Learning from Evaluative Feedback I

Human-Interactive Robot Learning (HIRL)
Silvia Tulli - Kim Baraka - Mohamed Chetouani

Learning Goals

By the end of this lecture, you should be able to:

- Explain the advantages and limitations of Learning from Evaluative Feedback (LfE)
- Mathematically formulate LfE with appropriate terminology
- Distinguish learning with and without a reward function
- Explain different shaping methods:
  - Reward Shaping
  - Policy Shaping
  - Value Shaping
- Master one basic LfE algorithm:
  - Training an Agent Manually via Evaluative Reinforcement (TAMER)

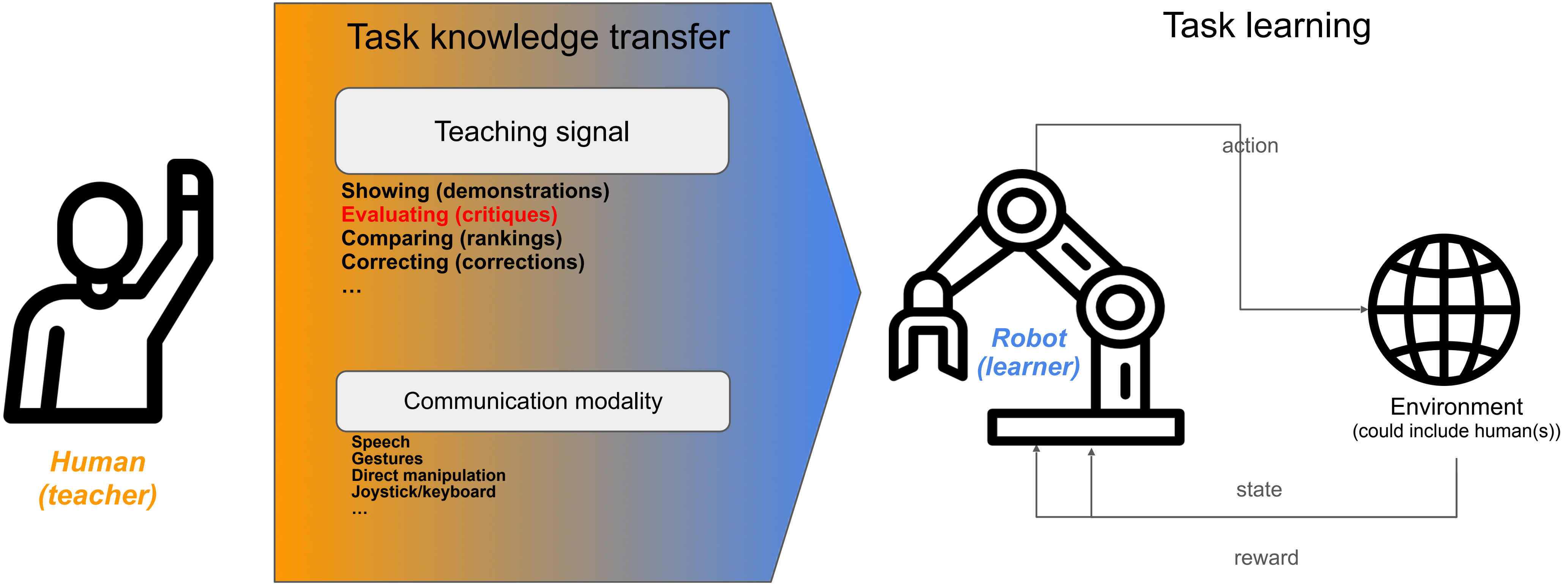
Scope

1. **Introduction**
    a. Learning from evaluative feedback: overview
    b. Examples: motivating examples & toy example
    c. What is a Human Feedback?
2. **Why learning from Human feedback?**
    a. Technical motivations
3. **Reward vs. Human feedback in RL**
    a. MDP\R
    b. Limitations
4. **Shaping methods**
    a. Reward shaping
    b. Value shaping
    c. Policy shaping

Scope

5. **Human feedback model**
    a. Reinforcement learning from human feedback (RLHF)
    b. TAMER: Training and Agent Manually via Evaluative Feedback
6. **Challenges & Limitations of RHLF**
    a. Amount of feedback
    b. Feedback distribution
    c. Reward hacking
    d. Overview of challenges

Introduction: Learning from evaluative feedback



Task knowledge transfer

Teaching signal

Showing (demonstrations)
Evaluating (critiques)
Comparing (rankings)
Correcting (corrections)
…

Communication modality

Speech
Gestures
Direct manipulation
Joystick/keyboard
…

Task learning

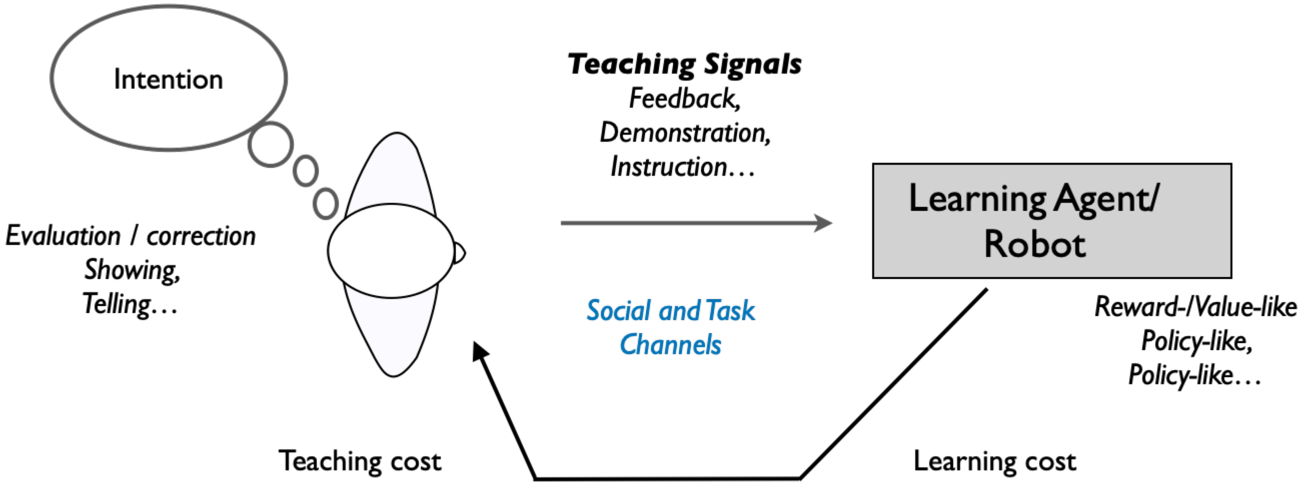
Human-driven: teacher decides when to intervene
Robot-driven: robot actively asks for human input
Mixed-initiative: both teacher and robot can take initiative

1. Introduction: Learning from evaluative feedback

Human feedback **serves the purpose of evaluating the performance of a robot's action,** providing **critique** in the process.

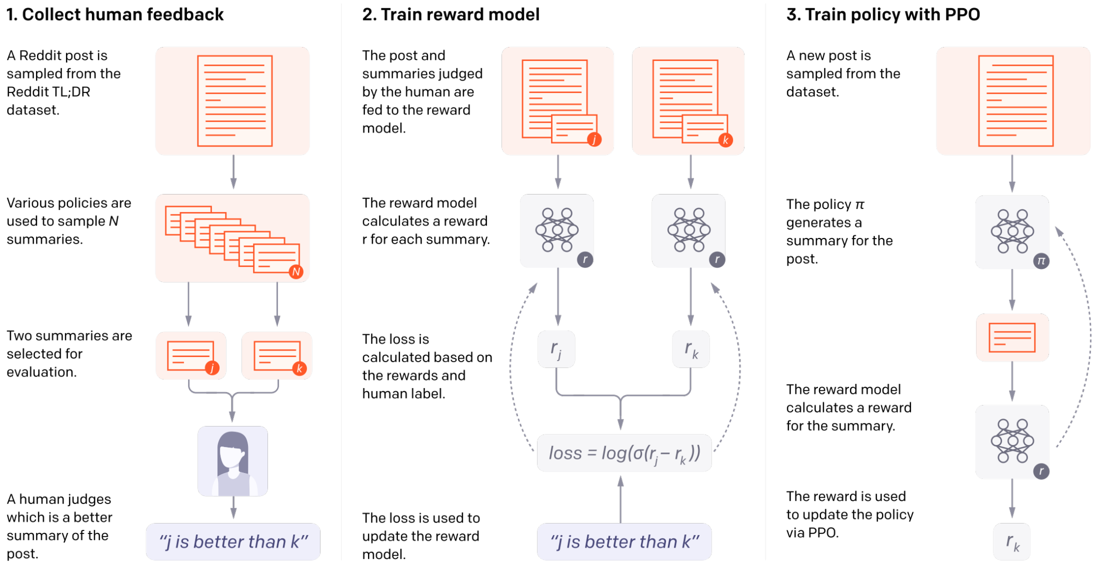
It is commonly **referred to as evaluative feedback,** as it entails **the human observer assessing the states at two consecutive time steps** (st and st+1) alongside the last action taken by the robot (a), **subsequently issuing a reward signal**.

The value of this evaluative feedback is **contingent upon the most recent action executed by the robot**.



1. Introduction: Motivating examples

Humans prefer summaries generated through “Human feedback” training based models



1. Collect human feedback

A Reddit post is sampled from the Reddit TL;DR dataset.

Various policies are used to sample *N* summaries.

Two summaries are selected for evaluation.

A human judges which is a better summary of the post.

“j is better than k”

2. Train reward model

The post and summaries judged by the human are fed to the reward model.

The reward model calculates a reward r for each summary.

The loss is calculated based on the rewards and human label.

The loss is used to update the reward model.

$loss = log(\sigma(r_j - r_k))$

“j is better than k”

3. Train policy with PPO

A new post is sampled from the dataset.

The policy π generates a summary for the post.

The reward model calculates a reward for the summary.

The reward is used to update the policy via PPO.

