp. 1–179. DOI: 10.1561/2300000053.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe (2022). "Training Language Models to Follow Instructions with Human Feedback". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 35, pp. 27730–27744. URL: https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Paduraru, Cosmin (2013). "Off-Policy Evaluation in Markov Decision Processes". McGill University. URL: https://escholarship.mcgill.ca/concern/theses/p8418r74h.

Palan, Malayandi, Gleb Shevchuk, Nicholas Charles Landolfi, and Dorsa Sadigh (2019). "Learning Reward Functions by Integrating Human Demonstrations and Preferences". In: *Proceedings of Robotics: Science and Systems (RSS)*. Vol. 15. ISBN: 978-0-9923747-5-4. URL: http://www.roboticsproceedings.org/rss15/p23.html.

Park, Jongjin, Younggyo Seo, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee (2022). "SURF: Semi-supervised Reward Learning with Data Augmentation for Feedback-efficient Preference-based Reinforcement Learning". In: *Proceedings of International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=TfhfZLQ2EJO.

Pinto, André Susano, Alexander Kolesnikov, Yuge Shi, Lucas Beyer, and Xiaohua Zhai (2023). "Tuning Computer Vision Models With Task Rewards". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 33229–33239. URL: https://proceedings.mlr.press/v202/susano-pinto23a.html.

Plackett, R. L. (1975). "The Analysis of Permutations". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 24.2, pp. 193–202. DOI: 10.2307/2346567. JSTOR: 2346567.

Pommeranz, Alina, Joost Broekens, Pascal Wiggers, Willem-Paul Brinkman, and Catholijn M. Jonker (2012). "Designing Interfaces for Explicit Preference Elicitation: A User-Centered Investigation of Preference Representation and Elicitation Process". In: *User Modeling and User-Adapted Interaction* 22.4, pp. 357–397. DOI: 10.1007/s11257-011-9116-6.

Poole, Benjamin and Minwoo Lee (2022). *Towards Intrinsic Interactive Reinforcement Learning*. arXiv: 2112.01575. preprint.

Precup, Doina, Richard S. Sutton, and Satinder P. Singh (2000). "Eligibility Traces for Off-Policy Policy Evaluation". In: *Proceedings of the International Conference on Machine Learning (ICML)*. Morgan Kaufmann Publishers Inc., pp. 759–766. ISBN: 978-1-55860-707-1.

Puiutta, Erika and Eric M. S. P. Veith (2020). "Explainable Reinforcement Learning: A Survey". In: *Proceedings of Machine Learning and Knowledge Extraction (CD-MAKE)*. Springer International Publishing, pp. 77–95. DOI: 10.1007/978-3-030-57321-8_5.

Qian, Junqi, Paul Weng, and Chenmien Tan (2023). "Learning Rewards to Optimize Global Performance Metrics in Deep Reinforcement Learning". In: *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. IFAAMAS, pp. 1951–1960. DOI: 10.5555/3545946.3598864.

Qing, Yunpeng, Shunyu Liu, Jie Song, Huiqiong Wang, and Mingli Song (2023). *A Survey on Explainable Reinforcement Learning: Concepts, Algorithms, Challenges*. arXiv: 2211.06665. preprint.

Racca, Mattia, Antti Oulasvirta, and Ville Kyrki (2019). "Teacher-Aware Active Robot Learning". In: *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 335–343. DOI: 10.1109/HRI.2019.8673300.

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn (2023). "Direct Preference Optimization: Your Language Model Is Secretly a Reward Model". In: Conference on Neural Information Processing Systems (NeurIPS). URL: https://openreview.net/forum?id=HPuSIXJaa9.

Rahtz, Matthew, Vikrant Varma, Ramana Kumar, Zachary Kenton, Shane Legg, and Jan Leike (2022). *Safe Deep RL in 3D Environments Using Human Feedback*. arXiv: 2201.08102. preprint.

Ramamurthy, Rajkumar, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi (2023). "Is Reinforcement Learning (Not) for Natural Language Processing: Benchmarks, Baselines, and Building Blocks for Natural Language Policy Optimization". In: *Proceedings of International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=8aHzds2uUyB.

