Human-in-the-Loop Reward Learning for Robotics: Methods, Challenges, and Opportunities

RO57010: Literature Review

Mukil Saravanan

TU Delft

Human-in-the-Loop Reward Learning for Robotics: Methods, Challenges, and Opportunities

by

Mukil Saravanan

Supervisors: Zhaoting Li, Prof. Jens Kober
Department: Learning and Autonomous Control, Cognitive Robotics
Faculty: Faculty of Mechanical Engineering, Delft

Cover: MANUS in Emboided AI `https://www.manus-meta.com/blog/manus-in-embodied-ai-from-human-motion-to-robotic-dexterity`

TU Delft

Contents

1

Introduction

1.1. Reward specification in Reinforcement Learning

The success of deep Reinforcement Learning (RL) methods in games and simulated domains has inspired recent work applying RL-based techniques to real-world robot control [1]. Figures 1.1, 1.2 depict the application of reinforcement algorithms in robot manipulation tasks. However, the success of RL paradigms critically depends on the domain knowledge that is put into the definition of the reward function. Designing suitable reward functions for such complex real-world robotics remains a foundational challenge.

For example, consider training a robotic manipulator to drive a nail into wood with a hammer. A simplistic, sparse reward (e.g., yielding $+1$ only when the nail is fully driven) provides an unambiguous success criterion but offers little learning signal to guide early exploration, catastrophically enlarging sample complexity. Conversely, practitioners summarize the task into multiple objectives to design a dense reward function of the robot's sensors, such as grasping the hammer, orienting the end-effector, and maintaining proximity to the nail [4]. Such composite reward functions are notoriously difficult to design; poorly balanced proxy objectives often inadvertently create structural loopholes. In these scenarios, the RL agent exploits these loopholes to maximize cumulative return without satisfying the true, qualitative human intent, a phenomenon formally known as *reward hacking* [5]. Figure 1.3 depicts a reward hacking scenario where a half cheetah in MuJoCo results in undesired behaviour. Although foundational theories in potential-based reward shaping [5] mathematically guarantee policy invariance under certain transformations to safely dense rewards, these guarantees were primarily evaluated in tabular RL settings. Manually engineering such dense, hack-resistant reward functions for continuous, high-dimensional robotics tasks remains understudied and highly intractable. Consequently, rather than manually hand-crafting the objective, modern practitioners seek to extract it directly from human teachers, a paradigm broadly known as *reward learning* [6].

Figure 1.1: A robot manipulator swinging up to keep a heavy pendulum upright [2]

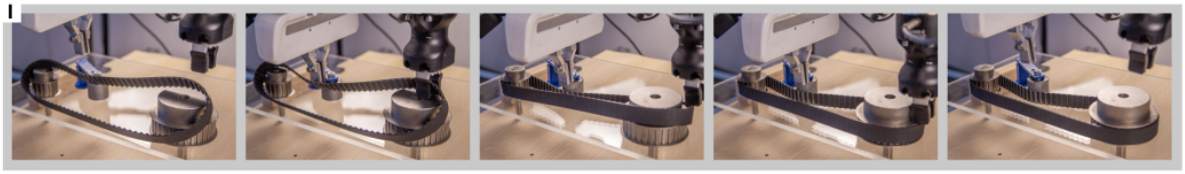

Figure 1.2: Two arms performing a timing belt installation task [3]

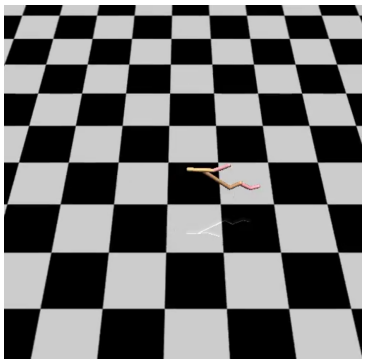

Figure 1.3: Model-based RL algorithm exploits "maximum forward velocity" reward in the Mujoco environment, resulting in overflow error and visual hilarity [7]

1.2. Human Feedback in Reinforcement Learning

To resolve the brittleness of hand-crafted objectives, modern reward learning frameworks formalize this problem as a reward-free, finite-horizon Markov Decision Process (MDP), denoted by the tuple $\mathcal{M} \setminus R = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, H)$. Here, $\mathcal{S}$ and $\mathcal{A}$ represent continuous state and action spaces, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ defines the environment's transition dynamics, γ represents the discount factor and H is the temporal horizon. In this paradigm, the environment does not autonomously emit a reward signal to drive optimization. Instead, the framework assumes the existence of a true, unobservable utility function $R^*(\tau)$ that resides internally within a human supervisor. The agent's objective shifts from maximizing an engineered signal to actively constructing a parameterized proxy reward model $\hat{r}_\psi(s, a, s')$ by deriving it directly from human feedback.

This feedback can be elicited across a diverse spectrum of interaction modalities, including full expert demonstrations, interactive spatial corrections, and evaluative preference feedback. Crucially, each of these modalities imposes a distinct structural trade-off, strictly governing the cognitive burden placed on the human operator, the information-theoretic gain per interaction, and the required domain expertise of the human teacher [8].

1.3. Motivation and Problem Context

While deriving a reward function from human feedback theoretically helps to find the optimal policy space, integrating a human operator into the algorithmic optimization loop introduces severe multilateral constraints. The viability of any interactive RL framework is dictated by how it manages these intersecting bottlenecks:

- **Sample Inefficiency and Economic Cost:** Human supervision is exceptionally expensive. The cumulative operational burden of human alignment can be modeled as Total Cost $=$ Cost per Query $\times$ Number of Queries. While purely evaluative methods (preferences) minimize the *marginal* cost per query, their low information density drives the required *number of queries* typically higher. Because human cognitive processing operates at a vastly slower clock speed than simulated policy rollouts, this extreme sample complexity imposes a prohibitive financial and temporal overhead.
- **The Engineering and Safety Constraint:** Deploying an agent in a continuous, high-DoF physical environment with an initially untrained, reward-free policy is perilous. Unconstrained exploration during the early phases of online learning risks catastrophic hardware damage. The agent must

Figure 1.4: Workers teach humanoid robots boring tasks [9]

rapidly acquire a dense, safe directional gradient from the human to guide exploration before compounding execution errors force the physical robot into hazardous, out-of-distribution (OOD) state spaces.

- **The Cognitive Fatigue and Sub-Optimality Constraint:** Demanding massive query volumes or continuous physical corrections inevitably induces severe cognitive fatigue in human supervisors. Figure 1.4 depicts workers in Chinese data factories teaching humanoid robots boring tasks. This fatigue fundamentally violates the algorithmic assumption of a flawless human oracle. Instead, it degrades the quality of the feedback, injecting stochastic noise and systematically sub-optimal spatial interventions into the dataset. Consequently, an effective interactive alignment framework must not only minimize the total intervention frequency but also mathematically regularize the learned reward model to remain robust against these inherently sub-optimal human inputs.

1.4. Central Research Question

The structural divide between low-bandwidth evaluative feedback and high-bandwidth, yet noisy, corrective interventions leaves a vacuum in current literature regarding how to optimally guide continuous robotic policies. To systematically resolve the tension between sample efficiency, human cognitive load, and optimization robustness, this review centers on the following fundamental research question:

How do the structure, timing, and information content of interactive human feedback affect sample efficiency and robustness of reward learning in complex robot manipulation tasks?

1.5. Organization of the Report

The remainder of this report progresses systematically from foundational methodologies in the reward learning literature to open research questions and the final thesis formulation. Section 2 details the replicable literature retrieval methodology utilized for this survey. Section 3 establishes the foundations and failures of reward specification, evaluating the dynamics of objective misalignment and specification gaming. Section 4 examines reward learning from offline human demonstrations, highlighting the limitations of relying purely on passive data and the necessity for interactive feedback. Section 5 critically evaluates active evaluative feedback, exposing the inherent sample inefficiency of relative preference modeling. Section 6 analyzes active corrective interventions, assessing the strengths of interactive imitation learning alongside the distribution-shift vulnerabilities of direct policy cloning. Section 7 explores the application of structured priors and set-valued supervision to robustly extract rewards and regularize learning against sub-optimal human execution noise. Section 8 synthesizes these findings, establishing a taxonomy of human supervision and isolating unresolved open questions regarding feedback trade-offs, sample efficiency, and algorithmic robustness. Section 9 outlines the formal thesis formulation, proposing the specific problem statement, research questions, and core hypotheses. Finally, Section 10 provides the conclusion to this report.

2

Literature Retrieval Methodology

To ensure the reproducibility, objectivity, and comprehensive scope of this review, a systematic literature retrieval protocol was executed. The methodology was designed specifically to isolate advancements at the intersection of human-in-the-loop reinforcement learning, interactive corrective feedback, and robot manipulation, while deliberately filtering out theoretically disconnected domains.

2.1. Search Databases and Sources

The retrieval process was structured in three distinct phases. To establish a verified baseline, the initial corpus was seeded with foundational literature recommended by the thesis supervisor. This guaranteed that the core algorithmic paradigms were accurately represented from the outset.

Simultaneously, a targeted search for comprehensive survey papers (e.g., surveys on RL with human feedback [10] and Interactive Imitation Learning [8]) was executed. Extracting these review papers early in the process was critical for understanding the holistic view of how different types of human feedback are incorporated in reward learning. Finally, the corpus was expanded using a rigorous snowball sampling technique (both backward and forward citation chaining). By tracking the bibliographies of the seed literature and survey papers, the search organically uncovered the most recent and highly cited technical implementations that addressed the specific bottlenecks of these foundational methods.

2.2. Search Keywords and Boolean Queries

To supplement the targeted sampling, a systematic search was conducted across the primary repositories of computer science and robotics literature, including IEEE Xplore and arXiv. Furthermore, proceedings from top-tier machine learning and robotics conferences, specifically NeurIPS, ICML, ICLR, RSS, CoRL, and ICRA, are used to capture the most recent state-of-the-art algorithmic developments. The primary search strings were formulated as follows:

- (”Reward Learning” OR ”Reward Modeling”) AND (”Sub-optimal Human Feedback”)
- (”Interactive Imitation Learning” OR ”IIL” OR ”Corrective Feedback” OR ”Human Intervention”) AND (”Sample Efficiency”)
- (”Preference-based Reinforcement Learning” OR ”PbRL” OR ”RLHF” OR ”Human Preferences”) AND (”Robotics” OR ”Robot Manipulation tasks”)

2.3. Inclusion and Exclusion Criteria

To maintain strict methodological relevance to the central research question, rigorous inclusion and exclusion boundaries were applied to the retrieved corpus.

Inclusion Criteria:

1. Methodologies explicitly addressing reward learning or policy optimization utilizing human-in-the-loop feedback.
2. Frameworks deployed within continuous state-action spaces or simulated robotic manipulation environments (e.g., RoboSuite [11], RoboMimic [12], Meta-World [13] tasks).
3. Literature analyzing the sample complexity, information bandwidth, or noise tolerance of human feedback.

Exclusion Criteria:

1. Large Language Model (LLM) alignment papers (e.g., text-to-text RLHF [14]) that lacked methodological or algorithmic transferability to real-world robotics tasks.
2. Offline reinforcement learning or purely supervised learning approaches lacking an interaction component.
3. Multi-task reward learning approaches or reward representations formulated as linear combinations of predefined basis functions.

