Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Survey Track

Reward Models in Deep Reinforcement Learning: A Survey

Rui Yu , Shenghua Wan , Yucen Wang , Chen-Xiao Gao ,
Le Gan , Zongzhang Zhang , De-Chuan Zhan

National Key Laboratory for Novel Software Technology, Nanjing University, China
School of Artificial Intelligence, Nanjing University, China
{yur,wansh,wangyc,gaocx}@lamda.nju.edu.cn, {ganle,zzzhang,zhandc}@nju.edu.cn

Abstract

In reinforcement learning (RL), agents continually interact with the environment and use the feedback to refine their behavior. To guide policy optimization, reward models are introduced as proxies of the desired objectives, such that when the agent maximizes the accumulated reward, it also fulfills the task designer's intentions. Recently, significant attention from both academic and industrial researchers has focused on developing reward models that not only align closely with the true objectives but also facilitate policy optimization. In this survey, we provide a comprehensive review of reward modeling techniques within the deep RL literature. We begin by outlining the background and preliminaries in reward modeling. Next, we present an overview of recent reward modeling approaches, categorizing them based on the source, the mechanism, and the learning paradigm. Building on this understanding, we discuss various applications of these reward modeling techniques and review methods for evaluating reward models. Finally, we conclude by highlighting promising research directions in reward modeling. Altogether, this survey includes both established and emerging methods, filling the vacancy of a systematic review of reward models in current literature.

1 Introduction

In recent years, deep reinforcement learning (DRL), a machine learning paradigm that combines RL with deep learning, has demonstrated its immense potential in applications across various domains. For example, AlphaGo [Silver *et al.*, 2016] showcased RL's capability of complex decision-making in game scenarios; InstructGPT [Ouyang *et al.*, 2022] marked the irreplaceable role of RL in aligning language models with human intents; agents trained via large-scale RL, such as OpenAI-o1 and DeepSeek-R1 [Guo *et al.*, 2025], demonstrated impressive reasoning intelligence that is comparable or even exceeds human capability. Unlike supervised learning (SL) where the agent is required to imitate and replicate the behavior recorded in the dataset, RL sets itself apart by enabling the agent to explore, adapt, and optimize its behavior based on the outcome of its actions, thereby achieving unprecedented levels of autonomy and capability.

A key component of reinforcement learning is the **reward**, which essentially defines the goal of interest in the task and guides the agents to optimize their behavior toward that intent [Sutton *et al.*, 1998]. Just as dopamine motivates and reinforces adaptive actions in biological systems, rewards in RL encourage exploration of the environment and guide intelligent agents towards desired behaviors [Glimcher, 2011]. However, while rewards are typically predefined in research environments [Towers *et al.*, 2024], they are often absent or difficult to specify in many real-world applications. In light of this, a significant portion of modern RL research focuses on how to extract effective rewards from various types of feedback, after which standard RL algorithms can be applied to optimize the policies of agents.

Despite the crucial role of reward modeling in RL, existing surveys [Arora and Doshi, 2021; Kaufmann *et al.*, 2023] are often embedded within specific subdomains such as inverse reinforcement learning (IRL) and reinforcement learning from human feedback (RLHF), with a limited focus on reward modeling as a standalone topic. To bridge this gap, we provide a systematic review of reward models, covering their foundations, key methodologies, and applications across diverse RL settings. We introduce a new categorization framework that addresses three fundamental questions: (1) *The source*: Where does the reward come from? (2) *The mechanism*: What drives the agent's learning? (3) *The learning paradigm*: How to learn the reward model from various types of feedback? Furthermore, we highlight recent advancements in reward models based on foundation models, such as large language models (LLMs) and vision-language models (VLMs), which have received relatively little attention in previous surveys. The framework of reward modeling we establish in this survey is illustrated in Figure 1. Specifically, this survey is organized as follows:

1. **Background of reward modeling (Section 2).** We first provide the necessary background on RL and reward models;
2. **Categorization of reward models.** We propose a classification framework for reward models, distinguishing them by three key factors: the *source* (**Section 3**), the *mechanism* that drives learning **(Section 4)**, and the

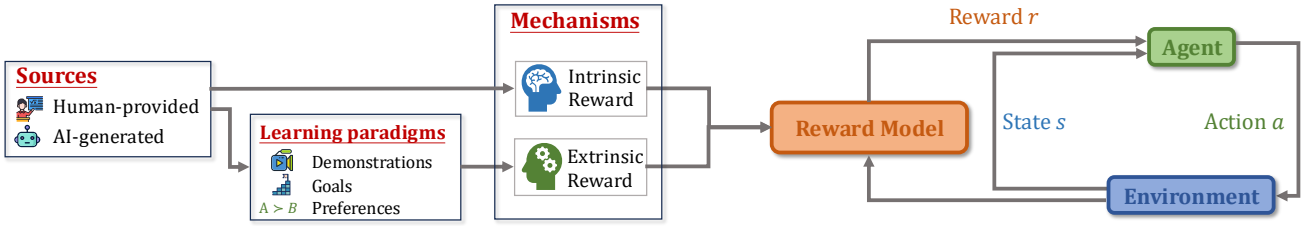

Figure 1: A framework for reward modeling in RL, categorizing reward models by their sources, feedback types, and mechanisms to provide a structured understanding of how rewards are derived and utilized in RL systems.

learning paradigm used to derive rewards **(Section 5)**. We also list recent publications about reward modeling and categorize them based on our hierarchy in Table 1.

3. **Applications and evaluation methods of reward models (Section 6 and Section 7).** We provide a discussion on the applications of reward models in practical scenarios, together with evaluation methods for these models.
4. **Prosperous directions and discussions (Section 8).** We summarize this survey by presenting potential future directions in this topic.