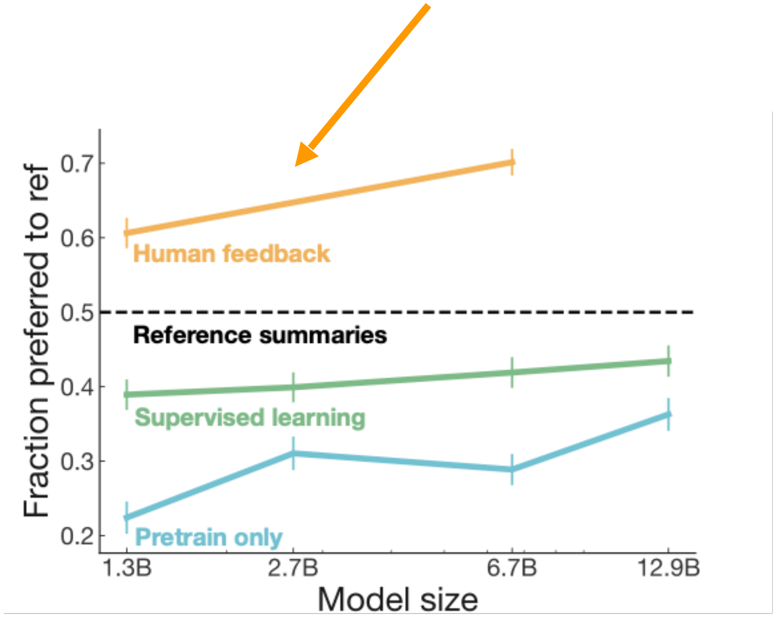


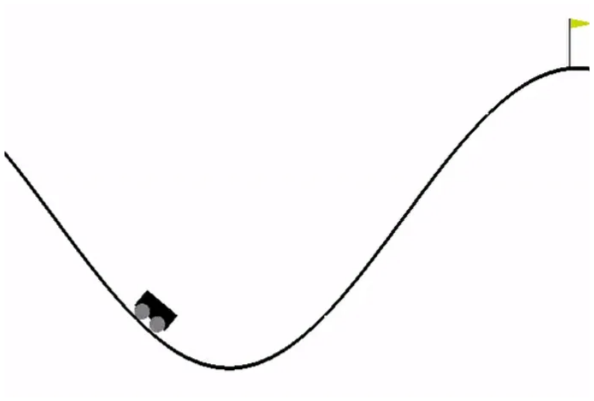
Figure 1: Fraction of the time humans prefer our models’ summaries over the human-generated reference summaries on the TL;DR dataset.[4] Since quality judgments involve an arbitrary decision about how to trade off summary length vs. coverage within the 24-48 token limit, we also provide length-controlled graphs in Appendix F; length differences explain about a third of the gap between feedback and supervised learning at 6.7B.

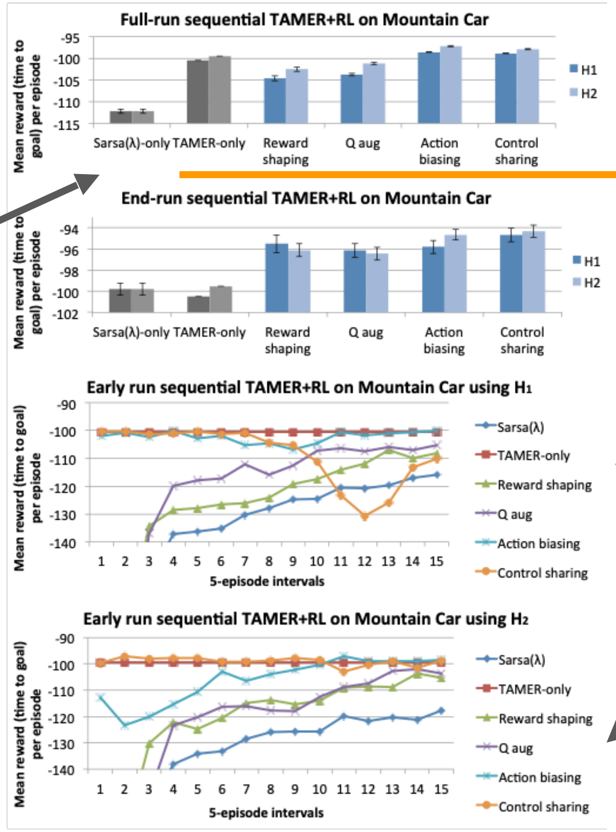
Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). Curran Associates Inc., Red Hook, NY, USA, Article 253, 3008–3021.

1. Introduction:
Toy example: Mountain car

Agents behavior can be **shaped through signals of approval and disapproval**, a natural form of human feedback.

Various approaches from reward shaping aiming to augment a **traditional reinforcement learning (MDP) agent with human feedback**





RL

Using Human Feedback

Two different Humans

Figure 1: Comparison of TAMER+RL techniques with SARSA(λ) and the TAMER-only policy on mountain car over 40 or more runs of 500 episodes. $\hat{H}_1$ and $\hat{H}_2$ are models from two different human trainers. The top chart considers reward over the entire run, and the second chart evaluates reward over the final 10 episodes. Error bars show standard error. The third and fourth charts display mean performance using $\hat{H}_1$ and $\hat{H}_2$ early in the run, during the first 75 episodes.

W. Bradley Knox and Peter Stone. 2012. Reinforcement learning from simultaneous human and MDP reward. In Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1 (AAMAS '12). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 475–482.

1. Introduction: Toy example: Sophie Kitchen

Examples of asymmetric intentions

Interactive Reinforcement Learning

- 10,000 states, and 2-7 actions available in each state
- Using the mouse, a **human trainer** can— at any point—**award a scalar reward signal, r ∈ [–1,1].**

- « The purpose of the experiment was to understand, when given a single reward channel, **how do people use it to teach the agent?** »

- People have **asymmetric intentions** they are communicating with their **positive and negative feedback**

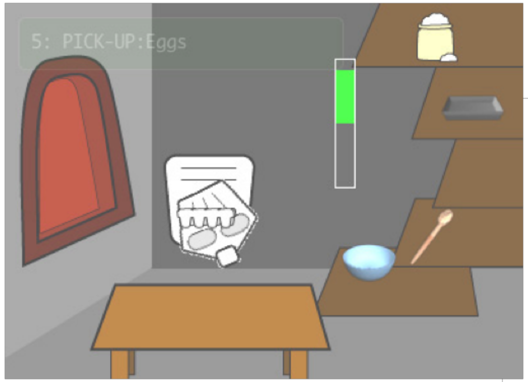


Fig. 2. *Sophie's Kitchen*: There are three locations (oven, table, shelf), and five baking objects (flour, eggs, pan, spoon, and bowl). The virtual robot, Sophie, learns to bake a cake via Q-Learning, and a human partner playing the game can contribute by issuing feedback with the mouse, creating a red or green bar for positive/negative feedback. The green bar seen here is the interactive human feedback.

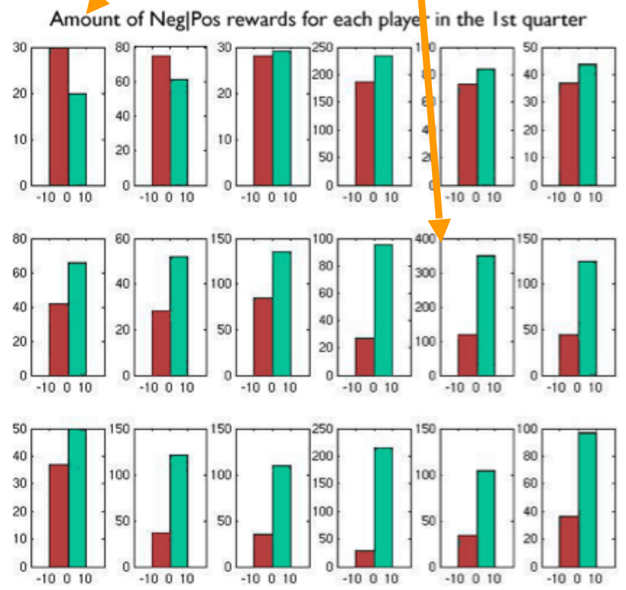


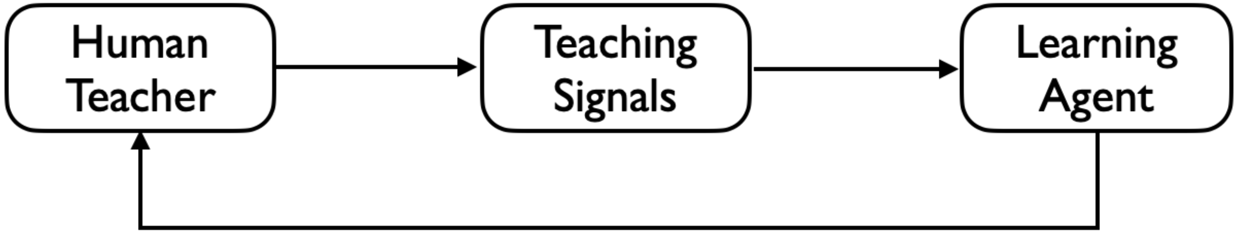
Fig. 3. Histograms of rewards for each individual in the first quarter of their session. The left column is negative rewards and the right is positive rewards. Most people even in the first quarter of training have a much higher bar on the right.

Thomaz, A.L.; Breazeal, C. Asymmetric Interpretations of Positive and Negative Human Feedback for a Social Learning Agent. In Proceedings of the 16th IEEE International Symposium Robot and Human Interactive Communication (RO-MAN 2007), Jeju, Korea, 26–29 August 2007; pp. 720–725, doi:10.1109/ROMAN.2007.4415180.

1. What is a Human Evaluative Feedback?

Feedback signals are **delivered by an observing human expert** as the **agent attempts to perform a task.**

The **human observes robot's action and states** and then provides a **teaching signal** called **evaluative feedback.**

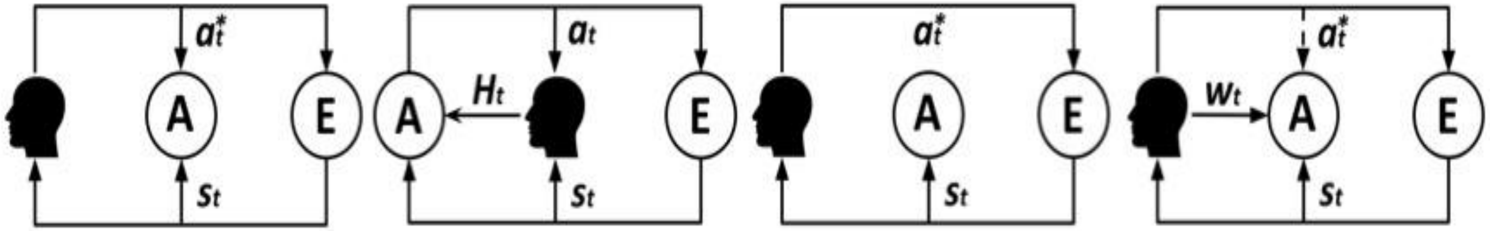


Human Teacher

Teaching Signals

Learning Agent

Evaluative feedback=> evaluation of robot's action
≠ Demonstrations



(a) Standard imitation learning (b) Evaluative feedback (c) Imitation from observation (d) Learning attention from human

Chetouani, Interactive Robot Learning: An overview, Springer (2023)
Ruohan Zhang, Faraz Torabi, Lin Guan, Dana H. Ballard, Peter Stone: Leveraging Human Guidance for Deep Reinforcement Learning Tasks. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence

1. What is a Human Evaluative Feedback?

Evaluative feedback signals are communicated **after robot's action.**

Human feedback H(s, a) is considered as an observation about the reward r(s, a).

Binary and Real-valued quantities have been considered in interactive reinforcement learning.