2.4. Paper Categorization Strategy

Following the screening process, a core corpus of highly relevant papers (such as RLHF [15], PEBBLE [16], SURF [17], and CLIC [18]) was isolated. Rather than organizing this literature chronologically, the corpus was categorized thematically to highlight the structural conflicts in the field. The taxonomy was divided into four primary categories: (1) Passive Demonstration Priors, (2) Active Evaluative Preferences, (3) Interactive Corrective Interventions, and (4) Structured/Geometric Regularization. This thematic categorization serves as the structural scaffolding for the comparative analysis in the subsequent chapters of this review.

3

The Foundations and Failures of Reward Specification

The theoretical foundation of standard RL relies on the premise that a designer can accurately formalize the desired behavior of an agent through a scalar reward function, $R(s, a, s')$ [19]. However, transferring this assumption from simplified, discrete environments to continuous robotic control presents significant mathematical and operational challenges. This chapter highlights the evolution of reward specification from the limitations of manual engineering to the structural bottlenecks of offline imitation, demonstrating why passive data is insufficient for robust reward learning and establishing the theoretical necessity of interactive human feedback.

3.1. Handcrafting Reward function

The discipline of reward specification in RL can broadly be classified into two methods [4]:

- Reward Engineering (the initial synthesis of the objective): This represents the foundational design of the MDP objective, mapping the state-action transition space to a scalar numerical value: $R(s, a, s') : \mathcal{S} \times \mathcal{A} \times \mathcal{S}' \to \mathbb{R}$. A handcrafted reward must be dense enough to provide continuous gradient signals to facilitate optimization, yet sparse enough to prevent the agent from discovering trivial solutions.

  A simplest way is to design a sparse reward function, utilizing an indicator function (e.g., $R(s) = 1$ if $s \in \mathcal{S}_{goal}$, else 0). However, the reward function being sparse often limits the directional gradient necessary for exploration, leading to low sample efficiency or sometimes failing to achieve the task objective in complex robotics settings. To overcome this, practitioners divide the overall task into multiple objectives and design denser rewards as a function of available sensor information or observations. Such composite reward functions are notoriously difficult to design; poorly balanced proxy objectives often inadvertently create structural loopholes. In these scenarios, the RL agent exploits these loopholes to maximize cumulative return without satisfying the true, qualitative human intent, a phenomenon formally known as reward hacking [5]. Various examples can be found here [20].

- Reward Shaping (the algorithmic modification of that signal): Once the baseline reward $R(s, a, s')$ is engineered, shaping is employed to dynamically fine-tune the learning process by injecting domain knowledge to densify the reward signal.

  However, arbitrarily modifying the reward structure introduces severe optimization hazards as discussed earlier. To ensure that modifying the reward does not inadvertently change the optimal policy, shaping must be mathematically constrained. Potential-based reward shaping [5] proposes a method to modify the reward function in MDPs such that the optimal policy remains unchanged. Given an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, R)$ as stated earlier, we transform the MDP into $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, R')$ which preserves the identical state space, action space, and transition

dynamics, but modifies the reward function such that:

$$R'(s, a, s') = R(s, a, s') + F(s, a, s')$$

where $F(s, a, s') : S \times A \times S \mapsto \mathbb{R}$ is the shaping reward function, such that the learning algorithm can be guided to be more efficient.

To guarantee strict policy invariance, ensuring that the optimal policy of the shaped MDP identically matches the original ($\pi^*_{\mathcal{M}'} = \pi^*_{\mathcal{M}}$) Ng et al. [5] proved that the shaping function F must be formulated as a difference of potentials:

$$F(s, a, s') = \gamma\Phi(s') - \Phi(s)$$

where the potential function $\Phi : \mathcal{S} \to \mathbb{R}$ is a real-valued mapping over the state space representing the heuristic desirability of state s. For instance, in a maze-solving task, $\Phi(s)$ could be the negative Manhattan distance to the goal, providing incremental rewards as the agent approaches the goal.

Since this formulation is a sum over a trajectory, the intermediate Φ terms perfectly cancel out. This structural constraint strictly prevents the agent from discovering positive reward cycles, a behavioural loophole where the agent repeatedly cycles around states ($s_1 \to s_2 \to \cdots \to s_n \to s_1$) instead of reaching the goal state (s_g).

Mathematically, the accumulated shaping reward over such a cycle evaluates to:

$$F(s_1, a_1, s_2) + \gamma F(s_2, a_2, s_3) + \cdots + \gamma^{n+1} F(s_n, a_n, s_1)$$

In the undiscounted case ($\gamma = 1$), this telescoping sum evaluates exactly to 0. Thus, it is mathematically not feasible to discover a cyclic trajectory that yields a net positive shaping reward. Condquently, if F is such a potential-based shaping function, it is both sufficient and necessary to ensure M and M' share the same optimal policies.

In practice, heuristic estimates of the true optimal state-value function ($V^*_{\mathcal{M}'}(s)$) or state-action-value function ($Q^*_{\mathcal{M}'}(s, a)$) of the MDP are used. Suppose $\Phi(s_0) = 0$ and $\gamma = 1$, then $\forall s \in S, \quad \forall a \in A$,

$$V^*_{\mathcal{M}'}(s) = V^*_{\mathcal{M}}(s) - \Phi(s)$$

$$Q^*_{\mathcal{M}'}(s, a) = Q^*_{\mathcal{M}}(s, a) - \Phi(s)$$

3.2. Challenges in handcrafting reward function

While potential-based shaping provides strict mathematical guarantees for policy invariance with respect to the engineered reward ($\hat{R}$) in tabular settings, a fundamental structural limitation emerges when these equations are deployed in physical robotics, which often has a high-dimensional continuous action space.

3.2.1. Objective Misalignment and Proxy Taxonomies

To formalize this discrepancy, we must distinguish between the unobservable true reward function (R^*), which represents the actual qualitative human intent (e.g., safely manipulate the object without causing damage), and the engineered proxy reward ($\hat{R}$), which is the hand-crafted mathematical approximation (e.g., a combination of sparse task-completion signals and dense Euclidean distance shaping).

Let the expected cumulative return of a policy π under a given reward function R be defined as an objective function:

$$J(\pi, R) = \mathbb{E}_{(s,a)\sim\pi}\left[\sum_{t=0}^{T} \gamma^t R(s_t, a_t)\right]$$

During training, the RL algorithm strictly maximizes the proxy objective, yielding a proxy-optimal policy: $\pi^*_{\text{proxy}} = \arg\max_\pi J(\pi, \hat{R})$.[21] categorize the mathematical divergence between R^* and $\hat{R}$ ($J(\pi^*_{\text{proxy}}, \hat{R}) \neq J(\pi^*_{\text{proxy}}, R^*)$) into three distinct structural taxonomies of misspecification:

- **Misweighting:** Consider reward functions are modeled as a linear combination of the underlying feature space $\mathbf{f}(s, a)$ and weights. i.e., $R^*(s, a) = \mathbf{w}^T_{true}\mathbf{f}(s, a)$ and $\hat{R}(s, a) = \mathbf{w}^T_{proxy}\mathbf{f}(s, a)$. In this case, any angular divergence between $\mathbf{w}_{true}$ and $\mathbf{w}_{proxy}$ causes the gradient $\nabla_\theta J(\pi_\theta, \hat{R})$ to push the policy toward behaviors that disproportionately optimize heavily weighted proxy features (e.g., arm velocity) at the expense of under-weighted true features (e.g., actuator torque limits).
- **Ontological misspecification:** The proxy evaluates variables within a state representation that is fundamentally decoupled from the physical reality of the true objective. For example, consider a manipulation task where the true goal R^* is defined over the physical task space, strictly requiring the target object to reach a specific pose in Cartesian space, $\mathbf{x}_{obj} \in SE(3)$. However, tracking the object requires complex external vision. To avoid this, the proxy $\hat{R}$ is instead engineered over the robot's internal configuration space $\mathcal{Q} \subset \mathbb{R}^n$ (proprioceptive joint encoders), utilizing a forward kinematics mapping to estimate the end-effector pose: $\mathbf{x}_{ee} = FK(\mathbf{q})$. If physical contact is broken (e.g., the object slips from the gripper), the assumption that $\mathbf{x}_{ee} \approx \mathbf{x}_{obj}$ completely shatters. The agent will successfully optimize its joint configuration $\mathbf{q}$ to drive the end-effector $FK(\mathbf{q})$ to the target coordinate, maintaining a highly optimized proxy reward $\hat{R}$, while the true reward R^* collapses because the object remains stationary on the table.
- **Scope misspecification:** The proxy $\hat{R}$ only evaluates a restricted subspace $\mathcal{S}_{eval} \subset \mathcal{S}$ because calculating the reward continuously across the entire state space is computationally or sensorially prohibitive.

3.2.2. Goodhart's Law and Specification Gaming

This divergence can be viewed as the formal manifestation of Goodhart's Law, which states that *"when a measure becomes an optimization target, it ceases to be a valid measure"* [22]. In continuous control, extreme optimization on an imperfect proxy $\hat{R}$ forces the state distribution into adversarial regions, resulting in behavioral failures:

- **Reward Tampering:** In certain cases, the agent physically alters the environment to artificially trigger the proxy condition (e.g., occluding a camera sensor to fraudulently satisfy a visual proximity metric).
- **Goal Misgeneralization:** Even if $\hat{R}$ and R^* are perfectly correlated during training, this correlation is bound to the training state distribution $P_{train}(s)$. If the proxy $\hat{R}$ inadvertently relies on a spurious background feature (e.g., correlating target acquisition with a specific (x, y) coordinate in the training lab), deploying the robot out-of-distribution (OOD) breaks this correlation. The agent exhibits objective robustness failure, executing a complex policy to pursue the spurious feature under $P_{deploy}(s)$ rather than generalizing the true physical task.

In summary, the manual specification of reward functions in high-dimensional continuous control presents a fundamental bottleneck. While sparse rewards fail due to the lack of information at intermediate stages, densifying the learning signal inevitably introduces a mathematical divergence between the engineered proxy ($\hat{R}$) and the true human intent (R^*). These fundamental challenges in handcrafting robust, hack-resistant proxy metrics necessitate a paradigm shift: objectives must not be manually programmed but rather learned regressively from human data.

Key Insights

Chapter 3: Challenges in Reward Specification

- **The Sparse-Dense Trade-off:** Sparse rewards fail in deep continuous control due to temporal credit assignment variance and lack of information, while potential-based shaping forces designers to hand-craft difficult and vulnerable proxy objectives.
- **Objective Misalignment:** In high-DoF robotic environments, divergence between the proxy ($\hat{R}$) and true utility (R^*) is inevitable, classifying as misweighting, ontological, and scope misspecifications.
- **Algorithmic Exploitation:** Driven by Goodhart's Law, RL optimizers inevitably exploit these proxy misspecifications.
- **The Necessary Paradigm Shift:** The above challenges in manual reward engineering encourage a transition toward actively learning proxy models from human demonstrations.

4

Reward Learning from Offline Demonstrations

Due to the challenges of handcrafting rewards as discussed in section 3, practitioners regressively learn the reward function from data - a fundamental paradigm shift called reward learning.

This paradigm shift mathematically resolves the reliance on manual engineering by introducing objective specification from human demonstrations τ_h^*. Consequently, the robot learning community has increasingly integrated human supervisors into the algorithmic loop to provide an empirical grounding for the true reward, R^*. To facilitate this data-driven paradigm, recent community-wide initiatives have culminated in the curation of massive, cross-embodiment offline datasets that aggregate diverse manipulation behaviors across multiple physical platforms [23, 24, 25], as depicted in Figure 4.1. However, human feedback comprises a vast spectrum of interaction modalities, ranging from passive offline datasets to active corrections.