2 Background

RL is typically formulated as a Markov Decision Process (MDP) $\langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state space and the action space, respectively. The transition function $T(\cdot|s, a)$ defines the distribution over the next states after taking action a at state s. The reward model $R(s, a)$ specifies the instantaneous reward that the agent will receive after taking action a at state s, and γ is the discount factor that balances the importance of future rewards. An RL agent aims to find the policy $\pi(a|s)$ maximizing the following expected discounted cumulative reward (a.k.a. return):

$$\mathcal{J}(\pi) = \mathbb{E}_{\pi,T}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]. \tag{1}$$

where the expectation is taken over the distribution of states and actions that the agent will encounter following π and T.

The fundamental objective of learning is to refine an agent's behavior to accomplish predefined goals or tasks. While supervised learning (SL) offers a principled approach by training agents on human-annotated datasets to mimic human behavior, this method is limited by the quantity and quality of available human demonstrations. Consequently, agents trained solely by SL may make irrational decisions when human behavior is missing or sub-optimal. Reinforcement learning instead offers another principled way that permits the agent to explore the environment autonomously and adapt its behavior based on the rewards it receives. Such a trial-and-error approach exempts the agent from the constraints of datasets and opens the possibility of achieving or even surpassing human-level performance.

Although $\mathcal{S}, \mathcal{A}$, and the transition model T are inherently defined by the environment, the reward model R must be carefully crafted by the task designer. This careful design is crucial to ensure that the specified rewards truly reflect the underlying objectives. In many applications, only descriptive guidelines or standards of the intended goals are available, and therefore we need to convert them into statistical reward models. This process is termed as *reward modeling* throughout this survey.

3 Sources of Rewards

In this section, we explore different sources of reward signals in RL. We categorize reward sources into two main types: human-provided rewards, which leverage human expertise and supervision, and AI-generated rewards, which rely on foundation models typically trained by self-supervised learning on internet-scale datasets.

3.1 Human-Provided Rewards

Manual Reward Engineering

Manual reward engineering refers to the process where researchers meticulously design reward functions to steer agents toward optimal policies. Take the walker task in Gym-MuJoCo [Towers *et al.*, 2024] as an example: its reward is manually designed as a combination of survival, forward movement, and control cost penalties. However, reward engineering requires human experts to translate ambiguous task objectives into precise statistical models. Such an undertaking can be both resource-intensive and perilous: if the reward function is inadequately crafted, the agent may suffer from reward hacking, leading to unpredictable behaviors [Kaufmann *et al.*, 2023].

Human-in-the-Loop Reward Learning

Instead of directly crafting the reward models, human-in-the-loop reward learning derives rewards from indirect human supervision, including demonstrations [Abbeel and Ng, 2004], goals [Liu *et al.*, 2022], and preferences [Kaufmann *et al.*, 2023]. Compared to manual reward engineering, asking human experts to provide demonstrations or feedback of such kind is much more straightforward. However, the reward learning process needs to be specifically designed to accommodate different kinds of supervision and ensure alignment with the intended task objectives.

3.2 AI-Generated Rewards

Foundation models, such as large language models (LLMs) and vision-language models (VLMs) pre-trained on internet-scale human-generated data, have demonstrated a remarkable

ability to interpret human intent and autonomously define reward models for RL. For instance, LLMs have been employed to design reward functions [Xie *et al.*, 2023] and generate feedback for reward learning [Klissarov *et al.*, 2023; Bai *et al.*, 2022; Lee *et al.*, 2024]. VLMs, in particular, are highly effective in specifying rewards and tasks within visually complex environments. Some studies [Fan *et al.*, 2022; Sontakke *et al.*, 2023] compute semantic similarity between agent states and task descriptions, enabling dense reward signals from visual observations. Others [Wang *et al.*, 2024] utilize VLMs to analyze visual inputs and generate preference-based feedback for reward model training. While certain approaches [Baumli *et al.*, 2023] leverage off-the-shelf foundation models for zero-shot reward specification, others [Fan *et al.*, 2022; Sontakke *et al.*, 2023] fine-tune these models on domain-specific datasets to improve reward design.

4 Reward Mechanisms

In this section, we focus on two different reward mechanisms that drive RL agent's learning.

4.1 Extrinsic Reward

Rewards are defined by incentives that drive the agent. The term *extrinsic reward* corresponds to incentives that arise from external sources and directly relate to the desired task objective, e.g., instructions or goals set by supervisors or employers. Defining extrinsic rewards requires the task designer to translate abstract goals into concrete, quantifiable rewards that can be incorporated into a standard RL pipeline. The approach to accomplish this is detailed in Section 5.

4.2 Intrinsic Motivation

In contrast to extrinsic rewards, intrinsic motivation (IM) captures an agent's innate motivation to explore and refine its behavior in the environment [Ryan and Deci, 2000]. [1950] observed that even without an extrinsic stimulus, monkeys have a spontaneous desire and curiosity to solve complex puzzles. Later [Barto *et al.*, 2004] introduced IM into the reward mechanism, leading to the application of intrinsic reward. Unlike extrinsic rewards, intrinsic rewards are often disentangled from specific task objectives; rather, they encapsulate the encouragement for beneficial behaviors for problem-solving, such as exploration.

To coordinate the intrinsic reward and extrinsic reward, one common approach is to compute the agent's reward r as a weighted sum of the intrinsic reward r_{int} and the extrinsic reward r_{ext}:

$$r = \lambda r_{\text{int}} + (1 - \lambda) r_{\text{ext}}, \tag{2}$$

where $0 \leq \lambda \leq 1$ is a coefficient that balances the intrinsic reward r_{int} and extrinsic reward r_{ext}.

Next, we introduce three widely used types of intrinsic motivation in reinforcement learning.

Exploration

IM has long been used to encourage exploration. By leveraging concepts such as surprise [Pathak *et al.*, 2017], epistemic uncertainty [Houthooft *et al.*, 2016], and disagreement [Pathak *et al.*, 2019; Sekar *et al.*, 2020], many methods quantify the strangeness of states as the prediction errors of state transition, and thus use the errors as intrinsic rewards to encourage the agent to explore unseen areas of the environment. The strangeness of states can also be quantified using the distillation error between randomly initialized networks [Burda *et al.*, 2018], which can be more flexible to implement.

