

# Mapping out the Space of Human Feedback for Reinforcement Learning: A Conceptual Framework

YANNICK METZ, University of Konstanz, Germany and ETH Zurich, Switzerland

DAVID LINDNER, ETH Zurich, Switzerland

RAPHAËL BAUR, ETH Zurich, Switzerland

MENNATALLAH EL-ASSADY, ETH Zurich, Switzerland

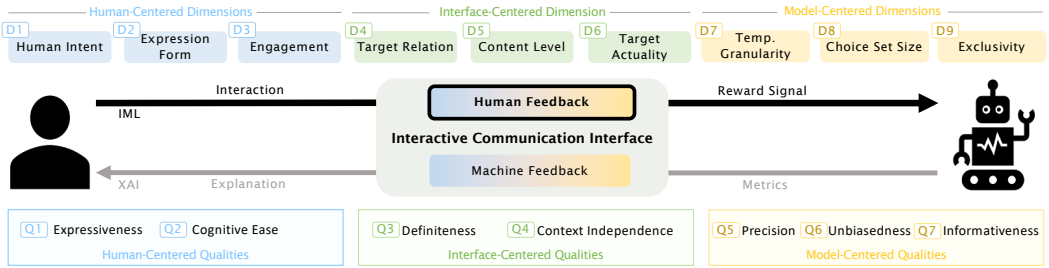


Fig. 1. In the context of *Reinforcement Learning*, we present a *conceptual framework* of human feedback based on nine dimensions: We identify human-centered, interface-centered and model-centered dimension. For different types of feedback, we also need to consider *human-centered*, *interface-centered* and *model-centered* qualities which influence how feedback can be collected and processed.

Reinforcement Learning from Human feedback (RLHF) has become a powerful tool to fine-tune or train agentic machine learning models. Similar to how humans interact in social contexts, we can use many types of feedback to communicate our preferences, intentions, and knowledge to an RL agent. However, applications of human feedback in RL are often limited in scope and disregard human factors. In this work, we bridge the gap between machine learning and human-computer interaction efforts by developing a shared understanding of human feedback in interactive learning scenarios. We first introduce a taxonomy of feedback types for reward-based learning from human feedback based on nine key dimensions. Our taxonomy allows for unifying human-centered, interface-centered, and model-centered aspects. In addition, we identify seven quality metrics of human feedback influencing both the human ability to express feedback and the agent's ability to learn from the feedback. Based on the feedback taxonomy and quality criteria, we derive requirements and design choices for systems learning from human feedback. We relate these requirements and design choices to existing work in interactive machine learning. In the process, we identify gaps in existing work and future research opportunities. We call for interdisciplinary collaboration to harness the full potential of reinforcement learning with data-driven co-adaptive modeling and varied interaction mechanics.

CCS Concepts: • **Computing methodologies** → *Reinforcement learning*; • **Human-centered computing** → *HCI theory, concepts and models*.

Authors' Contact Information: Yannick Metz, yannick.metz@uni-konstanz.de, University of Konstanz, Konstanz, Germany and ETH Zurich, Zurich, Switzerland; David Lindner, ETH Zurich, Zurich, Switzerland; Raphaël Baur, ETH Zurich, Zurich, Switzerland; Mennatallah El-Assady, ETH Zurich, Zurich, Switzerland.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM XXXX-XXXX/2025/2-ART

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

Additional Key Words and Phrases: Reinforcement learning, Human feedback, Human-in-the-loop learning

#### ACM Reference Format:

Yannick Metz, David Lindner, Raphaël Baur, and Mennatallah El-Assady. 2025. Mapping out the Space of Human Feedback for Reinforcement Learning: A Conceptual Framework. 1, 1 (February 2025), 65 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## 1 Introduction

The development of intelligent agents that can execute tasks, align with human values, and adhere to our preferences represents a significant long-term goal in artificial intelligence. Reinforcement Learning from Human Feedback (RLHF) and related approaches have emerged as promising tools for training agents, shaping them to follow instructions and align with user preferences [42, 79, 150].

While research on training intelligent agents with human input has a rich history—including approaches like imitation learning [78], inverse RL [145], and interactive RL [45] — RLHF specifically has gained prominence due to its success in fine-tuning language models [150]. Past research has explored diverse interaction modalities for human feedback, including demonstrations, evaluative and comparative feedback, and corrections. However, the field lacks a common classification of design dimensions for human feedback and a unified perspective that bridges human-computer interaction and machine learning.

Human feedback in real-world scenarios encompasses a wide array of communication strategies adapted to various contexts, tasks, and recipients [37]. In contrast, feedback commonly used in RLHF, in particular for the fine-tuning of generative models such as LLM, i.e., pairwise comparisons [42, 205] of outputs, represents just a very limited way to provide feedback. We therefore argue that the space for possible feedback in human-AI communication should be broadened. This paper aims to address two key research questions:

- What different kinds of human feedback exist?
- What constitutes good human feedback?

To answer the first question, we propose a taxonomy that structures human feedback along nine dimensions, linking these to the broad spectrum of existing work in human-AI communication. For the second question, we discuss seven quality metrics to evaluate the effectiveness of human feedback. The paradigm shift towards more expressive, targeted, and dynamic human feedback enables a flexible, multi-faceted training process for intelligent agents. In this new paradigm, humans may interact with agents at various stages of training and deployment while agents learn from diverse data sources, environments, and feedback types.

**Contributions** — Our contribution is threefold:

(i) We introduce a **taxonomy of nine dimensions** of human feedback towards AI agents, unifying model-centered, interface-centered, and user-centered aspects. This taxonomy provides a foundation for classifying existing feedback patterns and includes a formalization of human feedback.

(ii) We present **seven quality metrics** that summarize reported quality criteria from a broad range of literature, discussing their manifestations and measurement methods.

(iii) Based on the taxonomy and feedback qualities, we **derive requirements and design choices** for reinforcement learning systems integrating human feedback. We relate these to existing work across interactive machine learning and visual analytics, **identifying research opportunities** for dynamic multi-type human feedback.

## 2 Problem Statement and Background

We start by establishing the problem space and key concepts and review related work.

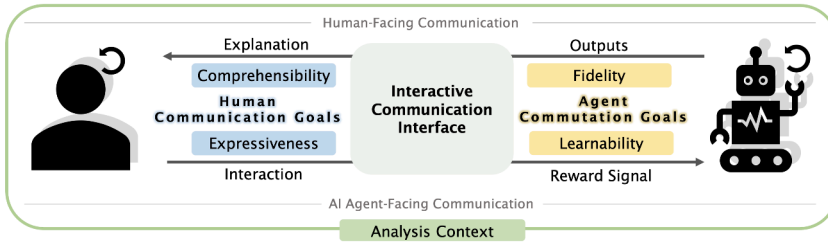


Fig. 2. Human-AI Interaction via Interactive Communication Interfaces: Modal Outputs are communicated to the humans via the interface and presented as comprehensible explanations that should preserve high fidelity. Human feedback is communicated via interactions that must be mapped to a reward signal suitable for RL training.

## 2.1 Problem Characterization

**Human-AI Communication Space** — In order to effectively bridge different fields like human-computer interaction (HCI) and machine learning, we develop our conceptual framework from first principles. We identify the key actors and goals in the human-AI communication process.

Throughout the paper, we will use two running examples: Due to its ubiquity in current RLHF research [17, 84, 150]; we will use the example of human feedback for a text generation model and outline potential feedback in such a use case. Especially beyond simple, verifiable, and single-objective generation tasks, including multi-step reasoning or domain-dependent task execution, more expressive feedback can be highly desirable.

Additionally, we consider a more advanced, future-oriented use case of a household robot deployed in a new household environment. Even if the robot has been pre-trained in general environments, e.g., via supervised [78] or fully-autonomous learning [89] in simulation, it needs to learn to act according to the personal preferences of its new owners. This includes stowing objects in desired locations, avoiding certain behaviors, and using location-specific tools. Ideally, we want intuitive ways to dynamically express our goals and preferences to the agent, such as demonstrating desired behavior or commenting on observed actions.

To achieve this type of teaching, expressive human-to-AI communication is crucial. Humans intuitively use various forms of feedback in social contexts [111], but these feedback utterances need to be translated for AI learning algorithms. Effective teaching also requires the learning agent to communicate back to the human teacher. We can, therefore, identify two directions for *Human-AI interaction*.

In Figure 2, we provide an overview of the *Human-AI communication space* [55]. The key elements are:

- **Humans** – Involved as teachers and potential users of the agent’s capabilities, humans have preferences about the agent’s behavior.
- **AI Agents** – Agents optimize their behavior using a reward model trained from human feedback.
- **H-AI Communication Interfaces** – A variety of interfaces exist, ranging from visual and numerical to natural language, gestures, or mimics [109].
- **Analysis Context** – The human-AI interaction is embedded in a surrounding context, determining available information and feedback communication methods. We must decide on a specific analysis context, which restricts the contextual information and stakeholders under consideration.

*Human feedback* is transmitted from human to agent. Humans benefit from high *expressiveness* (see section 4), allowing effective communication of internal knowledge, goals, and preferences. Conversely, agents require *learnable* input that can reasonably translate into model updates. Traditionally, the research field of interactive machine learning (IML) addresses this direction [6, 53]. In the opposite direction, agents provide *machine feedback* via model outputs or auxiliary data. These metrics are presented as explanations to humans. Machine feedback must balance *fidelity* and *comprehensibility*. High-fidelity outputs accurately describe model behavior [71], but this information must be comprehensible to humans via abstractions or filtering, which limits fidelity. This direction is addressed by explainable AI (XAI) [1]. We highlight these four high-level goals in Figure 1. In this work, we focus on the first direction.

The context of human-AI interaction directly influences human feedback [113]. Different types of context include the (*sociotechnical*) *personal context* (i.e., the human’s background) [113], the *task context* (the basis for the human-AI interaction, e.g., text generation vs. household tasks), and the *utterance context* (the immediate context within a potentially longer interaction process). Current RLHF approaches often fail to consider these contextual factors by not modeling them appropriately. Rich, diverse interactions via expressive human feedback can enable the capture of additional contextual information.

We use this space of human, interface, and model to categorize expressions of feedback and discuss feedback quality. We posit that human feedback can be categorized from a human perspective ("How can we categorize the human experience when giving different feedback?"), from an interface perspective ("Which types of interfaces accommodate different feedback?"), Moreover, from a model perspective ("How do we classify feedback for different reward models?"). Besides classifying feedback, we must also address the question "What makes for high-quality feedback?" from human, interface, and model perspectives.

## 2.2 Related Work

**Human-in-the-Loop Reinforcement Learning** — Human-in-the-loop reinforcement learning (HITL RL), also known as collaborative RL [109] or interactive RL [12], centers on incorporating human knowledge and feedback. Reinforcement learning, particularly in challenging environments, often suffers from high sample inefficiency and difficult credit assignments when reward signals are sparse. Human experts can leverage their prior knowledge about the environment to enhance training efficiency by guiding RL agents during exploration in large state spaces or when agents are stuck in local optima [104]. In HITL RL, human teachers use feedback to accelerate or improve an agent’s learning process. For instance, a human trainer might observe the learning agent’s behavior in an environment and provide positive or negative reward signals to inform the quality of its actions [86, 104, 131]. Existing surveys have primarily focused on input modalities and learning methods [12, 111]. We draw from work in this area and expand it to the domain of RLHF.

**Reinforcement Learning from Human Feedback** — In scenarios where no environment reward function is available, Reinforcement Learning from Human Feedback (RLHF), using pairwise preferences, has emerged as a viable learning technique [42]. Pairwise preferences have become so ubiquitous that RLHF is often synonymously with “RL from human *preferences*”. Existing surveys [36, 113] have highlighted limitations of current RLHF approaches, such as overly simplistic modeling of humans, misspecification, and limitations in expressiveness. Our paper lays out a framework to study and address these limitations by broadening the space of human feedback for RL systems.

**Existing Classifications of Feedback** — Another classification separates feedback into three categories [109, 111]:

*Explicit* feedback: Directly usable as reward model input and consciously given by a human observer. *Implicit* feedback: Indirectly processed information humans communicate, such as facial expressions or non-directed natural language. *Multi-modal* feedback: A combination of different feedback types, though this concept is not well-defined beyond using multiple methods simultaneously.

This classification, while helpful, is rather broad and lacks specificity in defining multi-modality.

Jeon et al. [80] present a unifying formalism for reward learning from various feedback sources. Their reward-rational (implicit) choice model represents different types of feedback—including demonstrations, corrections, comparisons, or direct rewards—as the human choosing an option from a choice set. This allows for a standard implementation of reward models across different feedback types. We extend this work by incorporating HCI aspects and developing a structured conceptual feedback framework.

A survey by Fernandes et al. [61] discusses a series of common formats for feedback in the context of natural language generation, including numerical and ranking feedback and natural language. The differentiation of feedback formats is only a minor part of the survey and is not integrated into a common shared framework or design space.

Concurrent work by Kaufmann et al. [84] presents a feedback classification along several dimensions but does not focus on HCI aspects and uses a more straightforward classification overall. Their classification is inspired by preliminary work [137].

### 3 Mapping out the Space of Human Feedback in Reinforcement Learning

While existing work in human-in-the-loop RL raises some crucial points, it does not provide a consistent view of feedback types, conflating different concepts and levels. This section proposes a novel, exhaustive categorization of human feedback types. The following sections discuss feedback quality criteria and derive actionable requirements for human-AI interaction systems.

#### 3.1 Methodology and Process

We propose a conceptual framework to structure and unify the space of different feedback mechanisms in related work and guide future exploration and research.

We designed the conceptual framework through an iterative survey and refinement process. In multiple iterations, we surveyed work on possible types of human feedback for training intelligent agents and then discussed potential dimensions for classifying human feedback. Discussions were conducted in a core group of four contributors, with an additional open-group session involving six experts. Based on these discussions, we defined dimensions and quality criteria. For the chosen dimensions and quality criteria, we specified the following requirements:

- Exhaustiveness: All surveyed papers describing types of human feedback for agent training must be classifiable with the given dimensions.
- Orthogonality: Significantly different approaches should be classified into separate dimensions.
- Relevance: The chosen dimensions should relate to the elements of the human-to-AI communication process and enable deriving quality metrics and requirements.

We begin with a limited survey of 18 papers on natural language feedback. We identified five initial types of human feedback (equivalent to the types specified in subsection 3.7) and defined a first candidate set of possible dimensions to classify feedback types. We then extended the candidate set with an additional 22 papers describing approaches for human feedback in reinforcement learning. We defined a set of 6 dimensions and seven quality criteria to classify the surveyed papers accurately.



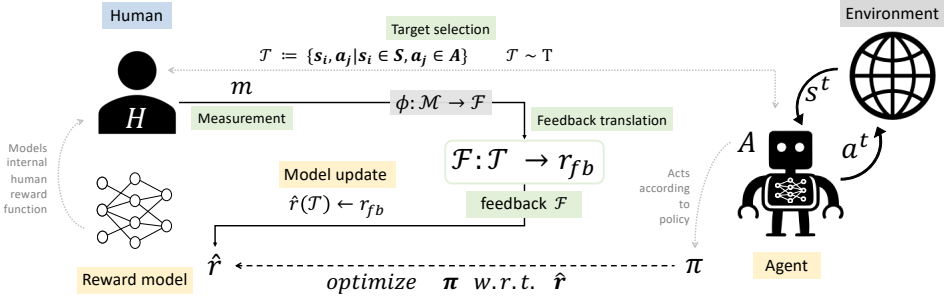


Fig. 4. The feedback process formalized: Humans generate feedback for agents that they observe acting in an environment. A feedback instance is based on a target and measured values. It is translated into a form that can be used to update a reward model. The policy of the agent, in turn, is optimized with respect to the updated reward model.

As in traditional reinforcement learning (RL), we consider an agent that acts in an environment over a sequence of steps. Concretely, at a timestep  $t$ , the agent chooses an action  $a^t \in \mathcal{A} \subseteq \mathbb{R}^M$  given an observation  $o^t \in \mathcal{O}$  of its current state  $s^t \in \mathcal{S} \subseteq \mathbb{R}^N$ . Here, we assumed, without the loss of generality, that we can map all actions  $a^t$  and states  $s^t$  to a real-valued vector. Contrary to traditional RL, however, we lack access to the *true*<sup>1</sup> **reward function**  $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ , that provides a reward  $r^t = r(s^t, a^t)$  for a state-action pair. In many real-world problems, defining a valid reward function is as hard as solving the problem itself. Instead, RLHF resorts to learning an estimate  $\hat{r} : \mathcal{S} \times \mathcal{A}$  of the true reward function based on human preferences transmitted via human feedback. The goal is to learn an optimal policy  $\pi^* : s^t \rightarrow a^t$  that maximizes the expected discounted rewards according to  $\hat{r}$ . In our example, the observations can be the robot's sensor readings and camera input to capture the state of the home environment. Actions can be low-level (motor controls of robot joints) or high-level (moving to a location, picking/placing objects, etc.). In this work, we do not assume a certain action granularity. However, the task of aligning available feedback with actions is a challenge in itself, aided by the availability of varied feedback.

In this paper, we assume human feedback to be based on a **measurement**  $m$ , obtained in the human-AI interaction process. A measurement is defined for a set of measurement variables  $m := v_1, \dots, v_k$ , which depend on the design of the human-AI interaction interface. Variables can be based on diverse interactions, including explicit inputs like buttons or sliders and sensors measuring implicit interactions. Language feedback is a versatile type of feedback that can be interpreted in different ways.

Any state information can be interpreted as a measurement as well. We may consider a measurement as a "snapshot" of the interaction process that can be arbitrarily complex based on the chosen analysis context—for example, a "goal state" like a desired arrangement of objects on shelf.

Based on the Human-AI communication space subsection 2.1, we define a **feedback state**  $f_s$  in which the interaction takes place and that a measurement is conditioned on. The feedback state can be decomposed into *three sub-states* mirroring the actors: (1) A human feedback state  $f_{s_h}$ , which might contain information like the knowledge state or exhaustion level of the human, (2) the interface feedback state  $f_{s_I}$ , i.e. in which state is the learning process, which information and interactions are available to the human, and which measurements are used, (3) an agent feedback

<sup>1</sup>In many scenarios, a true reward function might not exist, especially if the desired output can only be judged based on subjective, potentially inconsistent, preferences, such as aesthetic criteria.



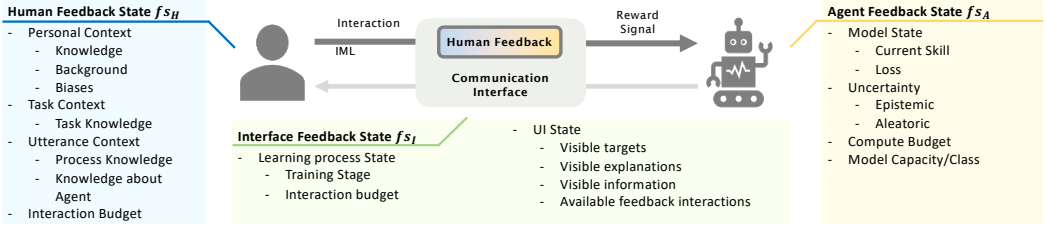


Fig. 5. The feedback state consists of three different sub-states: (1) The human feedback state, (2) the state of the interface, and (3) the model state.

state  $f_{s_A}$ , e.g. containing the model state and computation budget. The feedback state is summarized in Figure 5. The use of a human model  $H$  and an agent/robot model  $R$  to capture such aspects of human-AI communication has already been proposed in existing work [72]. However, in this work, we go beyond already proposed modeling approaches.

As part of the reward learning process, "raw" measurements need to be translated to a usable reward signal that can be used to optimize a reward model. To achieve this, a vector of measurements  $m$  must be assigned to a **target**  $\mathcal{T} \subseteq \Xi$  from the space of possible trajectories  $\Xi$ , i.e., all possible behaviors. On a high level, a target  $\mathcal{T}$  is nothing more than a set of features from the observation or action spaces (which are the input to the reward model):  $\mathcal{T} := \{s_i, a_j | s_i \in \mathcal{S}, a_j \in \mathcal{A}\}$ . As a special case of  $s_i = s, a_j = a$ , we can write a target as a set of state-action trajectories  $\mathcal{T} = \{(s^0, a^0), (s^1, a^1), \dots, (s^i, a^i), \dots\}$ , for example, a target can be an entire episode or even a set of episodes. This is related to the concept of grounding functions found in the reward rational (implicit) choice framework [80]. Special cases include a single state-action pair  $\mathcal{T} = (s, a)$  and the entire set of possible trajectories  $\mathcal{T} = \Xi$ . In our example, a target of feedback can be a single state-action pair, like placing a single object in a particular location based on the observation of the existing locations, or feedback for a whole trajectory of behavior, like cleaning an entire room.

Secondly, as a usable reward signal, the target  $t$  must be mapped to the inputs in the domain of the reward function  $r$ . The content of the utterance must be translated to a feedback reward value  $r_{fb} \in R(r)$ , with  $R(r)$  being the range of the reward function  $r$ , commonly the real numbers  $R(r) = \mathbb{R}$ . We define a processed **human feedback instance**  $\mathcal{F}$  as a mapping from a target  $\mathcal{T}$  to a feedback value  $r_{fb}$ :

$$\mathcal{F} : \mathcal{T} \rightarrow r_{fb} \in \mathbb{R}$$

To generate processed feedback, we need to design a **translation algorithm**  $\phi : m \rightarrow (\mathcal{T}, r_{fb})$ . A translation algorithm generates or predicts targets and **feedback values**, i.e., a feedback value encoding, from raw measurements. In the case of simple explicit measurements, like a slider value or buttons indicating preferences, the translation function might be the identity function  $\phi(m) : m$ . This setup is assumed in most existing work in machine learning [80].

Assuming more complex, implicit feedback, e.g., language-based feedback as in our running example, a simple translation algorithm mapping from language feedback to a value is a sentiment classifier [95], which maps the raw utterance, contained in  $m$  into a feedback value  $r_{fb} = -1, 0, +1$ .

Based on our previous discussions, a robust translation algorithm/reward modeling approach should take the feedback state into context [113]. Based on our definition, we may identify two types of measurement variables: (1) *Intrinsic* measurement variables, e.g., the primary user interaction, designed to receive a feedback value  $r_{fb}$  for a specified target  $\mathcal{T}$ ,  $m_{int}$ , (as specified above), and (2) *contextual* measurements  $m_{ctx}$ , that contain information about the feedback state  $f_s$ . Again, a translation algorithm translates these raw measurements  $m_{ctx}$  into a **context encoding**:  $\phi(m_{ctx}) \rightarrow$



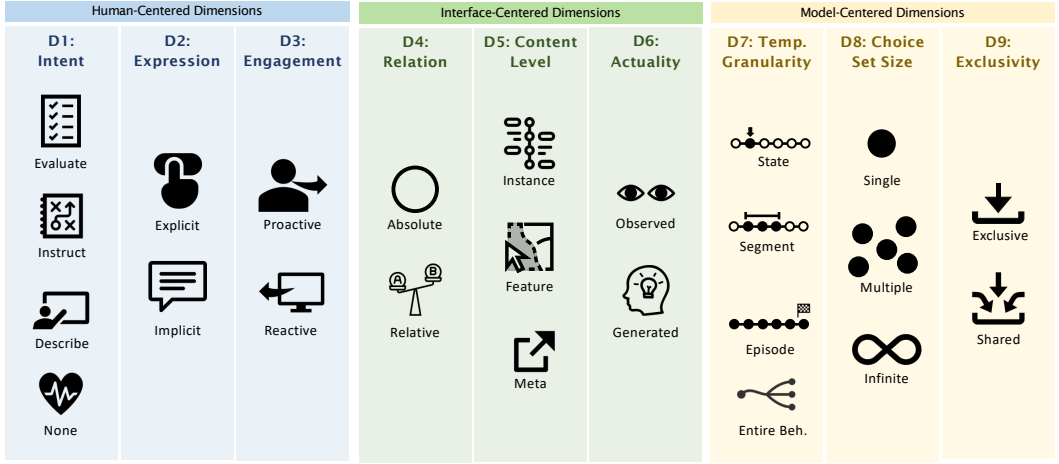


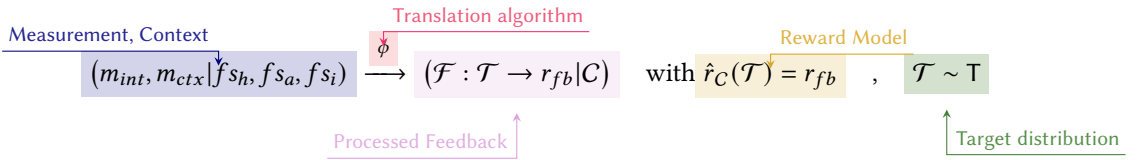
Fig. 6. In the context of *Reinforcement Learning*, we present a *conceptual framework of human feedback* based on nine dimensions: We identify human-centered, interface-centered and model-centered dimension.