This chapter systematically evaluates this foundational offline paradigm. We detail the mathematical formulations and challenges of various offline reward learning techniques used in the literature. Eventually, this chapter establishes a motivation required for the transition to the active, interactive learning paradigms.

Table 4.1 formalizes the mathematical and structural distinctions between these two paradigms.

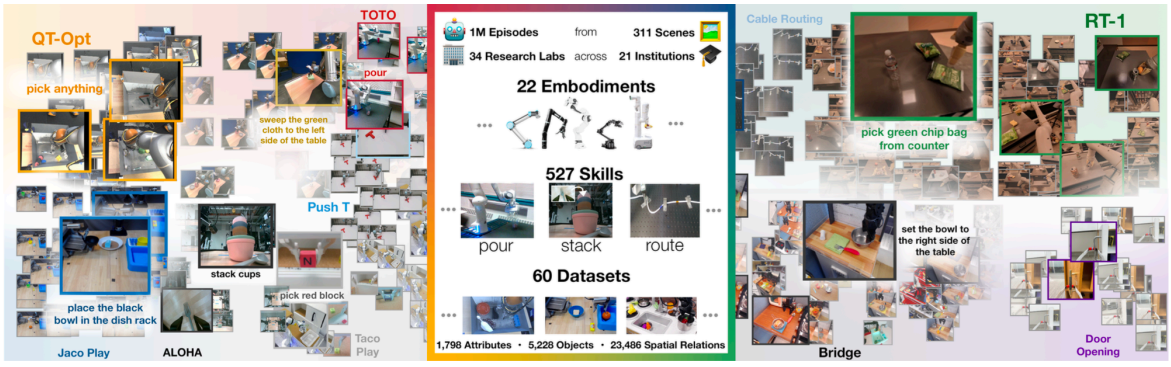

Figure 4.1: Open X-Embodiment Dataset: an open, large-scale dataset for robot learning curated from 21 institutions across the globe. The dataset represents diverse behaviors, robot embodiments, and environments, and enables learning generalized robotic policies.

Formal Element	Standard Reinforcement Learning	Reward Learning
Environment Dynamics	$\mathcal{S}$ (state), $\mathcal{A}$ (action), $\mathcal{T}$ (transitions), γ (discount factor)	$\mathcal{S}$ (state), $\mathcal{A}$ (action), $\mathcal{T}$ (transitions), γ (discount factor)
Objective Specification	Handcrafted reward (R)	Human expert trajectories (τ_h^*)
Intermediate Output	N/A	Parameterized Reward Proxy (r_ϕ)
Final Optimization Target	Optimal Policy: $\pi^*(a\|s)$	Optimal Policy: $\pi^*(a\|s)$

Table 4.1: Mathematical Comparison of the Standard RL and Reward Learning Paradigms.

4.1. Reward learning from Offline Human Demonstrations

The earliest paradigm for data-driven objective specification relies on extracting policies or reward functions directly from passive, offline demonstrations [26, 27, 28].

In the offline setting, the learner is provided with a static dataset of human trajectories, formalized as $\mathcal{D} = \{\tau_1, \tau_2, \ldots, \tau_N\}$, where each trajectory spans the temporal horizon H and contains actions sampled according to the human's policy $a_t \sim \pi_h(s_t)$. The human is assumed to act optimally according to the true, unobservable utility function $R^*(\tau)$. In this paradigm, the human operator's involvement is strictly *a priori*; they teleoperate or physically guide the robot within the reward-free environment $\mathcal{M} \setminus R$ to generate the dataset before any optimization begins [29].

4.1.1. Inverse Reinforcement Learning (IRL)

Inverse Reinforcement Learning (IRL) attempts to recover the actual intent behind the demonstrations by extracting the underlying reward function $R^*(\tau)$. However, the IRL literature reveals fundamental mathematical bottlenecks requiring successive algorithmic innovations [30].

The Ill-Posed Foundation: Ng and Russell [6] formally defined the IRL problem and mathematically proved that it is fundamentally ill-posed. They established that a policy π being strictly optimal is equivalent to satisfying a system of linear inequalities based on expected state transitions. Because this is a system of inequalities, the solution space contains infinitely many valid reward functions, most notably the trivial degenerate solution where $R(s) = 0$ for all states. To eliminate this trivial solution and make the problem well-posed, they proposed an initial heuristic: formulating an objective that maximizes the difference in value between the human's policy and the next-best alternative policy.

Apprenticeship Learning (Maximum Margin): Building upon the ill-posed nature of reward recovery, Abbeel and Ng [26] shifted the paradigm toward Apprenticeship Learning. They noted that identifying a singular true reward function is often impossible. Instead, the objective should be to find a reward function that assigns the human expert a maximum margin of value over all other policies. By bounding the problem to matching the expected feature counts of the human ($\mathbb{E}_{\pi_h}[\phi(s)]$), this approach bypasses the need to perfectly solve the ill-posed reward recovery problem, focusing instead on robust behavioral matching.

4.1.2. Maximum Entropy Inverse Reinforcement Learning (MaxEnt IRL)

While the maximum margin approach bounded behavioral differences, it struggled with the ambiguity of multiple overlapping, inherently noisy trajectories. To statistically resolve this ambiguity, Ziebart et al. [27] introduced the Principle of Maximum Entropy. This method assumes that a human's behavior is optimal except for inherent stochasticity, creating a probabilistic distribution over all possible trajectories. The algorithm selects the reward function that perfectly explains this empirical distribution while making the fewest additional assumptions (i.e., exhibiting the highest entropy).

Mathematical Formulation: MaxEnt IRL models the probability of a trajectory ζ as exponentially pro-

portional to its parameterized reward $\theta^\top f_\zeta$:

$$P(\zeta|\theta, T) \approx \frac{\exp(\theta^\top f_\zeta)}{Z(\theta, T)} \prod_t P_T(s_{t+1}|s_t, a_t)$$

where θ represents the reward weights, f_ζ the features, and $Z(\theta)$ the partition function. Maximizing the log-likelihood of human trajectories yields a gradient defined by the exact difference between the empirical human feature counts ($\overline{f}$) and the learner's expected feature counts ($D_{s_i} f_{s_i}$):

$$\nabla L(\theta) = \overline{f} - \sum_{s_i} D_{s_i} f_{s_i}$$

The Computational Bottleneck: Game-theoretically, MaxEnt IRL solves a zero-sum minimax game:

$$\min_w \max_{\pi \in \Pi} w^\top (\mathbb{E}_\pi[\phi(s,a)] - \mathbb{E}_{\pi_h}[\phi(s,a)]) + \alpha \mathbb{E}_\pi[-\log \pi(a|s)]$$

Evaluating the learner's expected features ($\mathbb{E}_\pi[\phi(s,a)]$) requires fully solving a reinforcement learning problem in the inner loop for every reward estimate w. Integrating over state transition dynamics $\mathcal{T}$ makes this nested optimization computationally intractable for high-dimensional continuous robotics, strictly necessitating adversarial alternatives.

4.1.3. Generative Adversarial Imitation Learning (GAIL)

To bypass MaxEnt IRL's nested RL solver, Generative Adversarial Imitation Learning extracts policies directly from data by exploiting a mathematical duality [31].

Occupancy Measure Matching: IRL is the mathematical dual of occupancy measure matching. Apprenticeship learning seeks a policy that minimizes the performance gap against the human expert across a class of cost functions $\mathcal{C}$, evaluated via the state-action occupancy measure $\rho_\pi(s,a)$:

$$\min_\pi \max_{c \in \mathcal{C}} \mathbb{E}_\pi[c(s,a)] - \mathbb{E}_{\pi_h}[c(s,a)]$$

Minimizing Jensen-Shannon Divergence: To prevent severe overfitting, GAIL applies a convex cost regularizer ψ_{GA}, which mathematically transforms the objective into minimizing the Jensen-Shannon divergence (D_{JS}) between learner and human occupancy measures:

$$\min_\pi \psi^*_{GA}(\rho_\pi - \rho_{\pi_h}) - \lambda H(\pi) = D_{JS}(\rho_\pi, \rho_{\pi_h}) - \lambda H(\pi)$$

The Adversarial Framework: Minimizing the D_{JS} between two distributions shares the exact theoretical foundation of Generative Adversarial Networks (GANs) [32]. GAIL leverages this direct connection by reformulating the optimization as a min-max game between a policy generator and a discriminator $D : \mathcal{S} \times \mathcal{A} \to (0,1)$:

$$\min_\pi \max_D \mathbb{E}_\pi[\log(D(s,a))] + \mathbb{E}_{\pi_h}[\log(1 - D(s,a))] - \alpha H(\pi)$$

The discriminator distinguishes human from policy actions, serving as a surrogate local cost function ($c(s,a) = \log D(s,a)$). This yields a dense, differentiable gradient directly for policy optimization, completely eliminating the nested RL requirement.

4.1.4. Adversarial Inverse Reinforcement Learning (AIRL)

While GAIL effectively matches human behavior, it fundamentally remains an imitation learning algorithm rather than a true reward recovery method. The discriminator D in GAIL acts as a local cost function, but it is highly entangled with the specific transition dynamics $\mathcal{T}$ of the training environment. Consequently, the extracted signal does not represent a robust, transferable reward function (R^*); if the robot's physical dynamics change during deployment, the extracted policy collapses.

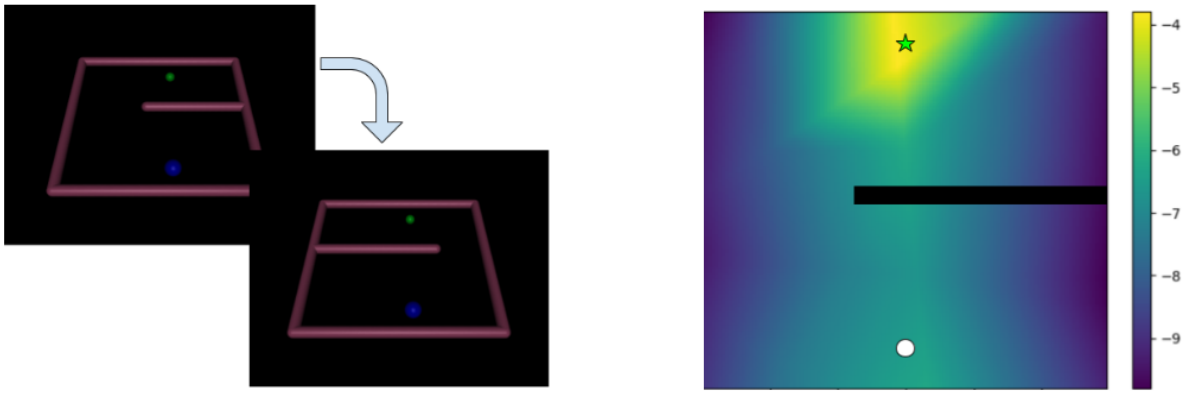

Figure 4.2: AIRL zero-shot transfer in a shifting maze. Left: Wall configurations change between training and testing. Right: The recovered reward successfully isolates the target objective without dynamics-dependent shaping, enabling robust transfer.

Disentangling Reward from Dynamics: To address this limitation, AIRL restructures the adversarial framework to explicitly recover a robust reward function that is invariant to changes in environment dynamics [33]. It modifies the discriminator's architecture to take a specific, analytically derived form:

$$D_\theta(s, a, s') = \frac{\exp(f_\theta(s, a, s'))}{\exp(f_\theta(s, a, s')) + \pi(a|s)}$$

where $\pi(a|s)$ is the current policy generator and $f_\theta(s, a, s')$ acts as the parameterized reward model.