Reddy, Siddharth, Anca Dragan, Sergey Levine, Shane Legg, and Jan Leike (2020). "Learning Human Objectives by Evaluating Hypothetical Behavior". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 8020–8029. URL: https://proceedings.mlr.press/v119/reddy20a.html.

Regan, Kevin and Craig Boutilier (2009). "Regret-Based Reward Elicitation for Markov Decision Processes". In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, pp. 444–451. ISBN: 978-0-9749039-5-8. URL: https://dl.acm.org/doi/10.5555/1795114.1795166.

Regan, Kevin and Craig Boutilier (2011). “Robust Online Optimization of Reward-Uncertain MDPs”. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI Press, pp. 2165–2171. DOI: 10.5591/978-1-57735-516-8/IJCAI11-361.

Ren, Zhizhou, Anji Liu, Yitao Liang, Jian Peng, and Jianzhu Ma (2022). “Efficient Meta Reinforcement Learning for Preference-based Fast Adaptation”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 35, pp. 15502–15515. URL: https://papers.nips.cc/paper_files/paper/2022/hash/63b2b056f48653b7cff0d8d233c96a4d-Abstract-Conference.html.

Russo, Daniel and Benjamin Van Roy (2013). “Eluder Dimension and the Sample Complexity of Optimistic Exploration”. In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 26. Curran Associates, Inc., pp. 2256–2264. URL: https://papers.nips.cc/paper/2013/hash/41bfd20a38bb1b0bec75acf0845530a7-Abstract.html.

Rust, John (1987). “Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher”. In: *Econometrica* 55.5, pp. 999–1033. DOI: 10.2307/1911259. JSTOR: 1911259.

Ryan, Mandy, Karen Gerard, and Mabel Amaya-Amaya, eds. (2008). *Using Discrete Choice Experiments to Value Health and Health Care*. Red. by Ian J. Bateman. Vol. 11. The Economics of Non-Market Goods and Resources. Springer Netherlands. DOI: 10.1007/978-1-4020-5753-3.

Sadigh, Dorsa, Anca Dragan, Shankar Sastry, and Sanjit Seshia (2017). “Active Preference-Based Learning of Reward Functions”. In: *Proceedings of Robotics: Science and Systems (RSS)*. Vol. 13. ISBN: 978-0-9923747-3-0. URL: http://www.roboticsproceedings.org/rss13/p53.html.

Saha, Aadirupa (2021). “Optimal Algorithms for Stochastic Contextual Preference Bandits”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34. Curran Associates, Inc., pp. 30050–30062. URL: https://proceedings.neurips.cc/paper/2021/hash/fc3cf452d3da8402bebb765225ce8c0e-Abstract.html.

Saha, Aadirupa, Aldo Pacchiano, and Jonathan Lee (2023). “Dueling RL: Reinforcement Learning with Trajectory Preferences”. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, pp. 6263–6289. URL: https://proceedings.mlr.press/v206/saha23a.html.

Saunders, William, Girish Sastry, Andreas Stuhlmüller, and Owain Evans (2018). “Trial without Error: Towards Safe Reinforcement Learning via Human Intervention”. In: *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. IFAAMAS, pp. 2067–2069. DOI: 10.5555/3237383.3238074.

Schoenauer, Marc, Riad Akrour, Michele Sebag, and Jean-Christophe Souplet (2014). “Programming by Feedback”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 1503–1511. URL: https://proceedings.mlr.press/v32/schoenauer14.html.

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov (2017). *Proximal Policy Optimization Algorithms*. arXiv: 1707.06347. preprint.

Schwarzer, Max, Ankesh Anand, Rishab Goel, R. Devon Hjelm, Aaron Courville, and Philip Bachman (2022). “Data-Efficient Reinforcement Learning with Self-Predictive Representations”. In: *Proceedings of International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=uCQfPZwRaUu.

Settles, Burr (2012). *Active Learning*. Morgan & Claypool Publishers. 114 pp. ISBN: 978-1-60845-725-0.