$C$ , which can be utilized for reward model training, e.g., by training multiple rewards models for different contexts (e.g., user groups) or condition a reward model on the context.

Lastly, a key aspect is the selection of targets that are shown to the human user: We assume that the selection of targets follows a probability distribution  $T$ , i.e.,  $\mathcal{T} \sim T$ , which we call the **target distribution**.  $T$  may simply be randomly sampled from the agent's behavior or follow more sophisticated strategies based on the feedback state, as found in many querying strategies [42, 214].

The entire process, including key terms, is shown in Figure 4.

To summarize, the elements we need to consider for human feedback are the choice of *measurement variables and domains*, models of *feedback state*, set of available *targets*, and *translation algorithm*  $\phi : m \rightarrow \mathcal{F}$ . The dimensions we present in this chapter enable the appropriate choice of elements. We can define the entire process as follows:



This basic model opens up an extensive design space of possible human feedback. In the following sections, we introduce nine dimensions that help to structure the space of possible design choices and help us to think about diverse human feedback in a systematic way.

### 3.3 Dimensions of Human Feedback

As shown in Figure 1, we classify human feedback along nine dimensions split into three categories: human-centered dimensions, interface-centered & structural dimensions, and model-centered dimensions. Despite this grouping, the full set of dimensions influences all design and interpretation stages.

### 3.4 Human-Centered Dimensions

These dimensions are directly connected to the design of human-centered interfaces. As user interface designers, we must ensure that the interfaces match the dimensions of the feedback we want to be able to collect from humans. The interfaces should allow humans to (1) provide feedback according to their *intention*, (2) *expressed* in a way that fits the ability and preferences of the human, (3) with the correct level of *engagement*.

#### Intent D1

*Definition:* The intentions or the "goal" of human feedback in a given situation.

*Attributes:*

<b>Evaluate</b>	Evaluative feedback can be in the form of <i>binary critique</i> [73, 88, 138, 212], <i>scalar ratings</i> [96, 115, 185, 201, 202, 212], <i>preferences</i> [14, 17, 42, 102, 205], etc. [12]
<b>Instruct</b>	Instructive feedback can be given via <i>action advice</i> [8, 64, 93], <i>demonstrations</i> [146, 182], <i>physical</i> [38, 123, 135] or <i>verbal corrections</i> [116]
<b>Describe</b>	Descriptive feedback can be in the form of <i>rules</i> [25, 97, 132] or <i>principles</i> [17], <i>constraints</i> [50, 75, 164, 203], <i>subgoal generation</i> [148, 208], among others [178].
<b>None</b>	Feedback can be given without intention, by <i>facial expressions</i> [9, 48, 103, 199], <i>error potentials in the brain</i> [85, 207], <i>gaze</i> [10, 219, 220] and other subconscious reactions.

While there is ample space for possible intentions in general, we focus on four attributes that are exhaustive in the human-AI context: evaluating the agent, i.e., assessing how well it is doing, instructing the agent, i.e., communicating a desired behavior or outcome, describing, i.e., providing supplementary information to complement other feedback. Finally, feedback can be provided without a clear interaction intention.

In existing work for LLMs, simple evaluative feedback (in particular preferences) are the dominant ways a human can express their intent [61]. However, we can imagine interactions for instructive feedback (e.g., providing a fragment of desired or corrected text output to the model). In the case of our household robot, we might want to evaluate the observed behavior ("*I did or did not like the way the robot cleaned some furniture*"), instruct it ("*Clean it using a different type of cleaner*"), give high-level descriptive feedback ("*These type of wooden surfaces should always be cleaned with a special cleaner*"). We might also provide feedback unintentionally (e.g., by grimacing when watching the robot cleaning incorrectly, but, e.g., when manually cleaning something after the robot is done without a clear intention to teach the robot).

**Formal Definition: Intent** — Recall the basic definition of the reward function  $r : (s, a) \rightarrow r$ . We can interpret feedback of different intentions to update our knowledge about the reward estimate  $r$ :

$$m \xrightarrow{\phi} \begin{cases} (\mathcal{T}, r_{fb}) & (\text{evaluative}) \\ (\mathcal{T} : [(s^0, a_H^1), \dots], r_{fb} = (0, 1]) & (\text{instructive}) \\ (\mathcal{T} \subseteq [(s_0, a_1), \dots], r_{fb}) & (\text{descriptive}) \\ \{(\mathcal{T}, r_{fb}), C\} & (\text{no intention}) \end{cases}$$

Here, we might assume that instructive feedback is optimal ( $r_{fb} = 1$ ) or has a certain optimality, e.g., confidence, assigned ( $r_{fb} < 1$ ). Feedback from other types, such as evaluative, can also contribute to the context  $C$ ; we omit it here for simplicity. For unintentional feedback, we receive a measurement

$m$ , which needs to be translated into a well-defined value range.

### Expression Form D2

*Definition:* Whether the human provides the feedback deliberately.

*Attributes:*

<b>Explicit</b>	Feedback can be directly articulated as numeric or ordinal values, via <i>buttons</i> [42, 88, 106], <i>sliders</i> [33, 130, 188, 202] and can be directly interpreted.
<b>Implicit</b>	Involves more subtle forms of expression, such as <i>gestures</i> [47, 127, 191], <i>natural language</i> [110, 116, 127, 178, 206] <i>movement</i> [38, 204] or <i>physiological signals</i> [85, 207]

Feedback can be communicated either explicitly, i.e., via a deliberate action by a human, or implicitly, i.e., through indirect interactions [109]. The contrast between explicit and implicit feedback has already been established in a series of previous works [12, 111, 113]. Explicit feedback is generated by direct interaction, e.g., pressing a button. Implicit feedback includes, e.g., gestures or mimics. Implicit feedback can but does not have to be unconscious. Unintended feedback is related to the concept of information leakage [80], i.e., information that can be inferred from observing human actors without direct interaction.

When interacting with our robot, we could, e.g., have an app that lets us rate the observed behavior of the robot explicitly ("*How would I rate the performance of the robot on a scale from 1 to 5?*"), but we could also choose implicit communication channels, like verbal feedback ("I was not satisfied with the behavior"). Unintentional feedback is, by definition, also a type of implicit feedback and can be a type of leaked information as described above (e.g., correcting the execution of the robot).

**Formal Definition: Expression** — We have introduced the translation function  $\phi : m \rightarrow \mathcal{F}$ . For explicit feedback, we receive a measurement  $\hat{m}$ , containing measurement variables that can be directly translated into usable feedback, i.e., via a deterministic, simple translation function  $\phi$ . For implicit feedback, we require an intermediate mechanism to process the measurement into a simplified measurement via a (statistical) model to interpret the measurement as a reward value.

$$\phi : m \rightarrow \mathcal{F} = \begin{cases} \phi(m) = m & (\text{explicit}) \\ \phi_\theta(m) \neq m & (\text{implicit}) \end{cases}$$

The reward model itself, in particular, if coupled with (deep) representation learning, can potentially map a measurement to a reward value directly.

### Engagement D3

*Definition:* How the human is engaged when giving the feedback

*Attributes:*

<b>Proactive</b>	Feedback is offered proactively by the user, without direct solicitation, via selection or manual intervention [108, 132, 144, 183].
<b>Reactive</b>	Feedback is provided in response to a query, initiated by the system or agent [63, 140, 216].

Feedback can be provided proactively, i.e., when the human decides to give feedback for a specific state, action, or feature out of a potentially large set of other states. On the other hand, feedback can

be provided reactively, i.e., when specifically requested or being queried by the system. Proactive feedback can include giving it only at specific points in time, but otherwise, not giving any feedback towards agent behavior. On the other hand, providing feedback on request also includes standard labeling of given samples, e.g., by crowd workers. One significant implication of user engagement is the interpretation of non-feedback. When not being able to give feedback, non-engagement can be interpreted as silent agreement, and in turn, providing feedback must be weighted appropriately.

When interacting with generative AI systems, humans may be able to give optional feedback, i.e., *thumbs up/thumbs down*, for a response to a query. When interpreting this feedback, we should consider how often a user decides to use this feedback, as it indicates either a very positive or very negative response to a query, compared to feedback received due to a direct query. When observing the household robot, we might either choose to pro-actively engage with the robot (*"Please stop the current action and use the vacuum cleaner first"*) or reactively, e.g., by being queried for feedback (*"Please rate the robot's behavior on a scale from 1 to 5"*).

**Formal Definition: Engagement** — For reactive/queried feedback, we can assume that feedback is sampled according to the known target distribution  $T$ . For proactive feedback, we cannot assume this fact. Instead, we assume that human feedback is a result of a human target distribution  $T_H$ . Considering this fact can be important when updating the reward model. Also, this type of "information leakage" might help us more accurately model the human's internal objectives.

$$\mathcal{T} \sim \begin{cases} T & (\text{reactive}) \\ T_H \neq T & (\text{proactive}) \end{cases}$$

### 3.5 Interface-Centered Dimensions

The following dimensions are relevant when processing human feedback as they determine which information can be encoded. The feedback can contain different **relations** between the selected targets, it can be given at different **levels of content**, and finally, it can contain information about an **actual** or a non-existing target.

#### Target Relation D4

*Definition:* The relation of the target of feedback to other targets.

*Attributes:*

<b>Absolute</b>	Feedback that is not comparative, focusing on a single action or behavior at a time, like <i>binary critique</i> [88, 138], <i>scalar feedback</i> [185, 201], or <i>demonstrations</i> [146, 182].
<b>Relative</b>	Feedback like <i>preferences</i> [42, 205], <i>rankings</i> [142, 222], or <i>corrections</i> [56, 116, 123, 181], comparing multiple actions or behaviors to guide choice or improvement.

Feedback can target a single unit, i.e., a state, episode, or input feature, which is an absolute way, i.e., standing on its own. For example, when we say if a state is good or bad, the relationship to other states is only given implicitly. Alternatively, we can give feedback on relations between units, e.g., comparisons or corrections. In these cases, the target relation is relative. In sequential settings, where we observe or interact with multiple instances after another, we could consider all feedback relative to previous instances. However, for our taxonomy, we only consider explicitly contrasted examples in an immediate shared context as relative feedback.

When interacting with the robot, we might comment on a single action ("*The way you handled the duster was not good, you could have damaged the delicate glasses*"), but potentially also in relation to another action or state ("*The way you used the duster was way better compared to the previous time*"). The same preferences can also be stated for states or features analogous to absolute feedback ("*I like the placement of pillows much better than the last time*", "*In general, I like blue pillows more than green ones*").

**Formal Definition: Target Relation** — Absolute feedback contains information about a single target associated with a scalar feedback value, whereas relative feedback contains information about multiple targets, with the feedback value being a relation like inequality or order.

$$(\mathcal{T}, r_{fb}) \quad \text{with} \quad \begin{cases} |\mathcal{T}| = 1, r_{fb} \in \mathbb{N}, \mathbb{R}, \{0, 1, -1\} & (\text{absolute}) \\ |\mathcal{T}| > 1, r_{fb} \in \{>, <, =\} & (\text{relative}) \end{cases}$$

with  $|\cdot|$  meaning the cardinality of a set.

#### Content Level D5

*Definition:* The levels of abstraction the feedback is given at.

*Attributes:*

<b>Instance</b>	Specific feedback related to particular instances or actions, such as <i>critique</i> [88, 138], <i>preferences</i> [42, 205] or <i>action advice/demonstrations</i> [64, 146].
<b>Feature</b>	Targeting detailed aspects of behavior, e.g. <i>feature preferences</i> [19, 166], <i>feature selection</i> [10, 219], <i>constraints</i> [75, 203], or <i>subgoal generation</i> [148].
<b>Meta-Level</b>	Feedback revealing contextual influence factors, e.g., distinguishability queries [60], skill assessment questions [170].

Feedback can be targeted at specific instances of behavior, e.g., environment states the agent encountered, or at more abstract features of the behavior. For example, a human actor might label a specific action as "good" or "bad", or comment on a particular feature of the action being desirable. Finally, we may receive meta-level feedback, which can be information that cannot be directly translated into a reward signal but instead reveals information about the feedback context and can, therefore, be used to improve the contextual translation and modeling of feedback.

In our running example, feedback can be given on an instance level ("*I rate this particular action with a score from 1 to 5*") or on a feature level ("*I like to use the yellow table cloth in the summer*"). Meta-level feedback could be a statement like *My eyesight is bad, so I might have trouble distinguishing small objects from each other*). Finally, measurements such as thinking time, etc., can be interpreted as meta-feedback, which indicates, e.g., the knowledge state of the human.

**Formal Definition: Content Level** — For instance-level feedback, we assume that the target contains whole states or actions. For feature-level feedback, a target is a set of features out of the state or action space that generally do not correspond to single states or actions. Meta-level feedback is translated to contextual information.

$$\phi(m) = \begin{cases} (\mathcal{T} : \{[(s^0, a^1), \dots] | s^k = (s_1, \dots, s_N), a^k = (a_0, \dots, a_M)\}, r_{fb}) & (\text{instance}) \\ (\mathcal{T} : \{s_i, a_j | s_i \subset \mathcal{S}, a_i \subseteq \mathcal{A}\}, r_{fb}) & (\text{feature}) \\ C & (\text{meta}) \end{cases}$$

### Target Actuality D6

*Definition:* Whether the feedback is targeted at something the human has observed.

*Attributes:*

**Observed** Feedback or instructions are based on actual observed behavior or actions, such as *critique* [88, 138], *preferences* [42, 205] or *corrections* [116, 123].

**Hypothetical** Feedback for scenarios that haven't occurred but are conceivable, such as *demonstrations* [182], *generalizable rules* [25, 97], or *imagined examples* [148, 209].

When giving feedback, a human actor needs to choose a target to which the given feedback is directed. The target can be something the human observed, e.g., a step or episode of agent behavior. Alternatively, the target can be hypothetical, e.g., when imagining a preferable sequence of actions an agent should have taken or a goal state an agent should have reached. Feedback can also be targeted at the training process or overall workflow, e.g., expressions of fatigue or commenting on an interaction mechanic. In a later section, we comment on using such meta feedback, e.g., for user modeling. In the context of observation actuality, we consider such feedback to have an observed target.

Our feedback can target a behavior observed from robot behavior ("*You made the carpet wet when placing the mop over it*") or be based on a hypothetical or even counterfactual scenario ("*You should walk around it to avoid dripping water*"). Descriptive feedback might both target observed and hypothetical scenarios ("*In case you have mop in your hand, avoid walking over the carpet*"). Given the language model use case, observed feedback is targeted at the generations of the language model, e.g., rating or correcting a text fragment generated by the LLM in response to a prompt. Hypothetical feedback can incorporate human-written answers to a prompt, as well as more high-level rules, such as in *constitutional AI* [17], that are not directly related to the immediate generations of an LLM-based agent.

**Formal Definition: Target Actuality** — As the name suggests, we are interested in the origin of feedback targets; we can assume that observed feedback is *in-distribution* with respect to the agent's policy, i.e., it is part of rollouts that result from following an agent's policy during training. In contrast, hypothetical feedback cannot be assumed to be in-distribution. Moreover, hypothetical feedback might be completely independent of a specific policy and instead describe general rules and preferences.

$$\mathcal{T} \sim \begin{cases} \mathbb{E}_{\mathcal{T} \sim \pi}[(s^0, a^0), \dots] & (\text{observed}) \\ \{s_i, a_j | s_i \in s, a_j \in a\}, \Xi & (\text{hypothetical}) \end{cases}$$

Observed feedback is sampled from policy rollouts, whereas hypothetical feedback can incorporate all trajectories from the trajectory space  $\Xi$  and all possible features from the state and action spaces.

### 3.6 Model-Centered Dimensions

The model-centered dimensions directly influence the design of the reward model, for example, for which **temporal granularity** the feedback is available, which **value granularity** it has, or if it is **exclusive** or given alongside other feedback.

**Temporal Granularity** D7

*Definition:* The temporal granularity of states, actions, or features the feedback is given at.

*Attributes:*

<b>State</b>	Feedback can be given for single states or state-action pairs [8, 64, 82, 165].
<b>Segment</b>	Feedback is also often assigned to segments of various length [9, 11, 42, 190].
<b>Episode</b>	Here, the reward is given to entire episodes in episodic RL tasks
<b>Entire Behavior</b>	Feedback might be directed at overall behavior, not at a specific state or action [110, 140, 203].

Feedback can have targets at different levels of granularity. Very granular feedback can be given about individual states, actions, or features, e.g., if a human recommends an alternative action in a particular state. Less granular is the typical scenario where humans rate segments or episodes. We can give even less granular feedback, e.g., about an agent's (current) overall behavior. Importantly, this granularity is a spectrum, and many different levels between individual actions and the agent's overall behavior are possible.

For generated text, we may give feedback for single words or even tokens [206], sentences and segments (e.g., single reasoning steps, phrases in a generated poem, etc.), as well as entire task executions (e.g., whether a LLM-based agent has solved a complex task correctly). When communicating feedback to our household robot, we could either comment on single actions (*"I mark the correctness of actions in this sequence via the app"*), on a segment level (*"The way you folded this was great"*), per episode if we define it as an episodic task (an episode might be an entire task) (*"Your performance today folding the clothes was better than last week overall"*). Finally, there are different ways to give feedback not directed at particular states (*"You should not fold shirts, they go directly to the coathanger"*, *"I have no complaints about your work"*).

**Formal Definition: Temporal Granularity** — Temporal granularity influences the type of targets available for a feedback instance, i.e., the set of targets to which a feedback value is assigned.

$$\mathcal{T} \in \begin{cases} (s, a) & (state) \\ [(s_k, a_k), (s_{k+1}, a_{k+1}), \dots, (s_{k+M}, a_{k+M})] & (segment) \\ [(s_0, a_0), (s_1, a_1), \dots, (s_T, a_T)] & (episode) \\ \{s_i, a_j | s_i \in s, a_j \in a\}, \Xi & (entire\ behavior) \end{cases}$$

with  $M$  being the length of a segment,  $s_0$  being a start state and  $T$  being a terminal state. The first three cases can be easily extended to sets of either single states/state-action pairs, segments, or episodes (e.g., for relative feedback).



**Choice Set Size** D8

*Definition:* The value granularity of states, actions, or features the feedback is given at.

*Attributes:*

<b>Binary</b>	Simple feedback indicating a binary choice, like good/bad [86, 165], binary preferences [42, 205], stop execution [80], hard constraints or rules in first-order logic [50, 97].
<b>Discrete</b>	Feedback can be given on whole number (Likert) scales [67], ranking of k elements [142, 168], choosing the best action out of a set [64].
<b>Infinite</b>	Feedback that exists on a continuous scale, such as demonstrative and corrective feedback in continuous action spaces or implicit feedback (e.g., gaze or physiological signals).

Feedback can have different value granularities determined by the interface that the user is provided with. For example, giving a thumbs up or down is interpreted as binary feedback. Similarly, an intervention can also be interpreted as a binary value (either intervening or not intervening). Similarly, a pairwise comparison is a type of binary-valued feedback. Additionally, we can interpret certain types of descriptive feedback, like hard constraints, as binary feedback. Feedback from a multi-step ranking or on a multi-point scale can be interpreted as discrete/integer-valued feedback. Continuous feedback might either be based on a very fine-grained rating scale or, more often, come from implicit modalities, e.g., physical corrections of a robot, or feedback based on human facial expressions might be available on a continuous scale.

When interacting with commercial LLM-based chat systems, we are often restricted to binary feedback, such as "thumbs up"/"thumbs down" or the choice between two alternative generations. However, we might extend these choices, e.g., selecting words to generate out of a set of multiple possible choices [174], or freely generating desired output, which in practice coincides with a virtually infinite choice set. When giving binary feedback to our robot, we give a positive or negative signal ("*Thumbs up/down for this*"), binary preference ("*I like this placement better*"), or fixed rule ("*The remote should always be put on the armrest*"). Integer-valued feedback could be on a scale ("I rate this execution on a 5 out of 10") or choose out a set of alternatives ("*I like this arrangement best, this arrangement second best, and the third one the least*"). Finally, continuous feedback can correspond to a target out of an infinite set, e.g., when applying a physical force to correct the movements of a robot. Also, we might choose from a "virtually" infinite space of possible alternative state-action trajectories when generating a demonstration.

**Formal Definition: Choice Set Size** — The choice set size indicates how many different options for a measured variable exist, i.e., from how many options a human can choose a feedback value.

$$r_{fb} \in \begin{cases} \{0, 1, >, <\} & (\text{binary}) \\ \mathbb{N} & (\text{discrete}) \\ \mathbb{R} & (\text{infinte}) \end{cases}$$

**Feedback Exclusivity** D9

*Definition:* Whether the utterance is the exclusive source of feedback or available besides other sources.

*Attributes:*

<b>Single</b>	Scenarios in which a type of human feedback is the sole source of a learning signal, e.g. in RLHF fine-tuning for LLMs [150, 223].
<b>Mixed</b>	Feedback that supplements other sources of feedback, e.g., multi-faceted feedback [206], shaping rewards [87] or simultaneous multi-modal feedback [63, 103].

When providing feedback, human feedback can serve as the exclusive single source of reward in a situation, e.g., when tuning an agent with personal preferences. However, feedback for the same target might also be given from multiple sources. These other sources can be environment reward signals (like in reward shaping), feedback from other human users for the same target, or even other AI models. Finally, feedback might also be provided by the same user via a different feedback channel, e.g., an additional feedback type.

When not provided with another reward source, the household robot purely relies on personal feedback to optimize its behavior. A different source can be other humans, e.g., other persons living in the household with potentially differing opinions (*"I want the book to be put on a different shelf"*), but also environment signals (e.g., *not damaging objects could be an objective irrespective of human supervision*) or even other AI models (*Besides the human feedback, the robot also uses a cloud-based reward model to update its policy*).

**Formal Definition: Exclusivity** — In our framework, the non-exclusivity of feedback means that multiple feedback values can be assigned to the same target, not caused by variance in repeated measurement but by expressing an actual underlying difference in feedback value.

$$\exists \mathcal{T} \quad \begin{cases} \exists! \hat{m}_A & (exclusive) \\ \exists \{\hat{m}_A, \hat{m}_B, \dots\} & (mixed) \end{cases} \quad \text{with } A, B, \dots \text{ different sources}$$

So  $\hat{m}_A, \hat{m}_B$  are measurements from different sources which can be potentially different, i.e., measurements from different sources can provide different values for the same target. We want to differentiate this from repeated measurements from the same source, which can be different (e.g., by changing preferences throughout training). We will comment on this in section 4.

### 3.7 Classifying Established Feedback Types

First, we want to relate the taxonomy to commonly used types of feedback. Here, we select feedback types that are either commonly used in practice [12, 137] or at least clearly described in theory [80]. We summarize the identified feedback types in Table 1. As mentioned above, in this work, we focus on feedback types for which algorithmic exploitation has been established in the literature to ensure applicability [79, 80, 113, 177].