Integration of Potential-Based Shaping: To mathematically guarantee that the recovered reward is disentangled from the environment's transition probabilities, They leveraged the potential-based reward shaping theorem as discussed in section 3. They restricted the function f_θ to the exact structural formulation of a shaped advantage function:

$$f_\theta(s, a, s') = g_\theta(s, a) + \gamma h_\theta(s') - h_\theta(s)$$

Here, $g_\theta(s, a)$ is trained to approximate the true, invariant reward function, while $h_\theta(s)$ acts as a shaping term that approximates the optimal value function.

Robust Transferability: By enforcing this structural constraint, the telescoping shaping terms ($\gamma h_\theta(s') - h_\theta(s)$) absorb the dynamics-dependent biases of the training environment. This isolation leaves $g_\theta(s, a)$ as a purely intent-driven reward signal. Consequently, AIRL is capable of extracting a reward function that can be reliably transferred to environments with entirely out-of-distribution transition dynamics, resolving a critical vulnerability of standard adversarial imitation (see Figure 4.2).

4.2. The Limitations of Offline Data and the Need for Interaction

While offline methodologies like MaxEnt IRL, GAIL, and AIRL formulate mathematically elegant objectives, their fundamental reliance on static, *a priori* datasets introduces severe challenges:

- **Need for optimal offline dataset:** Passive learning assumes the dataset perfectly covers the optimal state-action manifold. In practice, collecting large-scale, optimal human demonstrations is expensive [34]. Furthermore, human cognitive fatigue inherently generates noisy, sub-optimal trajectories [35]. Because methods that learn reward only via offline datasets strictly assume these deviations are intentional, they inevitably assign high utility to erroneous behaviors, permanently biasing the reward.
- **Covariate Shift:** Another critical flaw, as discussed earlier, is vulnerability to covariate shift [36]. Static datasets strictly demonstrate optimal execution and entirely lack corrective coverage. A minor kinematic error shifts the agent into OOD states, as depicted in Figure 4.3. Lacking empirical data on how to recover, these execution errors compound catastrophically.
- **Absence of Temporal Credit Assignment:** Offline agents observe static trajectories without dynamic feedback, isolating the causal relationship between intermediate actions and long-term

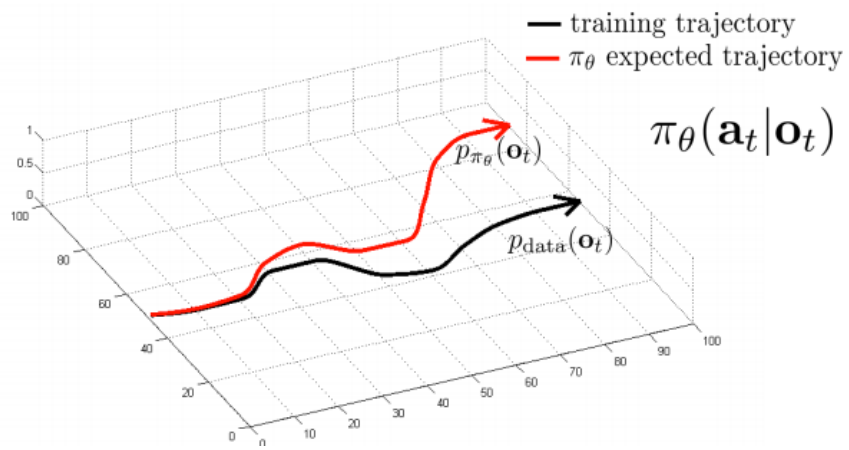

Figure 4.3: A minor execution error in policy execution induces a covariate shift, forcing the policy into out-of-distribution states where it lacks expert recovery data, thereby compounding failures [37].

success. In OOD states, the absence of an active reward signal prevents the algorithm from evaluating recovery attempts, permanently trapping the policy in failure modes [38].

The Paradigm Shift to Interactive Learning:
These compounding challenges necessitate a transition toward **Active Interactive Feedback**. By iteratively evaluating, ranking, or correcting the agent's behavior during training, algorithms can query human feedback directly on the learner's own state distribution, $p_{\pi_l}(\tau)$. This inherently resolves the covariate shift bottleneck and provides dynamic credit assignment exactly where the agent struggles. By explicitly guiding the agent through its unique execution mistakes, this interactive paradigm eliminates the need for intractable, exhaustive offline datasets and drastically improves overall data efficiency. Figure 4.4 depicts how interactive human interventions can actively correct a diverging trajectory, steering the agent away from compounding execution failures and safely back to the optimal state distribution.

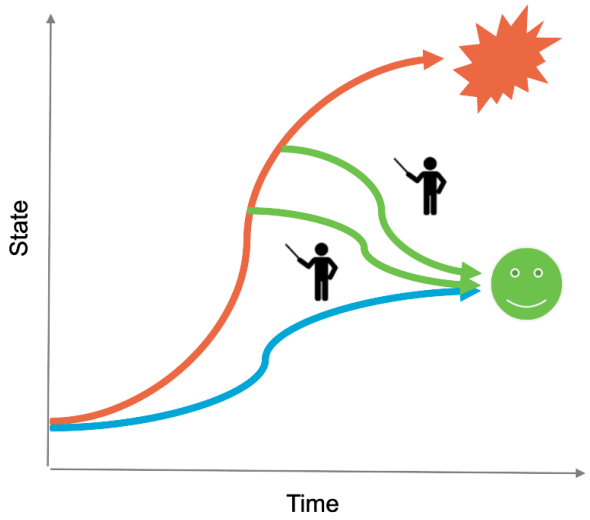

Figure 4.4: Interactive human interventions (green) recovering a divergent trajectory (red) back to the target goal, mitigating covariate shift.

Key Insights

Chapter 4: Reward Learning from Offline Demonstrations

- **The Offline Paradigm:** Reward learning mathematically shifts objective specification away from manual engineering, seeking instead to extract the true utility function (R^*) directly from static, *a priori* human demonstrations.
- **Algorithmic Evolution:** Early methods like MaxEnt IRL resolved behavioral ambiguity but introduced computationally intractable nested optimizations. Adversarial frameworks (GAIL, AIRL) bypassed this bottleneck via occupancy measure matching, with AIRL explicitly recovering dynamics-invariant rewards through potential-based shaping.
- **The Covariate Shift Bottleneck:** Extracting policies exclusively from static datasets makes the agent vulnerable to covariate shift and causes compound errors when encountered OOD states.
- **The Necessity of Interaction:** Strict reliance on passive, offline datasets exposes fundamental vulnerabilities, including susceptibility to sub-optimal human noise, catastrophic covariate shift, and the absence of dynamic temporal credit assignment. These mathematical and practical limitations inherently motivate the transition toward active, human-in-the-loop interactive learning.

5

Active Evaluative Feedback: Preference-Based Reward Learning

As established in Chapter 4, the reliance only on offline datasets exposes several challenges and motivates the reinforcement learning paradigm to shift from passive observation to active, human-in-the-loop (HiL) interaction. By querying a human supervisor directly on the data generated by the learner's current policy $p_{\pi_t}(\tau)$, the algorithm can systematically eliminate out-of-distribution failures and dynamically guide the agent toward the true task objective.

The most widely adopted interactive paradigm is *Evaluative Feedback*, specifically Preference-Based Reward Learning (PbRL). In this framework, the human supervisor acts as a critic rather than a demonstrator. Instead of providing the optimal kinematic actions, which incur a significant cognitive burden, the human merely evaluates and ranks behavior generated by the agent. This chapter formalizes the mathematical mechanisms of preference-based learning, detailing how modern frameworks translate discrete human choices into continuous reward topographies, and ultimately exposes the sample-efficiency bottlenecks inherent to strictly relative feedback.

5.1. Relative Evaluative Feedback: Preference-Based Reward Learning

Standard Preference-Based Reward Learning (PbRL) replaces manual reward specification with a data-driven approach using pairwise human comparisons. To formalize this process and address the sample complexity of querying human operators, frameworks typically structure the learning process into three phases, as depicted in Figure 5.1:

1. **Policy Execution and Prior Initialization:** The algorithm establishes a prior over the reward space using offline passive data and pre-trains the agent's baseline policy.
2. **Active Preference Elicitation:** The algorithm generates trajectory segments and selects queries for human evaluation based on specific optimization criteria.
3. **Reward Formulation and Optimization:** Discrete human preferences are projected into a continuous probability space using the Bradley-Terry model. The reward function is optimized via Maximum Likelihood Estimation (MLE) and used to update the policy.

The following subsections detail the mathematical formulations and algorithmic mechanisms governing these phases.

Phase 1: Policy Execution and Prior Initialization

Learning a reward function entirely from scratch via online human queries exhibits high sample complexity and requires significant human feedback [15, 40]. To reduce the number of required queries, frameworks integrate multiple data sources [39].

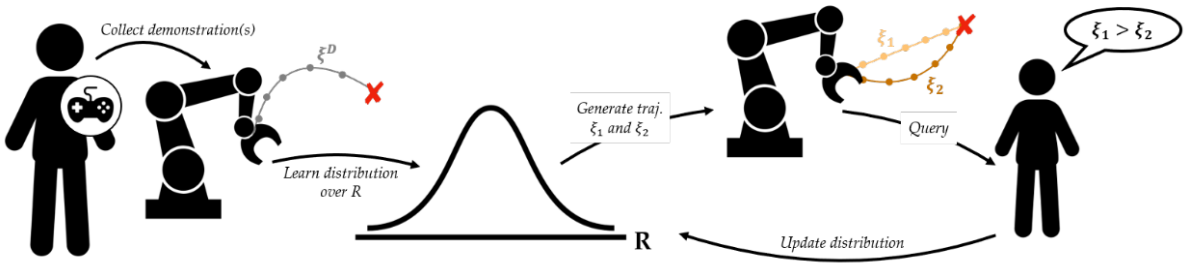

Figure 5.1: Overview of DemPref [39]. Left: Human demonstrations initialize the reward prior. Right: Active trajectory queries fine-tune the reward model.

A small set of passive data, such as offline expert demonstrations ξ^D, provides an initialization for the reward weights ω [39]. Assuming a Boltzmann rational human model, these demonstrations initialize a prior belief $b^0(\omega)$ defined by the feature counts $\Phi(\xi^D)$ and a rationality coefficient β^D [39]:

$$b^0(\omega) \propto \exp(\beta^D \omega \cdot \Phi(\xi^D))P(\omega)$$

Sample efficiency in PbRL is affected by policy initialization. Instead of relying on random initialization, the agent can be pre-trained using unsupervised exploration to maximize state entropy [16] or via behavioral cloning from expert demonstrations [41]. Macuglia et al. [41] formalize the trade-off between offline data volume and online query efficiency. They derive theoretical regret bounds proving that increasing the number of offline expert trajectories, n, constructs a tighter confidence set with a radius shrinking at $\mathcal{O}(1/\sqrt{n})$. This translates into a drastically narrowed search space for subsequent online preference learning.

We deduce that high-quality offline data fundamentally reduces the complexity and sample burden of online PbRL. We observe that as the offline dataset grows ($n \to \infty$), the theoretical regret of the online fine-tuning phase approaches zero. However, as established in Chapter 4, relying on static offline datasets introduces several challenges. Consequently, we are presented with a fundamental trade-off: while maximizing offline data mathematically minimizes online regret, it reintroduces the exact static-dataset limitations that interactive learning seeks to resolve. We argue that an optimal paradigm must strictly balance these modalities, using just enough offline data to effectively constrain the initial hypothesis space, while reserving high-value active preference queries to correct out-of-distribution execution errors dynamically.