Other works design intrinsic rewards for exploration through the lens of data diversity. Among them, count-based methods, such as the well-known upper confidence bound (UCB) [Lai and Robbins, 1985], maintain the state visitation counts and assign higher intrinsic rewards for less-visited states. Later, static hashing [Tang *et al.*, 2017] and density estimation [Bellemare *et al.*, 2016; Ostrovski *et al.*, 2017] are incorporated to extend count-based exploration to problems with larger or even continuous state spaces. On the other hand, [2021] and [2020] promote diversity by estimating the data entropy and using the entropy as the intrinsic rewards. In this way, they can encourage the agent to explore novel and diverse states.

Empowerment

Empowerment, an information-theoretic intrinsic motivation (IM) concept, motivates an agent to maximize its influence on the environment by seeking states where it possesses the greatest control over future outcomes [Klyubin *et al.*, 2005]. An intrinsic reward signal can then be formulated to guide the agent's exploration towards states that offer greater control and a wider diversity of achievable consequences. Many previous works leverage empowerment for skill discovery [Eysenbach *et al.*, 2018; Mazzaglia *et al.*, 2022]. These works aim to find a skill-conditioned policy $\pi(a|s, z)$ that maximizes the mutual information between the resulting trajectory and the latent variable z. The intrinsic reward is designed based on the decomposition of this mutual information. The agent is then encouraged to recover the latent z from the trajectory, implying that different z should produce distinctly different trajectories, thereby defining z as the skill. By providing an intrinsic reward based on the agent's potential to influence the environment, skill learning through empowerment enables more generalizable agent behaviors and facilitates rapid adaptation to new tasks.

Knowledge-Driven IM

Many approaches leverage high-level knowledge and structured reasoning to generate intrinsic rewards, bridging the gap between abstract understanding and low-level sensorimotor interactions. Some methods derive preferences from structured event descriptions, comparing pairs of observations to infer meaningful intrinsic signals [Klissarov *et al.*, 2023]. [2023] adopted a reward-shaping technique by treating valuable propositional logic knowledge as intrinsic rewards for the RL procedure. [2023] generates goal candidates based on an agent's current context and provides rewards for achieving those inferred objectives. In recent work [Klissarov *et al.*, 2023], large-scale models such as LLMs and VLMs have been employed to facilitate this process due to their broad knowledge and reasoning capabilities.

Source	Mechanism	Feedback	Method
human	intrinsic	-	[Pathak *et al.*, 2017; Houthooft *et al.*, 2016; Pathak *et al.*, 2019; Sekar *et al.*, 2020; Burda *et al.*, 2018; Bellemare *et al.*, 2016; Badia *et al.*, 2020; Liu and Abbeel, 2021; Eysenbach *et al.*, 2018; Mazzaglia *et al.*, 2022; Wan *et al.*, 2024]
AI	intrinsic	-	[Klissarov *et al.*, 2023; Xu *et al.*, 2023; Du *et al.*, 2023]
human	extrinsic	demonstration	[Abbeel and Ng, 2004; Ziebart *et al.*, 2008; Finn *et al.*, 2016a; Finn *et al.*, 2016b; Fu *et al.*, 2017; Jeon *et al.*, 2020]
human	extrinsic	goal	[Liu *et al.*, 2022; Nachum *et al.*, 2018; Mazzaglia *et al.*, 2024; Hartikainen *et al.*, 2019; Mendonca *et al.*, 2021; Park *et al.*, 2023; Myers *et al.*, 2024; Wang *et al.*, 2025]
AI	extrinsic	goal	[Sontakke *et al.*, 2023; Fan *et al.*, 2022; Rocamonde *et al.*, 2023]
human	extrinsic	preference	[Christiano *et al.*, 2017; Kim *et al.*, 2023; Verma and Metcalf, 2024; Knox *et al.*, 2022; Touvron *et al.*, 2023; Liu *et al.*, 2024a; Ouyang *et al.*, 2022; Köpf *et al.*, 2023; Rafailov *et al.*, 2023; Song *et al.*, 2024; Liu *et al.*, 2024b]
AI	extrinsic	preference	[Bai *et al.*, 2022; Lee *et al.*, 2024; Wang *et al.*, 2024]

Table 1: Summary of the algorithms mentioned in Section 3, Section 4, and Section 5.

5 Learning Paradigms

In this section, we focus on the paradigms of learning the reward model R_θ from different kinds of human feedback. Specifically, existing literature that involves reward learning can be broadly categorized into three paradigms, namely:

- **Learning from demonstrations,** which extracts reward models based on demonstrations provided by human experts. This is related to inverse RL (IRL) [Arora and Doshi, 2021].
- **Learning from goals,** which derives reward models from specified goal states. This is related to goal-conditional RL (GCRL) [Liu *et al.*, 2022].
- **Learning from preferences,** which extracts reward models from human preferences among two or more trajectory segments. This is related to preference-based RL (PbRL) and reinforcement learning from human feedback (RLHF) [Kaufmann *et al.*, 2023].

In each subsection, we will provide a brief overview of the established methods in each setting.