**Quantitative Feedback** — Humans can give agents numerical performance evaluation. Examples include a binary or scalar reward about an episode (sometimes called as *critique*[12]), or reward shaping, which relies on assigning individual scalar values to single steps or trajectory segments [86, 145]. It is expressed explicitly, i.e., via user interactions like buttons or sliders. We might call implicit feedback targeted at absolute observed instances as *reactions*. In many cases, quantitative feedback is *reactive*, so queried from a human. However, when combined with the interactive selection of targets, it can also be given *proactively* [215]. This feedback type always targets observed behavior

Table 1. Classifying Established Feedback Types: We structure feedback along the nine dimensions we introduce in subsection 3.4 to subsection 3.6. (✓) Black checkmarks refer to exclusive attributes (✓). Grey checkmarks indicate that feedback can have characteristics across different work. We can classify existing types of feedback according to our dimensions.

	D1	D2	D3	D4	D5	D6	D7	D8	D9	
	Intent	Expres.	Engag.	Rela.	Content	Tgt.Act.	Temp.Gra.	Choi.Set	Exclus.	
Feedback Type	None Describe Instruct Evaluate	Explicit	Implicit	Reactive Proactive	Relative Absolute	Instance Feature	Actual Hypothetical	Discrete Continuous Binary	Exclusive Augmenting	Publications
Critique ✓		✓	✓	✓	✓	✓	✓	✓	✓	[70]
Shaping ✓		✓	✓	✓	✓	✓	✓	✓	✓	[43, 86, 200]
Behavior Pref. ✓		✓	✓	✓	✓	✓	✓	✓	✓	[32, 42, 79]
Outcome Rating ✓		✓	✓	✓	✓	✓	✓	✓	✓	[151, 186]
Action Advice ✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	[3, 49, 64]
Demos ✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	[79, 146, 154]
Demos w/o acts. ✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	[118, 189]
Correction ✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	[123, 134, 164]
Feat. Selection ✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	[179, 186]
Feat. Saliency ✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	[70]
Goal Spec. ✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	[90, 148]
Goal Pref. ✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	[205]
Gaze ✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	[220]
Reactions ✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	[193]

and is not given at the instance level (i.e., does not comment on the merit of particular features). It is absolute, i.e., not referring to another target. In terms of temporal granularity, we can observe the use of step-wise, segment-wise, and episode-wise feedback, as the assignment of a quantitative value to a selected target is equivalent among granularities. Most often, we see binary or scale, i.e., multiple possible feedback values. Quantitative feedback can be used both exclusively and non-exclusively.

**Comparative Feedback** — This type of feedback provides the agent with information about how its performance compares to a reference, such as an alternative execution, benchmark, or a human expert. As such, it is still evaluative. The most known example of comparative feedback is rankings [32] or pairwise comparisons [42]. Existing forms of comparative feedback are generally targeted at observed instances, i.e., by contrasting the replay of agent behavior against one another. As such, this type of feedback represents relative target relations. We often see these comparisons between sequence segments or full episodes. Rankings give a discrete choice set, while pairwise preferences present a binary choice. Preferences have been used as the exclusive source of feedback to train reward models.

**Corrective Feedback** — Corrective feedback provides agents with specific information about what they did wrong and how they can correct their behavior and is, therefore, instructive. A main component of this type of feedback is the necessity of generative human processing, i.e., the human needs to imagine and demonstrate or describe a hypothetical sequence of actions or a hypothetical preferred goal state. For example, when recommending an alternative action compared

to the one taken by agents (often referred to as *action advice* [64, 111, 113]), we must imagine a hypothetical/counterfactual future in which the alternative action leads to a more desired outcome. For corrections, we provide feedback with a relative target relation, i.e., in direct reference to an observed instance. Corrective feedback is generally combined with other types of feedback, i.e., non-exclusive.

**Demonstrative Feedback** — Here, humans provide agents with information about the desired behavior by demonstrating the sequence of corrective actions in a particular context. Compared to corrective feedback, demonstrations are absolute feedback. If provided explicitly, e.g., by directly inputting the sequence of required actions, we call this type of feedback *demonstrations*. If only provided implicitly, e.g., by showing body movements to an observing embodied robot, we might call them *demonstrations without actions* [118, 189]. Further, compared to corrective feedback, demonstrative feedback almost always has a higher granularity, e.g., an entire episode, compared to just a small number of steps.

**Descriptive Feedback** — is a weakly defined term in the literature [177]. In our taxonomy, we relate it to feedback at the feature level, compared to all previous forms of feedback, which are primarily given at the instance level. Descriptive feedback generally refers to certain aspects of either the observed behavior or (hypothetical) outcomes. We do not explicitly distinguish between process-level descriptive feedback (e.g., a particular joint position of a robot at a certain point in time is undesirable) and outcome-level descriptive feedback (e.g., a goal state of the environment). Descriptive feedback is given as complementary non-exclusive feedback.

#### 4 Qualities of Human Feedback

We have now established a conceptual framework for feedback types. So, having answered the question *"What kind of human feedback is there?"*, we now want to look at the question *"What makes for good human feedback?"*.

To tackle this question, we choose a similar approach to identify relevant dimensions. From the surveyed set of papers, we identify quality criteria discussed in the literature. We find a large set of different quality measures and raised issues about human feedback in interactive RL and RLHF Table 2. We posit that the large variety of terms is because we can look at feedback qualities from three different view angles as outlined in subsection 2.1: We ask the question *"What is high-quality feedback from ...?"*

- a human perspective: How good is the human experience of giving feedback?
- an interface perspective: How well can the feedback be collected and processed?
- a model perspective: How well can a model be optimized based on the feedback?

In the following, we present seven quality criteria that can be used to evaluate the quality of human feedback. Based on these three perspective, we group the quality metrics into three categories equivalently to dimensions: human-centered, interface-centered, and model-centered feedback quality criteria.

We discuss how these seven qualities can be evaluated. We also discuss how we can measure each quality, i.e., which metrics exist.

##### 4.1 Human-Centered Qualities

From a human perspective, we identify two key desiderata with their associated qualities: feedback should be **expressive**, and it should require minimal **effort** to provide. These qualities directly relate to the human-centered dimensions of intent [\(D1\)](#), expression form [\(D2\)](#), and engagement [\(D3\)](#).

Table 2. Human feedback qualities used in the literature: We summarize a large set of quality measures from surveyed literature into seven main quality criteria to give a more condensed set of relevant quality criteria.

Expressiveness	Ease	Definiteness	Context Independence	Precision	Unbiasedness	Informativeness
Richness [36, 179]	Fatigue [65, 91]	Uncertainty [158]	Choice set dependence [65]	Accuracy [24, 26, 59, 158]	Reward bias [24, 88, 185]	Information gain [26, 91]
Open-Endedness [126, 179]	Availability [24]	Confidence [65, 158]	Policy dependence [130]	Repeatability [24]	Asymmetry [185]	Stability [135]
	Time requirements [26, 84]	Security [158]	Knowledge level [24]	Consistency [24, 47, 112, 158]	Cognitive bias [24]	Performance [91]
	Latency [24]	Decision problem uncertainty [100]	Independence [108]	Variation [24]	Concept drift [24]	Retrospective optimality [158]
	Response time [158]	Understandability [84]	Ambiguity [122]	Reliability [84, 217]	Scale difference [112]	Completeness [24, 59]
	Budgetary efficiency [57]	Verifiability [26, 130]	Hidden Context [171]	Noise [59, 91, 135]		
	Cognitive load [84, 91]	Perception quality [47]		Mislabeling probability [79]		

**Expressiveness** Q1

*Definition:* How expressive is the feedback in communicating human intentions?

*Attributes:* **Low Expressiveness** (Limited, unable to express according to intentions) – **High Expressiveness** (Flexible, able to express intentions)

Expressiveness can be understood as *adequacy* from the human perspective: Does the available feedback mechanism match user expectations? Is it appropriate for the task complexity? The perceived *expressiveness* of feedback heavily depends on the user’s current state and their intended feedback goals.

**How to evaluate Expressiveness?** — *Expressive* feedback must align with the user’s intention in any given situation. Users should be able to provide open implicit feedback when needed. As a human-centered quality, expressiveness is evaluated through *human-centered evaluation* methods [173].

*Subcategories:*

- **Open-Endedness:** tasks may be either *open-ended*, i.e., not having a clearly defined outcome (e.g., "Write a story..") or close-ended ("put the dirty dishes into the dishwasher"). Open-ended tasks are well served by expressive, diverse feedback to give users ways to express complex preferences. Simple, explicit feedback can be the right choice for close-ended, clearly defined tasks.
- **Satisfaction:** Users can rate their satisfaction with the feedback’s expressiveness during the interaction process [91].
- **Chosen Feedback:** When multiple feedback types are available, users’ choices can reveal their preferences [137]. These choices often indicate feedback appropriateness, though care should be taken to minimize confounding factors to obtain meaningful preference data.
- **Engagement Metrics:** Retention rates, frequency of additional user comments, and similar behavioral indicators can serve as proxy measures for perceived expressiveness [91].

*Optimizing expressiveness:* To optimize expressiveness, users must be able to effectively communicate their intent D1 in any given state. We hypothesize that instructive or descriptive feedback will be perceived as more expressive, especially when agents exhibit low skill levels or the human has expert knowledge. Additionally, open-ended, implicit or multi-modal feedback D2 options may enhance the user’s sense of expressiveness. Similarly, the opportunity for users to act proactively D3 for targets has the potential to increase perceived expressiveness. We will discuss further research opportunities in section 7.

**Ease** Q2

*Definition:* The level of cognitive effort and time required to provide feedback.

*Attributes:* **No Conscious Effort** (unconscious or reactive) – **Low Effort** (simple and brief interactions) – **High Effort** (requires concentration or creativity)

Different feedback types impose varying levels of cognitive demand on humans. Implicit feedback typically requires minimal or no conscious effort, making it well-suited for extended sessions. In contrast, explicit feedback generally demands higher cognitive engagement, especially for demonstrations and descriptions. Reactive feedback, initiated by user queries, enhances speed and ease of use. Pro-active feedback, however, requires greater user attention and additional effort to identify appropriate feedback opportunities.

**How to evaluate Ease?** — Ease is evaluated through *human-centered approaches*. The field of human-computer interaction offers numerous experimental tools for measuring users' cognitive workload [92, 173].

*Subcategories:*

- **Cognitive Load:** This can be assessed through surveys or ratings that measure experienced cognitive demand and exhaustion [84, 91]. Regular assessments during the interaction process can track exhaustion levels. Response time measurements can serve as proxy indicators for cognitive demand, particularly when analyzing how interaction times evolve throughout the process [158].
- **Temporal Demand:** This metric quantifies the time required to provide feedback [26]. Temporal demand typically correlates with cognitive load [24] and has significant implications for budgetary efficiency [57].
- **Knowledge Level:** A key consideration is the level of knowledge expected to give feedback for a task. Tasks might be either highly *domain-specific*, requiring rare expert knowledge, or be *domain-general*, i.e., can be evaluated or even solved by large parts of the population. We must collect feedback from domain experts for domain-specific tasks, e.g., coding tasks. Consequently, we both must be efficient with feedback, i.e., focusing on highly informative feedback and giving domain experts ways to express their knowledge by providing interactions for instructive or descriptive feedback. For domain-general tasks, simplicity of interactions and cost-effectiveness might be more important considerations.

*Optimizing ease:* User experience can be enhanced by aligning feedback mechanisms with human intent D1, avoiding ineffective or repetitive feedback. Leveraging implicit feedback D2 where appropriate, and striking an optimal balance between proactive and reactive feedback approaches D3.

## 4.2 Interface-Centered Qualities

From an interface or system perspective, good feedback needs to fulfill two criteria mainly: It should be *definite*, i.e., a faithful representation of the human expression and its associated uncertainty, and it ideally has a tractable *context*, i.e., the context necessary to correctly interpret the context is available and transparent. Alternatively, the influence of external context should be minimized. These qualities can be mainly influenced by choices in the design dimensions of relation D4, content level D5, and actuality D6.

### Definiteness Q3

*Definition:* How definite is the measurement of feedback, its associated context, and uncertainty is.

*Attributes:* **Low Definiteness** (Measurement quality unknown and user uncertainty unknown) – **High definiteness** (Measurement uncertainty is known and tractable)

Uncertainty of human feedback comes in various forms. We distinguish between user uncertainty, systematic uncertainty, and model uncertainty. We define user uncertainty as something that the user is either aware of (explicit) and could quantify or which is communicated via implicit interactions, e.g., hesitation or correcting feedback (implicit). Systematic uncertainty is inherent in the system; we refer to it as *aleatoric* or irreducible uncertainty of the reward learning process. Lastly, model uncertainty is the uncertainty of the reward model concerning human feedback. This uncertainty has a reducible and irreducible component: Some part of it, the *epistemic* uncertainty of the reward learning process, is reducible by acquiring more data about the human via feedback. However, the inherent uncertainty from the human and systematic uncertainty can not be reduced with more information.

**How to evaluate Definiteness?** — As an interface-centered quality, *definiteness* is evaluated more systematically, i.e., more akin to software testing.

*Subcategories:*

- **Measurement Complexity:** Which measurement variables are tracked (i.e., how well defined the *feedback state*  $f_s$  is) and whether uncertainties, inherent to the user or introduced in the measurement process (i.e., when translating implicit feedback), are quantified.
- **Completeness:** As part of the feedback measurement process, tracking a wide set of user interactions (such as mouse events, timings, interaction with visual interactions, etc.) enables more comprehensive modeling [24, 59].
- **Sensor Uncertainty:** In particular for implicit feedback, the quality of feedback might rely on the quality of sensor reading [47].
- **User Uncertainty:** As part of the feedback process, users might be able to quantify their uncertainty for a given target (e.g., by a slider or by expressing uncertainty in natural language). We might also estimate uncertainty based on additional contextual measurements.
- **Task Verifiability:** An agents attempt at a task can be *objectively* or *subjectively* evaluated. We may also call certain tasks *verifiable* or *non-verifiable*. Objective tasks, such as mathematical calculations, with a clear objective solution or measurable outcome, are a good fit for automated validation strategies and require less or no human evaluation. Here, human feedback might be better suited as a source to give feedback on the process or intermediate steps, i.e., descriptive or instructive.

Feedback with a high level of *definiteness* enables the system to accurately model the feedback state and uncertainty associated with a feedback instance.

*Optimizing definiteness:* Relative feedback, or absolute feedback D4 with ways to communicate confidence/uncertainty generally has higher definiteness. Measuring additional metrics beyond the explicit feedback interactions, including contextual factors like expertise level, background, and task familiarity, improves definiteness. We summarize many of these measurements as meta-level D5 feedback. Finally, feedback for observed D6 feedback has a significantly higher degree of definiteness because we can directly control the selection and definition of samples. On the flip side, generated feedback might reflect the background and user context in a way that would be hard to uncover with pre-defined targets.



### Context Independence Q4

*Definition:* To what extent does the feedback depend on the context of the interaction?

*Attributes:* **Low Context Independence** means that feedback is highly contextual, so specific to a human or task – **High Context Independence** means that feedback values are less dependent on influencing factors

A related aspect is context dependency, which means that feedback has to be interpreted either in relation to other elements inside the analysis context, e.g., other states, feedback, or the interfaces, which we call *internal context dependency*. On the other hand, *external context dependency* encompasses factors that are not explicitly modeled as part of the analysis context. For example, hypothetical absolute feedback, e.g., demonstrations, are often generated implicitly in response to the current state of training, e.g., a desire to instruct agent behavior based on recent observation. Similarly, isolated ratings of episodes might also be influenced by previous internal contexts. In relative feedback, we provide a reference as a controlled internal context. External context, which encompasses, e.g., the cultural or sociological context of the human and the surrounding of the interaction environment, can influence how certain types of feedback are communicated. In general, the less formalized feedback, e.g., hypothetical descriptive and implicit feedback, the higher the degree of external context-dependency can be expected. Gestures, mimics, or natural language expressions may vary significantly between humans, whereas external factors influence simple scores or comparisons less.

**How to evaluate Context Independence?** – Context-independence of feedback can be evaluated by tracking contextual factors alongside feedback values. These allow us to relate the feedback and identify both confounding factors and whether the feedback is robust to these.

*Subcategories:*

- **Task Dependency:** The solution for a given task might be highly *context-dependent*, or *context-independent*. In context-dependent tasks, correct agent behavior must be adapted to specific circumstances. For example, some tasks must consider a user's background, adapt behavior, and interpret feedback accordingly. Based on this, we may adapt the feedback to track contextual factors or ignore them during translation and modeling.
- **User Dependency:** Feedback may be heavily dependent on the user's knowledge (both in terms of global domain knowledge or the knowledge during a given moment of the training). Tracking the knowledge of a user when measuring a feedback instance allows us to measure these dependencies.
- **Interface Dependency:** Feedback values depend on design choices in the implementation of the feedback process, e.g., choice set dependency [65] ("From how many different choices can a user select?"). Thus, measuring the effects of design variations on human feedback allows us to model these effects.
- **Model Dependency:** Feedback depends not only on the agent behavior and task but also the current policy of an agent [130]. In particular, humans adapt their feedback to the policy they observe during training. Measuring feedback values during the training process can uncover these dependencies.
- **Feedback Interdependency:** Feedback may depend on previous feedback instances, i.e., the magnitude of values is influenced by previous values [108]. One must treat sequential feedback as interdependent instead of viewing each instance as isolated to measure this bias.

*Optimizing context independence:* From an interface perspective, feedback is more independent of external context if it contains internal contextual information, i.e., relative D4 feedback already

induces a context between multiple targets, whereas absolute feedback is potentially more context-dependent (this comparative stability of relative (preference) feedback is generally accepted in the literature [205]). However, other contextual factors might influence feedback, particularly if it is formulated more openly [113]. Similar to the previous quality, additional measurements (meta feedback **D5**) can be used to determine dependencies better. Enabling users to provide generated **D6** feedback beyond actually observed targets or choices also allows them to learn more about context dependencies, as the choice of generated feedback by the human user can also reveal contextual information.

### 4.3 Model-Centered Qualities

From a model perspective, we need feedback that is *learnable*, i.e., feedback that enables us to learn a faithful and successful reward model. Two qualities can be derived directly from machine learning fundamentals: Assuming there exists an internal, i.e., not directly observable, human reward model<sup>2</sup>, we want the feedback to have *low variance/high precision* and *low bias*, i.e., representing the human internal rewards accurately. Additionally, we want feedback to be *informative*, i.e., revealing as much as possible about the internal human reward model. Model-centered qualities should be evaluated quantitatively by analyzing the feedback values and model performance.

#### Precision **Q5**

*Definition:* If feedback stays consistent over multiple measurements, i.e. has low variance

*Attributes:* **Low Precision** (high variation or change over time) – **High Precision** (stable and temporally coherent)

Given a similar state to evaluate or a similar correction to propose, we ideally expect a high agreement between the given ratings or instructions. However, for different types of feedback, we expect varying levels of consistency over time. For example, we might expect feedback with a hypothetical component to be relatively inconsistent because generating a demonstration without a reference, e.g., absolute hypothetical feedback, requires choosing from a vast implicit space of possible state-action sequences. We, therefore, expect a high degree of variation. The feedback that is generated relative to an observed state, for example, a correction, can already be considered more consistent. Furthermore, feedback without a hypothetical component, like evaluation, has fewer degrees of freedom, which should increase consistency over time. Consequently, relative evaluative feedback for observed states, for example, rankings or comparisons, will likely have the highest degree of consistency.

**How to evaluate Precision?** — High precision/low variance of feedback values means that the measured values are close to an underlying "correct" feedback value (which is not directly observable). We can query feedback for the same targets multiple times to receive estimates of feedback variance. The observed variance might serve as an approximate measure of overall feedback variance.

*Subcategories:*

- **Variation:** Stable feedback for similar targets indicates high precision, i.e., low variance. This means that for similar, or even the same, targets, we expect similar or equivalent feedback values instead of a widely spread distribution of values [24].
- **Rates of Human Error:** Rates of misslabeling or human error are indicative of undesired variance [42, 59, 79].

<sup>2</sup>There is extensive discussion why this assumption is not generally valid [36]

- **Translation Accuracy:** Translating implicit feedback might introduce additional variance into feedback values, e.g., if the translation is also based on probabilistic modeling [47].

*Optimizing precision:* Because labeling error is a key factor in low precision, reducing error must be a key objective. Secondly, reducing the choice set [D8](#) might improve precision/consistency because choices are more distinct. Pairwise preference might, therefore, have a lower variance than score feedback. Generally, reducing ambiguity in both the task and available choices for feedback values reduces error probability and variance. Although multiple non-exclusive reward sources [D9](#) have advantages (see the following quality criteria), they might introduce additional variance, as different sources potentially disagree with each other. To counteract the increased variance, the total number of samples must be increased.

### Unbiasedness [Q6](#)

*Definition:* Different biases that are inherent in the feedback and negatively influence its quality.

*Attributes:* **Unbiased** (feedback has low systematic error) – **Biased** (feedback values exhibit systematic error)

Existing work has established multiple biases that might impact human feedback in reinforcement learning. Psychology has found that human judgment and extension feedback are highly susceptible to biases, such as base rates, priming, asymmetry, etc. We expect that these biases are reflected in human feedback. These biases become problematic if they hinder the model's learning of a correct reward estimate.

**How to evaluate Unbiasedness?** — Without any reference, i.e., if human feedback is the only reward source, it is difficult to measure and estimate biases. We can specify a reference set of pre-labeled target data (in experimental settings, we could use ground-truth reward data) and then compare the human feedback with this reference data to estimate reward bias. In future cases, we can apply these learned bias estimates to new human feedback.

*Subcategories:*

- **Distribution Bias:** A simple bias that has been described in the literature is *asymmetry bias*, specifically humans tending to give more positive than negative feedback for rating-based feedback [88, 185]. The pretense of individual biases, e.g., in questionnaires [41], is long-established. We should, therefore, expect such *distributional* biases for evaluative feedback in RLHF.
- **Contextual Bias:** As mentioned previously (see [Q4](#)), the context can introduce biases in feedback. The context can act as a confounding factor in feedback. By defining the context, we can measure its effect on the feedback [24].
- **Complexity Bias:** A specific type of bias, prominently observed for LLMs is a bias towards more complex, e.g., lengthy or expertly seeming outputs or behaviors [76, 152, 169].
- **Temporal Bias:** Finally, feedback in a particular situation can be biased by temporal dependencies, e.g., by other preceding targets, ordering of targets, or learning/adaption processes during the interaction [108, 157]. Humans might give different feedback for the same target at different stages of the human-AI process.

*Optimizing unbiasedness:* To improve unbiasedness, we must control for the influence of contextual factors on feedback values. As outlined above, tracking of contextual measurements can be used to adapt the translation or reward model to the estimated bias of the human user. In general, using a combination of non-exclusive reward sources [D9](#), e.g., from different intentions, granularities [D7](#) or choice set sizes [D8](#) can be used to correct for specific biases of single sources.

## Informativeness Q7

*Definition:* The amount of information the feedback provides about how to update the model.

*Attributes:* **Uninformative** (Introduces redundant or irrelevant information) – **Informative** (Introduces novel and relevant information)

The agents update their reward model based on the given feedback. Informative feedback, therefore, leads to better model updates. Different types of feedback carry different amounts of information; for example, instance-level feedback for a single step, e.g., action advice, can potentially update the reward function to a small degree (especially if the feedback aligns well with the previous estimate). Feature-level information, e.g., evaluating the value of a feature's presence, might generalize more broadly and could, therefore, lead to a more significant update of the reward model. This quality is essential for reward models; however, it might be difficult to judge if a human is giving feedback to an agent.

Finally, we have established that human feedback is assigned to a particular target granularity, i.e., a subset of the state of potential state-action pairs. However, given that, human feedback may only target a part of the subset under observation. For example, a human might evaluate an entire episode but direct the feedback only to a particularly salient sub-sequence of an episode. We call this partial coverage. Ideally, we have a one-to-one match between feedback and state-action-sequence. On the other hand, humans could give feedback applicable beyond the immediate context in mind, e.g., descriptive feedback about the quality of particular features.