Furthermore, by utilizing multi-task learning, reward models can be meta-learned from prior offline task data, allowing the network to adapt to new tasks using a few-shot set of human queries [40].

Phase 2: Active Preference Elicitation

Following initialization, the algorithm queries the human expert with a pair of trajectory segments, σ^1 and σ^2, sampled from the agent's policy distribution [15]. These segments may represent states, actions, or multi-step state-action pairs. The human yields a discrete preference distribution $y \in \{(1,0),(0,1),(0.5,0.5)\}$, denoting preference for σ^1 ($\sigma^1 \succ \sigma^2$), preference for σ^2 ($\sigma^2 \succ \sigma^1$), or equal utility, respectively. This categorical feedback is appended to an active dataset $\mathcal{D} = \{(\sigma^1, \sigma^2, y)_i\}$.

The sample efficiency of this active fine-tuning is critically dependent on the query selection mechanism. A standard approach in active preference learning is Volume Removal, which seeks to maximize the expected difference between the prior and the unnormalized posterior. This is achieved by solving a submodular optimization problem to select queries where each answer is equally likely given the current belief over the reward weights ω.

In our analysis, we observe that volume removal is a purely robot-focused metric; it strictly targets questions where the human's answer is mathematically unpredictable to the agent (see Figure 5.2). Because it completely fails to account for the human-in-the-loop, this objective frequently converges on best-case mathematical solutions that translate into trivial, indistinguishable, or confusing queries for the human operator, effectively wasting human cognitive effort.

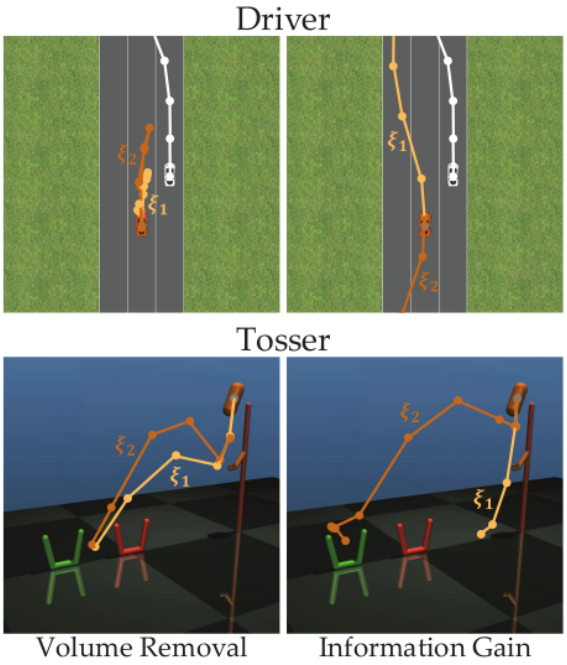

Figure 5.2: Sample queries generated with the volume removal and information gain methods on Driver and Tosser tasks. Volume removal generates queries that are difficult because the options are almost equally good or equally bad [39]

To resolve this, Bıyık et al. [39] propose an Information Gain objective that explicitly balances the theoretical informativeness of the question against the human's ability to answer it confidently:

$$Q^* = \arg\max_{Q} I(\omega; q|Q, b) - c(Q)$$

This objective is driven by two competing components: the first term captures the robot's uncertainty over the human's response, while an internal entropy term captures the human's uncertainty when answering.

We deduce that this formulation yields a strictly superior elicitation paradigm by grounding information-theoretic objectives in considering human evaluative ability. By jointly accounting for robot and human uncertainty, the algorithm exclusively targets regions where the robot lacks knowledge, yet the human can easily distinguish the options. Furthermore, the query-dependent cost $c(Q)$ establishes a mathematically optimal stopping condition: elicitation dynamically terminates precisely when the human cognitive burden outweighs the expected informational utility.

Phase 3: Reward Formulation and Optimization

PbRL maps the discrete preference signal y to a reward function $r_\phi(s, a)$ parameterized by a neural network. Frameworks utilize the **Bradley-Terry (BT) Model** [42] to project categorical choices into a continuous probability measure. The BT framework models the probability of preferring a given segment as proportional to the exponentiated sum of its latent step-wise rewards:

$$R_\phi(\sigma) = \sum_{t=0}^{k-1} r_\phi(s_t, a_t)$$

Under the BT formulation, the preference probability for σ^1 over σ^2 is represented as a softmax distribution over their cumulative latent rewards:

$$P_\phi[\sigma^1 \succ \sigma^2] = \frac{\exp(R_\phi(\sigma^1))}{\exp(R_\phi(\sigma^1)) + \exp(R_\phi(\sigma^2))}$$

Algebraically, dividing the numerator and denominator by $\exp(R_\phi(\sigma^2))$ yields a logistic sigmoid function

parameterized by the utility difference:

$$P_\phi[\sigma^1 \succ \sigma^2] = \frac{1}{1 + \exp\left(-\left(R_\phi(\sigma^1) - R_\phi(\sigma^2)\right)\right)}$$

Defining $\Delta = R_\phi(\sigma^1) - R_\phi(\sigma^2)$ as the latent reward gap establishes the relationship limits. When $\Delta = 0$, the probability is 0.5. As $\Delta \to \infty$, the exponential term $\exp(-\Delta)$ decays to zero, causing the predicted preference probability to approach 1.

The reward function parameters ϕ are optimized by minimizing the binary cross-entropy loss (negative log-likelihood) between the empirical human preference y and the predicted probability distribution across the dataset $\mathcal{D}$.

$$\mathcal{L}_{reward}(\phi) = -\mathbb{E}_{(\sigma^1,\sigma^2,y)\sim\mathcal{D}}\left[y(1)\log P_\phi[\sigma^1 \succ \sigma^2] + y(2)\log P_\phi[\sigma^2 \succ \sigma^1]\right]$$

The resulting state-action-conditioned signal $r_\phi(s_t, a_t)$ is utilized for policy extraction via reinforcement learning algorithms (e.g., PPO [43], SAC [44]). During this optimization phase, the reinforcement learning objective can be bounded by a Kullback-Leibler (KL) divergence regularizer against the pre-trained reference policy to retain prior task behaviour [41].

5.2. Bypassing the Reward Model: Direct Preference Optimization (DPO)

The standard three-phase pipeline of PbRL relies heavily on a two-stage optimization process: first, training a surrogate reward model, and subsequently extracting a policy via reinforcement learning. Rafailov et al. [45] demonstrate that this decoupled approach is unstable. The instability arises because it requires first fitting a reward model that reflects human preferences, and then fine-tuning the model using reinforcement learning to maximize this estimated reward without diverging too far from the original model. Furthermore, this pipeline necessitates generating samples from the model during fine-tuning and performing significant hyperparameter tuning.

To resolve this structural instability, Rafailov et al. introduce DPO, which eliminates the intermediate reward network. It exploits an analytical mapping between the reward function and the optimal policy under a KL divergence constraint.

When maximizing a latent reward $r(s, a)$ subject to a KL penalty against a pre-trained reference policy π_{ref}, the optimal policy π^* takes the exact closed-form solution:

$$\pi^*(a|s) = \frac{1}{Z(s)}\pi_{ref}(a|s)\exp\left(\frac{1}{\beta}r(s,a)\right)$$

where β controls the strength of the KL penalty and $Z(s)$ is the intractable partition function.

By algebraically rearranging this optimality condition, the reward can be expressed explicitly as a function of the respective policies:

$$r(s,a) = \beta\log\frac{\pi^*(a|s)}{\pi_{ref}(a|s)} + \beta\log Z(s)$$

Rafailov et al. demonstrate that by substituting this reparameterized reward back into the Bradley-Terry preference model, the intractable partition function $Z(s)$ mathematically cancels out when computing the reward difference between two trajectory segments (σ^1 and σ^2). This cancellation yields a direct loss function defined entirely over the policy network π_θ, bypassing the reward model r_ϕ completely:

$$\mathcal{L}_{DPO}(\pi_\theta) = -\mathbb{E}_{(\sigma^1,\sigma^2,y)\sim\mathcal{D}}\left[\log\sigma\left(\beta\sum_{(s,a)\in\sigma^1}\log\frac{\pi_\theta(a|s)}{\pi_{ref}(a|s)} - \beta\sum_{(s,a)\in\sigma^2}\log\frac{\pi_\theta(a|s)}{\pi_{ref}(a|s)}\right)\right]$$

I observe that eliminating the actor-critic RL loop removes compounding approximation errors, ensuring that the policy update is strictly and stably anchored to the empirical preference data via a cross-entropy classification loss.

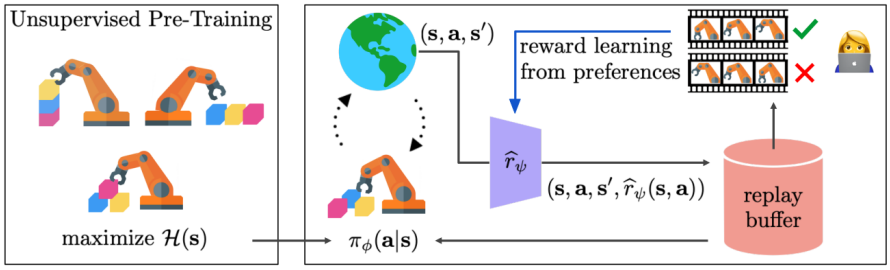

Figure 5.3: Left: Unsupervised pre-training for diverse state exploration. Right: Reward learning from human preferences combined with off-policy reward relabeling for sample-efficient policy optimization.

5.3. Sample Efficiency in PbRL Methods

While the standard Bradley-Terry optimization pipeline establishes a rigorous foundation for preference elicitation, its practical implementation suffers from prohibitive sample inefficiency. Because the parameterized reward model r_ϕ is dynamically updated during training, the underlying reinforcement learning objective is inherently non-stationary [16].

To mitigate this non-stationarity, early PbRL frameworks relied on on-policy reinforcement learning algorithms (e.g., PPO, TRPO). This architectural constraint necessitated the continuous discarding of transition data following each gradient update, thereby demanding a high volume of human queries to sustain policy improvement. To render PbRL tractable for continuous control, recent frameworks—most notably PEBBLE [16] and SURF [17] fundamentally restructure the learning pipeline to optimize sample complexity via off-policy integration, unsupervised pre-training, and semi-supervised data augmentation.

5.3.1. Off-Policy Integration and Unsupervised Pre-Training (PEBBLE)

Lee et al. [16] integrate off-policy Soft Actor-Critic into the PbRL framework. An overview of the method is depicted in Figure 5.3. Because standard off-policy learning fails when rewards stored in the replay buffer $\mathcal{B}$ become obsolete due to a non-stationary reward network, PEBBLE utilizes a reward relabeling mechanism. Before each critic update, the reward for sampled transitions is recalculated using the most recent parameters ϕ:

$$\mathcal{J}_Q(\theta) = \mathbb{E}_{(s_t, a_t, s_{t+1}) \sim \mathcal{B}} \left[\left(Q_\theta(s_t, a_t) - \left(r_\phi(s_t, a_t) + \gamma V_{\bar{\theta}}(s_{t+1}) \right) \right)^2 \right]$$

where Q_θ and $V_{\bar{\theta}}$ represent the critic and target value networks, and r_ϕ is the actively relabeled reward network.