5.1 Learning from Demonstrations

Maximum-Entropy Inverse Reinforcement Learning

Previous approaches to IRL iteratively optimize the reward model to maximize the performance margin between demonstrations and any other policy, such that the demonstrations appear optimal under the learned reward model [Abbeel and Ng, 2004]. However, the IRL problem is inherently ill-posed, because multiple distinct rewards may explain the same expert behavior. A common strategy for resolving this ambiguity is to incorporate additional regularization into the learning objective. As an example, the maximum-entropy IRL (MaxEnt-IRL) framework [Ziebart *et al.*, 2008] introduces entropy regularization such that the expert demonstrations are drawn from the Boltzmann distribution:

$$p_\theta(\tau) = \frac{\exp(R_\theta(\tau))}{Z_\theta}, \tag{3}$$

where $\tau = (s_1, a_1, \ldots, s_{|\tau|}, a_{|\tau|})$ denotes the demonstrated trajectory, and $R_\theta(\tau) = \sum_{t=1}^{|\tau|} R_\theta(s_t, a_t)$ is the cumulative reward along τ. The partition function Z_θ normalizes the distribution, and it can be computed via dynamic programming in small, discrete domains [Ziebart *et al.*, 2008] or approximated by importance sampling in continuous settings [Finn *et al.*, 2016b]. By parameterizing the reward model R_θ as linear models or neural networks, we can perform maximum likelihood training based on observed demonstrations and obtain the reward models that explain the demonstrations.

Adversarial Reward Learning

[2016a] demonstrated that the MaxEnt-IRL problem can be reformulated as a generative adversarial network (GAN) problem by employing a specifically structured discriminator. Let the generator of the trajectories and the reward model be $q_\psi(\tau)$ and $R_\theta(\tau)$ respectively, the discriminator is parameterized as:

$$D_\theta(\tau) = \frac{\frac{1}{Z}\exp(R_\theta(\tau))}{\frac{1}{Z}\exp(R_\theta(\tau)) + q_\psi(\tau)}, \tag{4}$$

where Z represents the partition function and can be estimated via importance sampling. The generator and the discriminator are trained via standard GAN losses:

$$\begin{aligned}
\mathcal{L}(\theta) &= \mathbb{E}_{\tau\sim\mathcal{D}_e}\left[-\log D_\theta(\tau)\right] + \mathbb{E}_{\tau\sim q}\left[-\log(1 - D_\theta(\tau))\right], \\
\mathcal{L}(\psi) &= \mathbb{E}_{\tau\sim q_\psi}\left[\log\frac{(1 - D_\theta(\tau))}{D_\theta(\tau)}\right] \\
&= \mathbb{E}_{\tau\sim q_\psi}\left[-R_\theta(\tau)\right] - \mathcal{H}(q_\psi) + \log Z,
\end{aligned} \tag{5}$$

where $\mathcal{D}_e$ denotes the expert demonstrations and $\mathcal{H}$ is the entropy. By optimizing (5), we can effectively optimize

the reward model R_θ. When the optimization converges, it follows from the maximum-entropy theory that $q^*(\tau) \propto \exp(R^*(\tau))$, which exactly recovers the MaxEnt-IRL problem in (3). However, conducting optimization over the trajectories incurs high variance, and therefore the adversarial inverse RL (AIRL) framework [Fu *et al.*, 2017] further decomposes the problem and operates on a state-action level:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathcal{D}_e}\left[-\log D_\theta(s,a)\right] + \mathbb{E}_{q_\psi}\left[-\log(1 - D_\theta(s,a))\right], \tag{6}$$

where $D_\theta(s,a) = \frac{\exp f_\theta(s,a)}{\exp(f_\theta(s,a)) + p_\psi(a|s)}$. Once training is complete, f_θ is shown to recover the optimal advantage function A^*, from which reward models may subsequently be extracted. Building on this foundation, the AIRL framework has been further extended – for instance, to encompass a broader class of regularizations [Jeon *et al.*, 2020] .

5.2 Learning from Goals

When our intended goals can be explicitly described or specified as a state $g \in \mathcal{S}$, the reward model can be conveniently defined based on whether the goal is achieved [Liu *et al.*, 2022]:

$$R(s,g) = \mathbb{1}(s \text{ accomplishes } g), \tag{7}$$

where $\mathbb{1}$ is the indicator function. However, this binary reward structure is extremely sparse and inefficient for policy optimization, because the agent only receives a reward upon reaching the goal state, without intermediate supervision. To address this sparsity, an alternative solution is to reshape the reward as the distance between the current and the desired goal:

$$R(s,g) = -d(\phi(s), \psi(g)), \tag{8}$$

where ϕ and ψ are mapping functions that transform the state s and the goal g to the same latent space, and $d(\cdot,\cdot)$ is a specific distance metric on that space. This distance-based reward provides a more nuanced measurement of the agent's progress toward the specified goal. In the below, we will introduce two commonly adopted distance metrics: spatial distance and temporal distance.

Spatial Distance

Spatial distance directly quantifies the similarity between states from the environment. Common approaches utilize measures such as the L2 distance [Nachum *et al.*, 2018], and cosine similarity [Mazzaglia *et al.*, 2024] to assess the proximity between states. These metrics may be computed either in the raw state space [Nachum *et al.*, 2018], or within a learned latent space [Mazzaglia *et al.*, 2024] which better captures and exploits the problem structure.

Temporal Distance

Other works focus on the notion of temporal distance, which conceptually assigns higher rewards to states that are *temporally* closer to the goal state. For instance, approaches like [Hartikainen *et al.*, 2019] and [2025] train a distance metric function d_θ, such that $d_\theta(s,g)$ approximates the number of time steps required for the agent to reach g from s. Using $R = -d_\theta$ as the reward model, the agent will be guided toward states that are in the proximity of the goal. Moreover, [Park *et al.*, 2023] frames temporal distance learning as a constrained optimization problem, maintaining a distance threshold between adjacent states while dispersing others. Recently, [Myers *et al.*, 2024] defines a temporal distance metric based on successor features and temporal contrastive learning, which is shown to satisfy the quasi-metric property. Temporal distance offers a more grounded reward signal by effectively reflecting the agent's progress toward the goal and capturing deeper task semantics beyond visual details.