**How to evaluate Informativeness?** — We can evaluate informativeness by considering the reward model state.

*Subcategories:*

- **Information gain:** For some types of problems, we can directly compute the information gain of an instance of feedback given the model parameters [24, 68, 91].
- **Epistemic uncertainty:** We can estimate uncertainty for a reward model, e.g., by quantifying the disagreement between sub-models of a model ensemble. In that case, we can estimate if a state-action pair is in or out of distribution. Thus, feedback for targets with high associated uncertainty is generally more informative. Similarly, we can analyze the drop of uncertainty for the reward model in response to a model update based on a feedback instance [42, 81, 214].
- **Learning Impact:** We can analyze learning speed/update strength during reward model training. If we can evaluate the trained reward model or agent learning with the reward model, we could estimate the effectiveness of feedback by looking at the learning curve [91].
- **Coverage:** Properties like the coverage, i.e., which part of the state-action space is covered by a feedback instance, could also be interpreted as measures of informativeness. The redundancy/non-redundancy of information relies on the state of reward model.

*Optimizing informativeness:* Based on metrics such as uncertainty or expected information gain, we can actively query the user for targets and feedback types [68, 80, 214]. Being able to adapt granularity D7 or choice set size D8 of a query improves the possibility of receiving targeted and varied feedback. Access to multiple non-exclusive reward sources D9 also opens up the space for informative queries and feedback.

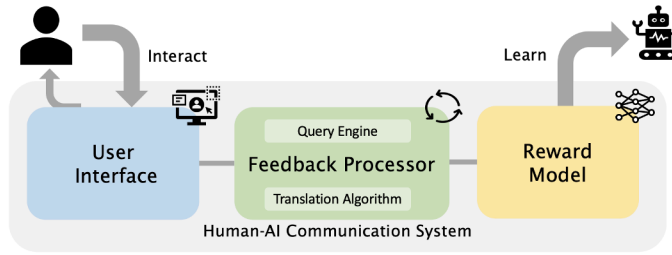


Fig. 7. The basic system components for reward learning: A user interface, feedback processor, and reward model.

## 5 Designing Systems for Diverse Human Feedback: Existing Approaches and Opportunities

In the previous sections, we have presented a conceptual framework of nine dimensions and seven qualities for human feedback. Now, we want to outline how we can operationalize the framework.

We begin by briefly introducing a basic architecture, i.e., the main components of a system learning from expressive reinforcement learning from human feedback. Then, we outline requirements and design choices for systems spanning the space of dimensions. We reference existing work in human-AI interaction and interactive RL for each dimension. Additionally, the framework allows us to identify research gaps and opportunities.

### 5.1 Basic System Design

In Figure 7, we summarize the identified core components of an RL-based human-AI interaction setup: A user interface, implementing feedback interactions, as well as providing a view of the two-sided communication, i.e., the targets feedback can be provided for and potentially additional information about the agent or training process. The feedback processor takes the data recorded in feedback interactions, potentially unstructured and highly diverse, and translates it into a standard format that can be passed onto the reward modeling stage. The reward model is optimized based on human feedback and must capture human preferences accurately and allow the AI agent to optimize its behavior to maximize the predicted reward.

The choice of feedback generally influences the whole stack of system components, i.e., implementing feedback of different types requires adaptations to the UI, feedback processor, and reward model. In the following chapter, we discuss the influence of feedback on all system components.

**User Interface** — How can humans communicate with the agent? Depending on the setting, we have specific interaction mechanics, like verbalization, movement, or using a GUI. Interactive visual user interfaces can enable expressive and situation-aware multi-type feedback. The user interface is responsible for recording measurements, intrinsic and contextual ones. Therefore, an ideal interface might go beyond just implementing the narrow explicit interactions and provide additional inputs, different types of feedback, and even multi-modal inputs like camera, voice, etc.

### User Interface Requirements

The user interface serves as the interface for communication between the human and the agent. Here, the human inputs feedback via interactions and receives information about the system and the environment. We infer the user interface requirements on human-centered qualities **Q1**, **Q2**, as well as interface-centered qualities **Q3**, **Q4**. However, UI design has huge implications for the reward model training as well:

- (UI.R1) Give users the ability to express feedback matching their intentions. Implement versatile input interactions for evaluative, instructive, or descriptive feedback should be available. **Q1**
- (UI.R2) User interfaces should consider cognitive demand and user fatigue, i.e., interactions should be efficient and the use of human input minimized. **Q2**
- (UI.R3) User interfaces should enable the measurement of uncertainty or additional metrics to verify the quality of feedback. **Q3**
- (UI.R4) User interfaces should track contextual factors. **Q4**

**Feedback Processor** — When going beyond simple, explicit feedback, which can be fed to a reward model directly, we must integrate a processing step that translates the measurements into "learnable" inputs for the reward model. The feedback processor implements the **translation algorithm**  $\phi$  specified above.

In our framework, we assign a second role to the feedback processor. The processor also serves as the **query engine**, which handles *target selection* ( $\mathcal{T} \sim T$ ), i.e., determines which targets are shown to the human user. This querying can be combined with varying levels of human involvement, i.e., manual or assisted selection of targets by the human user.

### Feedback Processing Requirements

The feedback processor parses the different expressions of human feedback. The processor has access to all relevant data in the analysis context and mediates bi-directional processes such as querying and user-aware modeling. We identify the following requirements, considering dimensions, based on qualities **Q3**, **Q4**, **Q5**, **Q6**, and **Q7**:

- (FP.R1) Define a consistent protocol to translate different forms of human feedback to a format that can be passed onto a reward model. **Q3**
- (FP.R2) Collect and process necessary information for reward learning, like user and task context, temporal dynamics, or uncertainty, to build the basis to adapt to the user and previous feedback. **Q3** **Q4**
- (FP.R3) Apply possible corrections for variance and biases based on measurements. **Q5** **Q6**
- (FP.R4) Implement a target querying strategy that facilitates high feedback quality. **Q2** **Q3** **Q4** **Q5** **Q7**

**Reward Model** — As introduced in subsection 2.1, the RL-based learner needs access to a reward function that assigns state-action(-state) pairs(triplets) a scalar-valued reward. This reward function can be implemented in a variety of ways: It can be explicitly supplied by the environment, e.g., by being implemented as a table stating the reward for a fixed set of state-action trajectories [180] or an in-game score [20]. However, this means the reward function is fixed or pre-specified. A *reward model* is a dynamic reward function that can be updated, e.g., with human feedback. A reward model is used to train the agent asynchronously. Because a model can predict reward for arbitrary state-action sequences, asking for human feedback after each step or episode is unnecessary. The choice of function class determines the expressiveness of the reward model, i.e., the complexity of factors that can be considered. Here, we identify a trade-off between different models: We may use complex high-fidelity models that can capture a large degree of variation and potential nuance,



e.g., by modeling the human as a separate agent with a partially observable state [72, 114, 163]. However, such models may require large amounts of data to provide “learnable” reward estimation, particularly in complex scenarios. On the other hand, we can choose simpler models that cannot capture certain aspects, e.g., fatigue or updating knowledge in humans. Such a model may be significantly more straightforward to train because there are fewer free parameters than need to adapt based on data [68, 80].

### Reward Model Requirements

The reward model learns a reward function that is used to perform internal training of the agent. Considering [Q1](#), [Q5](#), [Q6](#), and [Q7](#), we identify the following requirements:

(RM.R1) Learning joint reward models from diverse and expressive feedback [Q1](#) (RM.R2) Consider feedback bias and variance in the estimation [Q5](#) [Q6](#) (RM.R3) Use human feedback efficiently and provide a mechanism to estimate informativeness [Q7](#)

## 5.2 Systems for Feedback of Different Human-Centered Dimensions

In the following, we discuss proposed solutions for feedback on different intents, expression forms, and engagement. To move beyond existing work, we highlight research opportunities for human-centered human feedback.

### Human-Centered

Relevant Dimensions: Human Intent [D1](#), Expression [D2](#), Engagement [D3](#)

Relevant Qualities: Expressiveness [Q1](#), Ease [Q2](#)

**5.2.1 Human Intent [D1](#)** We can design systems to enable humans to express their feedback in a way that matches their intent, namely to evaluate, instruct, or describe. Furthermore, we can also implement mechanics that capture unintentional human feedback.

**Evaluate** — [VP](#) Human feedback is often obtained in experimental settings with simplistic user interfaces, e.g., where users elicit their preferences between pairs of trajectories shown in short videos by clicking a button [42, 190]. Often, it is also possible for humans to state no preference. This interface is simple, easy to understand, and therefore suited for large-scale collection of human preference data as is common for fine-tuning of LLMs [150].

[PI](#) A first notable example of interfaces going beyond simple interactions is by Thomaz and Breazeal [186]. Users provided feedback by clicking and dragging their mouse, where the length of the dragging motion determined the magnitude, and its direction decided the sign of the corresponding reward.

[TC](#) Recent examples of data labeling also approach fine-tuning large language models to generate more appropriate responses. Often, this feedback is comparative in that, e.g., users indicate which of two suggested answers is more harmful [66]. Recently, Zhang et al. [215] introduced *Time-Efficient Reward Learning* via visually assisted cluster ranking. This work utilizes scatterplot interactions with cluster candidates but is limited in scope and restricted to ranking-based feedback.

In another example, *Human Feedback in Continuous Actor-Critic Reinforcement Learning* employs a slider interface for approving or disapproving agent actions, enabling real-time preference indication.

[38] provides an overview of evaluative feedback methods tailored for robotics, demonstrating effectiveness in dynamic environments.



**Instruct** — There has been some work enhancing user interaction and feedback for demonstration generation. **CF** One example, *CueTip*, introduces a mixed-initiative tool for handwriting-to-text translation, where users can provide corrective feedback through multiple interface options [167]. Given the sequence-like nature of the text, some of these interface choices might translate well into episode editing.

**GG** Furthermore, *GestureScript* is a user interface designed for single-stroke gesture classification, allowing users to create example gestures and provide structural information about the gesture drawing [125]. The sequence-like nature of *Gesture Script* makes it comparable to episodes, and its interface offers an exciting way to structure feedback, potentially applicable for annotating demonstrations. **GA** In *ML-Agents*, users can record demonstrations directly in the game-like environment [83]. To that end, the developers implement a function that maps user input to agent actions.

**PC** Finally, an example of correcting the physical behavior of robots is to modify/correct a movement trajectory via physical interactions. For many physical problems, such direction interactions can be significantly more efficient than the use of a virtual user interface.

Instructional feedback involves interfaces where users guide agents by specifying actions directly. Studies like [8] employ a student-teacher framework, enabling a teacher to direct actions for optimized training outcomes.

In a similar approach, [93] maps linguistic instructions to actions within game environments, enhancing flexibility in human-robot communication.

**Describe** — **IL** One example for pixel-wise annotations, is the “Crayon method” [58], that enables to annotates image regions for pixel classification. Users are interactively queried to provide more annotations if needed.

**SL** *FeatureInsight* is a tool that allows users to classify websites based on their content in a process the authors dub *structured labeling* [98]. Users label websites by dragging them into self-created categories or opening a new category if needed. With that, it offers more flexibility and empowers the users’ decision-making.

**SI** *ReGroup* is a user interface that allows users to interactively create custom social networks by assigning members to groups [7]. The system simultaneously learns a probabilistic model that suggests new members to the user for inclusion, thus pruning irrelevant samples. RL training generates large amounts of unlabeled data, but the users’ attention and motivation are limited.

**FA** Annotation was not only for labeling instances but also to specify meaningful features for machine learning models. One such system is *Flock*, which uses human feedback to select a set of relevant features and then algorithmically weighs these features.

**Evaluate + Instruct** — Some interfaces allow both evaluative and instructional feedback. For example, [63] uses a multi-type feedback approach where users can both evaluate and instruct. Similarly, [134] provides a robotics interface where users can demonstrate or correct behaviors. Another notable example, [40], allows users to alternate between instructions and evaluations, facilitating flexible user-agent interactions.

**Evaluate + Instruct + Describe** — In cases where multiple types of feedback are needed, interfaces incorporate evaluative, instructional, and descriptive intents. An example is [179], which employs a game-like setting where users provide linguistic feedback that includes preferences, instructions, and descriptive elements, creating a robust, multi-faceted learning environment.

**Unintentional** — Unintentional feedback can be derived from spontaneous reactions or subconscious signals, providing insights that are indirectly informative. Studies like [193] leverage facial expression analysis to gather implicit satisfaction levels, which are used to enhance agent adaptation. Additionally, physiological signals, such as EEG potentials, can indicate user discomfort or attention

focus, as explored in [85]. Eye-tracking is another modality used for unintentional feedback, where gaze patterns reveal regions of interest, aiding agents in refining state representations.

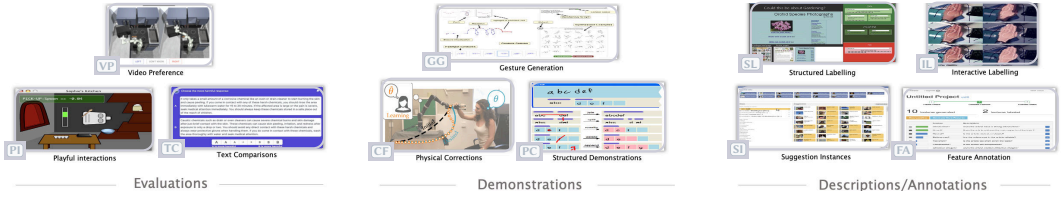


Fig. 8. **A selection of proposed solutions to match human intentions** – There have been several proposed solutions for user interfaces, the feedback processor, and the reward model. For user interfaces, we highlight (A) evaluative workflows, (B) demonstrations, and (C) descriptions.

*Opportunity 1. Matching Human Intent:* Even though some preliminary studies exist [44, 90, 134, 137], we see much potential in **combining feedback of multiple intents**. To enable expressive **Q1** feedback, it is worth going beyond simple interactions, e.g., ratings or pairwise preferences, and towards more complex interactions, enabling instructive or descriptive feedback if applicable. Giving the human user the ability to choose an appropriate interaction that matches their intent in a given situation can potentially help greatly improve the expressiveness of feedback. Adapting interactions to match human intent can also help improve ease **Q2** by reducing fatigue due to repetitive feedback or expressing feedback in a way that feels comfortable and requires an acceptable amount of cognitive effort in a given moment.

**5.2.2 Expression Form **D2**** On the topic of expression form, there has been much work in fields like human-robot interaction [40, 63], as well as interactive RL [111]. Due to the large body of existing work, we want to limit the discussion of implicit feedback in particular and point at related literature for inspiration [40, 111].

**Explicit Expression – **TC**** Explicit expression methods enable users to convey feedback with clarity and precision, often using textual interfaces or simple selection tools. For example, chat interfaces are commonly employed for preference ranking or qualitative feedback responses, as seen in Ouyang et al. [150] or Bai et al. [16]. Here, users can explicitly express preferences between responses, guiding model fine-tuning. Additionally, fine-grained feedback via the use of interactive text annotation interfaces has been explored [206], where explicit feedback is given directly on textual content, allowing for targeted improvements.

**Implicit Expression** – Implicit expression captures feedback indirectly, often through involuntary or less direct cues. Interfaces using eye-tracking technology, for example, capture the user's focus points to subtly guide agents in interpreting regions of interest [10]. Physiological signals, such as EEG measurements [207], can also provide feedback on user engagement, attention, or comfort, allowing systems to adapt based on real-time emotional states.

**Mixed-Modal Expression** – Combining explicit and implicit feedback interactions is still relatively rare.

*Opportunity 2. Expression Form:* Human-computer/human-robot interaction presents a huge opportunity to create novel feedback interactions. **Dynamic feedback interactions** can improve expressiveness by providing humans with natural and versatile input mechanics [Q1](#). Both explicit and implicit interactions can be designed with ergonomic principles to reduce cognitive load and stress, i.e., effort overall. [Q2](#). We see great potential in human-centered studies to investigate implicit user interactions for human feedback and optimize explicit interactions via user interfaces.

**5.2.3 Engagement** [D3](#) How to engage humans in RLHF is a central topic because we need to balance the quantity and quality of human feedback with human requirements such as fatigue and cognitive demand. As outlined, we can design proactive and reactive feedback strategies, i.e., letting humans select interesting instances or querying for feedback.

However, a promising middle ground is to combine selection and querying effectively: Instead of just improving querying strategies, users can be instructed to provide better (more AI-appropriate) teaching. Optimal teaching for humans can substantially differ from optimal teaching for AI [34]. In active learning, a type of semi-supervised learning, the primary goal is to achieve high accuracy with minimal manual labeling effort. This is achieved by a candidate selection strategy identifying the instances that potentially contribute most to the learning progress of the model. Several strategies have been proposed, including *uncertainty sampling*, *relevance-based selection*, *data-dependent sampling* or sampling based on *error reduction estimation* [21, 149, 160].

Reinforcement learning agents learn in a potentially large state space of unknown/unlabeled states, which must be reached by exploration. In active learning [21, 149, 160], exploration is undertaken by the human annotators, although often supported by automatic metrics, while it is generally fully automated in RL. Therefore, existing work in human-in-the-loop reinforcement learning has focused on integrating humans into the learning process to effectively guide the exploration of new states or the exploitation of good optimal states. Compared to supervised learning, giving feedback for state-action sequences (similar to unlabeled states) shapes the exploration of an acting agent, influencing the composition of the set of possible states. Selecting such high-impact states is still an unsolved question for on-policy learning and might, therefore, especially profit from human oversight and expertise.

Visual interactive labeling, established by [23], describes the unification of labeling across machine learning and visual analytics. The *VIAL* process introduces specific important steps, which we also find for reward-based learning from human feedback: (1) The necessity to visualize a model and its progress during training, (2) A visualization of results, i.e., the classification results. (3) Candidate suggestion, which needs to find a balance between instances that improve the model most while respecting a user's perspective and capacity. (4) A labeling interface appropriate for the used data instance and label types. (5) A way to interpret user feedback. Multiple works have explored visual interactive labeling, e.g., integrating gamification and explainable AI [162].

*Opportunity 3. Engagement:* Query mechanics for single feedback types have already been widely studied. However, querying for multiple feedback types has been explored to a small degree. Investigating this kind of **situational engagement** has great potential because querying for the right feedback type at the right time can improve expressiveness [Q1](#), reduce effort [Q2](#) while preserving informativeness [Q7](#). Finally, mixed strategies like visual interactive labeling could be highly effective for RLHF, and there is significant potential for exploring such approaches further. The opportunities for human-centered human feedback are summarized in Figure 9.

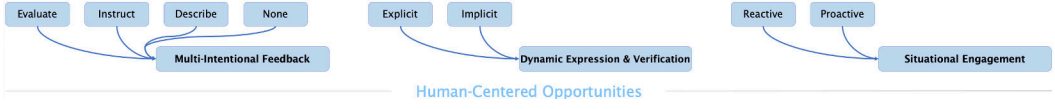


Fig. 9. A Summary of Opportunities for Human-Centered Dimensions

### 5.3 Systems for Feedback of Different Interface-Centered Dimensions

We want to discuss proposed solutions and opportunities for feedback on different relations, content levels, and actuality. Compared to the previous dimension, the choice of interface-centered dimensions is more dependent on the actual type of target (i.e., text, images, numerical values), so we will comment on this difference when appropriate.

#### Interface-Centered

Relevant Dimensions: Target Relation **D4**, Content Level **D5**, Target Actuality **D6**

Relevant Qualities: Definiteness **Q3**, Context Independence **Q4**

**5.3.1 Target Relation **D4**** The move from absolute to relative feedback, i.e., from approaches like interactive shaping to preference-based RL, has been established over the last few years. Preference-based RL is now the standard RLHF approach, particularly for fine-tuning LLMs [150, 223]. Binary preferences strike a favorable combination of simplicity, which is especially relevant for crowd-sourced feedback and stability of feedback.

So far, relative feedback, such as preferences or rankings, has been used almost exclusively for evaluations. Here, they are easily implemented, e.g., via simple button interfaces, and work for modalities that can be visualized (text, but also videos of execution). However, some work has, e.g., investigated the use of implicit feedback methods to extract relative feedback.

So far, relative feedback for instructions or descriptions has been investigated to a far smaller degree. Here, a potential example would be action advice with multiple actions (i.e., recommending a best, but also second-, third-best action, etc.). Similarly, users could be asked first to generate a few demonstrations and then rank them according to self-perceived quality, resulting in relative demonstrative feedback. Similarly, it might be possible to formulate relative descriptive statements or rules instead of absolute constraints or descriptive statements.

*Opportunity 4. Target Relation:* As relative feedback often has significant advantages in terms of context dependency **Q4**, a currently under-explored research direction is the use of instructive/descriptive feedback as relative feedback, e.g., by giving humans the ability to express preferences between demonstrations or descriptions like rules and constraints. By varying contextual feedback for the same target, we can use relative relationships between measurements to define influencing factors **Q3**.

**5.3.2 Content Level **D5**** A critical dimension to improving feedback's expressiveness is using feature-level feedback. Instance-level is the most dominant type of feedback because it can be implemented via simple user interactions (like sliders, buttons, mimics detection [42, 86, 105]). Even though user interactions for demonstrative feedback are generally more challenging to implement, instance-based feedback requires uni-directional interactions that let human users select or generate actions. On the other hand, feature-level feedback requires more complex user interfaces in order to let users select parts of the feature space (e.g., selecting regions of image-based input as salient features [137, 219]). Some feature-level feedback might be collected via implicit, non-conscious

interactions such as gaze tracking [220]. However, we might collect such feedback via explicit **annotations**, i.e., by allowing a human user to annotate behavior and states, selecting sub-features of the generated state-action sequences [217].

The differentiation between instance- and feature-level feedback for language models is more challenging because we could interpret each predicted token as an action of the *agent* language model. Therefore, annotating single words from a fully generated text could be interpreted as fine-grained feature-level feedback [206]. However, a better interpretation of feature-level feedback in the context of natural language might be the annotation of specific facts or generation patterns that occur over multiple generations (i.e., are not bound to a specific generated instance). User interfaces allowing these feature-level annotations might open novel ways to train language models.

Finally, meta-level feedback is an emerging technique in RLHF. Such feedback can be integrated into the normal interaction process, either before a session or during normal feedback interactions. One example is *skill assessment questions* [170], which uses additional contextual questions not directly related to feedback to assess the domain knowledge of a human labeler. This, in turn, allows us to calibrate the expected quality of reward for certain difficult queries. For such queries, feedback from domain experts should be rated higher than feedback from a lay user.

A second example is *distinguishability queries* [60], which combines pairwise preferences with queries to measure which segment pairs are easy or difficult to distinguish for labelers. In turn, we can trust preference pairs more highly distinguishable than pairs that are only hard to differentiate. We can also adopt our querying strategy to serve pairs specifically.

*Opportunity 5. Content Level:* In our mind, moving beyond instance-level feedback is a key step towards more expressive feedback. Feature-level feedback suits itself to define human values and preferences more fine-grained [Q3] and independent of the immediate feedback context [Q4]. Exploring ways to collect, encode, and process this feedback is crucial for more powerful human-AI communication. Meta-level feedback can be intensively valuable in understanding contextual factors of human feedback and calibrating the reward learning process accordingly. Combining instance-level feedback with content-level **annotations** and meta-level interactions is a way to combine learnability [Q5][Q6][Q7] with expressiveness [Q1], improving definiteness of feedback [Q3] and context dependence [Q4].

**5.3.3 Target Actuality [D6]** Generative feedback generally requires more complex user interactions compared to feedback for observed states. Feedback for observed states can be simple content-level feedback (such as preferences and scores) and feature-level feedback (annotations, descriptions). Such feedback, therefore, requires user interactions for **selection**, potentially at different temporal granularity (see [D7]), feature granularity. We need an interaction to select a correct action for demonstrative feedback, such as action advice.