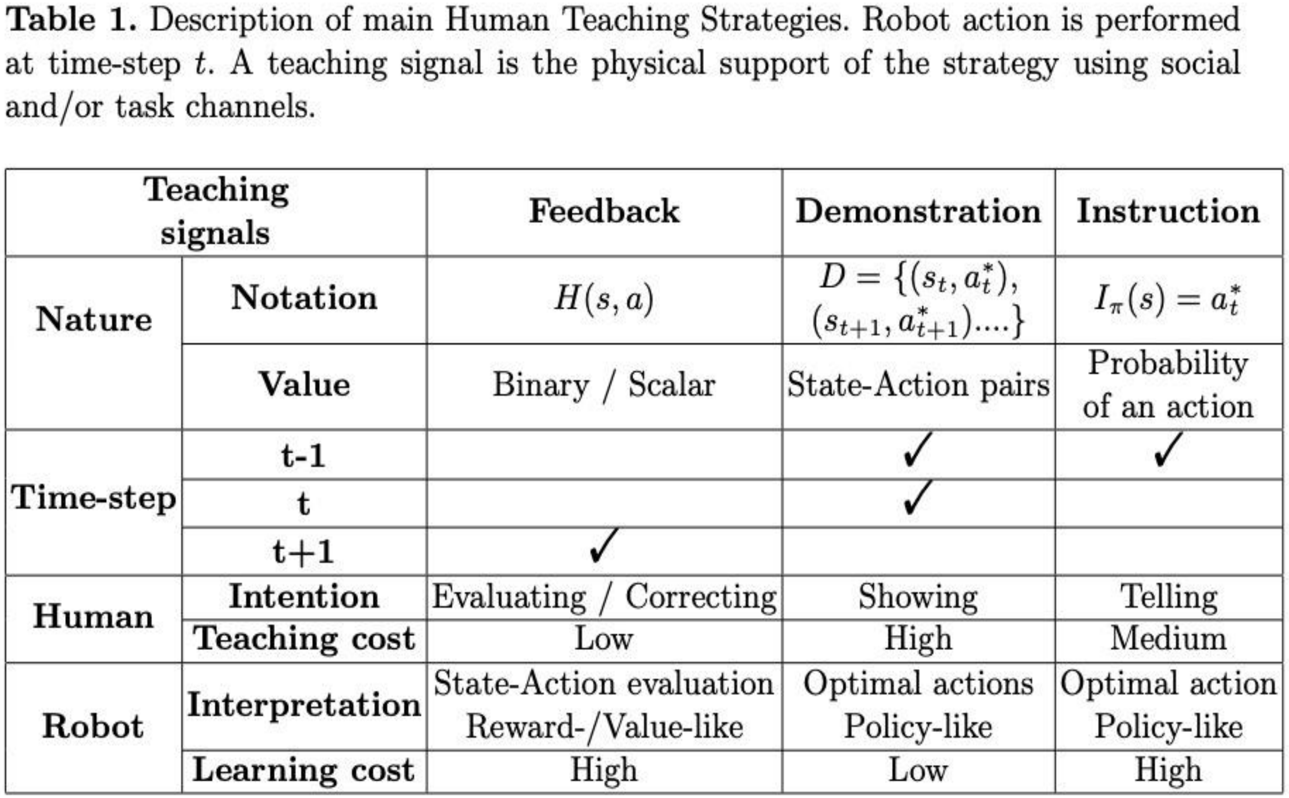


Table 1. Description of main Human Teaching Strategies. Robot action is performed at time-step t. A teaching signal is the physical support of the strategy using social and/or task channels.

Teaching signals		Feedback	Demonstration	Instruction
Nature	**Notation**	$H(s,a)$	$D = \{(s_t, a_t^*), (s_{t+1}, a_{t+1}^*)....\}$	$I_\pi(s) = a_t^*$
	Value	Binary / Scalar	State-Action pairs	Probability of an action
Time-step	**t-1**		✓	✓
	t		✓	
	t+1	✓		
Human	**Intention**	Evaluating / Correcting	Showing	Telling
	Teaching cost	Low	High	Medium
Robot	**Interpretation**	State-Action evaluation Reward-/Value-like	Optimal actions Policy-like	Optimal action Policy-like
	Learning cost	High	Low	High

Chetouani, Interactive Robot Learning: An overview, Springer (2023)
Hong Jun Jeon, Smitha Milli, and Anca Dragan. 2020. Reward-rational (implicit) choice: a unifying formalism for reward learning. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)

2. Why learning from Human feedback?

Guidance: Human feedback provides valuable guidance to the learning process of robots. It helps them understand what actions are desirable or undesirable in various situations, enabling them to improve their decision-making abilities.

Validation: Human feedback serves as a validation mechanism for the performance of robots. By receiving feedback from humans, robots can confirm whether their actions align with human expectations and requirements

Adaptability: Learning from human feedback allows robots to adapt to dynamic and complex environments. They can adjust their behavior based on the feedback received, improving their performance over time and in different scenarios.

2. Why learning from Human feedback?

Generalization: Human feedback facilitates the generalization of learned knowledge. By accurately modeling the fundamental principles inherent in human feedback, robots can apply similar concepts to new situations, even if they haven't encountered them before.

Efficiency: Human feedback can accelerate the learning process for robots. Instead of relying solely on trial and error, robots can leverage human guidance to more quickly identify effective strategies and avoid ineffective ones.

Robustness: Human feedback helps enhance the robustness of AI systems by providing insights into real-world scenarios and potential challenges. By learning from human evaluations, robots can adapt their behavior to various conditions, thus improving their resilience and ability to perform effectively in diverse environments.

2. Why learning from Human feedback?

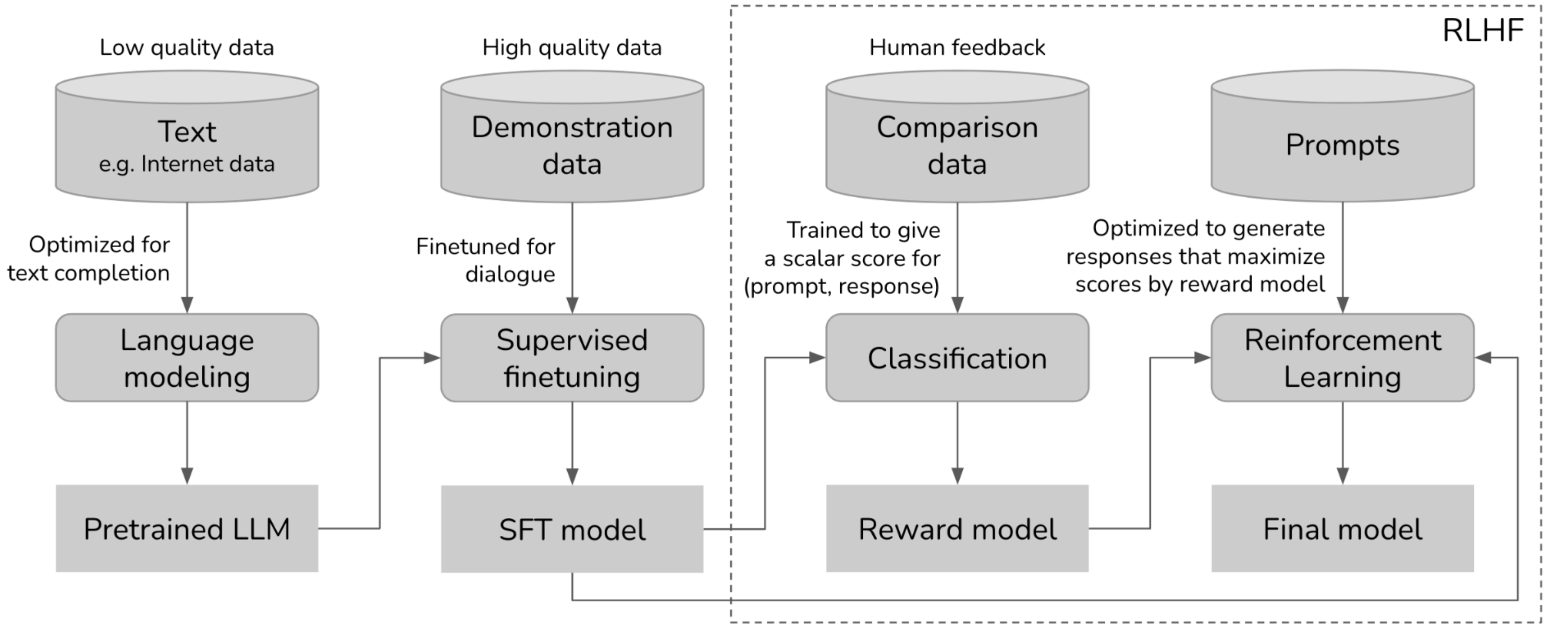
Safety: Incorporating human feedback contributes to the safety of AI and robotics applications. By considering human evaluations, robots can avoid actions that pose risks to themselves, humans, or the environment.

Efficiency in Exploration: Human feedback accelerates the exploration process by guiding robots towards more efficient learning strategies. By leveraging insights from human evaluations, robots can prioritize actions that are more likely to yield valuable information or lead to successful outcomes.

2. Why learning from Human feedback?

Ethical Considerations: Integrating human feedback is pivotal for addressing ethical considerations in AI and robotics. By learning from human evaluations, robots can steer clear of harmful or unethical actions and prioritize behaviors that align with human values and ethical standards.

2. Why learning from Human feedback?



RLHF

Low quality data

Text
e.g. Internet data

Optimized for
text completion

Language modeling

Pretrained LLM

High quality data

Demonstration data

Finetuned for
dialogue

Supervised finetuning

SFT model

Human feedback

Comparison data

Trained to give
a scalar score for
(prompt, response)

Classification

Reward model

Prompts

Optimized to generate
responses that maximize
scores by reward model

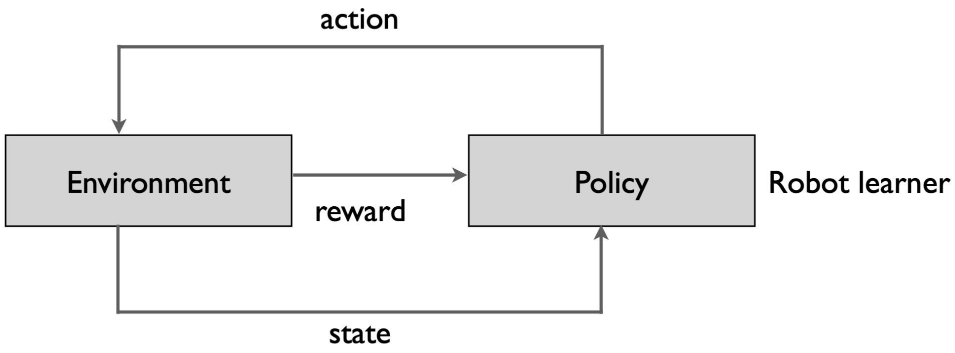
Reinforcement Learning

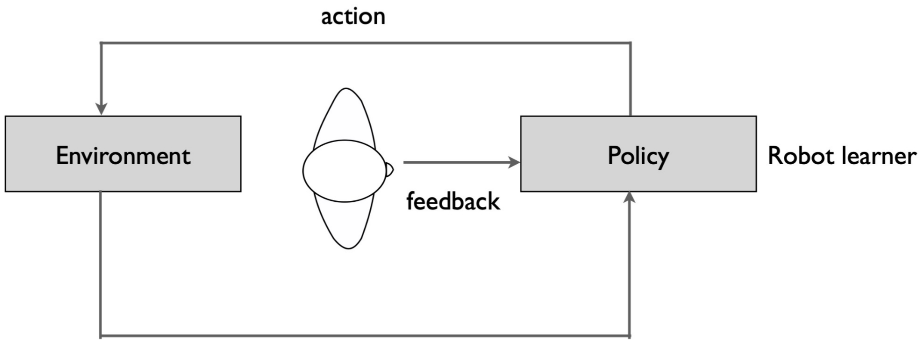
Final model

Scale May '23	**>1 trillion** tokens	**10K - 100K** (prompt, response)	**100K - 1M** comparisons (prompt, winning_response, losing_response)	**10K - 100K** prompts
Examples **Bolded**: open sourced	GPT-x, Gopher, **Falcon**, LLaMa, **Pythia**, **Bloom**, **StableLM**	**Dolly-v2, Falcon-Instruct**		InstructGPT, ChatGPT, Claude, **StableVicuna**

3. Reward R(s,a) vs. Human Feedback H(s,a)

- Using human feedback directly as the reward signal represents a naive approach in learning from evaluative feedback.
- In this approach, the reward signal provided by humans serves as the sole measure of success or failure for the robot's actions.