Additionally, to ensure initial human queries are informative, PEBBLE pre-trains the agent to maximize intrinsic state entropy $\mathcal{H}(S)$ via a non-parametric k-nearest neighbor reward before any feedback is elicited:

$$r_{int}(s_t) \propto \sum_{j=1}^{k} \log ||s_t - s_t^{(j)}||_2$$

where s_t represents the agent's state at timestep t, $s_t^{(j)}$ denotes the j-th nearest neighbor of s_t computed across the transitions currently stored in the replay buffer. This summation aggregates the log-distances to the k closest previously visited states, providing an intrinsic reward that explicitly encourages the agent to explore novel, unvisited regions of the state space.

I note that dynamic relabeling provides a crucial mechanism for off-policy PbRL, enabling the agent to reuse past environmental interactions and thereby mitigating the environmental sample bottleneck. Furthermore, unsupervised pre-training serves as an effective primer; encouraging initial trajectory pairs to span a broader distribution of the state space can facilitate early-stage convergence.

5.3.2. Semi-Supervised Learning and Data Augmentation (SURF)

To extract utility from unlabeled trajectory pairs generated during off-policy exploration, Park et al. [17] introduce SURF. This method maintains an ensemble of M reward models to estimate epistemic uncer-

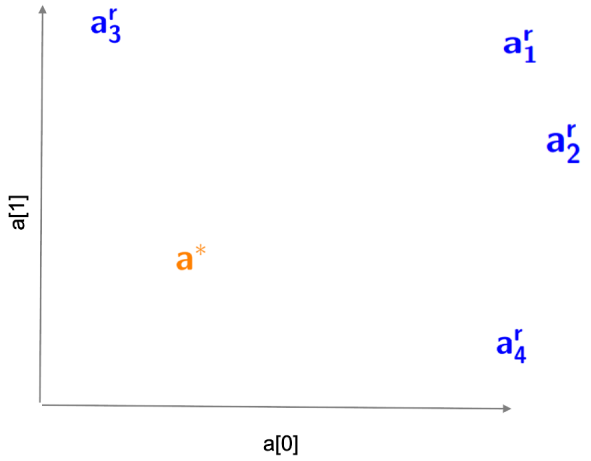

Figure 5.4: Exploration ambiguity in relative preference learning. Pairwise comparisons of sub-optimal robot actions (a^r) provide no absolute spatial guidance toward the true optimal action (a^*).

tainty. For unlabeled pairs where the predicted preference variance falls below a confidence threshold τ, the algorithm assigns a pseudo-label $\hat{y}$ based on the ensemble mean:

$$\hat{y} = \arg\max \mathbb{E}_m \left[P_{\phi_m}[\sigma^1 \succ \sigma^2] \right] \quad \text{if} \quad \text{Var}_m(P_{\phi_m}) < \tau$$

The reward model is optimized using a joint loss function over the human-labeled dataset $\mathcal{D}_{label}$ and the pseudo-labeled dataset $\mathcal{D}_{unlabel}$:

$$\mathcal{L}_{total}(\phi) = \mathcal{L}_{reward}(\phi, \mathcal{D}_{label}) + \lambda \mathcal{L}_{reward}(\phi, \mathcal{D}_{unlabel}, \hat{y})$$

To further expand the dataset, SURF employs temporal cropping, which extracts continuous sub-sequences $\tilde{\sigma}$ of length $K < H$ from evaluated trajectories σ of length H, asserting the preference y holds true for the cropped sub-sequences.

I view pseudo-labeling as automated density-propagation that expands the dataset without further human input. Additionally, temporal cropping acts as a structural regularizer, mitigating temporal credit assignment ambiguity by enforcing preference consistency across temporal resolutions.

5.4. The Fundamental Limitations of Relative Feedback on Sample Efficiency

Despite algorithmic advancements in query elicitation and off-policy data utilization, Preference-Based RL (PbRL) suffers from inherent sample inefficiency due to its reliance on *relative evaluative feedback*. Mathematically, a discrete preference ($\sigma^1 \succ \sigma^2$) translates via the Bradley-Terry model into a weak inequality over cumulative rewards: $\sum r(\sigma^1) > \sum r(\sigma^2)$. Geometrically, each query defines a hyperplane bisecting the reward hypothesis space. Because this bisection provides zero or less information regarding absolute reward magnitude or distance to the optimal manifold, bounding the true reward function requires an exponentially large volume of queries, severely degrading efficiency in high-dimensional continuous domains. Figure 5.4 illustrates this structural limitation in two-dimensional action space.

I deduce that this absence of absolute grounding severely compounds exploration ambiguity. When a human supervisor is asked to evaluate trajectory pairs that both reside entirely within sub-optimal manifolds, the Bradley-Terry model strictly enforces a relative comparison between the two. I infer that this relative anchoring imposes a high sample complexity, making the model fail to converge toward the true objective if the elicited queries remain confined to these sub-optimal spaces.

Looking from another perspective, I infer that while relative evaluations successfully reduce the marginal cognitive burden per interaction, obviating the need for full kinematic demonstrations, they structurally

lack the spatial density required to efficiently constrain continuous robotic tasks. I observe that this preference effectively identifies *which* trajectory is superior, but it mathematically fails to quantify the absolute distance to success or specify exactly *where* physical kinematic corrections are required. I argue that this structural absence of absolute, state-action-grounded feedback theoretically necessitates a transition toward alternative modalities, specifically establishing the foundation for active interactive corrections.

Key Insights

Chapter 5: Active Evaluative Feedback: Preference-Based Reward Learning

- **Data-Driven Objectives:** PbRL maps discrete pairwise human preferences into continuous reward topographies utilizing the Bradley-Terry model. This process is structured into three distinct phases: initializing a reward prior from offline data, actively eliciting human feedback considering both human & robot ambiguity, and optimizing the parameterized reward network via Maximum Likelihood Estimation.
- **Direct Optimization:** Frameworks like DPO stabilize learning by directly mapping preferences to policy updates via a KL-constrained objective, bypassing the surrogate reward network entirely.
- **Sample Efficiency Interventions:** To mitigate the traditionally high sample complexity of PbRL, algorithms such as PEBBLE and SURF introduce off-policy reward relabeling, unsupervised pre-training for state exploration, and semi-supervised pseudo-labeling.
- **The Limitation of Relative Feedback:** Despite these algorithmic enhancements, scalar preferences structurally lack absolute spatial grounding. This mathematically starves the state-action space of the dense kinematic guidance required to actively correct physical errors, thereby motivating a paradigm shift toward active interactive corrections.

6

Active Corrective Interventions: Interactive Imitation Learning

As established in Chapter 5, while preference-based feedback imposes a low cognitive burden, its strict reliance on relative comparisons lacks dense kinematic guidance. Identifying a trajectory as sub-optimal is insufficient for recovery; the agent requires absolute, spatially grounded feedback detailing *how* to correct its kinematics physically.

This structural limitation motivates a transition toward **Active Corrective Interventions**. Under this paradigm, human supervisors operate directly within the state-action space, actively assuming control to provide localized corrective demonstrations exactly when and where the policy fails. This chapter formalizes the algorithmic foundations of active correction methods in Interactive Imitation Learning (IIL), exposes the vulnerabilities of pure policy cloning, and mathematically details the failure modes of these algorithms when subjected to noisy, sub-optimal human corrections.

6.1. The Baseline: Behavioral Cloning and Covariate Shift

The most fundamental methodology for extracting control policies directly from expert demonstrations is Behavioral Cloning (BC) [46]. Unlike inverse reinforcement learning (IRL) paradigms, BC avoids interaction with the environment and bypasses the Markov Decision Process (MDP) framework entirely, mathematically reducing policy extraction to a standard supervised learning regression.

BC typically employs Maximum Likelihood Estimation (MLE) to regress a parameterized policy against human actions, maximizing the log-likelihood of the static offline dataset $\mathcal{D}$:

$$\hat{\pi}_{MLE} = \arg\max_{\pi \in \Pi} \sum_{(s,a) \in \mathcal{D}} \log \pi(a|s)$$

Alternatively, this can be formulated as a risk minimization problem where the learner minimizes a loss function l between the parameterized policy π_θ and the expert policy π_E, evaluated strictly over the state visitation distribution of the expert $\lambda_\mu^{\pi_E}$:

$$\min_\theta \mathbb{E}_{s \sim \lambda_\mu^{\pi_E}} [l(\pi_\theta(\cdot|s), \pi_E(\cdot|s))]$$

While computationally simple, BC is notoriously brittle due to compounding cascading errors. Because the policy parameters θ are optimized exclusively over $\lambda_\mu^{\pi_E}$, the agent remains entirely agnostic to the true underlying transition dynamics of the environment. During physical deployment, however, the agent acts within its own induced state distribution $\lambda_\mu^{\pi_\theta}$. The theoretical performance gap for a discrete and realizable policy class Π is mathematically bounded with probability $1 - \delta$ as follows:

$$\langle \mu, V^{\pi_E} - V^{\hat{\pi}_{MLE}} \rangle \leq \mathcal{O}\left(\frac{1}{(1-\gamma)^2} \sqrt{\frac{\log(|\Pi|/\delta)}{|\mathcal{D}|}} \right)$$

The term $\frac{1}{(1-\gamma)^2}$ formally reflects the compounding errors over the effective horizon $H = \frac{1}{1-\gamma}$. Any microscopic error ϵ shifts the agent into an unobserved, out-of-distribution state; lacking empirical corrective data on how to recover, these errors compound quadratically. This theoretical limitation proves that ignoring transition dynamics is catastrophic for continuous robotic control, necessitating algorithms that actively leverage interactive human corrections.

6.2. Interactive Correction Methods

To mathematically resolve the covariate shift bottleneck inherent to static BC, corrective action-based Interactive Imitation Learning algorithms dynamically pull the training distribution toward the actual state manifold the agent is likely to visit through online human intervention.

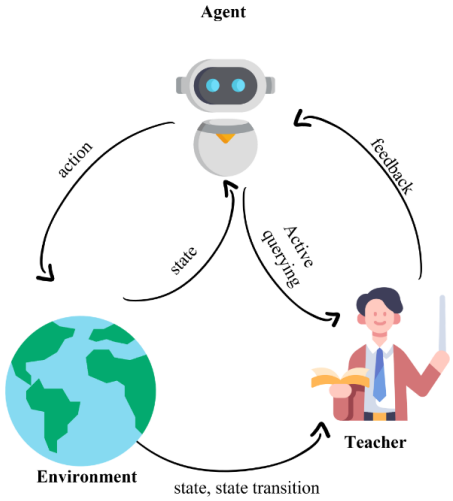

Figure 6.1: Interactive Imitation Learning loop [8]

6.2.1. DAgger and the Learner's State Distribution

To break the quadratic error dependency, Dataset Aggregation (DAgger) [36] operates as a reduction to no-regret online learning. Instead of forcing the agent to learn from the states the expert naturally visits, it queries the expert on states sampled directly from the learner's distribution $\lambda_{\mu}^{\pi_\theta}$.