For generative feedback, we need to provide **generative** user interactions, i.e., a way to choose and execute actions that generate demonstrations, but also sub-goal generation [208], creation of novel rules or constraints [97, 203]. For natural language targets, the generation of demonstrative feedback is simple in principle and can be implemented easily as normal text input. However, more complex user interactions can enable more expressive text-based feedback (such as descriptions).

Generating new targets without any reference might be overwhelming to users, especially if they are not experienced with certain AI systems. Providing stronger guidance in the form of references can improve the ability of humans to generate targeted and helpful feedback at the price of potentially biasing generation if the reference influences the user too strongly.

*Opportunity 6. Target Actuality:* Generative feedback, not reliant on observed behavior, can be highly informative and serve learning agents as strong guidance for exploration, especially at lower skill levels. Therefore, implementing systems that enable such generative feedback is a worthwhile challenge, especially in domains where explicit input or demonstrations are difficult (e.g., in robotics, the explicit specification of robot positions is difficult); we need to develop novel user interactions for generated feedback. Not losing information in this process (Q3) while maintaining a consistent context (Q4) remain important challenges. To not overwhelm the user, we can focus specifically on the **augmentation** of existing behavior, i.e., giving users a starting point (like an initial state-action sequence or another type of reference) to generate new targets for feedback.

The opportunities for human-centered human feedback are summarized in Figure 10.

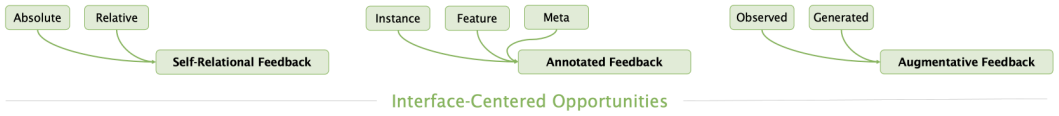


Fig. 10. A Summary of Opportunities for Interface-Centered Dimensions

## 5.4 Systems for Feedback of Different Model-Centered Dimensions

Finally, we want to discuss proposed solutions and opportunities for feedback on different granularities, choice set sizes, and exclusivity.

### Model-Centered

Relevant Dimensions: Temporal Granularity (D7), Choice Set Size (D8), Exclusivity (D9)

Relevant Qualities: Precision (Q5), Unbiasedness (Q6), Informativeness (Q7)

**5.4.1 Temporal Granularity (D7)** An element missing from simple approaches, e.g., in ranking-based reward learning [223], is a way to get a good overview of the candidate set, i.e., the set of possible instances available for feedback. For inspiration, we can look at existing visual interactive labeling approaches [22, 23, 23] that describe the unification of labeling across machine learning and visual analytics. The *VIAL* process introduces certain important steps, which we also find for reward-based learning from human feedback: (1) Visualizing a model and its progress during training. (2) Visualizing results, i.e., the classification results. (3) Visualizing candidate suggestion, which needs to find a balance between instances that improve the model most while respecting a user's perspective and capacity. (4) A labeling interface appropriate for the used data instance and label types. (5) A way to interpret user feedback. Multiple works have explored visual interactive labeling, e.g., integrating gamification and explainable AI [162]. Instead of instances where we want to assign a discrete or numeric label, we are interested in assigning rewards to state-action sequences of varying lengths. Compared to data types such as images, state-action sequences can have a high degree of heterogeneity, particularly due to different lengths and content.

We examine existing approaches to visualizing state-action sequences in interactive machine-learning systems. As previously outlined, these instances can include states, observations, actions, scalar quantities such as cumulative reward, and sequences thereof, for example, state-action sequences or more abstract objects like policies.

**Single States** — We can differentiate between two main approaches to state visualizations. In the first, states are visualized directly, and in the second, the visualization of the state is a function of the



state itself, which applies further processing or abstraction. Many popular reinforcement learning environments already provide visual representations of states that could convey the agent's state without further processing. **(DS)** In the *ATARI* domain [20], an observation itself is a raster image of the game state and is thus inherently visual. In other domains, however, visualizing states directly is not possible. **(IS)** In the *MuJoCo* domain [52], the state corresponds to a configuration of joint angles of robot skeletons, which, even for a small number of joints, is hard to visualize and understand intuitively. Instead, the state is shown by creating a 3D rendering of the robot in its environment.

**Single Actions** — Actions are often not visualized directly but are left implicit and need to be inferred by the user by comparing consecutive visualizations of states, which might be reasonable for small and discrete action spaces but becomes infeasible for large or continuous spaces, where the size of the action space does not permit this approach. **(AC)** A possible approach to visualizing distributions of chosen actions, divorced from the corresponding state, is exemplified with *DQNViz* on the *Breakout* game with a small, discrete action space [197]. The user interface provides a summary of chosen, discrete action distributions per episode as a pie chart and the temporal progression of these action distributions throughout training as a stacked bar chart.

**State-Action Segments** — States, chosen actions, and resulting states are highly interdependent, and as a result, existing visualizations often display them jointly. A widely-used approach to visualizing state-action sequences is to show them as videos, where each frame corresponds to the visualization of a single state in the sequence [15, 133]. As alluded to earlier, this approach has its limitations, as it does not explicitly show the actions taken by the agent.

**(SA)** A possible remedy can be found in Luo et al. [128], where state-actions sequences in a visual environment are displayed as grids consisting of visual observations, which have a different tint depending on which of the three possible actions the agent has taken.

**(ST)** *DQNViz* [197] also uses joint displays of states and actions in the *Breakout* environment by displaying the agent's state and actions in longitudinal scatter plots, where the x-coordinate corresponds to the time-step in the episode, and the y-coordinate corresponds to the horizontal position in the environment. The color of the marker encodes one of three actions, either *left*, *right*, or *idle*.

Similar approaches have been used for continuous action spaces, e.g., in Verma et al. [194], which plotted continuous actions against time steps.

**State-Action Distributions** — While visualizing the state or observation of an agent provides an intuitive understanding of single states, understanding the distribution of such states and their relationship to each other is challenging due to their often high-dimensional nature. Dimensionality reduction methods have been used to project state-action pairs into a low-dimensional latent space, providing an overview of a state-action distribution found, e.g., in replay buffers [51]. The authors suggest that RL users can obtain a quick overview of the state-action diversity of an agent. **(SD)** This approach was leveraged to highlight differences in state-action distribution between different RL agents [176].

**Policies** — Instead of visualizing the artifacts generated by policies, e.g., state-action distributions as described before, some authors have attempted to provide visual summaries of policies directly.

**(PO)** Lyu et al. [129] use *symbolic deep reinforcement learning* to formulate policies as symbolic plans, which they then draw into the environment.

---

*Opportunity 7. Temporal Granularity:* While many approaches for single granularities exist, there is still a gap in the simultaneous visualization of state-action sequences, which restricts effective exploration and selection by humans. While the ability to give feedback at different granularities also positively impacts expressiveness **(Q1)**,

---





Fig. 11. A selection of proposed visualizations for different granularities in RL scenarios – several proposed solutions for user interfaces display elements of a reinforcement learning process at different granularities.

**5.4.2 Choice Set Size (D8)** The available choice set is highly dependent on the task domain, but there are choices a designer can make for different feedback.

As with other dimensions, we find a trade-off in the choice set size: A larger choice set is potentially more expressive and informative. Having only a limited set of choices might feel restrictive, i.e., the human user feels like they lack control over the feedback. In terms of information, k-wise ranking compared to binary preferences [142, 222] can be more informative because a k-wise ranking contains multiple binary preferences. Similarly, the choice of the best action out of a larger set of available actions could be interpreted as ruling out more sub-optimal actions.

Conversely, a larger set of possible choices can lead to user fatigue and higher cognitive demand. Additionally, a larger choice set can also lead to a higher variance in the given values, although it might exhibit a larger bias if the given choices do not match the true desired internal best choice.

*Opportunity 8. Choice Set Size:* The choice set size is partially dependent on the chosen task, and there are apparent designs based on other chosen dimensions (e.g., for a demonstration, the choice set size is dependent on the action space, i.e., from which set of actions I can choose for the demonstration). However, there is potential in varying the choice set to find an optimal setup. For example, rankings might have advantages over pairwise comparisons; similarly, instead of generating a novel trajectory in a continuous space, we might want to select from a set of candidate trajectories instead to improve efficiency (Q2). We might find that a more extensive choice set size might positively influence unbiasedness (Q6), because the human is not forced to provide values in a restricted range. On the other hand, a large choice set size might lead to a higher variance (Q6). Finally, a small choice set size might increase learning speed because the model can be simpler, but choosing from a large choice set could be more precise and, therefore, informative (Q7). A potential remedy is using **adaptive choice feedback**, which dynamically adapts the choice size during the feedback generation process according to human and model qualities.

**5.4.3 Exclusivity (D9)** Training a new reward model for each human, estimating irrationality or the connection between internal reward and observable behavior is infeasible for most learning scenarios. To overcome this, we can draw inspiration from recommender systems that are challenged with the so-called “cold-start” problem, i.e., the initialization of models without any data [2, 195]. We can utilize meta-learning (ML) strategies to quickly adapt models to the given use case. A way to achieve this would be to train *population* models that, e.g., capture certain feedback types’ general irrationality. Based on this model, we could train a *user* model that captures the particularities of the user, e.g., specific ways implicit feedback is communicated. Finally, we might then use a *session* or *task* model that learns a reward function specific to the given task and can also consider session-specific feedback quality.

*Opportunity 9. Exclusivity:* Experimenting with the use of **multiple feedback sources** is a promising avenue. When dealing with feedback from multiple human labels, as for large-scale post-training of LLMs [150], we are natively confronted with non-exclusive reward sources and potential problems like labeler disagreement. This can lead to problems like decreased precision [Q5](#).

Different reward sources might exhibit different biases (like rating behavior based on different internal scales) or different preferences. Modeling different non-exclusive reward sources directly might help to alleviate some issues, actually correcting individual human biases [Q6](#), and detecting the quality of sources in terms of informativeness [Q7](#).

The opportunities for human-centered human feedback are summarized in Figure 12.



Fig. 12. A Summary of Opportunities for Model-Centered Dimensions

## 6 Discussion and Outlook

The goal of training with tight human-AI interaction is to create powerful, aligned, and trustworthy agents. Considering our running example, Our household robot might now be able to perform various additional tasks, act according to our usual preferences, and be predictable and reliable in its movements and actions.

As we have outlined in this paper, we think that diverse and complementary feedback types are essential for humans to give expressive feedback, require appropriate effort, enable reasoning about contextual information and uncertainty, and are learnable and informative. Powerful and natural human-AI interactions may only be achieved if diverse interactions and feedback types are blended dynamically and flexibly. Thus, human-AI communication incorporating multiple types of human feedback, expressive reward modeling, querying, and translation are part of a conceptual model for the next steps in RL with human feedback. Future work across different areas can contribute towards the common goal of human-aligned intelligent agents:

- (1) **Explore the Space of Human Feedback** Widening the space of human feedback types that RL-based systems can use enables more expressive and accessible systems to train AI agents. By using different interactions for feedback, humans have more ways to express their goals and preferences to AI systems, which can improve safety, robustness, and satisfaction on the users' side.
- (2) **Holistic Visual Interactive Systems** As we show in our survey, although many isolated solutions exist, there is huge potential in systems that implement a holistic approach for RL from diverse human feedback. Researchers with HCI or visualization backgrounds can contribute systems to evaluate the utility of different feedback types. In Figure 13, we show a prototype of a possible interface integrating several types of feedback into a common interface.
- (3) **Towards Expressive Reward Models** To capture the full fidelity of human feedback, including factors like uncertainty, we need reward models aware of human-centered and structured dimensions of feedback and the user or task context. To achieve this, these models need to receive additional information as input, for example, user and session information, potentially model temporal dynamics, and need the capacity to learn the irrationality or uncertainty of different feedback types. Meta-learning approaches can support adapting expressive models to new users and keep the demand for data and calibration in check.

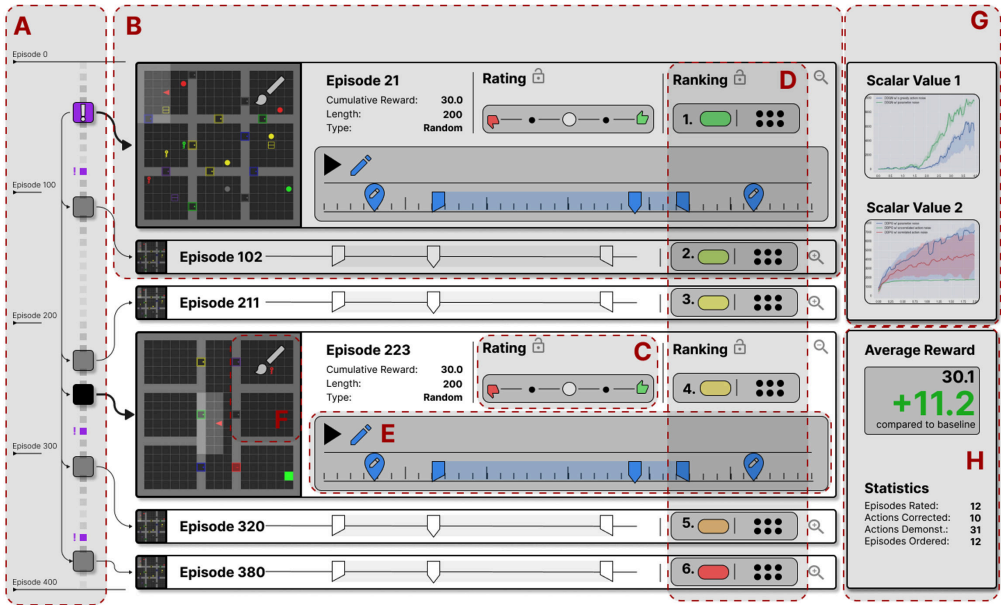


Fig. 13. Prototype implementation of an integrated system for multi-type human feedback in reinforcement learning. (A) is a scrollable list of episodes: The user can be queried for feedback by visual indicators, selected episodes are visible, and the user can visit previous episodes to investigate the effect of feedback. In the center column, we have an episode view with multiple feedback interactions (B): The user evaluates the episodes via ratings (C), ranks via drag & drop (D), select episode segments for correction (E), or brush to highlight important features (F). Furthermore, the human annotator is informed of global model training via output metrics (G). Shown is *BabyAI* [39].

**(4) Empirical Investigation of Human Feedback** When estimating the quality of feedback, like informativeness or consistency, we should not rely on gut feeling or scale alone. Instead, we should ground such considerations on resulting modeling decisions on rigorous experimentation with human subjects across diverse backgrounds and tasks [173]. Versatile visual interactive systems that allow for experimentation can be crucial tools to enable this experimentation. We find a critical current research gap in investigating the trade-off between natural language feedback, which was proposed in recent papers [74, 178]. However, as has been argued in the space of explainable AI [161], natural language has its limit, and we might instead view other modalities as complementary. For example, language alone can make it difficult to refer, e.g., to a certain granularity, feature, etc.

The given conceptual framework should enable thinking about human feedback in a structured and systematic way and inspire future system design for expressive, efficient, and highly effective learning of reward models. However, there are many learning factors from human feedback that are still not covered, e.g., temporal dynamics, or this framework does not fully discuss co-adaptation. Furthermore, although we developed the framework in an iterative process, there might be other ways to frame the problem space for different use cases.

## 7 Conclusion

This paper discusses a conceptual framework for human feedback for reinforcement learning in human-AI collaboration. We have presented nine dimensions to classify feedback types, encompassing human-centered, interface-centered, and model-centered dimensions. Additionally, we have presented seven quality criteria for human feedback. We outline the design requirements for interactive systems using reinforcement learning from human feedback for the three core components. We argue that systems using a rich set of possible feedback with expressive reward models, supported by targeted visual interactive interfaces, can lead toward human-aligned intelligent agents.

## References

- [1] A. Adadi and M. Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.
- [2] M Mehdi Afsar, Trafford Crump, and Behrouz Far. 2022. Reinforcement learning based recommender systems: A survey. *Comput. Surveys* 55, 7 (2022), 1–38.
- [3] Sharath Chandra Akkaladevi, Matthias Plasch, Andreas Pichler, and Markus Ikeda. 2019. Towards Reinforcement based Learning of an Assembly Process for Human Robot Collaboration. *Procedia Manufacturing* 38 (2019), 1491–1498. <https://doi.org/10.1016/j.promfg.2020.01.138> 29th International Conference on Flexible Automation and Intelligent Manufacturing ( FAIM 2019), June 24–28, 2019, Limerick, Ireland, Beyond Industry 4.0: Industrial Advances, Engineering Education and Intelligent Manufacturing.
- [4] Sharath Chandra Akkaladevi, Matthias Plasch, Andreas Pichler, and Markus Ikeda. 2019. Towards reinforcement based learning of an assembly process for human robot collaboration. *Procedia Manufacturing* 38 (2019), 1491–1498.
- [5] Riad Akrou, Marc Schoenauer, and Michèle Sebag. 2012. April: Active preference learning-based reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24–28, 2012. Proceedings, Part II* 23. Springer, 116–131.
- [6] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.
- [7] Saleema Amershi, James Fogarty, and Daniel Weld. 2012. ReGroup: Interactive Machine Learning for On-Demand Group Creation. *Conference on Human Factors in Computing Systems - Proceedings* (05 2012). <https://doi.org/10.1145/2207676.2207680>
- [8] Ofra Amir, Ece Kamar, Andrey Kolobov, and Barbara J. Grosz. 2016. Interactive Teaching Strategies for Agent Training. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) (IJCAI'16). AAAI Press, 804–811.
- [9] Riku Arakawa, Sosuke Kobayashi, Yuya Unno, Yuta Tsuboi, and Shin-ichi Maeda. 2018. Dqn-tamer: Human-in-the-loop reinforcement learning with intractable feedback. *arXiv preprint arXiv:1810.11748* (2018).
- [10] Reuben M Aronson and Henny Admoni. 2022. Gaze complements control input for goal prediction during assisted teleoperation. In *Robotics science and systems*.
- [11] Christian Arzate Cruz and Takeo Igarashi. 2020. Mariomix: Creating aligned playstyles for bots with interactive reinforcement learning. In *Extended abstracts of the 2020 annual symposium on computer-human interaction in play*. 134–139.
- [12] Christian Arzate Cruz and Takeo Igarashi. 2020. A survey on interactive reinforcement learning: Design principles and open challenges. *DIS 2020 - Proceedings of the 2020 ACM Designing Interactive Systems Conference* (July 2020), 1195–1209. <https://doi.org/10.1145/3357236.3395525> arXiv: 2105.12949 Publisher: Association for Computing Machinery, Inc ISBN: 9781450369749.
- [13] Christian Arzate Cruz and Takeo Igarashi. 2020. A survey on interactive reinforcement learning: Design principles and open challenges. In *Proceedings of the 2020 ACM designing interactive systems conference*. 1195–1209.
- [14] Amanda Askill, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A General Language Assistant as a Laboratory for Alignment. *CoRR abs/2112.00861* (2021). arXiv:2112.00861 <https://arxiv.org/abs/2112.00861>
- [15] Akanksha Atrey, Kaleigh Clary, and David Jensen. 2020. Exploratory Not Explanatory: Counterfactual Analysis of Saliency Maps for Deep Reinforcement Learning. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=rkl3m1BFDB>

- [16] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).
- [17] Yuntao Bai, 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. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL] <https://arxiv.org/abs/2212.08073>
- [18] Chandrayee Basu, Erdem Biyik, Zhixun He, Mukesh Singhal, and Dorsa Sadigh. 2019. Active learning of reward dynamics from hierarchical queries. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 120–127.
- [19] Chandrayee Basu, Mukesh Singhal, and Anca D Dragan. 2018. Learning from richer human guidance: Augmenting comparison-based learning with feature queries. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 132–140.
- [20] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. 2013. The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research* 47 (jun 2013), 253–279.
- [21] Jürgen Bernard, Marco Hutter, Matthias Zeppelzauer, Dieter Fellner, and Michael Sedlmair. 2017. Comparing visual-interactive labeling with active learning: An experimental study. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 298–308.
- [22] Jürgen Bernard, Matthias Zeppelzauer, Markus Lehmann, Martin Müller, and Michael Sedlmair. 2018. Towards User-Centered Active Learning Algorithms. *Computer Graphics Forum* 37, 3 (June 2018), 121–132. <https://doi.org/10.1111/cgf.13406>
- [23] Jürgen Bernard, Matthias Zeppelzauer, Michael Sedlmair, and Wolfgang Aigner. 2018. VIAL: a unified process for visual interactive labeling. *The Visual Computer* 34 (2018), 1189–1207.
- [24] Adam Bignold, Francisco Cruz, Richard Dazeley, Peter Vamplew, and Cameron Foale. 2021. An Evaluation Methodology for Interactive Reinforcement Learning with Simulated Users. *Biomimetics* 6, 1 (2021). <https://doi.org/10.3390/biomimetics6010013>
- [25] Adam Bignold, Francisco Cruz, Richard Dazeley, Peter Vamplew, and Cameron Foale. 2021. Persistent rule-based interactive reinforcement learning. *Neural Computing and Applications* (2021), 1–18.
- [26] Adam Bignold, Francisco Cruz, Richard Dazeley, Peter Vamplew, and Cameron Foale. 2023. Human engagement providing evaluative and informative advice for interactive reinforcement learning. *Neural Computing and Applications* 35, 25 (2023), 18215–18230.
- [27] Erdem Biyik, Nicolas Huynh, Mykel J Kochenderfer, and Dorsa Sadigh. 2023. Active preference-based Gaussian process regression for reward learning and optimization. *The International Journal of Robotics Research* (2023), 02783649231208729.
- [28] Andreea Bobu, Marius Wiggert, Claire Tomlin, and Anca D Dragan. 2021. Feature expansive reward learning: Rethinking human input. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 216–224.
- [29] W. Bradley Knox and Peter Stone. 2008. TAMER: Training an Agent Manually via Evaluative Reinforcement. In *2008 7th IEEE International Conference on Development and Learning*. 292–297. <https://doi.org/10.1109/DEVLRN.2008.4640845>
- [30] Satchuthananthavale RK Branavan, Harr Chen, Luke Zettlemoyer, and Regina Barzilay. 2009. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 82–90.
- [31] Daniel S Brown, Yuchen Cui, and Scott Niekum. 2018. Risk-aware active inverse reinforcement learning. In *Conference on Robot Learning*. PMLR, 362–372.
- [32] Daniel S. Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. 2019. Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations. <http://arxiv.org/abs/1904.06387> arXiv:1904.06387 [cs, stat].
- [33] Maya Cakmak, Siddhartha S Srinivasa, Min Kyung Lee, Jodi Forlizzi, and Sara Kiesler. 2011. Human preferences for robot-human hand-over configurations. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1986–1993.
- [34] Maya Cakmak and Andrea L. Thomaz. 2014. Eliciting good teaching from humans for machine learners. *Artificial Intelligence* 217 (2014), 198–215. <https://doi.org/10.1016/j.artint.2014.08.005>