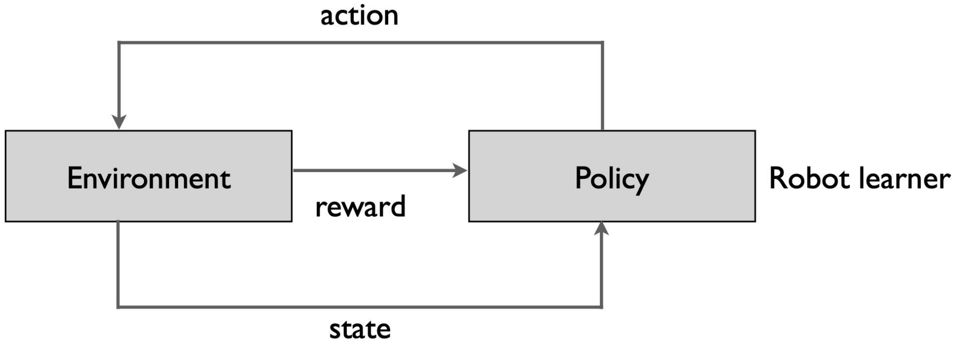


Chetouani, Interactive Robot Learning: An overview, Springer (2023)

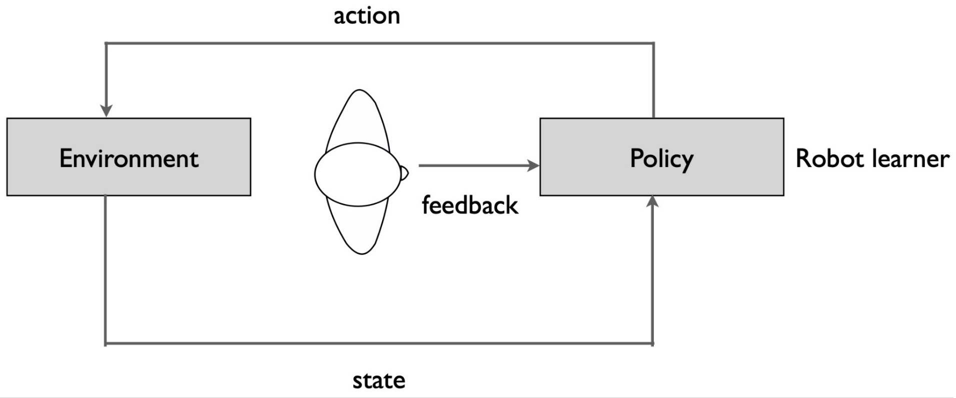
3. Reward R(s,a) vs. Human Feedback H(s,a)

Limitations:

- **Subjectivity:** Human feedback can be subjective and prone to biases. Different evaluators may provide varying assessments for the same action, leading to inconsistencies in the learning process.
- **Limited Scalability:** Relying solely on human feedback for rewards limits the scalability of the learning process. As the complexity of tasks increases or the number of actions grows, it becomes impractical for humans to provide continuous and detailed feedback



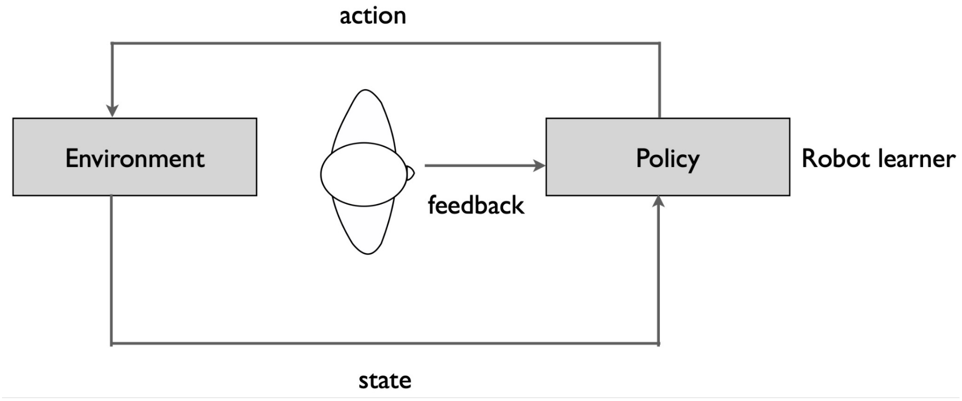
Chetouani, Interactive Robot Learning: An overview, Springer (2023)



3. Reward R(s,a) vs. Human Feedback H(s,a)

Limitations:

- **High Cost:** Collecting human feedback can be time-consuming and expensive, especially for tasks requiring extensive training data. This high cost makes it impractical for large-scale or real-time applications.
- **Limited Autonomy:** Depending solely on human feedback restricts the autonomy of the robot. Instead of learning directly from interactions with the environment, the robot relies heavily on external guidance, which may hinder its ability to adapt to new or unforeseen situations.

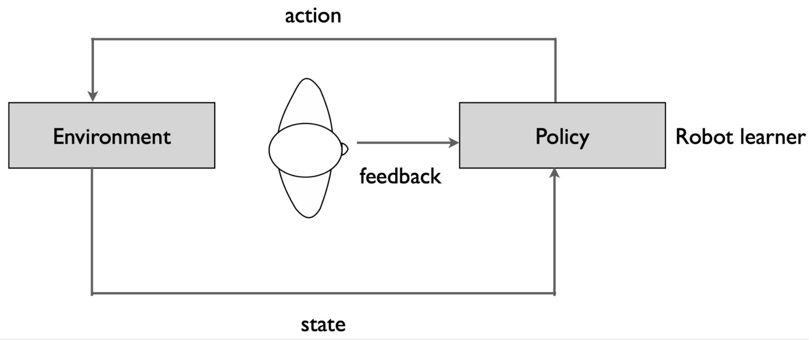


Chetouani, Interactive Robot Learning: An overview, Springer (2023)

3. Reward R(s,a) vs. Human Feedback H(s,a)

Attributes:

- **Arity:** This attribute describes whether a single instance is evaluated in isolation (unary) or relative to other instances (binary, n-ary).
- **Involvement:** The labeler may either passively observe an instance, actively generate it, or coactively participate in its generation (co-generation).
- **Granularity:** This ranges from whole episode recordings over partial segments to feedback on individual steps (i.e., states, actions, or state-action pairs).

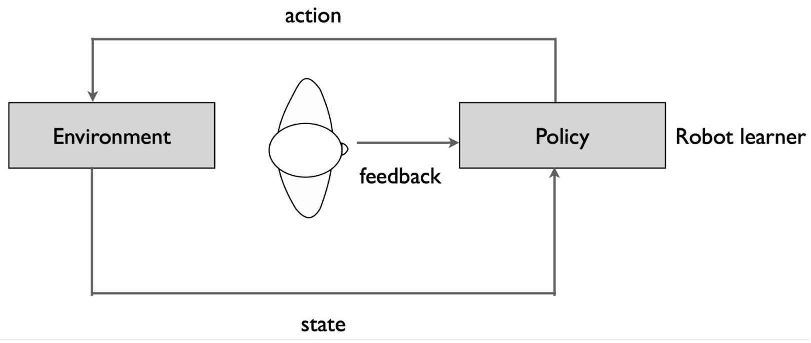


Timo Kaufmann and Paul Weng and Viktor Bengs and Eyke Hüllermeier (2024), A Survey of Reinforcement Learning from Human Feedback, arXiv:2312.14925

3. Reward R(s,a) vs. Human Feedback H(s,a)

Attributes:

- **Abstraction:** This describes whether feedback is given directly on raw instances, e.g., behavior recordings (see granularity) or on abstract features of the instances.
- **Explicitness:** Humans may communicate explicitly for the purposes of feedback or implicitly as a side-effect of actions directed at other purposes.
- **Intent:** A human may be *evaluative*, *instructive*, or *descriptive* in their explicit feedback, while they are generally *literal* in their implicit feedback.



Timo Kaufmann and Paul Weng and Viktor Bengs and Eyke Hüllermeier (2024), A Survey of Reinforcement Learning from Human Feedback, arXiv:2312.14925

4. Shaping methods: Reward shaping

The concept of **shaping rewards** aims to expedite the learning process by **providing additional guidance to the learning algorithm (Ng. et al. 1999)**

Objective: To learn a policy for MDP = (S, A, γ, gamma, R)

Where S represents the set of states, A the set of actions, T the transition function, γ the discount factor, and R the original reward function. The objective is to learn a policy that maximizes cumulative rewards over time.