Shah, Rohin, Cody Wild, Steven H. Wang, Neel Alex, Brandon Houghton, William Guss, Sharada Mohanty, Anssi Kanervisto, Stephanie Milani, Nicholay Topin, Pieter Abbeel, Stuart Russell, and Anca Dragan (2021). *The MineRL BASALT Competition on Learning from Human Feedback*. arXiv: 2107.01969. preprint.

Shin, Daniel, Anca Dragan, and Daniel S. Brown (2023). “Benchmarks and Algorithms for Offline Preference-Based Reward Learning”. In: *Transactions on Machine Learning Research*. ISSN: 2835-8856. URL: https://openreview.net/forum?id=TGuXXlbKsn.

Shivaswamy, Pannaga and Thorsten Joachims (2012). “Online Structured Prediction via Coactive Learning”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. Omnipress, pp. 59–66. ISBN: 978-1-4503-1285-1. URL: http://icml.cc/2012/papers/717.pdf.

– (2015). “Coactive Learning”. In: *Journal of Artificial Intelligence Research* 53, pp. 1–40. DOI: 10.1613/jair.4539.

Siddique, Umer, Abhinav Sinha, and Yongcan Cao (2023). “Fairness in Preference-based Reinforcement Learning”. In: ICML 2023 Workshop on The Many Facets of Preference-Based Learning. URL: https://openreview.net/forum?id=ColATVnkEl.

Silver, David, Satinder Singh, Doina Precup, and Richard S. Sutton (2021). “Reward Is Enough”. In: *Artificial Intelligence* 299, p. 103535. DOI: 10.1016/j.artint.2021.103535.

Singh, Avi, Larry Yang, Chelsea Finn, and Sergey Levine (2019). “End-To-End Robotic Reinforcement Learning without Reward Engineering”. In: *Proceedings of Robotics: Science and Systems (RSS)*. Vol. 15. ISBN: 978-0-9923747-5-4. URL: http://www.roboticsproceedings.org/rss15/p73.html.

Skalse, Joar Max Viktor and Alessandro Abate (2022a). “The Reward Hypothesis Is False”. In: NeurIPS 2022 Workshop on ML Safety. URL: https://openreview.net/forum?id=5l1NgpzAfH.

Skalse, Joar Max Viktor and Alessandro Abate (2023a). "Misspecification in Inverse Reinforcement Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.12 (12), pp. 15136–15143. DOI: 10.1609/aaai.v37i12.26766.

Skalse, Joar Max Viktor, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Adam Gleave, and Alessandro Abate (2023b). *STARC: A General Framework For Quantifying Differences Between Reward Functions*. arXiv: 2309.15257. preprint.

Skalse, Joar Max Viktor, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave (2023c). "Invariance in Policy Optimisation and Partial Identifiability in Reward Learning". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 32033–32058. URL: https://proceedings.mlr.press/v202/skalse23a.html.

Skalse, Joar Max Viktor, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger (2022b). "Defining and Characterizing Reward Gaming". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 35, pp. 9460–9471. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html.

Song, Ziang, Tianle Cai, Jason D. Lee, and Weijie J. Su (2023). "Reward Collapse in Aligning Large Language Models: A Prompt-Aware Approach to Preference Rankings". In: ICML 2023 Workshop on The Many Facets of Preference-Based Learning. URL: https://openreview.net/forum?id=dpWxK6aqIK.

Stiennon, Nisan, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano (2020). "Learning to Summarize with Human Feedback". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. Curran Associates, Inc., pp. 3008–3021. URL: https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

Sugiyama, Hiroaki, Toyomi Meguro, and Yasuhiro Minami (2012). "Preference-Learning Based Inverse Reinforcement Learning for Dialog Control". In: *Proceedings of Interspeech*. ISCA, pp. 222–225. DOI: 10.21437/Interspeech.2012-72.

Sui, Yanan, Vincent Zhuang, Joel W. Burdick, and Yisong Yue (2017). "Multi-Dueling Bandits with Dependent Arms". In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press. URL: http://auai.org/uai2017/proceedings/papers/155.pdf.

Sui, Yanan, Masrour Zoghi, Katja Hofmann, and Yisong Yue (2018). "Advancements in Dueling Bandits". In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, pp. 5502–5510. DOI: 10.24963/ijcai.2018/776.