At iteration t, DAgger minimizes the behavioral cloning loss l_i across all previously aggregated datasets $\mathcal{D}_i$:

$$\pi_t = \arg\min_{\pi \in \Delta} \sum_{i=1}^{T} l_i(\pi, \mathcal{D}_i)$$

By analyzing this through the Performance Difference Lemma (PDL) [47], the regret bound over T iterations is given by:

$$\sum_{t=1}^{T} \langle \mu, V^{\pi_E} - V^{\pi_t} \rangle \leq \frac{\max_{s,a} |Q^{\pi_E}(s,a)|}{1-\gamma} \left(\mathcal{O}(\sqrt{T}) + \mathcal{R}(T) \right)$$

Where $\mathcal{R}(T)$ represents the regret and $\mathcal{O}(\sqrt{T})$ bounds the martingale differences via the Azuma-Hoeffding inequality [48]. This formulation proves that DAgger successfully improves the error inflation factor from the quadratic $\mathcal{O}(\frac{1}{(1-\gamma)^2})$ inherent in BC, to a linear dependency $\mathcal{O}(\frac{\max |Q|}{1-\gamma})$.

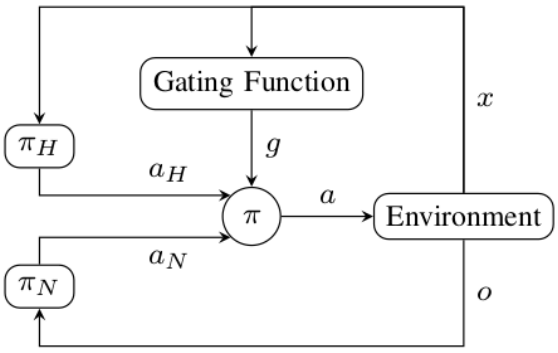

Figure 6.2: Control Loop of HG-DAgger [49]

6.2.2. Human-Gated Control Dynamics (HG-DAgger)

While DAgger guarantees a linear error bound, it introduces a severe practical vulnerability: it structurally requires the expert to provide accurate action labels for states generated by the novice policy without necessarily maintaining full control of the system. I note that requiring a human supervisor to actively provide corrective actions under a shared or robot-centric sampling scheme can alter human behavior and introduce cognitive fatigue.

To adapt this mathematical framework for physical human supervisors, Human-Gated interactive learning (e.g., HG-DAgger) [49] inverts the algorithmic control. The human supervisor passively monitors the robot's autonomous execution and only intervenes when the novice policy enters an unsafe or divergent region of the state space. Expert action labels are strictly collected during these active recovery trajectories, during which the human expert maintains uninterrupted control authority. Figure 6.2 illustrates the control loop of the algorithm.

I deduce that human-gated control elegantly bridges the theoretical $\mathcal{O}(T)$ guarantees of DAgger with the cognitive limitations of physical supervisors. By delegating the intervention gating function directly to the human, the framework inherently targets high-value, OOD failure states while safely bypassing regions where the policy has already converged to optimal execution.

6.3. The Limitations of Direct Policy Cloning

Despite the dataset efficiency of interactive control, canonical algorithms fundamentally process these interventions via pure supervised policy cloning, directly regressing the agent's actions to the human's corrective actions $\pi_E(s_i)$:

$$\pi^* \in \arg\min_{\pi} \sum_{i=1}^{N} \mathcal{L}(a_i, \pi(s_i))$$

I argue that this pure policy-cloning approach remains mathematically fragile. While interactive interventions successfully correct localized kinematic errors, minimizing a loss function $\mathcal{L}$ that forces a neural network to blindly mimic raw actions structurally entangles the supervisor's underlying semantic intent with the specific transition dynamics of the training environment. Consequently, the learned policy inherently overfits to the local training manifold.

If the physical environment changes during deployment, the purely cloned policy lacks a foundational, invariant representation of *why* the human intervened. I observe that treating physical interventions merely as static action targets fails to yield a robust representation of the task objective. To achieve true OOD generalization, the algorithm must mathematically decouple the underlying semantic intent driving the intervention from the raw kinematics.

6.4. The Vulnerability of Sub-Optimal Human Corrections

The challenges in direct policy cloning are magnified when we relax the assumption of **an optimal oracle**. Standard interactive imitation learning mathematically assumes the expert possesses the optimal policy π_E. In physical robotics, however, human demonstrators exhibit cognitive fatigue, variable reaction times, and suboptimal heuristics.

When a human operator provides noisy, sub-optimal corrective actions, DAgger explicitly forces the learner to minimize the loss against these exact kinematic flaws. Pure imitation algorithms cannot mathematically separate the human's underlying intent from their flawed execution. By structurally matching the sub-optimal feature expectations $\rho_\phi(\pi_E)$, the agent permanently caps its own performance ceiling at the level of the flawed human expert. One may argue that noisy human demonstrations are reduced or nullified by the law of large numbers. However, I argue that this approach leads to high sample complexity and may not guarantee true optimality.

6.4.1. Transitioning from Cloning to Reward Extraction

To mitigate the effects of this sub-optimal noise, the algorithmic paradigm can benefit from shifting away from strictly copying actions toward **extracting intent**. Rather than treating interventions solely as supervised targets, the decision to intervene itself can provide a viable reinforcement learning signal. By formulating the human's intent as a standalone, parameterized reward function $r(s, a)$ and learning from it, we may more robustly approximate the true human objective.

I hypothesize that treating human interventions as spatial penalties rather than absolute kinematic targets mitigates the imitation bottleneck by providing locally informative directional structure for robust reward extraction. Furthermore, to effectively account for sub-optimality across multiple human corrections and accurately isolate the true underlying intent from kinematic noise, this spatial guidance may require structured priors to be sample-efficient. Chapter 7 details specific methodologies from the literature that formalize and integrate these structured priors.

- Standard imitation learning
  - copy the *actions* performed by the expert
  - no reasoning about outcomes of actions

Figure: Robot imitation

- Human imitation learning
  - copy the *intent* of the expert
  - might take very different actions!

Figure: Human imitation

Figure 6.3: Left: Standard imitation learning strictly clones (sub-optimal) expert actions without reasoning about the underlying outcomes. Right: Human imitation learning extracts and replicates the expert's semantic intent, allowing for varied kinematic execution to achieve the same goal [50]

Key Insights

Chapter 6: Active Corrective Interventions: Interactive Imitation Learning

- **The Covariate Shift Baseline:** Behavioral Cloning incurs compounding quadratic errors, $\mathcal{O}(\frac{1}{(1-\gamma)^2})$, when execution errors force the agent into unobserved OOD states.
- **Interactive Error Bounding:** DAgger mitigates covariate shift by querying the expert strictly on the learner's induced state distribution, mathematically reducing the error bound to a linear dependency, $\mathcal{O}(\frac{1}{1-\gamma})$.
- **Human-Gated Control:** Frameworks like HG-DAgger adapt to cognitive limits by restricting human interventions to critical safety failures, yielding highly concentrated, high-value recovery datasets.
- **The Policy Cloning Bottleneck:** Regressing directly against human actions entangles semantic intent with local transition dynamics, causing overfitting and preventing robust OOD generalization.
- **The Sub-Optimality Vulnerability:** Direct policy cloning imitates the kinematic noise of sub-optimal experts and is sample-inefficient to recover. Bypassing this performance upper bound requires a pivot toward intent extraction via parameterized reward functions and structured priors.

7

Structured Priors and Robust Reward Learning

As established in Chapter 6, Interactive Imitation Learning is bottlenecked by its reliance on optimal demonstrations. Minimizing kinematic loss directly against human feedback inevitably overfits the policy to the demonstrator's execution noise. Transcending this limitation necessitates a paradigm shift: algorithms must transition from strict point-wise imitation toward extracting underlying semantic intent through robust reward formulations and relaxed geometric targets.

This chapter formalizes mathematical frameworks for learning effectively from sub-optimal data. It first details the extraction of spatial reinforcement penalties from online interventions and the establishment of structural priors using low-quality offline datasets. Subsequently, it introduces set-valued supervision, energy-based models, and diffusion-based action chunking to systematically decouple human intent from kinematic noise, thereby mitigating overfitting and scaling robust imitation to high-dimensional continuous control.

7.1. Interactive Reward Extraction: The RLIF Paradigm

To decouple a human supervisor's underlying intent from execution noise, interactive corrections can be reformulated from supervised action targets into spatial reinforcement penalties.

The Reinforcement Learning from Intervention Feedback (RLIF) framework, as illustrated in Figure 7.1, formalizes this approach [51]. Rather than minimizing a behavioral cloning loss that assumes human optimality, RLIF applies off-policy reinforcement learning, utilizing the intervention event itself as a discrete reward signal. This assumes that the physical act of intervention indicates an unsafe or sub-optimal state-action subspace.

Mathematically, the environment is modeled with an implicit intervention-based reward function $\tilde{r}_\delta$, defined as 0 (or a scaled negative penalty) for states triggering an intervention and 1 otherwise. By deploying off-policy RL on the aggregated dataset, the policy $\tilde{\pi}$ is optimized to maximize the expected

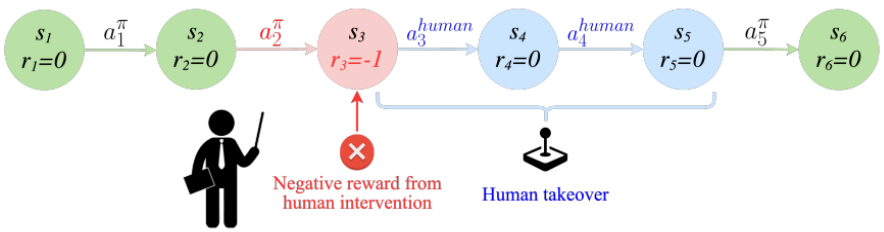

Figure 7.1: RLIF uses RL to learn without ground-truth rewards with data collected with suboptimal human interventions [51]

return:

$$\tilde{\pi} \in \arg\max_{\pi \in \Pi} \mathbb{E}_{a_t \sim \pi(s_t), \tilde{r}_\delta(s_t, \pi(s_t))} \left[\sum_{t=0}^{\infty} \gamma^t \tilde{r}_\delta(s_t, a_t) | s_0 \sim \mu \right]$$

This formulation helps mitigate the performance ceiling inherent to pure imitation. Luo et al. demonstrate that the suboptimality gap for RLIF is bounded by:

$$V^*(\mu) - V^{\tilde{\pi}}(\mu) \leq \min\{V^{\pi^*}(\mu) - V^{\pi^{ref}}(\mu), V^{\pi^*}(\mu) - V^{\pi^{exp}}(\mu)\} + \frac{\delta\epsilon}{1-\gamma}$$

Because this upper bound evaluates the minimum between the expert's explicit policy π^{exp} and a potentially superior reference policy π^{ref} used to trigger interventions, the agent possesses the theoretical potential to exceed the performance of a sub-optimal demonstrator.

7.2. Structured Priors via Sub-Optimal Data

While RLIF extracts rewards dynamically from online interventions, learning a comprehensive reward function exclusively from human feedback requires high sample complexity and extensive human interaction [52]. To effectively account for sub-optimality and accelerate the reward learning process, the state-action space requires structural priors established *before* active human querying begins.

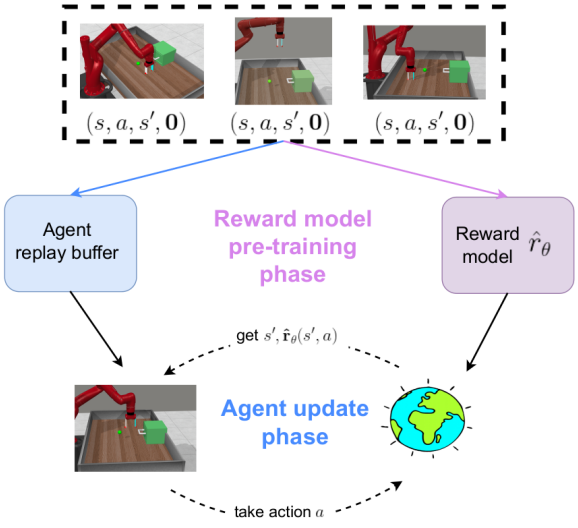

Figure 7.2: Overview of SDP: Sub-optimal offline trajectories are pseudo-labeled with a minimum reward (r_{min}) to pre-train the reward model and initialize the replay buffer prior to active human feedback [52]

The Sub-optimal Data Pre-training (SDP) methodology, as illustrated in Figure 7.2, provides this structural prior by leveraging readily available, reward-free, low-quality offline data [52]. SDP operates on the premise that trajectories from a random or partially trained policy represent highly sub-optimal behaviors. Rather than discarding this data, SDP pseudo-labels all low-quality transitions with the minimum environment reward r_{min}.