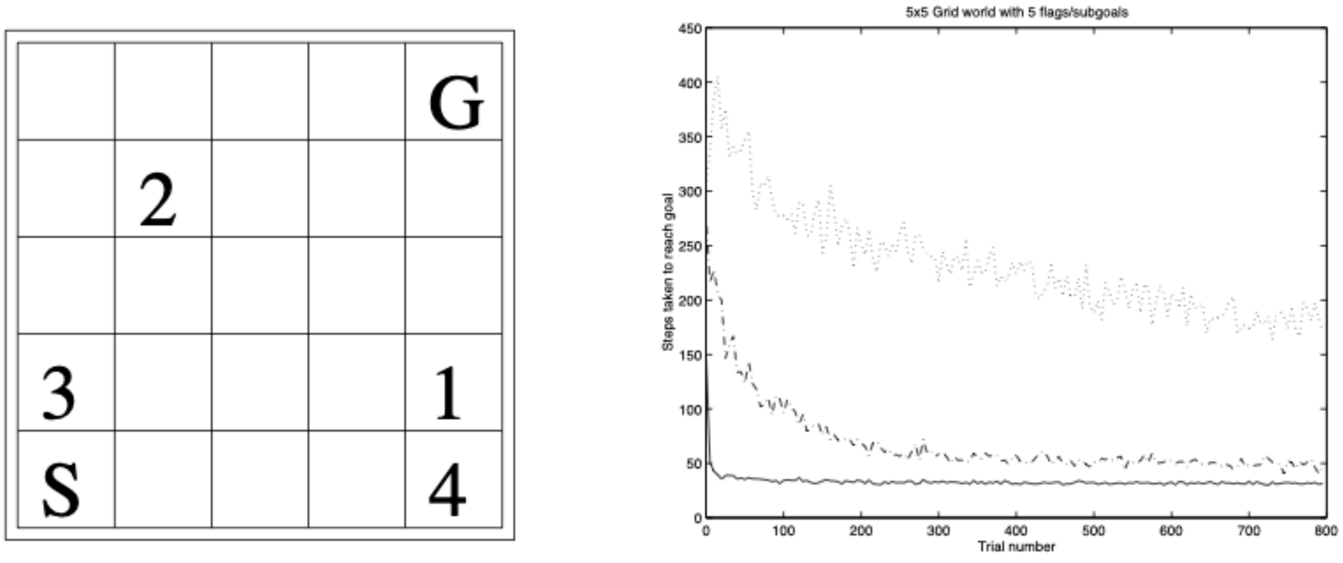
Method:

- **Help the learning RL algorithm by giving it additional "shaping' rewards"**
- These rewards are supplementary to the original rewards provided by the environment and are designed to steer the learning algorithm towards more efficient exploration of the state-action space.
- Instead of directly applying the RL algorithm to M, the authors propose working with a transformed MDP, denoted as M' = (S, A, T, γ, R'), **where R' = R + F**.
- **R' represents the augmented reward function that includes the original rewards (R) and the shaping rewards (F).**

The **shaping reward function,** denoted as F: S x A x S -> ℝ, provides additional rewards based on the current state, action, and next state transitions. This function is **carefully designed to encourage desirable behaviors or discourage undesirable ones**.

A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in Proc. of the 16th ICML, pp. 341–348, 1999.

4. Shaping methods: Reward shaping



(a) (b)

Figure 2: (a) 5x5 grid-world with 5 subgoals (including goal state), which must be visited in order $1, 2, 3, 4, G$. (b) Experiment with 5x5 grid-world with subgoals. Plot of steps taken to goal vs. trial number. Dot is no shaping, dot-dash is $\Phi = \Phi_0$, solid is $\Phi = \Phi_1$.

A. Y. Ng, D. Harada, and S. Russell, “Policy invariance under reward transformations: Theory and application to reward shaping,” in Proc. of the 16th ICML, pp. 341–348, 1999.

4. Shaping methods: Reward shaping

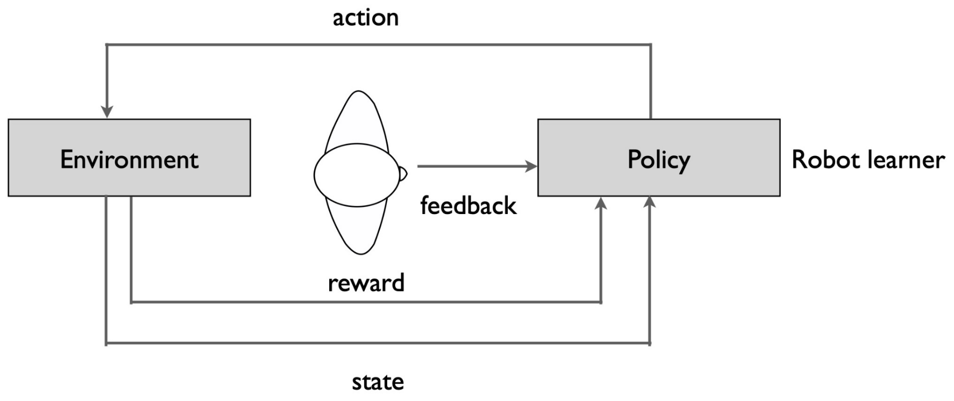
When the RL agent has access to a predefined reward function (R), a new reward function (R') is formed by adding both forms of reward (R + Rh), where Rh represents the human-delivered reward.

This shaping method is **model-free**, as the **numerical values from the human teacher directly augment the reward function.**

The revised reward function $R'(s,a)$is computed by adding a scaled version of the human-delivered reward $\hat{H}(s,a)$ to the original reward function $R(s,a)$.

This is expressed as: $\mathbf{R'(s,a)=R(s,a)+\beta \cdot \hat{H}(s,a)}$

Here, β represents a scaling factor that adjusts the influence of the human-delivered reward $\hat{H}(s,a)$ on the overall reward.



Najar A and Chetouani M (2021) Reinforcement Learning With Human Advice: A Survey. Front. Robot. AI 8:584075.

4. Shaping methods: Reward Shaping - Value Shaping - Policy Shaping

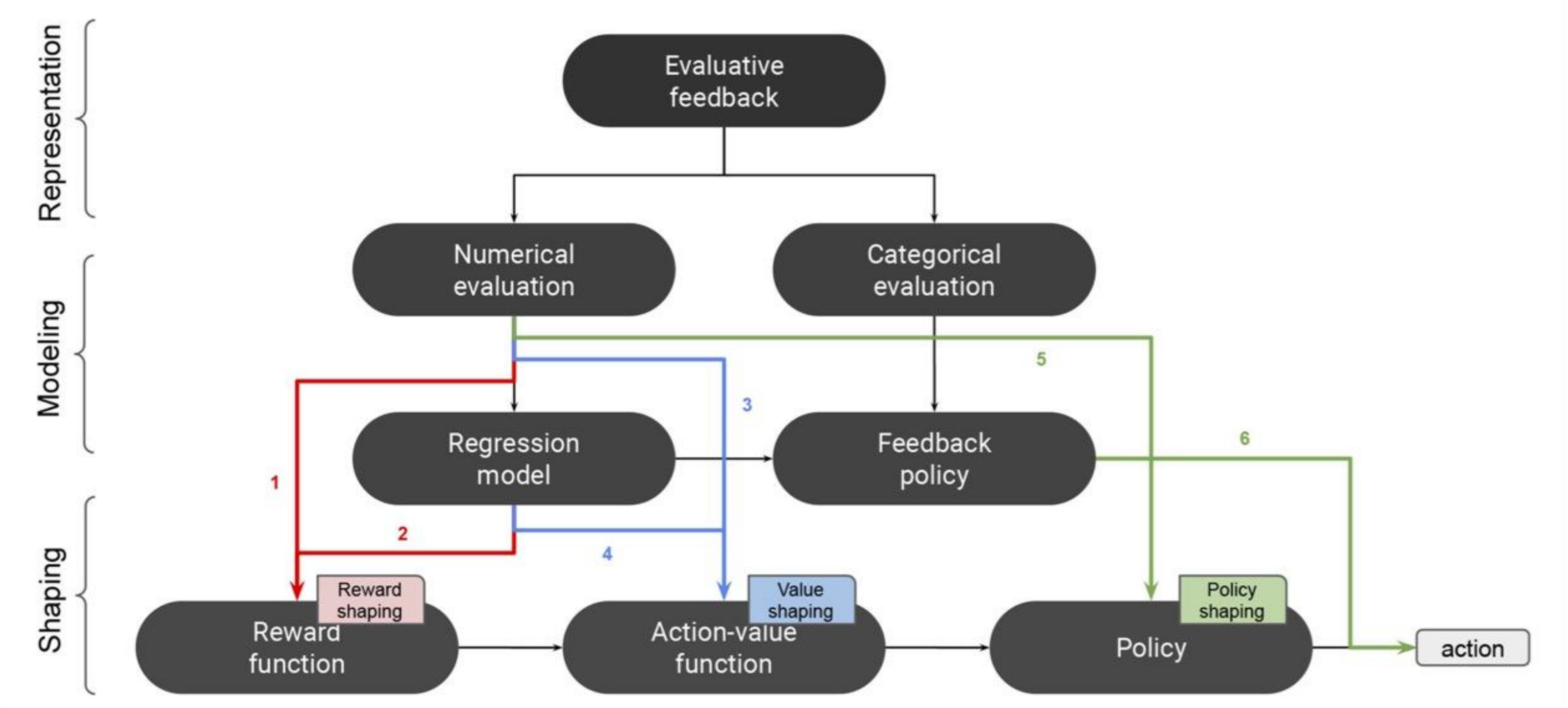


FIGURE 2 | Shaping with evaluative feedback. 1: model-free reward shaping. 2: model-based reward shaping. 3: model-free value shaping. 4: model-based value shaping. 5: model-free policy shaping. 6: model-based policy shaping.

Najar A and Chetouani M (2021) Reinforcement Learning With Human Advice: A Survey. Front. Robot. AI 8:584075. doi: 10.3389/frobt.2021.584075

4. Shaping methods: Value shaping

Human feedback is considered as a **human value function**, typically represented as a set of ratings corresponding to the agent's actions.

These ratings reflect the human evaluator's assessment of the agent's action in relation to anticipated future behavior. To incorporate this feedback into the learning process, various shaping methods have been explored in the literature.

One such method involves augmenting the **action-value function Q(s,a) with the human feedback signal Ĥ(s,a)**, formulated as follows:

$$Q'(s,a)=Q(s,a)+\beta\cdot\hat{H}(s,a)$$

Here, β represents a decaying weight factor that modulates the influence of the human feedback on the **action-value function**

Najar A and Chetouani M (2021) Reinforcement Learning With Human Advice: A Survey. Front. Robot. AI 8:584075.

4. Shaping methods: Policy shaping

This feedback is utilized to directly influence the agent's policy, impacting its **decision-making process**. Two primary methods have been explored in the literature:

1. **Action Biasing**: In this method, shaping occurs solely during the decision-making phase. The value function remains unaffected by the human feedback augmentation. **The agent selects actions by maximizing the combined score of the original action-value function Q(s,a) and the scaled human feedback $\hat{H}(s,a)$, using the formula:**

$$a_{*}=\mathrm{argmax}[Q(s,a)+\beta\cdot\hat{H}(s,a)]$$

Here, β represents a weighting factor determining the influence of human feedback on action selection.

2. **Control Sharing**: This approach involves **arbitrating between the Markov Decision Process (MDP) policy and the human value function**. The human-derived policy, based on feedback, is incorporated into action selection using a probability factor δ. T**he probability of selecting actions based on human feedback** is calculated as:

$$\Pr[\ a=\mathrm{argmax}\ \hat{H}(s,a)\ ]=\min(\delta,1)$$

Najar A and Chetouani M (2021) Reinforcement Learning With Human Advice: A Survey. Front. Robot. AI 8:584075.