Sutton, Richard S. and Andrew G. Barto (2018). *Reinforcement Learning: An Introduction*. Second edition. Adaptive Computation and Machine Learning Series. The MIT Press. 526 pp. ISBN: 978-0-262-03924-6.

Thurstone, Louis Leon (1927). "A Law of Comparative Judgment". In: *Psychological Review* 34, pp. 273–286. DOI: 10.1037/h0070288.

Tien, Jeremy, Jerry Zhi-Yang He, Zackory Erickson, Anca Dragan, and Daniel S. Brown (2023). "Causal Confusion and Reward Misidentification in Preference-Based Reward Learning". In: *Proceedings of International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=R0Xxvr_X3ZA.

Train, Kenneth E. (2009). *Discrete Choice Methods with Simulation*. 2nd ed. Cambridge University Press. DOI: 10.1017/CBO9780511805271.

Tran, Thi Ngoc Trang, Alexander Felfernig, and Nava Tintarev (2021). "Humanized Recommender Systems: State-of-the-art and Research Issues". In: *ACM Transactions on Interactive Intelligent Systems* 11.2, 9:1–9:41. DOI: 10.1145/3446906.

Tucker, Maegan, Kejun Li, Yisong Yue, and Aaron D. Ames (2022). *POLAR: Preference Optimization and Learning Algorithms for Robotics*. arXiv: 2208.04404. preprint.

Valiant, L. G. (1984). "A Theory of the Learnable". In: *Communications of the ACM* 27.11, pp. 1134–1142. DOI: 10.1145/1968.1972.

Vamplew, Peter, Benjamin J. Smith, Johan Källström, Gabriel Ramos, Roxana Rădulescu, Diederik M. Roijers, Conor F. Hayes, Fredrik Heintz, Patrick Mannion, Pieter J. K. Libin, Richard Dazeley, and Cameron Foale (2022). "Scalar Reward Is Not Enough: A Response to Silver, Singh, Precup and Sutton (2021)". In: *Autonomous Agents and Multi-Agent Systems* 36.2, p. 41. DOI: 10.1007/s10458-022-09575-5.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 30. Curran Associates, Inc., pp. 5998–6008. URL: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Verma, Mudit, Siddhant Bhambri, and Subbarao Kambhampati (2023a). "Exploiting Unlabeled Data for Feedback Efficient Human Preference Based Reinforcement Learning". In: AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI.

Verma, Mudit and Subbarao Kambhampati (2023b). “A State Augmentation Based Approach to Reinforcement Learning from Human Preferences”. In: AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI.

– (2023c). “Data Driven Reward Initialization for Preference Based Reinforcement Learning”. In: AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI.

Verma, Mudit and Katherine Metcalf (2022). *Symbol Guided Hindsight Priors for Reward Learning from Human Preferences*. arXiv: 2210.09151. preprint.

Wang, Haoran, Qiuye Jin, Shiman Li, Siyu Liu, Manning Wang, and Zhijian Song (2023a). *A Comprehensive Survey on Deep Active Learning and Its Applications in Medical Image Analysis*. arXiv: 2310.14230. preprint.

Wang, Yuanhao, Qinghua Liu, and Chi Jin (2023b). “Is RLHF More Difficult than Standard RL? A Theoretical Perspective”. In: Conference on Neural Information Processing Systems (NeurIPS). URL: https://openreview.net/forum?id=sxZLrBqg50.

Wang, Zizhao, Junyao Shi, Iretiayo Akinola, and Peter Allen (2020). “Maximizing BCI Human Feedback Using Active Learning”. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10945–10951. DOI: 10.1109/IROS45743.2020.9341669.

Weng, Paul and Bruno Zanuttini (2013). “Interactive Value Iteration for Markov Decision Processes with Unknown Rewards”. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI Press, pp. 2415–2421. ISBN: 978-1-57735-633-2. URL: https://www.ijcai.org/Proceedings/13/Papers/355.pdf.

Wilde, Nils and Javier Alonso-Mora (2023). “Do We Use the Right Measure? Challenges in Evaluating Reward Learning Algorithms”. In: *Proceedings of the Conference on Robot Learning (CoRL)*. PMLR, pp. 1553–1562. URL: https://proceedings.mlr.press/v205/wilde23a.html.