Semantic Similarity

Semantic similarity-based rewards measure how closely the agent's current state aligns with a given goal in a shared representation space. RoboCLIP [Sontakke *et al.*, 2023] computes the reward as the dot product between the text embedding of a language-specified goal and the video embedding of the agent's observed trajectory. MineCLIP [Fan *et al.*, 2022] computes rewards as $R = \max\left(P_G - \frac{1}{N_T}, 0\right)$, where P_G is the probability of the observation video matching the goal description against negatives, and $\frac{1}{N_T}$ serves as a baseline to filter out uncertain estimates. These embeddings can be obtained from VLMs, which map multimodal inputs into a common space, allowing the agent to learn from high-level instructions or demonstrations.

5.3 Learning from Preferences

In many applications, obtaining human evaluations is comparatively cost-effective compared to collecting demonstrations or identifying the goal states. Consider training language models to follow instructions as an example, it is both tedious and time-consuming to require human annotators to generate template responses for every request. On the contrary, comparing agent-generated responses using metrics such as helpfulness, harmlessness, and truthfulness is considerably more straightforward. In this section, we therefore investigate methods for deriving rewards from human-annotated preferences among candidate options.

In this framework, annotators are asked to label their preferences y between a pair of trajectories (τ^0, τ^1), where $\tau = (s_1, a_1, \ldots, s_{|\tau|}, a_{|\tau|})$. A label $y = 0$ means τ^0 is preferred over τ^1 (denoted as $\tau^0 \succ \tau^1$), and $y = 1$ implies the opposite. To build the connection between observed preferences and reward models, we need *preference models*. A widely used example is *Bradley-Terry* (BT) models [Bradley and Terry, 1952], which posit that the probability of preference can be described by a Boltzmann distribution applied to the cumulative reward:

$$P_{\mathrm{BT}}(\tau^0 \succ \tau^1; \theta) = \frac{\exp(\sum_{(s_t^0, a_t^0) \in \tau^0} R_\theta(s_t^0, a_t^0))}{\sum_{j \in \{0,1\}} \exp(\sum_{(s_t^j, a_t^j) \in \tau^j} R_\theta(s_t^j, a_t^j))}. \tag{9}$$

To optimize the reward model R_θ, we can maximize the likelihood of the observed preferences:

$$\mathcal{L}(\theta) = - \sum_{(\tau^0, \tau^1, y) \in \mathcal{D}} (1-y) \log P(\tau^0 \succ \tau^1; \theta) + y \log P(\tau^1 \succ \tau^0; \theta), \tag{10}$$

where P is defined according to the preference model in (9). After training, we can label the reward of each transition pair and subsequently employ any RL algorithm to optimize the policies [Christiano *et al.*, 2017]. Alternatively, we can also directly train the policy via (10) by reparameterizing the reward model through the policy in certain circumstances [Rafailov *et al.*, 2023].

Preference Models

Despite its popularity in PbRL literature, BT models may not align with reality [Kim *et al.*, 2023]. Consequently, several studies have proposed alternative preference models that more closely reflect the mechanisms underlying human preferences. Preference Transformer [Kim *et al.*, 2023] introduces importance weights over state-action pairs to account for the dependence on certain critical states in the trajectory:

$$P_{\mathrm{PT}}(\tau^0 \succ \tau^1; \theta) = \frac{\exp(\sum_{(s_t^0, a_t^0) \in \tau^0} w_t^0 R_\theta(s_t^0, a_t^0))}{\sum_j \exp(\sum_{(s_t^j, a_t^j) \in \tau^j} w_t^j R_\theta(s_t^j, a_t^j))}, \tag{11}$$

where $j \in \{0, 1\}$ and the weights w_t^j are the average attention weights of the pair $(s_t^j, a_t^j) \in \tau^j$ calculated by a bi-directional attention layer. Similarly, [2024] replaced weights in (11) with attention weights from a transformer-based transition model, thereby incorporating state importance priors from the perspective of transition models. Besides, the *regret-based* models [Knox *et al.*, 2022] propose to model human preferences by the sum of optimal advantages along the trajectory, rather than the rewards:

$$\begin{aligned} P_{\mathrm{Reg}}(\tau^0 \succ \tau^1) &= \frac{\exp(-\mathrm{Regret}(\tau^0))}{\exp(-\mathrm{Regret}(\tau^0)) + \exp(-\mathrm{Regret}(\tau^1))}, \\ \mathrm{Regret}(\tau) &= \sum_{t=1}^{|\tau|} [Q_R^*(s_t, a_t) - V_R^*(s_t)], \end{aligned} \tag{12}$$

with V_R^* and Q_R^* being the optimal state value function and Q-value function for the reward model R, respectively. [2022] demonstrated that this approach may better predict real human preference and the learned reward model may achieve superior performance in practice.

Extension to Ordinal Feedback

Ordinal feedback generalizes binary feedback by requiring annotators to additionally specify the strengths of their preferences (e.g., *slightly better* or *significantly better*). To integrate this more nuanced information, existing studies modify BT models by incorporating soft margins [Touvron *et al.*, 2023] or soft labels $y_i \in [0, 1]$ [Liu *et al.*, 2024a], where the margin or the label reflects the strength of the preference.

Beyond Pairwise Comparisons

Human feedback can also be provided in the form of rankings among multiple candidates [Ouyang *et al.*, 2022; Köpf *et al.*, 2023]. Although such listwise comparisons put a greater burden on annotators, they also carry richer information than pairwise comparisons. To accommodate rankings, *Plackett-Luce* (PL) models [Plackett, 1975] generalize BT models by extending the comparison to K candidates:

$$\begin{aligned} &P_{\mathrm{PL}}(\tau^1 \succ \tau^2 \succ \ldots \succ \tau^K) \\ &= \prod_{k=1}^{K} \frac{\exp(\sum_{(s_t^k, a_t^k) \in \tau^k} R(s_t^k, a_t^k))}{\sum_{j=k}^{K} \exp(\sum_{(s_t^j, a_t^j) \in \tau^j} R(s_t^j, a_t^j))}, \end{aligned} \tag{13}$$

where $(\tau^1 \succ \tau^2 \succ \ldots \succ \tau^K)$ is the observed ranking. Substituting (13) into (10) yields the objective of learning rewards from rankings [Rafailov *et al.*, 2023; Song *et al.*, 2024]. Another straightforward approach to rankings is breaking the ranking into pairs by selecting two candidates from the list and assigning the label according to their ranks, thereby reducing the problem of applying BT models to all possible pairwise comparisons [Ouyang *et al.*, 2022; Liu *et al.*, 2024b].