4. Shaping methods: Reward Shaping - Value Shaping - Policy Shaping

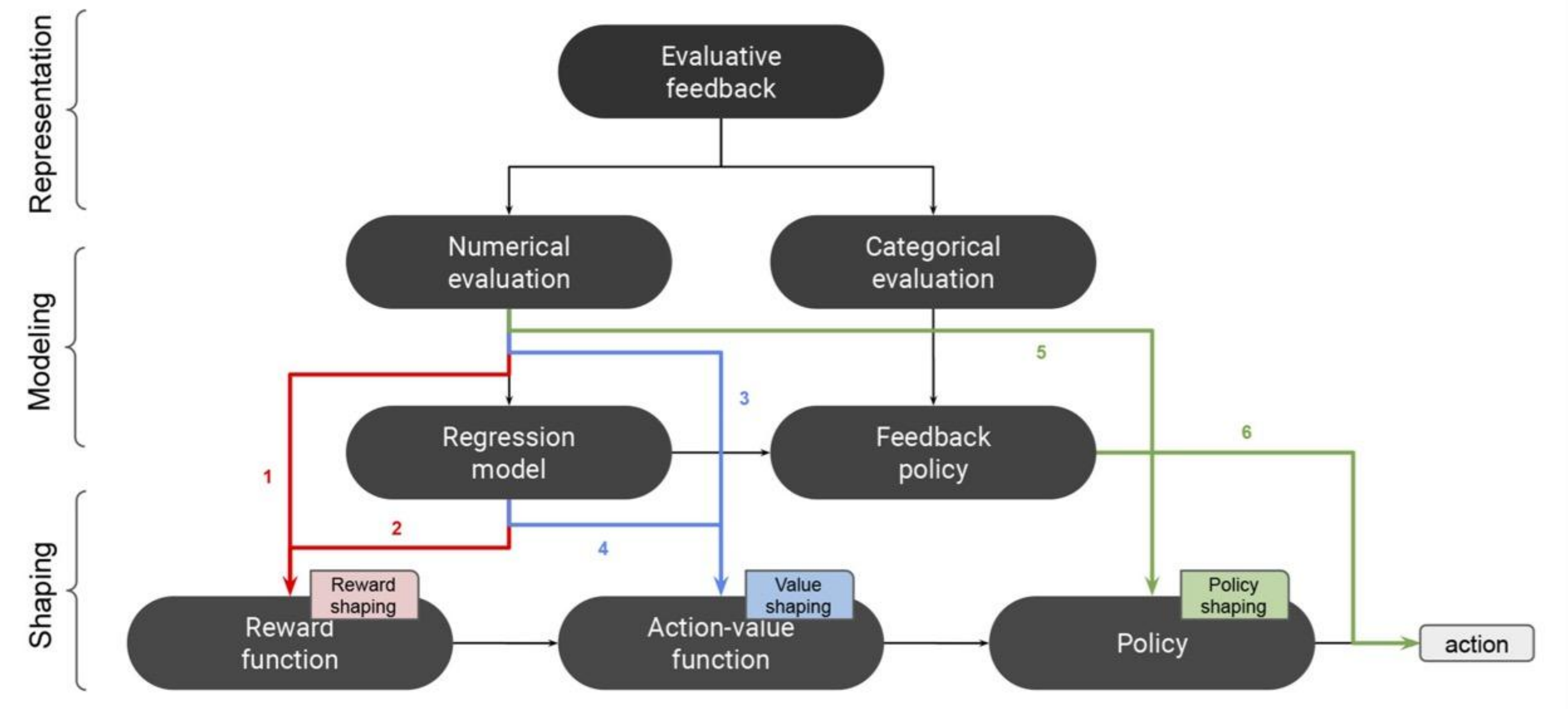


FIGURE 2 | Shaping with evaluative feedback. 1: model-free reward shaping. 2: model-based reward shaping. 3: model-free value shaping. 4: model-based value shaping. 5: model-free policy shaping. 6: model-based policy shaping.

Najar A and Chetouani M (2021) Reinforcement Learning With Human Advice: A Survey. Front. Robot. AI 8:584075. doi: 10.3389/frobt.2021.584075

5. Human feedback model:

Agent alignment problem:

“How can we create agents that behave in accordance with the user’s intentions?”

Alignment via reward modeling:

1. Learning a reward function from the feedback of the user that captures theirs intentions.
2. Training a policy with RL to optimize the learned reward function.

Approach: Reward modeling

Separate what to achieve (the “What?”) from learning how to achieve it (“The How?”)

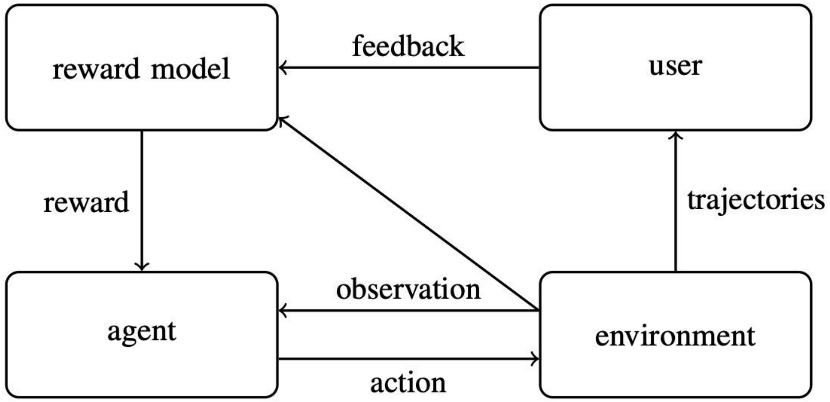


Figure 1: Schematic illustration of the reward modeling setup: a reward model is trained with user feedback; this reward model provides rewards to an agent trained with RL by interacting with the environment.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). Scalable agent alignment via reward modeling: a research direction. arXiv preprint arXiv:1811.07871.

5. Human feedback model:

Bootstrapping problem:

- Humans generally do poorly at training RL agents providing scalar reward directly.
- Which form or combination of feedback works well for which domain is currently an open question.
- How to adapt to the way humans provide feedback?
- **Problem:** “How do we train an algorithm that learns to interpret feedback, if itself does not already know how to interpret feedback?”

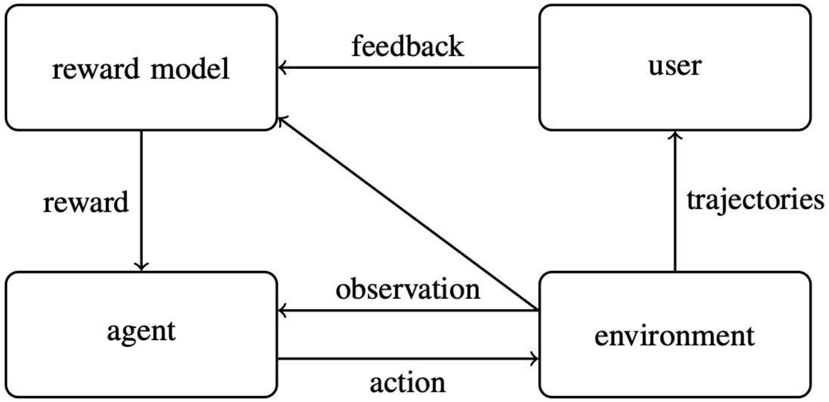
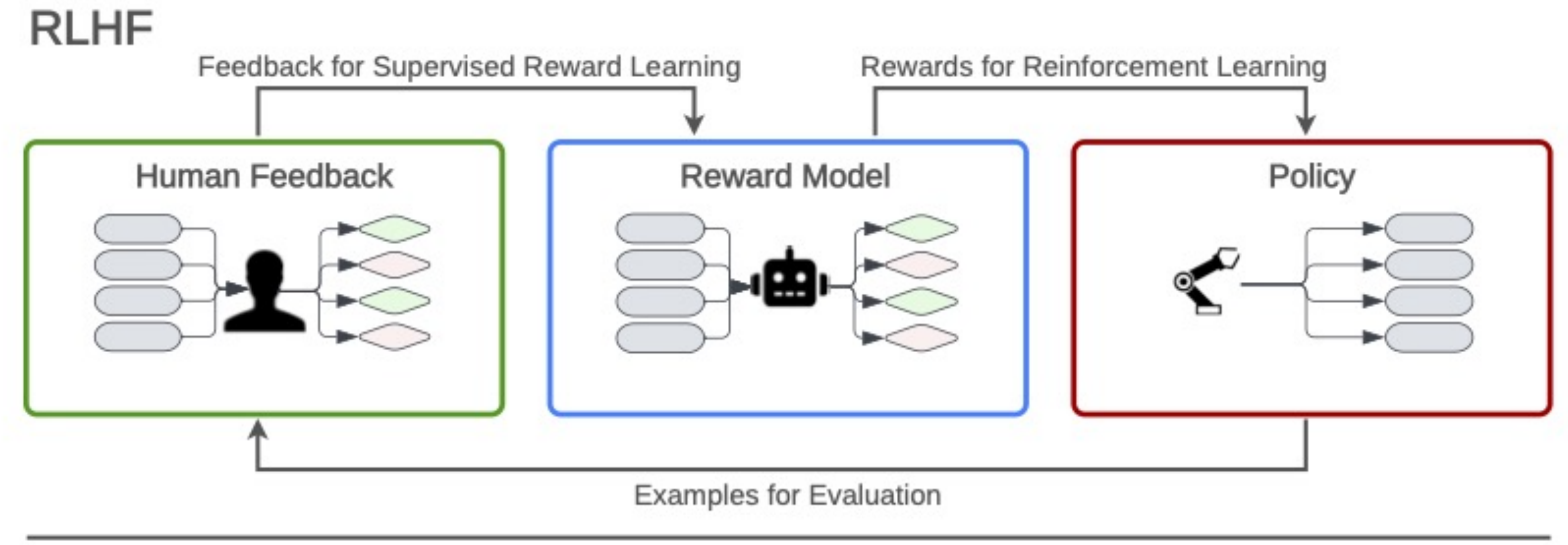


Figure 1: Schematic illustration of the reward modeling setup: a reward model is trained with user feedback; this reward model provides rewards to an agent trained with RL by interacting with the environment.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). Scalable agent alignment via reward modeling: a research direction. arXiv preprint arXiv:1811.07871.

5. Reinforcement learning from human feedback (RHLF)


RLHF
Feedback for Supervised Reward Learning
Rewards for Reinforcement Learning
Human Feedback
Reward Model
Policy
Examples for Evaluation

Casper, Stephen, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman et al. "Open problems and fundamental limitations of reinforcement learning from human feedback." arXiv preprint arXiv:2307.15217 (2023).

5. TAMER: Training an Agent Manually via Evaluative Feedback

TAMER follows a **Reward Modeling Approach**

TAMER exploits supervised learning techniques to construct a model of a human's reward function.