- [35] Kate Candon, Jesse Chen, Yoony Kim, Zoe Hsu, Nathan Tsoi, and Marynel Vázquez. 2023. Nonverbal Human Signals Can Help Autonomous Agents Infer Human Preferences for Their Behavior. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 307–316.
- [36] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217* (2023).
- [37] Colin Cherry. 1966. On human communication. (1966).
- [38] Mohamed Chetouani. 2021. Interactive Robot Learning: An Overview. *ECCAI Advanced Course on Artificial Intelligence* (2021), 140–172.
- [39] Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. 2018. BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning. *7th International Conference on Learning Representations, ICLR 2019* (Oct. 2018). <https://arxiv.org/abs/1810.08272v4> arXiv: 1810.08272 Publisher: International Conference on Learning Representations, ICLR.
- [40] Vivienne Bihe Chi and Bertram F Malle. 2022. Instruct or evaluate: how people choose to teach norms to social robots. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 718–722.
- [41] Bernard CK Choi and Anita WP Pak. 2004. A catalog of biases in questionnaires. *Preventing chronic disease* 2, 1 (2004), A13.
- [42] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. 30 (2017), 4299–4307. <https://proceedings.neurips.cc/paper/2017/file/d5e2c0adad503c91f91df240dcd4e49-Paper.pdf>
- [43] Yun-Shiuan Chuang, Xuezhou Zhang, Yuzhe Ma, Mark K. Ho, Joseph L. Austerweil, and Xiaojin Zhu. 2021. Using Machine Teaching to Investigate Human Assumptions when Teaching Reinforcement Learners. arXiv:2009.02476 [cs.LG]
- [44] Laure Crochepierre, Lydia Boudjeloud-Assala, and Vincent Barbesant. 2022. Interactive reinforcement learning for symbolic regression from multi-format human-preference feedbacks. In *31st International Joint Conference on Artificial Intelligence (IJCAI 2022)*.
- [45] Christian Arzate Cruz and Takeo Igarashi. 2021. Interactive explanations: Diagnosis and repair of reinforcement learning based agent behaviors. In *2021 IEEE Conference on Games (CoG)*. IEEE, 01–08.
- [46] Francisco Cruz, Adam Bignold, Hung Son Nguyen, Richard Dazeley, and Peter Vamplew. 2022. Broad-persistent Advice for Interactive Reinforcement Learning Scenarios. *arXiv preprint arXiv:2210.05187* (2022).
- [47] Francisco Cruz, German I. Parisi, and Stefan Wermter. 2018. Multi-modal Feedback for Affordance-driven Interactive Reinforcement Learning. In *2018 International Joint Conference on Neural Networks (IJCNN)*. 1–8. <https://doi.org/10.1109/IJCNN.2018.8489237>
- [48] Yuchen Cui, Qiping Zhang, Brad Knox, Alessandro Allievi, Peter Stone, and Scott Niekum. 2021. The empathic framework for task learning from implicit human feedback. In *Conference on Robot Learning*. PMLR, 604–626.
- [49] Felipe Leno Da Silva, Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. 2020. Uncertainty-aware action advising for deep reinforcement learning agents. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 5792–5799.
- [50] Joris De Winter, Albert De Beir, Ilias El Makrini, Greet Van de Perre, Ann Nowé, and Bram Vanderborght. 2019. Accelerating interactive reinforcement learning by human advice for an assembly task by a cobot. *Robotics* 8, 4 (2019), 104.
- [51] Shuby V. Deshpande, Jeff Schneider, Deepak Pathak, David Held, and Benjamin Eysenbach. 2020. *Towards Interpretable Reinforcement Learning, Interactive Visualizations to Increase Insight*. Master’s thesis. Carnegie Mellon University.
- [52] Yan Duan, Xi Chen, Rein Houthoofd, John Schulman, and Pieter Abbeel. 2016. Benchmarking Deep Reinforcement Learning for Continuous Control. arXiv:1604.06778 [cs.LG]
- [53] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 8 (jun 2018), 37 pages. <https://doi.org/10.1145/3185517>
- [54] Layla El Asri, Bilal Piot, Matthieu Geist, Romain Laroche, and Olivier Pietquin. 2016. Score-based Inverse Reinforcement Learning. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (Singapore, Singapore) (AAMAS ’16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 457–465.
- [55] Mennatallah El-Assady and Caterina Moruzzi. 2022. Which Biases and Reasoning Pitfalls Do Explanations Trigger? Decomposing Communication Processes in Human–AI Interaction. *IEEE Computer Graphics and Applications* 42, 6 (2022), 11–23.
- [56] Ahmed Elgohary, Christopher Meek, Matthew Richardson, Adam Fourney, Gonzalo Ramos, and Ahmed Hassan Awadallah. 2021. NL-EDIT: Correcting Semantic Parse Errors through Natural Language Interaction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard,

- Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 5599–5610. <https://doi.org/10.18653/v1/2021.naacl-main.444>
- [57] Anestis Fachantidis, Matthew E Taylor, and Ioannis Vlahavas. 2017. Learning to teach reinforcement learning agents. *Machine Learning and Knowledge Extraction* 1, 1 (2017), 21–42.
  - [58] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces* (Miami, Florida, USA) (IUI '03). Association for Computing Machinery, New York, NY, USA, 39–45. <https://doi.org/10.1145/604045.604056>
  - [59] Taylor A Kessler Faulkner, Elaine Schaertl Short, and Andrea L Thomaz. 2020. Interactive reinforcement learning with inaccurate feedback. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 7498–7504.
  - [60] Xuening Feng, Zhaohui JIANG, Timo Kaufmann, Eyke Hüllermeier, Paul Weng, and Yifei Zhu. 2024. Comparing Comparisons: Informative and Easy Human Feedback with Distinguishability Queries. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*. <https://openreview.net/forum?id=4TujKMpSKH>
  - [61] Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, and André F. T. Martins. 2023. Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation. *Transactions of the Association for Computational Linguistics* 11 (2023), 1643–1668. [https://doi.org/10.1162/tacl\\_a\\_00626](https://doi.org/10.1162/tacl_a_00626)
  - [62] Pierre W Ferrez and José del R Millán. 2005. You are wrong!—automatic detection of interaction errors from brain waves. In *Proceedings of the 19th international joint conference on artificial intelligence*.
  - [63] Tesca Fitzgerald, 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 *Conference on Robot Learning*. PMLR, 2241–2250.
  - [64] Spencer Frazier and Mark Riedl. 2019. Improving deep reinforcement learning in minecraft with action advice. In *Proceedings of the AAAI conference on artificial intelligence and interactive digital entertainment*, Vol. 15. 146–152.
  - [65] Rachel Freedman, Rohin Shah, and Anca D. Dragan. 2021. Choice Set Misspecification in Reward Inference. *CoRR* abs/2101.07691 (2021). arXiv:2101.07691 <https://arxiv.org/abs/2101.07691>
  - [66] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, and Saurav Kadavath et al. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv:2209.07858 [cs.CL]
  - [67] Yang Gao, Christian M Meyer, and Iryna Gurevych. 2018. APRIL: Interactively learning to summarise by combining active preference learning and reinforcement learning. *arXiv preprint arXiv:1808.09658* (2018).
  - [68] Gaurav R Ghosal, Matthew Zurek, Daniel S Brown, and Anca D Dragan. 2022. The Effect of Modeling Human Rationality Level on Learning Rewards from Multiple Feedback Types. *arXiv preprint arXiv:2208.10687* (2022).
  - [69] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems* 26 (2013).
  - [70] Lin Guan, Mudit Verma, Suna Sihang Guo, Ruohan Zhang, and Subbarao Kambhampati. 2021. Widening the pipeline in human-guided reinforcement learning with explanation and context-aware data augmentation. *Advances in Neural Information Processing Systems* 34 (2021), 21885–21897.
  - [71] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science robotics* 4, 37 (2019), eaay7120.
  - [72] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems* 29 (2016).
  - [73] Daniel Harnack, Julie Pivin-Bachler, and Nicolás Navarro-Guerrero. 2023. Quantifying the effect of feedback frequency in interactive reinforcement learning for robotic tasks. *Neural Computing and Applications* 35, 23 (2023), 16931–16943.
  - [74] Brent Harrison, Upol Ehsan, and Mark O Riedl. 2017. Guiding reinforcement learning exploration using natural language. *arXiv preprint arXiv:1707.08616* (2017).
  - [75] Frederick Hayes-Roth, Philip Klahr, and David J Mostow. 2013. Advice taking and knowledge refinement: An iterative view of skill acquisition. In *Cognitive skills and their acquisition*. Psychology Press, 231–253.
  - [76] Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human Feedback is not Gold Standard. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=7W3GLNImfS>
  - [77] Eric Hsiung, Eric Rosen, Vivienne Bihe Chi, and Bertram F Malle. 2022. Learning reward functions from a combination of demonstration and evaluative feedback. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 807–811.
  - [78] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. 2017. Imitation Learning: A Survey of Learning Methods. 50, 2, Article 21 (apr 2017), 35 pages. <https://doi.org/10.1145/3054912>
  - [79] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems* 31 (2018).



- [80] Hong Jun Jeon, Smitha Milli, and Anca D. Dragan. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. <http://arxiv.org/abs/2002.04833> arXiv:2002.04833 [cs].
- [81] Kaixuan Ji, Jiafan He, and Quanquan Gu. 2024. Reinforcement Learning from Human Feedback with Active Queries. arXiv:2402.09401 [cs.LG] <https://arxiv.org/abs/2402.09401>
- [82] Kshitij Judah, Saikat Roy, Alan Fern, and Thomas Dietterich. 2010. Reinforcement learning via practice and critique advice. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 24. 481–486.
- [83] Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, and Danny Lange. 2020. Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627* (2020).
- [84] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925* (2023).
- [85] Su Kyoung Kim, Elsa Andrea Kirchner, Arne Stefes, and Frank Kirchner. 2017. Intrinsic interactive reinforcement learning—Using error-related potentials for real world human-robot interaction. *Scientific reports* 7, 1 (2017), 1–16.
- [86] W Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the fifth international conference on Knowledge capture*. 9–16.
- [87] W Bradley Knox and Peter Stone. 2010. Combining manual feedback with subsequent MDP reward signals for reinforcement learning. In *AAMAS*. 5–12.
- [88] W Bradley Knox and Peter Stone. 2012. Reinforcement learning from human reward: Discounting in episodic tasks. In *2012 IEEE RO-MAN: The 21st IEEE international symposium on robot and human interactive communication*. IEEE, 878–885.
- [89] Jens Kober, J Andrew Bagnell, and Jan Peters. 2013. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research* 32, 11 (2013), 1238–1274.
- [90] Dorothea Koert, Maximilian Kircher, Vildan Salikutluk, Carlo D’Eramo, and Jan Peters. 2020. Multi-channel interactive reinforcement learning for sequential tasks. *Frontiers in Robotics and AI* 7 (2020), 97.
- [91] Pallavi Koppol. 2023. *Interactive Machine Learning from Humans: Knowledge Sharing via Mutual Feedback*. Ph.D. Dissertation. Carnegie Mellon University.
- [92] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W Woźniak. 2023. A survey on measuring cognitive workload in human-computer interaction. *Comput. Surveys* 55, 13s (2023), 1–39.
- [93] Samantha Krening. 2018. Newtonian action advice: Integrating human verbal instruction with reinforcement learning. *arXiv preprint arXiv:1804.05821* (2018).
- [94] Samantha Krening and Karen M Feigh. 2018. Interaction algorithm effect on human experience with reinforcement learning. *ACM Transactions on Human-Robot Interaction (THRI)* 7, 2 (2018), 1–22.
- [95] Samantha Krening, Brent Harrison, Karen M Feigh, Charles Lee Isbell, Mark Riedl, and Andrea Thomaz. 2016. Learning from explanations using sentiment and advice in RL. *IEEE Transactions on Cognitive and Developmental Systems* 9, 1 (2016), 44–55.
- [96] Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can Neural Machine Translation be Improved with User Feedback?. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, Srinivas Bangalore, Jennifer Chu-Carroll, and Yunyao Li (Eds.). Association for Computational Linguistics, New Orleans - Louisiana, 92–105. <https://doi.org/10.18653/v1/N18-3012>
- [97] Gregory Kuhlmann, Peter Stone, Raymond Mooney, and Jude Shavlik. 2004. Guiding a reinforcement learner with natural language advice: Initial results in RoboCup soccer. In *The AAAI-2004 workshop on supervisory control of learning and adaptive systems*. San Jose, CA.
- [98] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured Labeling for Facilitating Concept Evolution in Machine Learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) (*CHI '14*). Association for Computing Machinery, New York, NY, USA, 3075–3084. <https://doi.org/10.1145/2556288.2557238>
- [99] Minae Kwon, Siddharth Karamcheti, Mariano-Florentino Cuellar, and Dorsa Sadigh. 2021. Targeted data acquisition for evolving negotiation agents. In *International Conference on Machine Learning*. PMLR, 5894–5904.
- [100] Cassidy Laidlaw and Stuart Russell. 2021. Uncertain decisions facilitate better preference learning. *Advances in Neural Information Processing Systems* 34 (2021), 15070–15083.
- [101] Matthew V Law, Zhilong Li, Amit Rajesh, Nikhil Dhawan, Amritansh Kwatra, and Guy Hoffman. 2021. Hammers for Robots: Designing Tools for Reinforcement Learning Agents. In *Designing Interactive Systems Conference 2021*. 1638–1653.
- [102] Kimin Lee, Laura Smith, Anca Dragan, and Pieter Abbeel. 2021. B-pref: Benchmarking preference-based reinforcement learning. *arXiv preprint arXiv:2111.03026* (2021).

- [103] Guangliang Li, Hamdi Dibeklioğlu, Shimon Whiteson, and Hayley Hung. 2020. Facial feedback for reinforcement learning: a case study and offline analysis using the TAMER framework. *Autonomous Agents and Multi-Agent Systems* 34 (2020), 1–29.
- [104] Guangliang Li, Randy Gomez, Keisuke Nakamura, and Bo He. 2019. Human-Centered Reinforcement Learning: A Survey. *IEEE Transactions on Human-Machine Systems* 49, 4 (Aug. 2019), 337–349. <https://doi.org/10.1109/THMS.2019.2912447> Publisher: Institute of Electrical and Electronics Engineers Inc..
- [105] Guangliang Li, Bo He, Randy Gomez, and Keisuke Nakamura. 2018. Interactive reinforcement learning from demonstration and human evaluative feedback. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1156–1162.
- [106] Guangliang Li, Hayley Hung, Shimon Whiteson, and W Bradley Knox. 2013. Using informative behavior to increase engagement in the tamer framework. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. 909–916.
- [107] Kejun Li, Maegan Tucker, Erdem Bıyık, Ellen Novoseller, Joel W Burdick, Yanan Sui, Dorsa Sadigh, Yisong Yue, and Aaron D Ames. 2021. Roial: Region of interest active learning for characterizing exoskeleton gait preference landscapes. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3212–3218.
- [108] Mengxi Li, Alper Canberk, Dylan P Losey, and Dorsa Sadigh. 2021. Learning human objectives from sequences of physical corrections. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2877–2883.
- [109] Zhaoxing Li, Lei Shi, Alexandra I. Cristea, and Yunzhan Zhou. 2021. A Survey of Collaborative Reinforcement Learning: Interactive Methods and Design Patterns. In *Designing Interactive Systems Conference 2021 (Virtual Event, USA) (DIS '21)*. Association for Computing Machinery, New York, NY, USA, 1579–1590. <https://doi.org/10.1145/3461778.3462135>
- [110] Jessy Lin, Daniel Fried, Dan Klein, and Anca Dragan. 2022. Inferring rewards from language in context. *arXiv preprint arXiv:2204.02515* (2022).
- [111] Jinying Lin, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li. 2020. A review on interactive reinforcement learning from human social feedback. *IEEE Access* 8 (2020), 120757–120765.
- [112] Zhiyu Lin, Brent Harrison, Aaron Keech, and Mark O Riedl. 2017. Explore, exploit or listen: Combining human feedback and policy model to speed up deep reinforcement learning in 3d worlds. *arXiv preprint arXiv:1709.03969* (2017).
- [113] David Lindner and Mennatallah El-Assady. 2022. Humans are not Boltzmann Distributions: Challenges and Opportunities for Modelling Human Feedback and Interaction in Reinforcement Learning. (June 2022). <https://doi.org/10.48550/arxiv.2206.13316> arXiv: 2206.13316.
- [114] David Lindner, Rohin Shah, Pieter Abbeel, and Anca Dragan. 2021. Learning what to do by simulating the past. *arXiv preprint arXiv:2104.03946* (2021).
- [115] Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. 2018. Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 2060–2069. <https://doi.org/10.18653/v1/N18-1187>
- [116] Huihan Liu, Alice Chen, Yuke Zhu, Adith Swaminathan, Andrey Kolobov, and Ching-An Cheng. 2023. Interactive Robot Learning from Verbal Correction. *arXiv preprint arXiv:2310.17555* (2023).
- [117] Huihan Liu, Soroush Nasiriany, Lance Zhang, Zhiyao Bao, and Yuke Zhu. 2022. Robot Learning on the Job: Human-in-the-Loop Autonomy and Learning During Deployment. *arXiv preprint arXiv:2211.08416* (2022).
- [118] YuXuan Liu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. 2018. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1118–1125.
- [119] Robert Loftin, Bei Peng, James MacGlashan, Michael L Littman, Matthew E Taylor, Jeff Huang, and David L Roberts. 2014. Learning something from nothing: Leveraging implicit human feedback strategies. In *The 23rd IEEE international symposium on robot and human interactive communication*. IEEE, 607–612.
- [120] Robert Loftin, Bei Peng, James MacGlashan, Michael L Littman, Matthew E Taylor, Jeff Huang, and David L Roberts. 2016. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Autonomous agents and multi-agent systems* 30 (2016), 30–59.
- [121] Zvezdan Lončarević, Aleš Ude, Bojan Nemec, Andrej Gams, et al. 2018. User feedback in latent space robotic skill learning. In *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*. IEEE, 270–276.
- [122] Manuel Lopes, Thomas Cederbourg, and Pierre-Yves Oudeyer. 2011. Simultaneous acquisition of task and feedback models. In *2011 IEEE International Conference on Development and Learning (ICDL)*, Vol. 2. 1–7. <https://doi.org/10.1109/DEVLRN.2011.6037359>
- [123] Dylan P Losey, Andrea Bajcsy, Marcia K O'Malley, and Anca D Dragan. 2022. Physical interaction as communication: Learning robot objectives online from human corrections. *The International Journal of Robotics Research* 41, 1 (2022),

20–44.

- [124] Dylan P Losey and Marcia K O'Malley. 2018. Including uncertainty when learning from human corrections. In *Conference on Robot Learning*. PMLR, 123–132.
- [125] Hao Lü, James A. Fogarty, and Yang Li. 2014. Gesture Script: Recognizing Gestures and Their Structure Using Rendering Scripts and Interactively Trained Parts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) (*CHI '14*). Association for Computing Machinery, New York, NY, USA, 1685–1694. <https://doi.org/10.1145/2556288.2557263>
- [126] Jelena Luketina, 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 Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 6309–6317. <https://doi.org/10.24963/ijcai.2019/880>
- [127] P Lukowicz et al. 2023. “Who’s a Good Robot?!” Designing Human-Robot Teaching Interactions Inspired by Dog Training. In *HHAI 2023: Augmenting Human Intellect: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*, Vol. 368. IOS Press, 310.
- [128] Jieliang Luo, Sam Green, Peter Feghali, George Legrady, C , Etin, and Kaya Koç. 2018. Visual Diagnostics for Deep Reinforcement Learning Policy Development. (Sept. 2018). <https://arxiv.org/abs/1809.06781v2> arXiv: 1809.06781 ISBN: 1809.06781v2.
- [129] Daoming Lyu, Fangkai Yang, Bo Liu, and Steven Gustafson. 2019. SDRL: interpretable and data-efficient deep reinforcement learning leveraging symbolic planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2970–2977.
- [130] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. 2017. Interactive learning from policy-dependent human feedback. In *International conference on machine learning*. PMLR, 2285–2294.
- [131] Richard Maclin, Jude Shavlik, Lisa Torrey, Trevor Walker, and Edward Wild. 2005. Giving advice about preferred actions to reinforcement learners via knowledge-based kernel regression. In *AAAI* 819–824.
- [132] Richard Maclin and Jude W Shavlik. 1996. Creating advice-taking reinforcement learners. *Machine Learning* 22 (1996), 251–281.
- [133] Sean McGregor, Hailey Buckingham, Thomas G. Dietterich, Rachel Houtman, Claire Montgomery, and Ronald Metoyer. 2015. Facilitating testing and debugging of Markov Decision Processes with interactive visualization. In *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 281–282. <https://doi.org/10.1109/VLHCC.2015.7357227>
- [134] Shaunak A Mehta and Dylan P Losey. 2022. Unified learning from demonstrations, corrections, and preferences during physical human-robot interaction. *arXiv preprint arXiv:2207.03395* (2022).
- [135] Shaunak A Mehta, Forrest Meng, Andrea Bajcsy, and Dylan P Losey. 2024. StROL: Stabilized and Robust Online Learning from Humans. *IEEE Robotics and Automation Letters* (2024).
- [136] Marcel Menner, Lukas Neuner, Lars Lünenburger, and Melanie N Zeilinger. 2020. Using human ratings for feedback control: A supervised learning approach with application to rehabilitation robotics. *IEEE Transactions on Robotics* 36, 3 (2020), 789–801.
- [137] Yannick Metz, David Lindner, Raphaël Baur, Daniel Keim, and Mennatallah El-Assady. 2023. RLHF-Blender: A Configurable Interactive Interface for Learning from Diverse Human Feedback. *arXiv preprint arXiv:2308.04332* (2023).
- [138] Cristian Millán, Bruno JT Fernandes, and Francisco Cruz. 2019. Human feedback in continuous actor-critic reinforcement learning.. In *ESANN*.
- [139] Cristian C Millan-Arias, Bruno JT Fernandes, Francisco Cruz, Richard Dazeley, and Sergio Fernandes. 2021. A robust approach for continuous interactive actor-critic algorithms. *IEEE Access* 9 (2021), 104242–104260.
- [140] Sören Mindermann, Rohin Shah, Adam Gleave, and Dylan Hadfield-Menell. 2018. Active inverse reward design. *arXiv preprint arXiv:1809.03060* (2018).
- [141] Ithan Moreira, Javier Rivas, Francisco Cruz, Richard Dazeley, Angel Ayala, and Bruno Fernandes. 2020. Deep reinforcement learning with interactive feedback in a human–robot environment. *Applied Sciences* 10, 16 (2020), 5574.
- [142] Vivek Myers, Erdem Biyik, Nima Anari, and Dorsa Sadigh. 2022. Learning multimodal rewards from rankings. In *Conference on Robot Learning*. PMLR, 342–352.
- [143] Anis Najjar and Mohamed Chetouani. 2021. Reinforcement learning with human advice: a survey. *Frontiers in Robotics and AI* 8 (2021), 584075.
- [144] Benjamin A Newman, Christopher Jason Paxton, Kris Kitani, and Henny Admoni. 2023. Towards Online Adaptation for Autonomous Household Assistants. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 506–510.