6 Applications

Reward model designing constitutes an indispensable step before any practical applications of RL. Therefore in this section, we briefly review successful applications of reward models in deep RL, including control problems, generative model finetuning, and other fields.

6.1 Control Problems

Reward models play a pivotal role in control problems, as a fundamental mechanism for guiding decision-making in dynamic environments. [Christiano *et al.*, 2017] demonstrated their effectiveness in facilitating policy learning across diverse domains, including game-playing and simulated continuous control tasks. In gameplay scenarios, [Fan *et al.*, 2022] leveraged generated rewards to enhance learning in Minecraft tasks. In robotics, [Sontakke *et al.*, 2023] employed reward models to train agents across various robotic tasks. Similarly, in autonomous driving, the design of reward functions remains a critical aspect of training intelligent agents [Knox *et al.*, 2023].

6.2 Generative Model Post-training

Modern generative models typically feature a two-stage training procedure, where the *pre-training* stage involves unsupervised learning on internet-scale data, and the *post-training* stage fine-tunes the models and fits them for downstream tasks. A prominent example is InstructGPT [Ouyang *et al.*, 2022], which employs RL to optimize model outputs based on human preference data. Specifically, it trains a reward model on human-ranked responses and fine-tunes the language model to maximize this reward. This approach has become a standard method for enhancing the helpfulness, harmlessness [Dai *et al.*, 2023], and general task-solving capabilities of LLMs [Abramson *et al.*, 2022]. In mathematical problem-solving, golden rewards can be defined by comparing the model-generated answers with ground-truth answers [Luong *et al.*, 2024] or by verifying the correctness using formal solvers [Xin *et al.*, 2024]. Some works also use LLM-based verifiers [Zhang *et al.*, 2024], further leveraging the in-context learning ability provided by LLMs.

6.3 Other Fields

In recommendation systems, [2023] trained reward models to allow RL recommendation systems to learn from users' historical behaviors. [2020] used a reward model to automate peer-to-peer (P2P) energy trading, while [2019] designed a reward model to improve the management of healthcare resources.

7 Evaluating Reward Models

Once reward models are developed, reliable evaluation techniques are essential for comparing or selecting models for downstream policy optimization. However, due to the ambiguous link between reward models and final policy performance, relying on a single evaluation perspective is often insufficient [Arora and Doshi, 2021]. We categorize commonly used reward evaluation techniques into the following three types, which are often used in combination to achieve a more comprehensive assessment of reward models.

7.1 Evaluation via Policy Performance

Reward model quality can be evaluated by measuring the performance of policies trained with it. Primary metrics include ground-truth reward, task success rate, and training efficiency, with superior reward models yielding higher values across these measures. This approach is widely adopted in reinforcement learning literature to assess the alignment between reward models and actual objectives [Christiano *et al.*, 2017]. However, these metrics are sensitive to policy optimization algorithms and environmental stochasticity, potentially limiting their ability to independently reflect the true performance of the reward model itself.

7.2 Evaluation via Distance Metrics

To evaluate and compare reward models, another approach is to design distance metrics that accurately reflect the behavioral differences between the policies induced by these rewards. The pioneering work EPIC [Gleave *et al.*, 2020] introduces *canonically shaped rewards* to remove ambiguity and invariances from reward models and proposes to use the Pearson coefficient between two canonically shaped rewards as a measure of the reward similarity. The EPIC distance between two reward models is demonstrated to upper-bound the performance difference between the induced policies. Lower EPIC distances to the ground truth reward indicate superior reward modeling capability.

Based on EPIC, [2022] further incorporates the dynamics information when considering the invariant reward shaping and introduces the DARD metric, which is more predictive and accurate in quantifying the differences in rewards. Furthermore, [2023] presented a general framework for designing such distance metrics. The STARC metrics provided in this framework are shown to induce both the upper bound and the lower bound of the performance differences, and any other metrics that possess the same property must be equivalent to the STARC metrics up to bilipschitz scaling. When datasets containing ground-truth rewards are available, distance metrics are particularly suitable for offline evaluation, circumventing the necessity for policy learning.

7.3 Evaluation via Interpretable Representations

Although the evaluation of reward models may not be straightforward, we can transform them equivalently into interpretable representations. [2022] proposed to transform reward models with potential-based shaping and visualize the shaped reward instead. Since potential-based shaping preserves the optimal policy, characteristics of the shaped reward may also apply to the original reward model. Alternatively, we can evaluate the reward model through the behavior of the induced policy [Rocamonde *et al.*, 2023].

8 Conclusions

Recently, reward models have become a highly motivating area of research, driven by both theoretical challenges and practical needs across various domains. We consider the development of reward models as a significant step before the application of RL to real-world problems, and we hope this survey can offer valuable insights for both researchers and practitioners. Although our study provides a comprehensive overview of the topic, the design and variations of reward models still extend beyond the scope of this discussion. Interested readers can also refer to other survey papers [Eschmann, 2021; Liu *et al.*, 2022; Arora and Doshi, 2021; Kaufmann *et al.*, 2023] that focus on RL subfields closely related to reward modeling.