This learned model is then employed to select actions that are anticipated to yield the highest reward based on the provided feedback

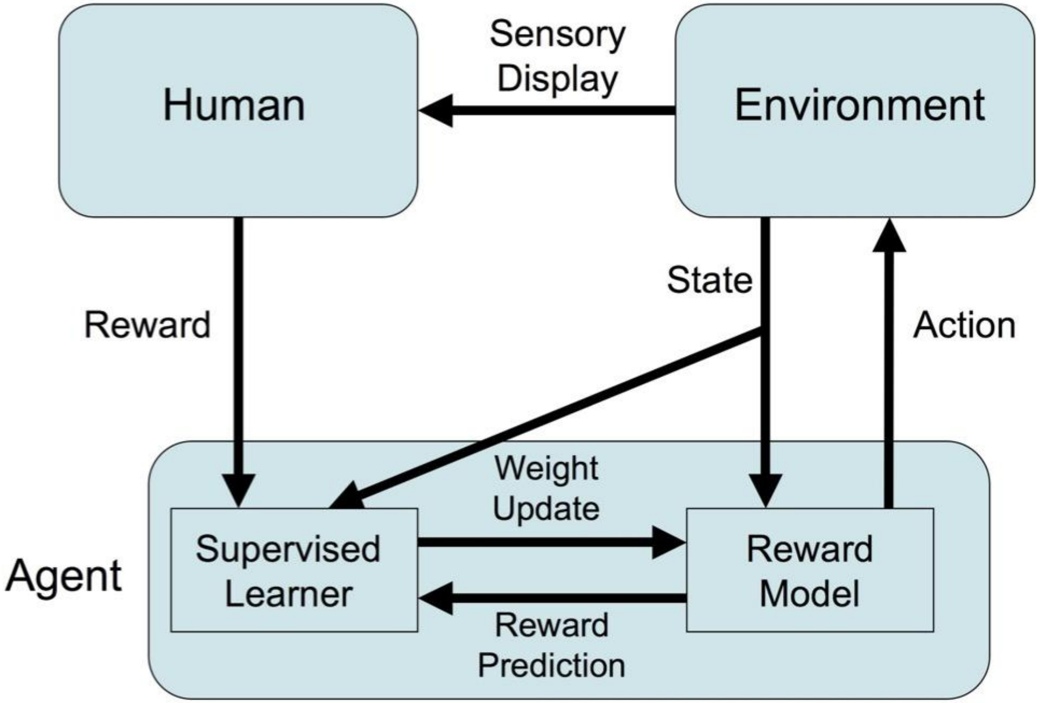


Fig. 2. Framework for Training an Agent Manually via Evaluative Reinforcement (TAMER).

5. TAMER: Training an Agent Manually via Evaluative Feedback

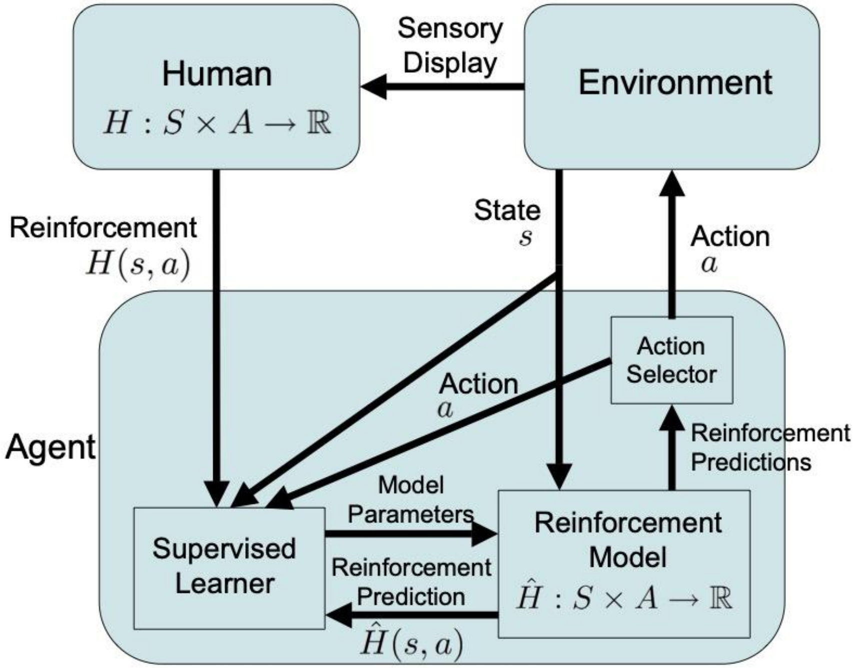
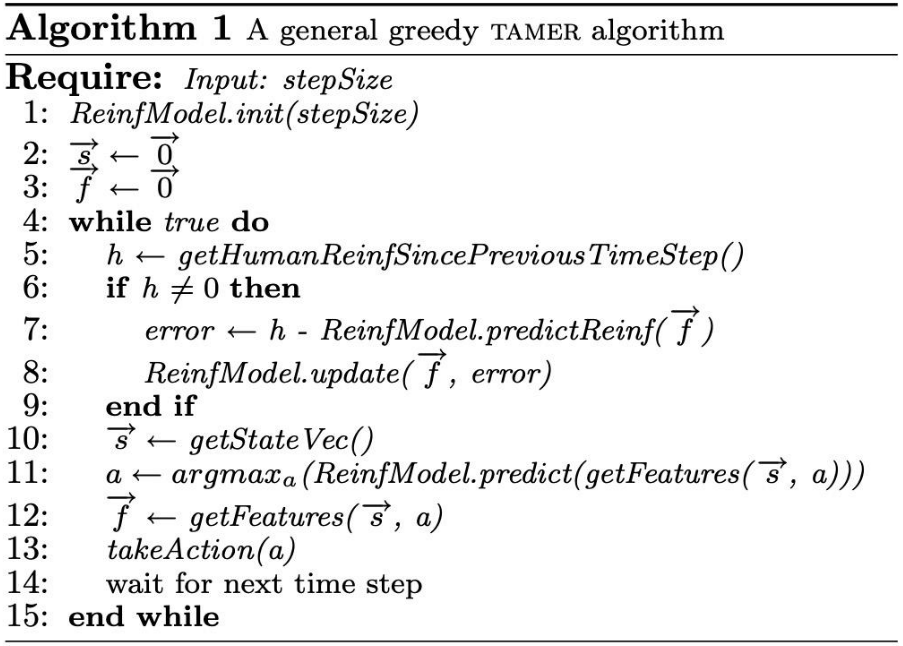


Figure 1: Framework for Training an Agent Manually via Evaluative Reinforcement (TAMER).

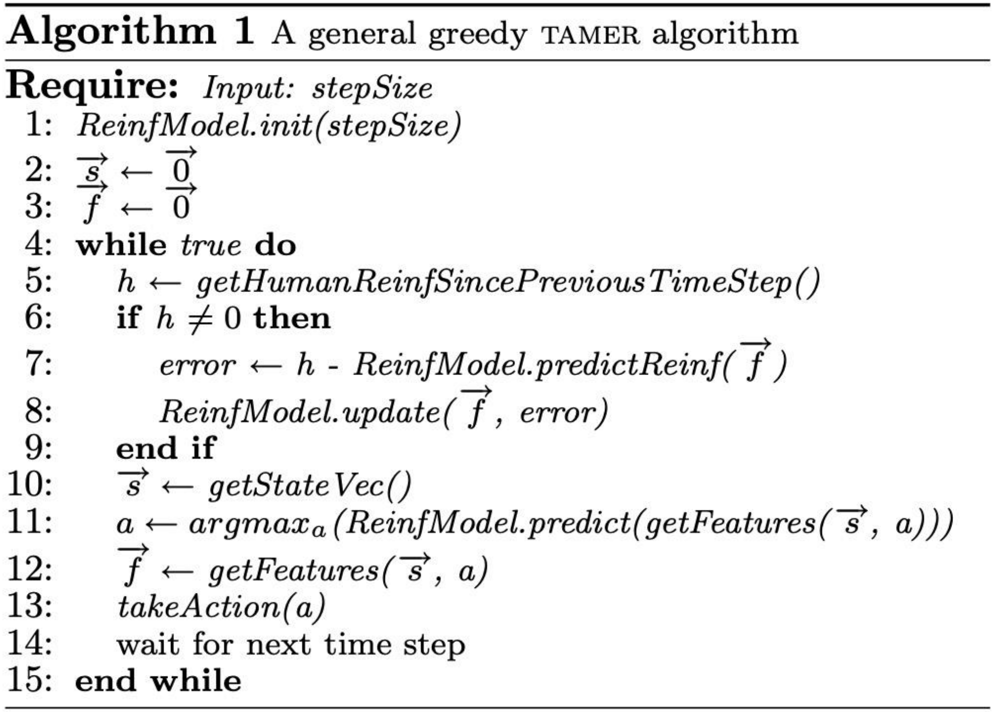


Algorithm 1 A general greedy TAMER algorithm

Require: *Input: stepSize*

```
ReinfModel.init(stepSize)
$\vec{s} \leftarrow \vec{0}$
$\vec{f} \leftarrow \vec{0}$
while true do
    $h \leftarrow$ getHumanReinfSincePreviousTimeStep()
    if $h \neq 0$ then
        $error \leftarrow h$ - ReinfModel.predictReinf($\vec{f}$)
        ReinfModel.update($\vec{f}$, error)
    end if
    $\vec{s} \leftarrow$ getStateVec()
    $a \leftarrow argmax_a$(ReinfModel.predict(getFeatures($\vec{s}$, a)))
    $\vec{f} \leftarrow$ getFeatures($\vec{s}$, a)
    takeAction(a)
    wait for next time step
end while
```

5. TAMER: Training an Agent Manually via Evaluative Feedback



Algorithm 1 A general greedy TAMER algorithm

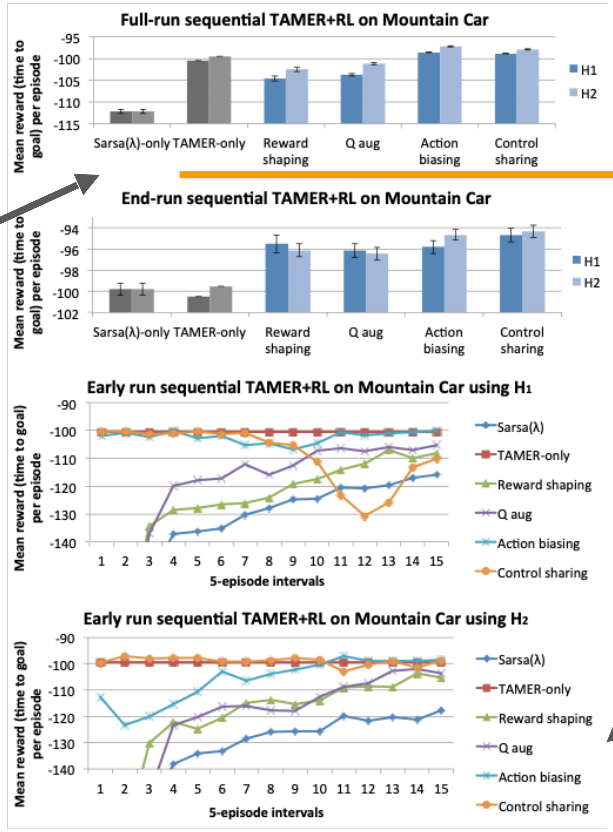
```
Require: Input: stepSize
ReinfModel.init(stepSize)
$\vec{s} \leftarrow \vec{0}$
$\vec{f} \leftarrow \vec{0}$
while true do
    $h \leftarrow getHumanReinfSincePreviousTimeStep()$
    if $h \neq 0$ then
        $error \leftarrow h - ReinfModel.predictReinf(\vec{f})$
        $ReinfModel.update(\vec{f}, error)$
    end if
    $\vec{s} \leftarrow getStateVec()$
    $a \leftarrow argmax_a(ReinfModel.predict(getFeatures(\vec{s}, a)))$
    $\vec{f} \leftarrow getFeatures(\vec{s}, a)$
    takeAction(a)
    wait for next time step
end while
```

- **Initialization**:
  - Initialize the reinforcement model (ReinfModel).
  - Initialize the state vector (s).
  - Initialize the feature vector (f).
- **Time Step Loop**:
  - Enter a loop that iterates once per time step.
  - Obtain a scalar measurement of the human trainer's reinforcement (h).
  - If h is nonzero:
    - Calculate the error as the difference between h and the predicted reinforcement from the agent's model.
    - Update the reinforcement model using the calculated error and the previous feature vector.
  - Obtain the new state description (s_new).
  - Choose an action (a):
    - If h is nonzero, greedily select the action that maximizes the predicted reinforcement according to the human reinforcement model.
    - If h is zero, choose a default action.
  - Calculate the feature vector for the new state-action pair (f_new).
  - Take the chosen action.
  - Update the state vector (s) and feature vector (f) for the next time step.