- [145] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, Vol. 99. Citeseer, 278–287.
- [146] Andrew Y Ng, Stuart Russell, et al. 2000. Algorithms for inverse reinforcement learning. In *Icml*, Vol. 1. 2.
- [147] Phillip Odom and Sriraam Natarajan. 2015. Active advice seeking for inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.
- [148] Takato Okudo and Seiji Yamada. 2021. Subgoal-based reward shaping to improve efficiency in reinforcement learning. *IEEE Access* 9 (2021), 97557–97568.
- [149] Fredrik Olsson. 2009. A literature survey of active machine learning in the context of natural language processing. (2009).
- [150] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens Amanda Askell, Peter Welinder Paul Christiano, Jan Leike, and Ryan Lowe. [n. d.]. Training language models to follow instructions with human feedback. ([n. d.]). arXiv: 2203.02155v1 ISBN: 2203.02155v1.
- [151] Rok Pahič, Zvezdan Lončarević, Aleš Ude, Bojan Nemec, and Andrej Gams. 2018. User Feedback in Latent Space Robotic Skill Learning. In *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*. 270–276. <https://doi.org/10.1109/HUMANOIDS.2018.8624972>
- [152] Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159* (2024).
- [153] Miriam Punzi, Nicolas Ladeveze, Huyen Nguyen, and Brian Ravenet. 2022. ImCasting: Nonverbal Behaviour Reinforcement Learning of Virtual Humans through Adaptive Immersive Game. In *27th International Conference on Intelligent User Interfaces*. 62–65.
- [154] Nikaash Puri, Sukriti Verma, Piyush Gupta, Dhruv Kayastha, Shripad Deshmukh, Balaji Krishnamurthy, and Sameer Singh. 2020. Explain Your Move: Understanding Agent Actions Using Specific and Relevant Feature Attribution. In *Intl. Conf. on Learning Representations*. <https://openreview.net/forum?id=SJgzLkBKPB>
- [155] Syed Ali Raza, Benjamin Johnston, and Mary-Anne Williams. 2016. Reward from demonstration in interactive reinforcement learning. In *The Twenty-Ninth International Flairs Conference*.
- [156] Syed Ali Raza and Mary-Anne Williams. 2020. Human feedback as action assignment in interactive reinforcement learning. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* 14, 4 (2020), 1–24.
- [157] Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity Bias in Preference Labeling by Large Language Models. arXiv:2310.10076 [cs.CL] <https://arxiv.org/abs/2310.10076>
- [158] Lisa Scherf, Cigdem Turan, and Dorothea Koert. 2022. Learning from Unreliable Human Action Advice in Interactive Reinforcement Learning. In *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*. IEEE, 895–902.
- [159] Mariah L Schrum, Erin Hedlund-Botti, Nina Moorman, and Matthew C Gombolay. 2022. Mind meld: Personalized meta-learning for robot-centric imitation learning. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 157–165.
- [160] Burr Settles. 2009. Active learning literature survey. (2009).
- [161] Rita Sevastjanova, Fabian Beck, Basil Ell, Cagatay Turkay, Rafael Henkin, Miriam Butt, Daniel A Keim, and Mennatallah El-Assady. 2018. Going beyond visualization: Verbalization as complementary medium to explain machine learning models. In *Workshop on Visualization for AI Explainability at IEEE VIS*.
- [162] Rita Sevastjanova, Wolfgang Jentner, Fabian Sperrle, Rebecca Kehlbeck, Jürgen Bernard, and Mennatallah El-assady. 2021. QuestionComb: A Gamification Approach for the Visual Explanation of Linguistic Phenomena through Interactive Labeling. *ACM Transactions on Interactive Intelligent Systems* 11, 3-4 (Dec. 2021), 1–38. <https://doi.org/10.1145/3429448>
- [163] Rohin Shah, Pedro Freire, Neel Alex, Rachel Freedman, Dmitrii Krashennnikov, Lawrence Chan, Michael D Dennis, Pieter Abbeel, Anca Dragan, and Stuart Russell. 2020. Benefits of assistance over reward learning. (2020).
- [164] Pratyusha Sharma, Balakumar Sundaralingam, Valts Blukis, Chris Paxton, Tucker Hermans, Antonio Torralba, Jacob Andreas, and Dieter Fox. 2022. Correcting robot plans with natural language feedback. *arXiv preprint arXiv:2204.05186* (2022).
- [165] Isaac Sheidlower, Allison Moore, and Elaine Short. 2022. Keeping Humans in the Loop: Teaching via Feedback in Continuous Action Space Environments. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 863–870.
- [166] Alvin Shek, Bo Ying Su, Rui Chen, and Changliu Liu. 2023. Learning from physical human feedback: An object-centric one-shot adaptation method. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 9910–9916.
- [167] Michael Shilman, Desney S. Tan, and Patrice Simard. 2006. CueTIP: A Mixed-Initiative Interface for Correcting Handwriting Errors. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology*

- (Montreux, Switzerland) (UIST '06). Association for Computing Machinery, New York, NY, USA, 323–332. <https://doi.org/10.1145/1166253.1166304>
- [168] Hilson Shrestha, Kathleen Cachel, Mallak Alkathlan, Elke Rundensteiner, and Lane Harrison. 2022. FairFuse: Interactive Visual Support for Fair Consensus Ranking. In *2022 IEEE Visualization and Visual Analytics (VIS)*. IEEE, 65–69.
  - [169] Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716* (2023).
  - [170] Shivam Singhal, Cassidy Laidlaw, and Anca Dragan. 2024. Scalable Oversight by Accounting for Unreliable Feedback. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*. <https://openreview.net/forum?id=Noy5wbyiCS>
  - [171] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2023. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. *arXiv preprint arXiv:2312.08358* (2023).
  - [172] Patrik D Sørensen, Jepph M Olsen, and Sebastian Risi. 2016. Breeding a diversity of super mario behaviors through interactive evolution. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 1–7.
  - [173] F. Sperrle, M. El-Assady, G. Guo, R. Borgo, D. Horng Chau, A. Endert, and D. Keim. 2021. A Survey of Human-Centered Evaluations in Human-Centered Machine Learning. *Computer Graphics Forum* 40, 3 (2021), 543–568. <https://doi.org/10.1111/cgf.14329> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14329>
  - [174] Thilo Spinner, Rebecca Kehlbeck, Rita Sevastianova, Tobias Stähle, Daniel A Keim, Oliver Deussen, and Mennatallah El-Assady. 2024. -generAltor: Tree-in-the-loop Text Generation for Language Model Explainability and Adaptation. *ACM Transactions on Interactive Intelligent Systems* 14, 2 (2024), 1–32.
  - [175] Kaushik Subramanian, Charles L Isbell Jr, and Andrea L Thomaz. 2016. Exploration from demonstration for interactive reinforcement learning. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*. 447–456.
  - [176] Felipe Petroski Such, Vashisht Madhavan, Rosanne Liu, Rui Wang, Pablo Samuel Castro, Yulun Li, Jiale Zhi, Ludwig Schubert, Marc G. Bellemare, Jeff Clune, and Joel Lehman. 2019. An Atari Model Zoo for Analyzing, Visualizing, and Comparing Deep Reinforcement Learning Agents. arXiv:1812.07069 [cs.NE]
  - [177] Theodore R. Sumers, Robert D. Hawkins, Mark K. Ho, Thomas L. Griffiths, and Dylan Hadfield-Menell. 2022. How to talk so AI will learn: Instructions, descriptions, and autonomy. <http://arxiv.org/abs/2206.07870> arXiv:2206.07870 [cs].
  - [178] Theodore R. Sumers, Robert D. Hawkins, Mark K. Ho, Thomas L. Griffiths, and Dylan Hadfield-Menell. 2022. Linguistic communication as (inverse) reward design. (April 2022). <http://arxiv.org/abs/2204.05091> arXiv: 2204.05091.
  - [179] Theodore R Sumers, Mark K Ho, Robert D Hawkins, Karthik Narasimhan, and Thomas L Griffiths. 2021. Learning rewards from linguistic feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6002–6010.
  - [180] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.
  - [181] Niket Tandon, Aman Madaan, Peter Clark, and Yiming Yang. 2022. Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 339–352. <https://doi.org/10.18653/v1/2022.findings-naacl.26>
  - [182] Matthew E Taylor, Halit Bener Suay, and Sonia Chernova. 2011. Integrating reinforcement learning with human demonstrations of varying ability. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. 617–624.
  - [183] Ana C Tenorio-Gonzalez, Eduardo F Morales, and Luis Villasenor-Pineda. 2010. Dynamic reward shaping: training a robot by voice. In *Advances in Artificial Intelligence—IBERAMIA 2010: 12th Ibero-American Conference on AI, Bahía Blanca, Argentina, November 1-5, 2010. Proceedings 12*. Springer, 483–492.
  - [184] Christopher Thierauf, Ravenna Thielstrom, Bradley Oosterveld, Will Becker, and Matthias Scheutz. 2023. “Do this instead”—Robots that Adequately Respond to Corrected Instructions. *ACM Transactions on Human-Robot Interaction* (2023).
  - [185] Andrea L Thomaz and Cynthia Breazeal. 2007. Asymmetric interpretations of positive and negative human feedback for a social learning agent. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 720–725.
  - [186] Andrea L. Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence* 172, 6 (2008), 716–737. <https://doi.org/10.1016/j.artint.2007.09.009>
  - [187] Andrea Lockerd Thomaz, Cynthia Breazeal, et al. 2006. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *Aaai*, Vol. 6. Boston, MA, 1000–1005.
  - [188] Andrea Lockerd Thomaz, Guy Hoffman, and Cynthia Breazeal. 2005. Real-time interactive reinforcement learning for robots. In *AAAI 2005 workshop on human comprehensible machine learning*, Vol. 3. Citeseer.
  - [189] Faraz Torabi, Garrett Warnell, and Peter Stone. 2018. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954* (2018).

- [190] Marcel Torne, Max Balsells, Zihan Wang, Samedh Desai, Tao Chen, Pulkit Agrawal, and Abhishek Gupta. 2023. Bread-crumbs to the Goal: Goal-Conditioned Exploration from Human-in-the-Loop Feedback. *arXiv preprint arXiv:2307.11049* (2023).
- [191] Susanne Trick, Franziska Herbert, Constantin A Rothkopf, and Dorothea Koert. 2022. Interactive reinforcement learning with Bayesian fusion of multimodal advice. *IEEE Robotics and Automation Letters* 7, 3 (2022), 7558–7565.
- [192] Sanne Van Waveren, Christian Pek, Jana Tumova, and Iolanda Leite. 2022. Correct me if I'm wrong: Using non-experts to repair reinforcement learning policies. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 493–501.
- [193] Vivek Veeriah, Patrick M Pilarski, and Richard S Sutton. 2016. Face valuing: Training user interfaces with facial expressions and reinforcement learning. *arXiv preprint arXiv:1606.02807* (2016).
- [194] Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. 2019. Programmatically Interpretable Reinforcement Learning. *arXiv:1804.02477 [cs.LG]*
- [195] Chunyang Wang, Yanmin Zhu, Haobing Liu, Tianzi Zang, Jiadi Yu, and Feilong Tang. 2022. Deep Meta-learning in Recommendation Systems: A Survey. *arXiv preprint arXiv:2206.04415* (2022).
- [196] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291* (2023).
- [197] Junpeng Wang, Liang Gou, Han-Wei Shen, and Hao Yang. 2018. DQNViz: A Visual Analytics Approach to Understand Deep Q-Networks. *IEEE Trans Vis Comput Graph* (Sept. 2018).
- [198] Xiaofei Wang, Kimin Lee, Kourosh Hakhamaneshi, Pieter Abbeel, and Michael Laskin. 2022. Skill preferences: Learning to extract and execute robotic skills from human feedback. In *Conference on Robot Learning*. PMLR, 1259–1268.
- [199] Zhaodong Wang and Matthew E. Taylor. 2019. Interactive Reinforcement Learning with Dynamic Reuse of Prior Knowledge from Human and Agent Demonstrations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 3820–3827. <https://doi.org/10.24963/ijcai.2019/530>
- [200] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. 2017. Deep TAMER: Interactive Agent Shaping in High-Dimensional State Spaces. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018* (Sept. 2017), 1545–1553. <https://doi.org/10.48550/arxiv.1709.10163> arXiv: 1709.10163 Publisher: AAAI press ISBN: 9781577358008.
- [201] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. 2018. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
- [202] Nils Wilde, Erdem Byrk, Dorsa Sadigh, and Stephen L Smith. 2021. Learning reward functions from scale feedback. *arXiv preprint arXiv:2110.00284* (2021).
- [203] Nils Wilde, Alexandru Blidaru, Stephen L Smith, and Dana Kulić. 2020. Improving user specifications for robot behavior through active preference learning: Framework and evaluation. *The International Journal of Robotics Research* 39, 6 (2020), 651–667.
- [204] Nialah Jenae Wilson-Small, David Goedicke, Kirstin Petersen, and Shiri Azenkot. 2023. A Drone Teacher: Designing Physical Human-Drone Interactions for Movement Instruction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 311–320.
- [205] Christian Wirth, Riad Akrou, Gerhard Neumann, Johannes Fürnkranz, et al. 2017. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research* 18, 136 (2017), 1–46.
- [206] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. *arXiv preprint arXiv:2306.01693* (2023).
- [207] Duo Xu, Mohit Agarwal, Ekansh Gupta, Faramarz Fekri, and Raghupathy Sivakumar. 2021. Accelerating Reinforcement Learning using EEG-based implicit human feedback. *Neurocomputing* 460 (2021), 139–153.
- [208] Kelvin Xu, Zheyuan Hu, Ria Doshi, Aaron Rovinsky, Vikash Kumar, Abhishek Gupta, and Sergey Levine. 2023. Dexterous manipulation from images: Autonomous real-world rl via substep guidance. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5938–5945.
- [209] Eric Yeh, Melinda Gervasio, Daniel Sanchez, Matthew Crossley, and Karen Myers. 2018. Bridging the gap: Converting human advice into imagined examples. *Advances in Cognitive Systems* 6 (2018), 1168–1176.
- [210] Lance Ying, Tan Zhi-Xuan, Vikash Mansinghka, and Joshua B Tenenbaum. 2023. Inferring the goals of communicating agents from actions and instructions. In *Proceedings of the AAAI Symposium Series*, Vol. 2. 26–33.
- [211] J Yow, Neha Priyadarshini Garg, Manoj Ramanathan, Wei Tech Ang, et al. 2024. ExTraCT-Explainable Trajectory Corrections from language inputs using Textual description of features. *arXiv preprint arXiv:2401.03701* (2024).
- [212] Hang Yu, Reuben M Aronson, Katherine H Allen, and Elaine Schaeertl Short. 2023. From “Thumbs Up” to “10 out of 10”: Reconsidering Scalar Feedback in Interactive Reinforcement Learning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4121–4128.

- [213] Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, et al. 2023. Language to Rewards for Robotic Skill Synthesis. *arXiv preprint arXiv:2306.08647* (2023).
- [214] Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. 2023. How to Query Human Feedback Efficiently in RL? *Interactive Learning with Implicit Human Feedback Workshop at ICML 2023* (2023).
- [215] David Zhang, Micah Carroll, Andreea Bobu, and Anca Dragan. 2022. Time-Efficient Reward Learning via Visually Assisted Cluster Ranking. (Nov. 2022). <http://arxiv.org/abs/2212.00169> arXiv:2212.00169 [cs].
- [216] Jenny Zhang, Samson Yu, Jiafei Duan, and Cheston Tan. 2023. Good Time to Ask: A Learning Framework for Asking for Help in Embodied Visual Navigation. In *2023 20th International Conference on Ubiquitous Robots (UR)*. IEEE, 503–509.
- [217] Qiping Zhang, Austin Narcomey, Kate Candon, and Marynel Vázquez. 2023. Self-Annotation Methods for Aligning Implicit and Explicit Human Feedback in Human-Robot Interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 398–407.
- [218] Ruohan Zhang, Dhruva Bansal, Yilun Hao, Ayano Hiranaka, Jialu Gao, Chen Wang, Roberto Martín-Martín, Li Fei-Fei, and Jiajun Wu. 2023. A Dual Representation Framework for Robot Learning with Human Guidance. In *Conference on Robot Learning*. PMLR, 738–750.
- [219] Ruohan Zhang, Faraz Torabi, Lin Guan, Dana H Ballard, and Peter Stone. 2019. Leveraging human guidance for deep reinforcement learning tasks. *arXiv preprint arXiv:1909.09906* (2019).
- [220] Ruohan Zhang, Calen Walshe, Zhuode Liu, Lin Guan, Karl Muller, Jake Whritner, Luxin Zhang, Mary Hayhoe, and Dana Ballard. 2020. Atari-head: Atari human eye-tracking and demonstration dataset. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 6811–6820.
- [221] Michelle D Zhao, Reid Simmons, and Henny Admoni. 2023. Learning Human Contribution Preferences in Collaborative Human-Robot Tasks. In *Conference on Robot Learning*. PMLR, 3597–3618.
- [222] Banghua Zhu, Michael Jordan, and Jiantao Jiao. 2023. Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 43037–43067. <https://proceedings.mlr.press/v202/zhu23f.html>
- [223] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-Tuning Language Models from Human Preferences. *CoRR* abs/1909.08593 (2019). arXiv:1909.08593 <http://arxiv.org/abs/1909.08593>
- [224] Matt Zucker, J Andrew Bagnell, Christopher G Atkeson, and James Kuffner. 2010. An optimization approach to rough terrain locomotion. In *2010 IEEE International Conference on Robotics and Automation*. IEEE, 3589–3595.



A Full Survey Results

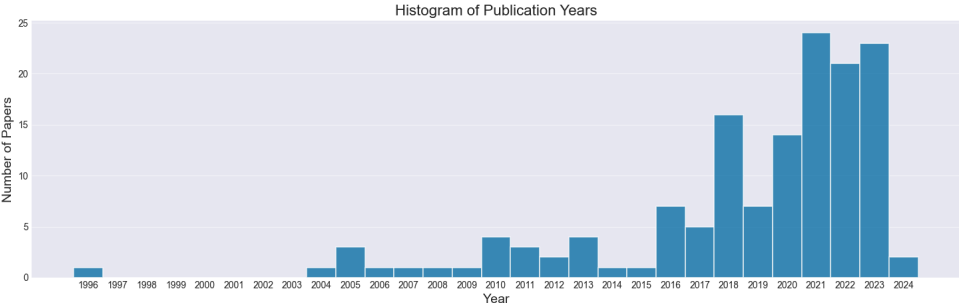


Fig. 14. The publication years of surveyed publications.

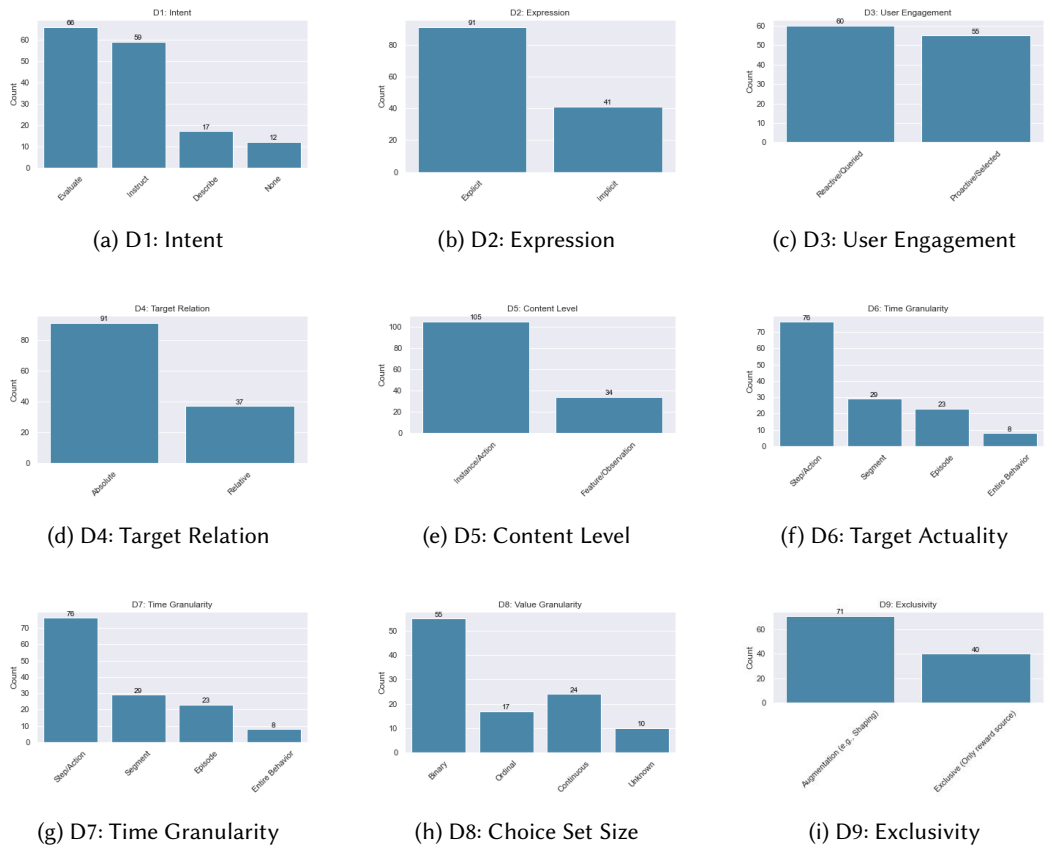


Fig. 15. Counts of surveyed papers for each attribute in Dimensions D1-D9

	D1	D2	D3	D4	D5	D6	D7	D8	D9			
	Intent	Expres.	Engag.	Rela.	Content	Tgt-Act.	Tmp.Gra.	Choi.Set	Excls.			
Publication	None	Explicit	Proactive	Reactive	Relative	Feature	Hypothetical	Discrete	Augmenting	Description	Year	Venue
	Describe			Absolute	Instance	Actual		Binary	Exclusive			
Maclin et al. [132]	✓	✓	✓	✓	✓	✓	✓		✓	Suggestions via Instructions	1996	Machine Learning
Kuhlmann et al. [97]	✓	✓	✓	✓	✓	✓	✓		✓	If-Then-Rules	2004	AAAI Workshop
Ferrez et al. [62]		✓		✓		✓		✓	✓	Interaction error-related potentials	2005	IJCAI
Maclin et al. [131]	✓		✓	✓	✓	✓		✓		Preference over actions	2005	AAAI 2005
Thomaz et al. [188]	✓	✓	✓	✓	✓	✓		✓	✓	Natural interaction	2005	AAAI
Thomaz et al. [187]	✓	✓		✓	✓	✓		✓		Action Preference (Wirth 2018)	2006	
Thomaz et al. [185]	✓	✓	✓	✓	✓	✓	✓	✓		Positive/Negative Scalar Feedback	2007	IEEE RO-MAN
Bradley Knox et al. [29]	✓	✓	✓	✓	✓	✓		✓	✓	Scalar Reward	2008	IEEE ICDL
Branavan et al. [30]	✓	✓	✓	✓	✓	✓	✓		✓	Action plans	2009	ACL
Zucker et al. [224]	✓	✓		✓	✓	✓		✓		State Preference (Wirth 2018)	2010	
Knox et al. [87]	✓	✓	✓		✓	✓	✓	✓		Shaping	2010	
Tenorio-Gonzalez et al. [183]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Voice Commands (Good/Bad, Action Adv.)	2010	IBERAMIA
Judah et al. [82]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Critique Advice	2010	AAAI
Cakmak et al. [33]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Examples for Bad and Good Hand-Overes	2011	IROS 2011
Taylor et al. [182]	✓	✓	✓		✓	✓	✓	✓	✓	Demonstrations, Interactive Action-Advice	2011	AAMAS
Cakmak et al. [33]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Preferences over Configurations	2011	IROS 2011
Knox et al. [88]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Shaping	2012	IEEE RO-MAN
Akrour et al. [5]	✓	✓	✓	✓	✓	✓	✓	✓		Preference	2012	ECML/PKDD
Hayes-Roth et al. [75]	✓	✓	✓	✓	✓	✓	✓	✓	✓	General constraints	2013	
Li et al. [106]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Feedback	2013	AAMAS
Griffith et al. [69]					✓	✓					2013	
Griffith et al. [69]	✓	✓	✓		✓	✓	✓	✓		Shaping	2013	NeurIPS
Loflin et al. [119]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Feedback signals	2014	IEEE ROMAN
Odom et al. [147]	✓	✓	✓	✓	✓	✓	✓			Advice	2015	AAAI
Amir et al. [8]	✓	✓	✓	✓	✓	✓	✓	✓	✓	(Action) Advice	2016	IJCAI 2016
Subramanian et al. [175]	✓	✓	✓	✓	✓	✓	✓			Demonstrations	2016	
Raza et al. [155]	✓	✓	✓	✓	✓	✓	✓		✓	Demonstrations/action advice	2016	AAAI
Loflin et al. [120]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Discrete Feedb. with diff. training strategies	2016	AAMAS

Table 3. Summary table of the literature survey. Surveyed work is classified according to the conceptual framework (✓) Black checkmarks refer to exclusive attributes, (✓) Grey checkmarks indicate that feedback can have characteristics across different work, (✓) Blue checkmarks indicate necessary simultaneous presence of two attributes. We can classify existing types of feedback according to our dimensions.

Table 3. Summary table of the literature survey. Surveyed work is classified according to the conceptual framework (✓) Black checkmarks refer to exclusive attributes, (✓) Grey checkmarks indicate that feedback can have characteristics across different work, (✓) Blue checkmarks indicate necessary simultaneous presence of two attributes. We can classify existing types of feedback according to our dimensions.