8.1 Future Directions

Efficient and accurate reward modeling is a valuable research direction with significant application prospects. It combines increasingly mature technologies such as large models and diffusion models with reward design and generation in reinforcement learning to provide behavioral feedback for agents in perception, planning, decision-making, and navigation. Although there is no definitive conclusion on which route can achieve efficient reward modeling, research on various technologies in recent years has effectively promoted the development of this field. With the continuous development of machine learning and reinforcement learning, reward modeling has many valuable research directions in the future, including:

1. **Vectorized rewards**: Constructing vectorized rewards to replace scalarized single rewards, dynamically balancing multiple competitive reward signals to provide agents with more comprehensive feedback.
2. **Interpretating reward models**: Improving the transparency of reward functions and explaining the decision-making logic behind reward models.
3. **Ethical alignment and social value constraints**: Quantifying ethical principles and embedding them into reward functions while avoiding potential side effects during the optimization process.
4. **Reward foundation models**: Similar to constructing a general representation space, consider training a foundation reward model that can obtain general reward values based on diverse inputs (such as limb movements).

Acknowledgements

We thank the anonymous reviewers for their insightful feedback. This work was supported by the National Science and Technology Major Project (Grant No. 2022ZD0114805) and Young Scientists Fund of the National Natural Science Foundation of China (PhD Candidate) (Grant No. 624B200197).

References

[Abbeel and Ng, 2004] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[Abramson *et al.*, 2022] Josh Abramson, Arun Ahuja, Federico Carnevale, Petko Georgiev, Alex Goldin, Alden Hung, Jessica Landon, Jirka Lhotka, Timothy Lillicrap, Alistair Muldal, et al. Improving multimodal interactive agents with reinforcement learning from human feedback. *arXiv preprint arXiv:2211.11602*, 2022.

[Arora and Doshi, 2021] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.

[Badia *et al.*, 2020] Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andew Bolt, et al. Never give up: Learning directed exploration strategies. *arXiv preprint arXiv:2002.06038*, 2020.

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

[Barto *et al.*, 2004] Andrew G Barto, Satinder Singh, Nuttapong Chentanez, et al. Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd International Conference on Development and Learning*, volume 112, page 19. Citeseer, 2004.

[Baumli *et al.*, 2023] Kate Baumli, Satinder Baveja, Feryal Behbahani, Harris Chan, Gheorghe Comanici, Sebastian Flennerhag, Maxime Gazeau, Kristian Holsheimer, Dan Horgan, Michael Laskin, et al. Vision-language models as a source of rewards. *arXiv preprint arXiv:2312.09187*, 2023.

[Bellemare *et al.*, 2016] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

[Bradley and Terry, 1952] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[Burda *et al.*, 2018] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

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

[Dai *et al.*, 2023] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

[Du *et al.*, 2023] Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. Guiding pretraining in reinforcement learning with large language models. In *International Conference on Machine Learning*, pages 8657–8677. PMLR, 2023.

[Eschmann, 2021] Jonas Eschmann. Reward function design in reinforcement learning. *Reinforcement learning algorithms: Analysis and Applications*, pages 25–33, 2021.

[Eysenbach *et al.*, 2018] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

[Fan *et al.*, 2022] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362, 2022.

[Finn *et al.*, 2016a] Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852*, 2016.

[Finn *et al.*, 2016b] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning*, pages 49–58. PMLR, 2016.

[Fu *et al.*, 2017] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

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

[Glimcher, 2011] Paul W Glimcher. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108:15647–15654, 2011.

[Guo *et al.*, 2025] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[Harlow, 1950] Harry F Harlow. Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys. *Journal of comparative and physiological psychology*, 43(4):289, 1950.

[Hartikainen *et al.*, 2019] Kristian Hartikainen, Xinyang Geng, Tuomas Haarnoja, and Sergey Levine. Dynamical distance learning for semi-supervised and unsupervised skill discovery. *arXiv preprint arXiv:1907.08225*, 2019.

[Houthooft *et al.*, 2016] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

[Jenner and Gleave, 2022] Erik Jenner and Adam Gleave. Preprocessing reward functions for interpretability. *arXiv preprint arXiv:2203.13553*, 2022.

[Jeon *et al.*, 2020] Wonseok Jeon, Chen-Yang Su, Paul Barde, Thang Doan, Derek Nowrouzezahrai, and Joelle Pineau. Regularized inverse reinforcement learning. *arXiv preprint arXiv:2010.03691*, 2020.

[Kaufmann *et al.*, 2023] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 10, 2023.

[Kim and Lee, 2020] Jin-Gyeom Kim and Bowon Lee. Automatic p2p energy trading model based on reinforcement learning using long short-term delayed reward. *Energies*, 13(20):5359, 2020.

[Kim *et al.*, 2023] Changyeon Kim, Jongjin Park, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. Preference transformer: Modeling human preferences using transformers for rl. *arXiv preprint arXiv:2303.00957*, 2023.

[Klissarov *et al.*, 2023] Martin Klissarov, Pierluca D'Oro, Shagun Sodhani, Roberta Raileanu, Pierre-Luc Bacon, Pascal Vincent, Amy Zhang, and Mikael Henaff. Motif: Intrinsic motivation from artificial intelligence feedback. *arXiv preprint arXiv:2310.00166*, 2023.

[Klyubin *et al.*, 2005] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 ieee congress on evolutionary computation*, volume 1, pages 128–135. IEEE, 2005.

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

[Knox *et al.*, 2023] W Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. Reward (mis) design for autonomous driving. *Artificial Intelligence*, 316:103829, 2023.

[Köpf *et al.*, 2023] Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36:47669–47681, 2023.

[Lai and Robbins, 1985] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[Lee *et al.*, 2024] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *International Conference on Machine Learning*, pages 26874–26901. PMLR, 2024.

[Liu and Abbeel, 2021] Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pre-training. *Advances in Neural Information Processing Systems*, 34:18459–18473, 2021.

[Liu *et al.*, 2022] Minghuan Liu, Menghui Zhu, and Weinan Zhang. Goal-conditioned reinforcement learning: Problems and solutions. *arXiv:2201.08299*, 2022.

[Liu *et al.*, 2024a] Shang Liu, Yu Pan, Guanting Chen, and Xiaocheng Li. Reward modeling with ordinal feedback: Wisdom of the crowd. *arXiv preprint arXiv:2411.12843*, 2024.