5. TAMER: Training an Agent Manually via Evaluative Feedback

Agents behavior can be **shaped through signals of approval and disapproval**, a natural form of human feedback.

Various approaches from reward shaping aiming to augment a **traditional reinforcement learning (MDP) agent with human feedback**



RL

Different evaluative feedback approaches

Two different Humans

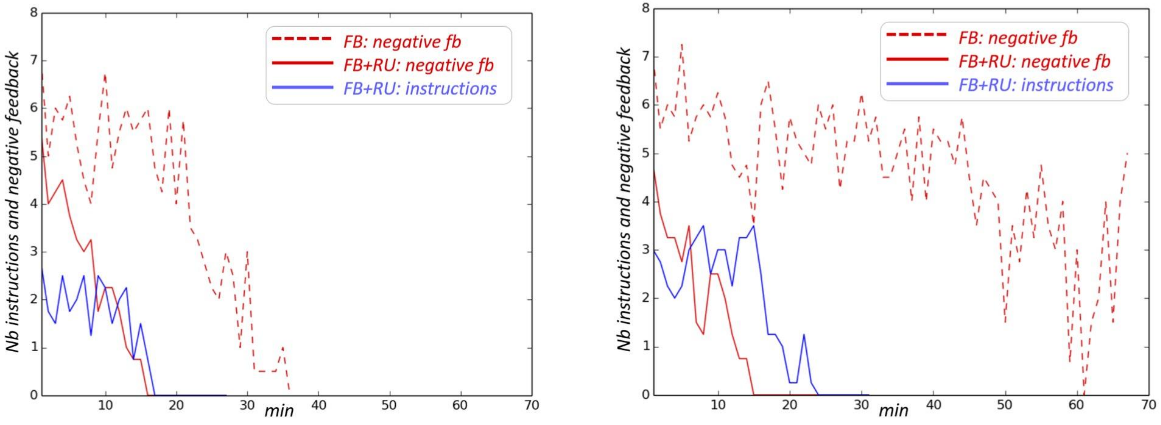
Figure 1: Comparison of TAMER+RL techniques with SARSA(λ) and the TAMER-only policy on mountain car over 40 or more runs of 500 episodes. $\hat{H}_1$ and $\hat{H}_2$ are models from two different human trainers. The top chart considers reward over the entire run, and the second chart evaluates reward over the final 10 episodes. Error bars show standard error. The third and fourth charts display mean performance using $\hat{H}_1$ and $\hat{H}_2$ early in the run, during the first 75 episodes.

W. Bradley Knox and Peter Stone. 2012. Reinforcement learning from simultaneous human and MDP reward. In Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1 (AAMAS '12). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 475–482.

6. Challenges & Limitations of RLHF

Amount of feedback

Obtaining a sufficient amount of high-quality feedback can be challenging, particularly in scenarios where human evaluators are scarce or the task complexity requires extensive feedback. Limited feedback may hinder the learning process and result in slower convergence or suboptimal performance.



(a) Small state space.

(b) Large state space.

Figure 7: Number of instructions (blue) and negative feedback (red) over time.

. Najar, A., Sigaud, O. & Chetouani, M. Interactively shaping robot behaviour with unlabeled human instructions. Auton Agent Multi-Agent Syst 34, 35 (2020). https://doi.org/10.1007/s10458-020-09459-6

6. Challenges & Limitations of RLHF

Feedback distribution

The distribution of feedback across different states and actions may be uneven, leading to biases in the learned policies. Unequal feedback distribution can result in overfitting to certain states or actions while neglecting others.

This leads to the distributional shift problem

In addition, Positive and Negative Feedback are not necessarily balanced.

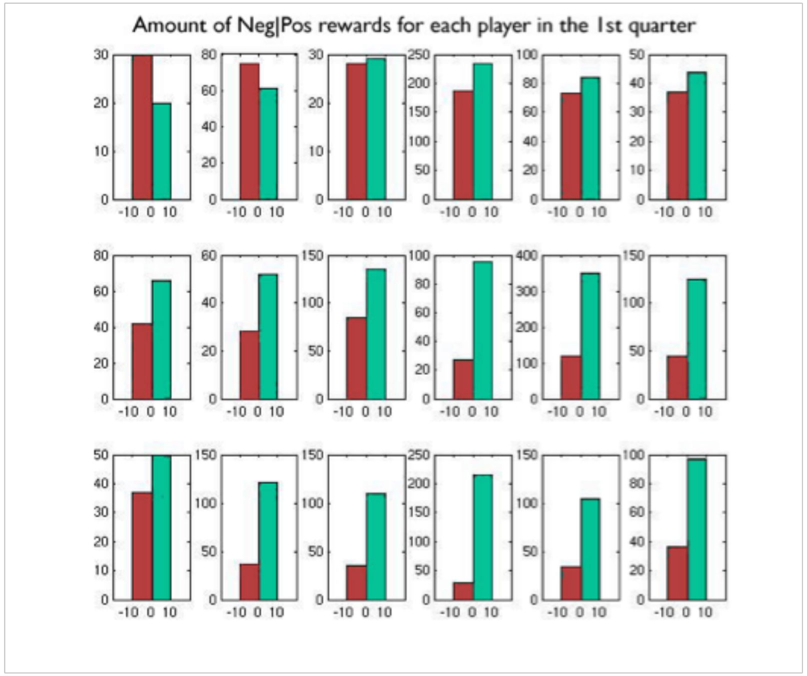


Fig. 3. Histograms of rewards for each individual in the first quarter of their session. The left column is negative rewards and the right is positive rewards. Most people even in the first quarter of training have a much higher bar on the right.

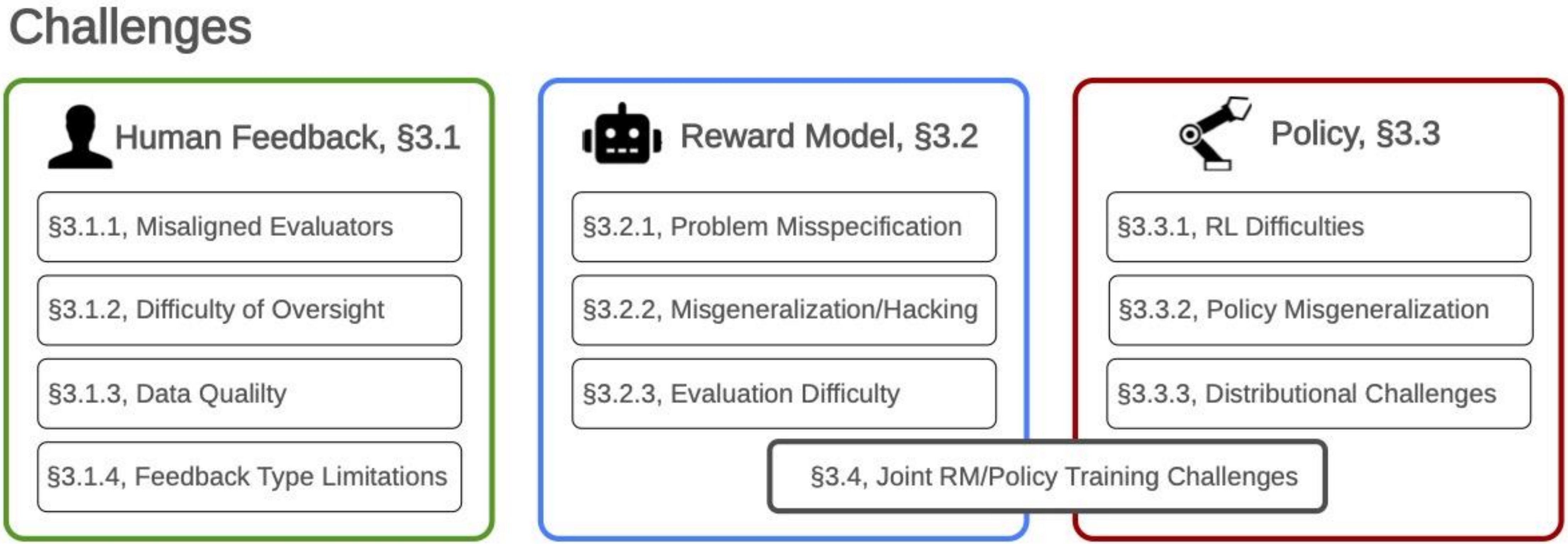
Thomaz, A.L.; Breazeal, C. Asymmetric Interpretations of Positive and Negative Human Feedback for a Social Learning Agent. In Proceedings of the 16th IEEE International Symposium Robot and Human Interactive Communication (RO-MAN 2007), Jeju, Korea, 26–29 August 2007; pp. 720–725, doi:10.1109/ROMAN.2007.4415180.

6. Challenges & Limitations of RLHF

Reward hacking

Reward hacking refers to a phenomenon where the agent achieves more reward than intended by exploiting loopholes in the reward determination process. This occurs when the agent introduces unforeseen actions to garner positive human feedback, deviating from the intended behavior or objectives of the task. Essentially, reward hacking allows the agent to manipulate the reward system to its advantage, potentially leading to undesired outcomes or undermining the integrity of the learning process.

6. Challenges & Limitations of RHLF


Challenges

Human Feedback, §3.1

§3.1.1, Misaligned Evaluators

§3.1.2, Difficulty of Oversight

§3.1.3, Data Qualilty

§3.1.4, Feedback Type Limitations

Reward Model, §3.2

§3.2.1, Problem Misspecification

§3.2.2, Misgeneralization/Hacking

§3.2.3, Evaluation Difficulty

Policy, §3.3

§3.3.1, RL Difficulties

§3.3.2, Policy Misgeneralization

§3.3.3, Distributional Challenges

§3.4, Joint RM/Policy Training Challenges

HUMANE AI NET

SORBONNE
UNIVERSITÉ

VU UNIVERSITY
AMSTERDAM