Wilde, Nils, Erdem Bıyık, Dorsa Sadigh, and Stephen L. Smith (2022). “Learning Reward Functions from Scale Feedback”. In: *Proceedings of the Conference on Robot Learning (CoRL)*. PMLR, pp. 353–362. URL: https://proceedings.mlr.press/v164/wilde22a.html.

Wilde, Nils, Dana Kulić, and Stephen L. Smith (2018). “Learning User Preferences in Robot Motion Planning Through Interaction”. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 619–626. DOI: 10.1109/ICRA.2018.8460586.

– (2020). “Active Preference Learning Using Maximum Regret”. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10952–10959. DOI: 10.1109/IROS45743.2020.9341530.

Wilson, Aaron, Alan Fern, and Prasad Tadepalli (2012). “A Bayesian Approach for Policy Learning from Trajectory Preference Queries”. In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 25. Curran Associates, Inc., pp. 1142–1150. URL: https://proceedings.neurips.cc/paper/2012/hash/16c222aa19898e5058938167c8ab6c57-Abstract.html.

Wirth, Christian, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz (2017). “A Survey of Preference-Based Reinforcement Learning Methods”. In: *Journal of Machine Learning Research* 18.136, pp. 1–46. ISSN: 1533-7928. URL: http://jmlr.org/papers/v18/16-634.html.

Wirth, Christian and Johannes Fürnkranz (2013a). “A Policy Iteration Algorithm for Learning from Preference-Based Feedback”. In: *Advances in Intelligent Data Analysis (IDA)*. Springer, pp. 427–437. DOI: 10.1007/978-3-642-41398-8_37.

– (2013b). “EPMC: Every Visit Preference Monte Carlo for Reinforcement Learning”. In: *Proceedings of the Asian Conference on Machine Learning (ACML)*. PMLR, pp. 483–497. URL: https://proceedings.mlr.press/v29/Wirth13.html.

Wirth, Christian, Johannes Fürnkranz, and Gerhard Neumann (2016). “Model-Free Preference-Based Reinforcement Learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 2222–2228. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12247.

Wu, Runzhe and Wen Sun (2023). *Making RL with Preference-based Feedback Efficient via Randomization*. arXiv: 2310.14554. preprint.

Wu, Xingjiao, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He (2022). “A Survey of Human-in-the-Loop for Machine Learning”. In: *Future Generation Computer Systems* 135, pp. 364–381. DOI: 10.1016/j.future.2022.05.014.

Wulfe, Blake, Logan Michael Ellis, Jean Mercat, Rowan Thomas McAllister, and Adrien Gaidon (2022). “Dynamics-Aware Comparison of Learned Reward Functions”. In: *Proceedings of International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=CALFyKVs87.

Xiao, Baicen, Qifan Lu, Bhaskar Ramasubramanian, Andrew Clark, Linda Bushnell, and Radha Poovendran (2020). “FRESH: Interactive Reward Shaping in High-Dimensional State Spaces Using Human Feedback”. In: *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. IFAAMAS, pp. 1512–1520. DOI: 10.5555/3398761.3398935.

Xie, Annie, Avi Singh, Sergey Levine, and Chelsea Finn (2018). "Few-Shot Goal Inference for Visuomotor Learning and Planning". In: *Proceedings of the Conference on Robot Learning (CoRL)*. PMLR, pp. 40–52. URL: https://proceedings.mlr.press/v87/xie18a.html.

Xu, Jiazheng, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong (2023). "ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation". In: Conference on Neural Information Processing Systems (NeurIPS). URL: https://openreview.net/forum?id=JVzeOYEx6d.

Xu, Yichong, Ruosong Wang, Lin Yang, Aarti Singh, and Artur Dubrawski (2020). "Preference-Based Reinforcement Learning with Finite-Time Guarantees". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. Curran Associates, Inc., pp. 18784–18794. URL: https://proceedings.neurips.cc/paper/2020/hash/d9d3837ee7981e8c064774da6cdd98bf-Abstract.html.