[Liu *et al.*, 2024b] Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, et al. Lipo: Listwise preference optimization through learning-to-rank. *arXiv preprint arXiv:2402.01878*, 2024.

[Luong *et al.*, 2024] Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*, 2024.

[Mazzaglia *et al.*, 2022] Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt, Alexandre Lacoste, and Sai Rajeswar. Choreographer: Learning and adapting skills in imagination. *arXiv preprint arXiv:2211.13350*, 2022.

[Mazzaglia *et al.*, 2024] Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt, Aaron Courville, and Sai Rajeswar. Genrl: Multimodal-foundation world models for generalization in embodied agents. *Advances in neural information processing systems*, 37:27529–27555, 2024.

[Mendonca *et al.*, 2021] Russell Mendonca, Oleh Rybkin, Kostas Daniilidis, Danijar Hafner, and Deepak Pathak. Discovering and achieving goals via world models. *Advances in Neural Information Processing Systems*, 34:24379–24391, 2021.

[Myers *et al.*, 2024] Vivek Myers, Evan Ellis, Sergey Levine, Benjamin Eysenbach, and Anca Dragan. Learning to assist humans without inferring rewards. *arXiv preprint arXiv:2411.02623*, 2024.

[Nachum *et al.*, 2018] Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

[Ostrovski *et al.*, 2017] Georg Ostrovski, Marc G Bellemare, Aäron Oord, and Rémi Munos. Count-based exploration with neural density models. In *International conference on machine learning*, pages 2721–2730. PMLR, 2017.

[Oueida *et al.*, 2019] Soraia Oueida, Moayad Aloqaily, and Sorin Ionescu. A smart healthcare reward model for resource allocation in smart city. *Multimedia tools and applications*, 78:24573–24594, 2019.

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

[Park *et al.*, 2023] Seohong Park, Oleh Rybkin, and Sergey Levine. Metra: Scalable unsupervised rl with metric-aware abstraction. *arXiv preprint arXiv:2310.08887*, 2023.

[Pathak *et al.*, 2017] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

[Pathak *et al.*, 2019] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International conference on machine learning*, pages 5062–5071. PMLR, 2019.

[Plackett, 1975] Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

[Rafailov *et al.*, 2023] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[Rocamonde *et al.*, 2023] Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-language models are zero-shot reward models for reinforcement learning. *arXiv preprint arXiv:2310.12921*, 2023.

[Ryan and Deci, 2000] Richard M Ryan and Edward L Deci. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1):54–67, 2000.

[Sekar *et al.*, 2020] Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International conference on machine learning*, pages 8583–8592. PMLR, 2020.

[Silver *et al.*, 2016] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[Skalse *et al.*, 2023] Joar Skalse, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Adam Gleave, and Alessandro Abate. Starc: A general framework for quantifying differences between reward functions. *arXiv preprint arXiv:2309.15257*, 2023.

[Song *et al.*, 2024] Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998, 2024.

[Sontakke *et al.*, 2023] Sumedh Sontakke, Jesse Zhang, Séb Arnold, Karl Pertsch, Erdem Bıyık, Dorsa Sadigh, Chelsea Finn, and Laurent Itti. Roboclip: One demonstration is enough to learn robot policies. *Advances in Neural Information Processing Systems*, 36:55681–55693, 2023.

[Sutton *et al.*, 1998] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[Tang *et al.*, 2017] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Towers *et al.*, 2024] Mark Towers, Ariel Kwiatkowski, Jordan Terry, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv:2407.17032*, 2024.

[Verma and Metcalf, 2024] Mudit Verma and Katherine Metcalf. Hindsight priors for reward learning from human preferences. *arXiv preprint arXiv:2404.08828*, 2024.

[Wan *et al.*, 2024] Shenghua Wan, Hai-Hang Sun, Le Gan, and De-Chuan Zhan. Moser: learning sensory policy for task-specific viewpoint via view-conditional world model. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 5046–5054, 2024.

[Wang *et al.*, 2024] Yufei Wang, Zhanyi Sun, Jesse Zhang, Zhou Xian, Erdem Biyik, David Held, and Zackory Erickson. Rl-vlm-f: reinforcement learning from vision language foundation model feedback. In *Proceedings of the 41st International Conference on Machine Learning*, pages 51484–51501, 2024.

[Wang *et al.*, 2025] Yucen Wang, Rui Yu, Shenghua Wan, Le Gan, and De-Chuan Zhan. Founder: Grounding foundation models in world models for open-ended embodied decision making. In *International Conference on Machine Learning*, 2025.

[Wulfe *et al.*, 2022] Blake Wulfe, Ashwin Balakrishna, Logan Ellis, Jean Mercat, Rowan McAllister, and Adrien Gaidon. Dynamics-aware comparison of learned reward functions. *arXiv preprint arXiv:2201.10081*, 2022.

[Xie *et al.*, 2023] Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. Text2reward: Automated dense reward function generation for reinforcement learning. *arXiv preprint arXiv:2309.11489*, 2023.

[Xin *et al.*, 2024] Huajian Xin, ZZ Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, et al. Deepseek-prover-v1. 5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. *arXiv preprint arXiv:2408.08152*, 2024.

[Xu *et al.*, 2023] Jiacheng Xu, Chao Chen, Fuxiang Zhang, Lei Yuan, Zongzhang Zhang, and Yang Yu. Internal logical induction for pixel-symbolic reinforcement learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2825–2837, 2023.

[Xue *et al.*, 2023] Wanqi Xue, Qingpeng Cai, Zhenghai Xue, Shuo Sun, Shuchang Liu, Dong Zheng, Peng Jiang, Kun Gai, and Bo An. Prefrec: Recommender systems with human preferences for reinforcing long-term user engagement. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2874–2884, 2023.

[Zhang *et al.*, 2024] Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*, 2024.

[Ziebart *et al.*, 2008] Brian D Ziebart, Andrew Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd national conference on Artificial intelligence-Volume 3*, pages 1433–1438, 2008.