	Publication	D1 Intent	D2 Expres.	D3 Engag.	D4 Rela.	D5 Content	D6 Tgt Act.	D7 Temp Gra.	D8 Choi Set	D9 Excls.	Description	Year	Venue
		None Describe Instruct Evaluate	Implicit	Reactive Proactive	Relative Absolute	Instance	Feature	Actual	Hypothetical	Ent.Beh. Episode Segment Step/Action	Continuous Discrete Binary	Exclusive	Augmenting
El Asri et al. [54]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Sorensen et al. [172]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Veeriah et al. [193]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fachantidis et al. [57]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MacGlashan et al. [130]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Kim et al. [85]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Christiano et al. [42]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Lin et al. [112]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Gao et al. [67]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Li et al. [105]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Krening et al. [93]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Torabi et al. [189]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Krening et al. [94]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ibarz et al. [79]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Arakawa et al. [9]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mindermann et al. [140]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Brown et al. [31]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Losey et al. [124]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Yeh et al. [209]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Cruz et al. [47]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Basu et al. [19]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Lontcarević et al. [121]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ibarz et al. [79]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Warmell et al. [201]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Millán et al. [138]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Basu et al. [18]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
De Winter et al. [50]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Akkaladevi et al. [4]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Wang et al. [199]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Frazier et al. [64]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 4. Summary table of the literature survey. Surveyed work is classified according to the conceptual framework (✓) Black checkmarks refer to exclusive attributes, (✓) Grey checkmarks indicate that feedback can have characteristics across different work, (✓) Blue checkmarks indicate necessary simultaneous presence of two attributes. We can classify existing types of feedback according to our dimensions.

Publication	Intent		Expres.		D2	D3	D4	D5	D6	D7	D8	D9	Description	Year	Venue
	Describe	Evaluate	Engag.	Reactive	Relative	Instance	Feature	Actual	Hypothetical	Ent.Beh.	Discrete	Continuous			
Zhang et al. [219]	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2019	IJCAI
Arzate Cruz et al. [12]	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2020	2020 ACM DIS
Arzate Cruz et al. [12]	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2020	2020 ACM DIS
Wilde et al. [203]	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2020	The Int. Journ. of Robotics Res
Faulkner et al. [59]	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2020	IEEE ICRA
Arzate Cruz et al. [11]	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2020	CHI PLAY '20
Koert et al. [90]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2020	Frontiers Robotics/AI
Arzate Cruz et al. [12]	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2020	2020 ACM DIS
Li et al. [103]	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2020	Auton. Agents and M-A Sys.
Moreira et al. [141]	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2020	Applied Sciences
Zhang et al. [220]	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2020	AAAI 2020
Arzate Cruz et al. [12]	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2020	2020 ACM DIS
Raza et al. [156]	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2020	ACM TAAS
Memner et al. [136]	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2020	IEEE Transactions on robotics
Wilde et al. [203]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2020	The Int. Journ. of Robotics Res
Bignold et al. [25]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	Neural Comp. and App.
Chetouani et al. [38]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	
Najar et al. [143]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	Frontiers Robotics/AI
Li et al. [108]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	ICRA 2021
Najar et al. [143]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	Frontiers Robotics/AI
Li et al. [107]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	ICRA 2021
Najar et al. [143]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	Frontiers Robotics/AI
Sumers et al. [179]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	Frontiers Robotics/AI
Najar et al. [143]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	AAAI 2021
Chetouani et al. [38]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	Frontiers Robotics/AI
Xu et al. [207]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	Neurocomputing
Law et al. [101]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	DIS21
Cui et al. [48]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	CoRL
Millan-Arias et al. [139]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	HAI/IEEE Access
Najar et al. [143]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	Frontiers Robotics/AI
Najar et al. [143]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	Frontiers Robotics/AI
Wilde et al. [202]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	2021	CoRL

Table 5. Summary table of the literature survey. Surveyed work is classified according to the conceptual framework (✓) Black checkmarks refer to exclusive attributes, (✓) Grey checkmarks indicate that feedback can have characteristics across different work, (✓) Blue checkmarks indicate necessary simultaneous presence of two attributes. We can classify existing types of feedback according to our dimensions.

Publication	D1 Intent	D2 Expres.	D3 Engag.	D4 Rela.	D5 Content	D6 Tgt.Act.	D7 Imp.Gra.	D8 Choi.Set	D9 Excls.	Description	Year	Venue
Okudo et al. [148]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Subgoal Generation	2021	IEEE Access
Bobu et al. [28]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Feature Traces	2021	HRI '21
Chetouani et al. [38]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Teaching Signals: Demonstration Preferences	2021	
Lee et al. [102]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Expert Annotations	2021	ICML 2021
Kwon et al. [99]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Corrective Feedback	2021	Frontiers Robotics/AI
Najar et al. [143]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Repairing via interactive explanations	2021	2021 IEEE CoG
Cruz et al. [45]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Nonverbal behaviour training	2022	IUI
Punzi et al. [153]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Correcting one response out of a set responses	2022	ArXiv
Bai et al. [16]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Corrections: adding constraints/specifying sub-goals	2022	Robotics: Science and Systems
Sharma et al. [164]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Action Instructions and Evaluative Scenar Feedback	2022	ACM/IEEE HRI '22
Hsiung et al. [77]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Broad Persistent Advice	2022	IROS2022 Workshop
Cruz et al. [46]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Preference cat./pairs, Expression Sugg.	2022	IJCAI 22
Crochepierre et al. [44]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Interventions via Shared Control	2022	RSS
Liu et al. [117]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Teaching by instruction or evaluation	2022	IEEE RO-MAN
Chi et al. [40]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Open-Language, Block-Based Non-Expert Feeds	2022	ACM/IEEE HRI '22
Van Waveren et al. [192]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Demonstrations, Corrections, and Preferences	2022	ACM THRI
Mehra et al. [134]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Natural Eye Gaze and Joystick Input	2022	Robotics Science and Systems
Aronson et al. [10]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Skill preferences, mark. of positive trajectory seg.	2022	CoRL
Wang et al. [198]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Action Advice	2022	IEEE-RAS
Scherf et al. [158]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Corrective Actions	2022	HRI '22
Schrum et al. [159]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Multimodal Advice	2022	IEEE RA-L
Trick et al. [191]	✓	✓	✓	✓	✓	✓	✓	✓	✓		2022	IROS2022
Sheidlower et al. [165]	✓	✓	✓	✓	✓	✓	✓	✓	✓		2022	ACL 2022
Lin et al. [110]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Natural language utterance (in context)	2022	The Int. Journ. of Robotics Res.
Loscy et al. [123]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Intentional Physical Corrections	2022	IUI
Punzi et al. [153]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Voting	2022	CoRL 2021
Myers et al. [142]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Rankings	2023	HRI '23
Wilson-Small et al. [204]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Drones-Physical Instructive Feedback	2023	ACM/IEEE HRI '23
Zhang et al. [217]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Survey feedb., self-annot. via facial express./label val.	2023	CoRL
Zhang et al. [218]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Human Guidance via evaluation or preference	2023	ACM THRI
Thierauf et al. [184]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Corrected Instructions	2023	CoRL
Byrk et al. [27]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Preference Queries	2023	UR 2023
Zhang et al. [216]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Object-in-view feedback	2023	

Table 6. Summary table of the literature survey. Surveyed work is classified according to the conceptual framework (✓) Black checkmarks refer to exclusive attributes, (✓) Grey checkmarks indicate that feedback can have characteristics across different work, (✓) Blue checkmarks indicate necessary simultaneous presence of two attributes. We can classify existing types of feedback according to our dimensions.

	D1 Intent	D2 Expres.	D3 Engag.	D4 Rela.	D5 Content	D6 Tgt Act.	D7 Ttmp Gra.	D8 ChoiSel	D9 Excls.	Description	Year	Venue
Publication	None Describe Instruct Evaluate	Implicit Explicit	Reactive Proactive	Relative Absolute	Instance Feature	Actual Hypothetical	Ent.Beh. Episode Segment Step/Action	Discrete Binary	Continuous Exclusive			
Newman et al. [144]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Corrections	2023	HRI '23
Torne et al. [190]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Binary Preferences	2023	ArXiv
Bignold et al. [26]	✓	✓	✓	✓	✓	✓	✓	✓	✓	(Eval.) Binary Advice, (Informative) Advice Scalar Feedback	2023	Neural Comp. and Appl. IROS
Yu et al. [212]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Instructions and Actions for Goal Inference Visual Milestones/Transitions	2023	AAAI 2023 Fall Symposium ICRA 2023
Ying et al. [210]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Object Preferences	2023	ICRA 2022
Xu et al. [208]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Reinforcement	2023	Frontiers in AI and App.
Shek et al. [166]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Demos, Prefs., Corrections/Binary Feedb.	2023	CoRL 2023
Lukowicz et al. [127]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Verbal correction	2023	CoRL 2023 Workshop
Fitzgerald et al. [63]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Language instructions/corrections	2023	ArXiv
Liu et al. [116]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Gameplay Interact. for intermediate goal states	2023	ArXiv
Yu et al. [213]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Fine-grained annotations	2023	ArXiv
Wang et al. [196]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Nonverbal reactions	2023	AAMAS 2023
Wu et al. [206]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Actions	2023	CoRL 2023
Candon et al. [35]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Visual Inspection of Intermediate Results	2023	TMLR
Zhao et al. [221]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Corrections	2024	IEEE RA-L
Wang et al. [196]	✓	✓	✓	✓	✓	✓	✓	✓	✓	Language correction	2024	ArXiv
Mehta et al. [135]	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Yow et al. [211]	✓	✓	✓	✓	✓	✓	✓	✓	✓			

Table 7. Summary table of the literature survey. Surveyed work is classified according to the conceptual framework (✓) Black checkmarks refer to exclusive attributes, (✓) Grey checkmarks indicate that feedback can have characteristics across different work, (✓) Blue checkmarks indicate necessary simultaneous presence of two attributes. We can classify existing types of feedback according to our dimensions.

## B Feedback User Interfaces

In this appendix, we provide a collection of proposed user interfaces for the collection of feedback.

### B.1 User Interface: Language

The following works mention language-based user interactions for feedback collection:

Publication	Description
Kuhlmann et al.[97]	State-dependent if-then-else rules to guide agent
Tenorio-Gonzalez et al.[183]	Using language commands (e.g., action to take, evaluation, confirmation of reached goal state)
Krening et al.[93]	Language instructions mapped to actions in a grid world domain
Krening et al.[94]	Language instructions/binary critique for grid world domain
De Winter et al. [50]	Verbally telling constraints to the system in natural language, then translated into first-order logic; input can be temporal action sequences
Bignold et al. [25]	When giving a rating for a state, the user can also formulate a general rule; multiple rules are processed with probabilistic policy reuse
Sumers et al.[179]	Simple game-like object collecting example; linguistic feedback is grouped into three categories and produced for observed episodes
Okudo et al.[148]	Generate subgoals for a more complex overall behavior in a grid-world domain (points to reach via a UI drawing style); for a robot (verbal description of subgoals to reach)
Millan-Arias et al. [139]	No interactive experiment discussed; using simulated oracles for cart-pole language-based actions
Chi et al.[40]	Humans are queried to either teach via instruction/action advice or evaluation on a five-point scale for predicted action; simplified medical care setting
Van Waveren et al.[192]	From natural language and "block-based" (i.e., programming language) feedback, infer action refinement, alternative actions, and forbidden actions
Bai et al.[16]	Chat interface; ranking of two entries with confidence (one or the other is better)
Trick et al.[191]	Giving action advice (encoding by mentioning an object for the robot to grip) via language (with keyword spotting) or by interpreting pointing gestures; robotics grasping task
Lin et al. [110]	Flight booking use case; users can give natural language feedback to an agent; reward function is inferred from free-form feedback
Ying et al.[210]	First, humans utter an instruction, then perform a series of actions; goal is inferred from both; grid world domain
Lukowicz et al.[127]	Speech and visual cues to penalize/reinforce (anticipated) behavior
Liu et al.[116]	User interrupts robot behavior to utter verbal corrections

Table 8. User Interface: Language



Publication	Description
Yu et al.[213]	Robotics domain; language is translated into rewards; demonstrated as a chat-like interface where humans can give multiple text feedback utterances, e.g., comments on quality, what to do instead/focus on
Wu et al.[206]	Feedback for text; interactive text annotations on sentence-granularity, separate annotations for missing information, relevance, and factuality
Yow et al. [211]	Robot object grasping scenarios; correct trajectories of robots in space via language corrections (e.g., move closer to, go down, etc.)

Table 9. User Interface: Language (continuation)

**B.2 User Interface: Visual**

The following works mention visual user interactions for feedback collection:

Publication	Description
Veeriah et al. [193]	Using facial features as input in the observations and human-generated explicit rewards. Facial values are used to optimize rewards
Torabi et al. [189]	Instead of giving demonstrations via an agent-compatible interface, agent actions are observed and inferred with an inverse dynamics model
Arakawa et al.[9]	Using facial gestures to infer reward/preferences
Li et al.[103]	Simultaneous binary feedback and facial feedback (classified as good and bad)
Zhang et al.[220]	Eye-tracking for Atari games; gaze can be used as additional information (e.g., attention guidance) during human play
Cui et al.[48]	Mapping facial gestures (without clear intention) to a reward distribution; using calibration phase to learn mapping
Okudo et al.[148]	Generate subgoals for a more complex overall behavior in a grid-world domain (points to reach via a UI drawing style); for a robot (verbal description of subgoals to reach)
Aronson et al.[10]	Combine joystick inputs with gaze tracking in a robot teleoperation setting (picking objects) to identify goals
Trick et al.[191]	Giving action advice (encoding by mentioning an object for the robot to grip) via language (with keyword spotting) or by interpreting pointing gestures; robotics grasping task
Zhang et al.[217]	Photo-taking task (optimal frame); explicit evaluative survey feedback, followed by a review of the feedback with both facial tracking and interaction changing of values
Xu et al.[208]	Users provide visual milestones (successful subgoal states as image observations) and move the robot to new positions
Candon et al. [35]	Measure multiple features of nonverbal human behavior during co-execution of an autonomous agent in a video game setting

Table 10. User Interface: Visual

### B.3 User Interface: Buttons/Controls

The following works mention visual buttons/other types of controls for feedback collection:

Publication	Description
Thomaz et al.[185, 188]	Sophie's Kitchen (interactive interface) with on-screen scales
Judah et al.[82]	Critique session where the end-user can analyze trajectories of the current policy and label actions as good or bad
Cakmak et al.[33]	Four good and four bad examples for hand-over; robot controlled by human via GUI sliders
Taylor et al.[182]	Interactive football simulation with multiple players; human provides "action/-pass" command at appropriate times
Knox et al.[88]	Buttons for positive/negative feedback
Li et al.[106]	Button presses to give positive/negative feedback for an action
Loftin et al.[119]	Investigating human feedback strategies
Amir et al.[8]	In student-teacher learning, the teacher can give action advice
Raza et al.[155]	Action advice for reward shaping instead of evaluative reward, Gridworld domain, small user study (2 participants)
Loftin et al.[120]	Simple video game environment with buttons for feedback
El Asri et al.[54]	Users rate trajectories (dialogs) with scores from 1 to 10
Sørensen et al. [172]	Super Mario; select most promising behavior to guide evolution
Fachantidis et al.[57]	Teaching task with finite advice budget; teacher can choose advice as action or no advice
MacGlashan et al.[130]	Slider with simple environment rendering
Christiano et al. [42]	Two videos side-by-side with feedback buttons
Li et al.[105]	Grid world domain; demonstration, then evaluative feedback during learning
Basu et al.[19]	Using preferences between observed episodes induced by different reward functions and specifying different salient features
Lončarević et al.[121]	Robotics; basketball throw, users can rate throws on different scales (ordinal scale: further than goal, hit, closer than goal, etc.)
Ibarz et al.[79]	Atari games; two phases: initial demo generation, then iterative updates with preference reward model
Warnell et al.[201]	Atari games; Button-based UI for feedback
Akkaladevi et al.[4]	Robotics co-assembly task; feedback via GUI, users can choose the best action from a list of possible actions
Wang et al.[199]	Train a supervised policy based on human-generated data to support learning in interactive RL; feedback not specified
Frazier et al.[64]	Minecraft domain, goal-seeking task; simulated action advice (with low frequency) from oracle, no actual user studies

Table 11. User Interface: Button/Controls



Publication	Description
Arzate Cruz et al.[13]	Provide agent with action they believe is optimal + Scalar-valued rating
Wilde et al.[203]	Visual interface showing 2D trajectories; users asked for preferred trajectories, list of violations and speed, refinement of annotated zones
Arzate Cruz et al.[11]	Super Mario; segmented level view
Koert et al.[90]	Multiple ways for human interaction with robot learning systems; initial reward subgoals, evaluative feedback or action advice after performed actions, state manipulation helping the robot
Arzate Cruz et al.[13]	Gridworld; give +1/-1 rating for a single action
Moreira et al.[141]	User interface showing the available actions with associated predicted probabilities; users must select best actions during a consecutive 100-step phase
Arzate Cruz et al.[13]	The human user specifies the goal object(s) in the environment
Raza et al.[156]	Human teacher "relabels" the selection actions, i.e., chooses best action but agents act according to learned policy
Menner et al.[136]	Therapists rate gait of subjects on a scale from 1-3 in four categories to train a reward model in a supervised fashion
Wilde et al.[203]	GUI to specify zones indicating traffic rules in a 2D warehouse; zones represent avoidance, speed limits, desired roads, etc.
Law et al.[101]	Redesign agents in a loop based on performance
Wilde et al.[202]	Combine preference-based and scalar reward by using a slider to show preference on a spectrum
Hsiung et al.[77]	Users can choose to use evaluative or action advice to teach robot agents
Cruz et al.[46]	Humans can give "action advice" for steps if they want; do not give for each step; discusses efficient use of samples
Crochepierre et al.[44]	Symbolic regression domain; first group expressions into three groups, then generate preferences for pairs; optionally create "corrected" expression
Chi et al.[40]	Humans are queried to either teach via instruction/action advice or evaluation on a five-point scale for predicted action; simplified medical care setting
Wang et al.[198]	Two phases: label trajectories from an offline robotics dataset into preferable and non-preferable; label segments of trajectories having made more progress towards the goal
Bai et al.[16]	Chat interface; ranking of two entries with confidence (one or the other is better)
Scherf et al.[158]	During action advice giving, users are recorded and behavioral cues like response time and facial expressions are recorded
Sheidlower et al.[165]	Binary feedback with "toggle" approach (i.e., states are rewarded as long as the feedback direction is not actively changed)
Punzi et al.[153]	Immersive VR environment; voting for the best actor out of a pool of actors via button presses

Table 12. User Interface: Button/Controls (continuation)

Publication	Description
Myers et al.[142]	Robotics; Lunar Lander; rankings over observed trajectories
Zhang et al.[218]	Robotics/simulated robotics interfaces; robotics task
Bıyık et al.[27]	Preferences between segments in driving, tossing, and robotics tasks
Torne et al.[190]	Robotics domain; use binary preferences to guide exploration; asynchronous feedback
Bignold et al.[26]	Cartpole; view steps and actions, possibility to overwrite/rate given action if disagreement
Yu et al.[212]	Showing that scalar feedback can be better than binary feedback if considering the labelers
Ying et al.[210]	First, humans utter an instruction, then perform a series of actions; goal is inferred from both; grid world domain
Fitzgerald et al.[63]	Rare example of a multi-type feedback approach; studies how to query; no actual implementation with human feedback, just with oracles
Zhao et al. [221]	Selection of an object according to a personal reward function and an overall reward function; users select one of several objects assigned a weight and color
Li et al.[108]	Robotics application; applies corrections to robot movements to avoid goal misspecification; effect from multiple corrections is accumulated

Table 13. User Interface: Button/Controls (continuation)

B.4 User Interface: Physical

The following works mention using physical user interfaces for feedback collection, especially relevant in the domain of human-robot interaction:

Publication	Description
Li et al.[107]	Gait/walking support; requests ordinal feedback (good, very good, etc.) and preference compared to previous setup
Bobu et al. [28]	Sequence of states achieved by human manipulation of robot; representing decreasing value of a requested feature (e.g., distance to an object)
Liu et al.[117]	Human operations can intervene in robot behavior based on tele-operation equipment
Mehta et al.[134]	Robotics application; multi-stage approach starting with demonstrations, then corrections, and finally preferences
Aronson et al.[10]	Combine joystick inputs with gaze tracking in a robot teleoperation setting (picking objects) to identify goals
Schrum et al. [159]	Users can choose to provide correction to race car driving towards a goal; learned to be interpreted as action labels
Losey et al.[123]	In a human-robot co-working scenario, intentional physical corrections to the robot actions (like pushing/pulling, etc.) can be interpreted as feedback, changing the current trajectory online
Wilson-Small et al.[204]	Humans can guide drone behavior by physical touch/interactions with the drones
Shek et al. [166]	Users can correct the position of a specific object that the robot handles with its actuators (so the robot arm can be repositioned, but just indirectly by adjusting the robot position)
Mehta et al. [135]	Robotics; correction for moving robot trajectories via physical guidance

Table 14. User Interface: Physical

**B.5 User Interface: Programming**

The following works mention using programming-like interfaces (i.e. directly using programming, formal or domain-specific languages) for feedback collection:

Publication	Description
Maclin et al. [132]	Give instructions/rules in inductional logic/programming language (PROLOG)
Hayes-Roth et al.[75]	General constraints as LISP expressions to specify rules, heuristics, etc.
Yeh et al.[209]	Minecraft; humans can give template-based advice, then generates more fine-grained reward via a buffer
Van Waveren et al.[192]	From natural language and "block-based" (i.e., programming language) feedback, infer action refinement, alternative actions, and forbidden actions
Lin et al. [110]	Flight booking use case; users can give natural language feedback to an agent; reward function is inferred from free-form feedback

Table 15. User Interface: Other



B.6 User Interface: Other

The following works mention using other, miscellaneous types of user input for feedback collection:

Publication	Description
Ferrez et al.[62]	Using EEG error potentials to measure error in command interpretation (if the command was correctly understood by the interface)
Cakmak et al.[33]	Preferences over hand-overs executed by a robot (hand-over of objects)
Odom et al.[147]	Group features and request feedback for entire groups of features
Subramanian et al.[175]	Demo generation interface
Branavan et al. [30]	Provide action plans for reward function evaluation
Kim et al.[85]	Measure EEG for a robot imitating human gestures
Lin et al.[112]	Only simulated; noisy oracle for action advice
Losey et al.[124]	Not implemented, but potentially drawing of new trajectory/leading a robot
Cruz et al.[47]	Simulated robotics environment
Millán et al. [138]	Simulated user giving "advice", agreeing/disagreeing with a particular action
Zhang et al.[219]	Using human gaze as an additional feedback signal
Faulkner et al.[59]	Robot learning scenario; same value range as ground-truth reward function, i.e., scalar reward function, given via oracle
Najar et al.[143]	Context-dependent with respect to the current state of the task; examples: evaluative feedback, corrective feedback, guidance, contextual instructions
Najar et al.[143]	Communicated non-interactively prior to learning; not context-dependent, e.g., general constraints or general instructions
Najar et al.[143]	Feedback informs about past actions
Xu et al.[207]	Measure error potentials from EEG to detect suboptimal behavior
Najar et al.[143]	Guidance informs about future actions
Najar et al.[143]	Action advice given before an action; differentiates between high-level and low-level instructions (e.g., a single action or a goal similar to options)
Chetouani et al.[38]	Not specified, but assumed to be through the "social channel"
Chetouani et al.[38]	Through the "task channel," e.g., via kinesthetic, teleoperation, and observation
Harnack et al.[73]	Fully simulated feedback; binary signal based on whether state has improved based on action or not
Thierauf et al.[184]	Correction of natural language instructions during training and inference
Newman et al. [144]	Correcting the goal position of household items in a (simulated) dishwasher

Table 16. User Interface: Other