Xue, Wanqi, Bo An, Shuicheng Yan, and Zhongwen Xu (2023a). *Reinforcement Learning from Diverse Human Preferences*. arXiv: 2301.11774. preprint.

Xue, Wanqi, Qingpeng Cai, Zhenghai Xue, Shuo Sun, Shuchang Liu, Dong Zheng, Peng Jiang, Kun Gai, and Bo An (2023b). "PrefRec: Recommender Systems with Human Preferences for Reinforcing Long-term User Engagement". In: *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. Association for Computing Machinery, pp. 2874–2884. DOI: 10.1145/3580305.3599473.

Yannakakis, Georgios N. and Héctor P. Martínez (2015). "Ratings Are Overrated!" In: *Frontiers in ICT* 2, p. 13. DOI: 10.3389/fict.2015.00013.

Yue, Yisong and Thorsten Joachims (2009). "Interactively Optimizing Information Retrieval Systems as a Dueling Bandits Problem". In: *Proceedings of the International Conference on Machine Learning (ICML)*. Association for Computing Machinery, pp. 1201–1208. DOI: 10.1145/1553374.1553527.

Zhan, Huixin, Feng Tao, and Yongcan Cao (2021). "Human-Guided Robot Behavior Learning: A GAN-Assisted Preference-Based Reinforcement Learning Approach". In: *IEEE Robotics and Automation Letters* 6.2, pp. 3545–3552. DOI: 10.1109/LRA.2021.3063927.

Zhan, Wenhao, Masatoshi Uehara, Nathan Kallus, Jason D. Lee, and Wen Sun (2023a). "Provable Offline Reinforcement Learning with Human Feedback". In: ICML 2023 Workshop on Interactive Learning with Implicit Human Feedback. URL: https://openreview.net/forum?id=fffH7DRz9X.

Zhan, Wenhao, Masatoshi Uehara, Wen Sun, and Jason D. Lee (2023b). "How to Query Human Feedback Efficiently in RL?" In: ICML 2023 Workshop on The Many Facets of Preference-Based Learning. URL: https://openreview.net/forum?id=kW6siW4EB6.

Zhang, David, Micah Carroll, Andreea Bobu, and Anca Dragan (2022). *Time-Efficient Reward Learning via Visually Assisted Cluster Ranking*. arXiv: 2212.00169. preprint.

Zhang, Guoxi and Hisashi Kashima (2023a). "Learning State Importance for Preference-Based Reinforcement Learning". In: *Machine Learning*. DOI: 10.1007/s10994-022-06295-5.

Zhang, Ruohan, Faraz Torabi, Garrett Warnell, and Peter Stone (2021). "Recent Advances in Leveraging Human Guidance for Sequential Decision-Making Tasks". In: *Autonomous Agents and Multi-Agent Systems* 35.2, p. 31. DOI: 10.1007/s10458-021-09514-w.

Zhang, Tianjun, Fangchen Liu, Justin Wong, Pieter Abbeel, and Joseph E. Gonzalez (2023b). "The Wisdom of Hindsight Makes Language Models Better Instruction Followers". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 41414–41428. URL: https://proceedings.mlr.press/v202/zhang23ab.html.

Zhao, Yao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu (2023). *SLiC-HF: Sequence Likelihood Calibration with Human Feedback*. arXiv: 2305.10425. preprint.

Zhou, Wangchunshu and Ke Xu (2020). "Learning to Compare for Better Training and Evaluation of Open Domain Natural Language Generation Models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 9717–9724. DOI: 10.1609/aaai.v34i05.6521.

Zhu, Banghua, Michael Jordan, and Jiantao Jiao (2023). "Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons". In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, pp. 43037–43067. URL: https://proceedings.mlr.press/v202/zhu23f.html.

Ziebart, Brian D., Andrew Maas, J. Andrew Bagnell, and Anind K. Dey (2008). "Maximum Entropy Inverse Reinforcement Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 1433–1438. ISBN: 978-1-57735-368-3. URL: https://cdn.aaai.org/AAAI/2008/AAAI08-227.pdf.

Ziegler, Daniel M., Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving (2020). *Fine-Tuning Language Models from Human Preferences*. arXiv: 1909.08593. preprint.