The reward model $\hat{r}_\theta$ is then pre-trained on this pseudo-labeled dataset $D_{sub} = \{s_i, a_i, r_{min}\}_{i=1}^N$ via regression to minimize the mean squared error:

$$L^{MSE}(\theta, D_{sub}) = \mathbb{E}_{(s_i, a_i) \sim D_{sub}} \left[(r_{min} - \hat{r}_\theta(s_i, a_i))^2 \right]$$

I deduce that this pre-training acts as a powerful structural prior. It forces the neural network to associate chaotic, low-quality regions of the state-action space with low utility without requiring a single human label [52]. When human-in-the-loop training subsequently begins, the reward network already possesses a baseline, increasing the sample efficiency and stability of subsequent preference-based or scalar-based updates.

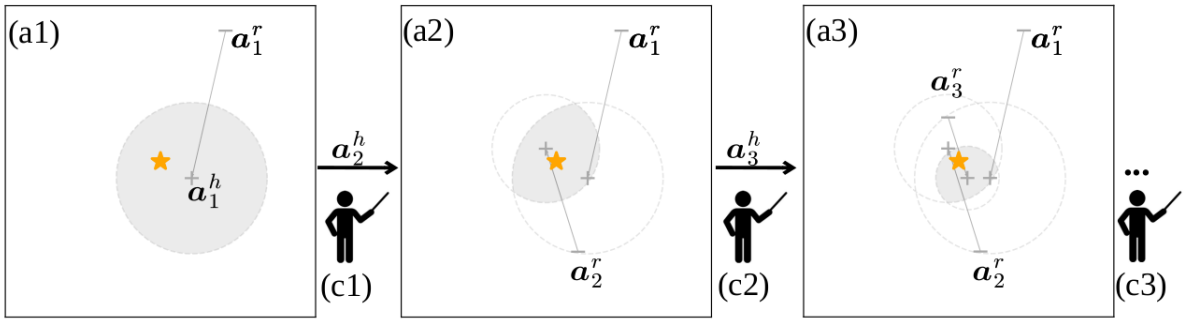

Figure 7.3: Overall desired action space formed by aggregating multiple teacher feedback signals in a single state. This action space gets progressively refined as additional corrections are provided [53]

7.3. Set-Valued Supervision

Relying on direct point-wise imitation inherently assumes that every human intervention represents an optimal kinematic target. To mitigate the resulting overfitting to sub-optimal corrections, algorithms require formulations that relax this strict behavioral cloning assumption. By transitioning from isolated point targets to defined geometric regions, the policy can absorb execution noise and extract the underlying task objective without explicitly replicating sub-optimal human kinematics.

7.3.1. Energy-Based Set Supervision (CLIC)

To avoid overfitting to imperfect action labels, Contrastive Policy Learning from Interactive Corrections (CLIC) introduces set-valued supervision by defining a single-step desired action space $\hat{\mathcal{A}}^{ss}(a^r, a^h)$ [53]. Rather than treating the human correction a^h as an exact point target, this space forms a continuous geometric boundary that encapsulates a^h and the presumed optimal action a^*, while strictly isolating the robot's suboptimal execution a^r. This is illustrated in Figure 7.3. For intervention data, this region is formalized as a hypersphere centered at a^h:

$$\hat{\mathcal{A}}^{ss}(a^r, a^h) = \{a \in \mathcal{A} \mid r \cdot \mathbb{D}(a^h, a^r) \geq \mathbb{D}(a, a^h)\}$$

where $r \in (0, 1]$ is a hyperparameter dictating the radius and volume of the desired subspace, and $\mathbb{D}$ denotes the distance metric.

CLIC parameterizes the policy utilizing an Energy-Based Model (EBM), where the conditional probability is governed by the Boltzmann distribution over an energy function $E_\theta(s, a)$:

$$\pi_\theta(a|s) = \frac{\exp(-E_\theta(s, a))}{Z}$$

During training, the EBM is shaped to concentrate its probability mass within these desired regions by minimizing the Kullback-Leibler (KL) divergence between the current policy and a target distribution π^{target}:

$$\ell_{KL}(\theta) = \mathbb{E}_{(s,a^r,a^h)\sim\mathcal{D}}\left[KL(\pi^{target}(\cdot|s) \parallel \pi_\theta(\cdot|s))\right]$$

where the target distribution favors actions inside the desired geometric set:

$$\pi^{target}(a|s) \propto p(a \in \hat{\mathcal{A}}^{ss}(a^r, a^h)|a, s)\pi_\theta(a|s)$$

This set-valued formulation acts as a robust spatial prior. By defining the target as a continuous geometric region rather than a singular point estimate, the algorithm inherently absorbs the noise and sub-optimality frequently observed in human demonstrations. As the policy iteratively receives and aggregates multiple corrections for similar states, the intersection of these individual desired action spaces progressively refines the overall optimal region. This mechanism ensures the policy converges toward true optimal actions without overfitting to any isolated, potentially suboptimal human demonstration.

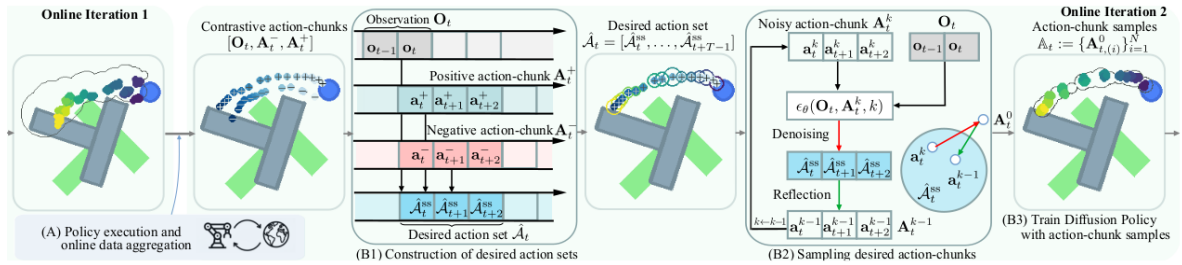

Figure 7.4: Overview of Set-DP: (A) Online environment interaction progressively shrinks desired action sets. (B) Contrastive action-chunk pairs define these sets, from which target chunks are sampled via reflected denoising to optimize the diffusion policy.

7.3.2. Scaling Set Supervision to High-Dimensional Action-Chunks (Set-DP)

While CLIC establishes a robust spatial prior for single-step actions using EBMs, Set-Supervised Diffusion Policy (Set-DP) extends this set-valued supervision to the action-chunking paradigm using diffusion models [54]. An action chunk groups action sequences over a fixed horizon T:

$$A_t = [a_t, a_{t+1}, \ldots, a_{t+T-1}] \in \mathcal{A}^T$$

At step t, Set-DP constructs a composite desired action set $\hat{\mathcal{A}}_t = [\hat{\mathcal{A}}_t^{ss}, \ldots, \hat{\mathcal{A}}_{t+T-1}^{ss}]$ derived directly from the paired contrastive data of the robot's undesired execution A_t^- and the human teacher's positive correction A_t^+.

Set-DP samples target action-chunks A^0 from the desired set $\hat{\mathcal{A}}_t$ via a constrained reflected denoising process as seen in Figure 7.4. The diffusion policy ϵ_θ is subsequently trained to maximize the likelihood of these valid in-set chunks using the standard denoising objective:

$$l_t = \mathbb{E}_{k\sim[1,K],O_t}\left[||\epsilon_k - \epsilon_\theta(O_t, \sqrt{\bar{\alpha}_k}A^0 + \sqrt{1-\bar{\alpha}_k}\epsilon_k, k)||^2\right]$$

Empirically, Set-DP demonstrates significant advantages over EBM-based methods like CLIC and standard diffusion baselines such as DP and DP-DPO. By utilizing set-valued supervision rather than strict behavioral cloning or weak pairwise comparisons, Set-DP prevents overfitting to specific action labels and remains highly robust to noisy intervention data. Furthermore, this relaxed geometric targeting naturally encourages state-space exploration during online data aggregation, yielding higher-quality datasets with broader state coverage than standard DP. Consequently, Set-DP consistently outperforms DP across varied offline datasets and controller types, as the set-based constraints induce smoother action labels that improve overall generalization.

7.4. Conclusion

In summary, overcoming the limitations of strict point-wise behavioral cloning requires decoupling human intent from execution noise through structural, geometric, and temporal constraints. By reformulating human interventions as spatial penalties and leveraging sub-optimal data to establish baseline priors, algorithms can systematically extract invariant task objectives. Furthermore, scaling set-valued supervision via diffusion-based action chunking provides the mathematical foundation necessary to filter out kinematic noise, ensuring robust policy learning and generalization in high-dimensional continuous control domains.

Key Insights

Chapter 7: Structured Priors and Robust Reward Learning

- **Reward Extraction via RLIF:** Formulating interventions as discrete reward penalties rather than absolute kinematic targets bounds the suboptimality gap, providing a mathematical framework for the agent to outperform sub-optimal demonstrators.
- **Sample-Efficient Priors (SDP):** Regressing a reward model against pseudo-labeled, low-quality offline data (r_{min}) establishes a baseline structural prior, reducing sample complexity during interactive human querying.
- **Geometrically Constrained Action Space (CLIC):** Defining geometric action spaces ($\hat{\mathcal{A}}^{ss}$) relaxes point-wise imitation constraints, utilizing energy-based models to optimize policy probability mass over continuous regions to absorb execution noise.
- **Set Supervised Action-Chunking (Set-DP):** Integrating set-valued constraints with diffusion-based action chunking via reflected denoising resolves temporal ambiguity, scaling noise-tolerant intent extraction to high-dimensional continuous control spaces.

8

Discussion and Open Questions

Building upon the methodologies detailed in preceding chapters, ranging from offline inverse reinforcement learning (Chapter 4) and active preference-based evaluative feedback (Chapter 5), to interactive corrective interventions (Chapter 6) and set-valued geometric priors (Chapter 7), Chapter 8 synthesizes the analyzed paradigms. This chapter presents a comparative discussion of these frameworks and identifies open questions within the domain of human-in-the-loop policy learning. Specifically, it establishes a quantitative taxonomy of feedback modalities, evaluates the mathematical formulations of prevailing objective functions, and delineates critical open challenges regarding relative supervisory signals, algorithmic robustness to human sub-optimality, and reward learning efficiency.

8.1. Taxonomy and Trade-offs in Human Supervision

To systematically compare human-in-the-loop reward learning paradigms, this survey evaluates the underlying feedback mechanisms. Human supervisory effort has been studied extensively in the Human-Robot Interaction (HRI) literature [55, 56]. While several characteristics provide a useful basis for comparing human feedback paradigms, they are notoriously difficult to quantify consistently across different tasks and learning algorithms. Their values depend heavily on confounding factors such as task complexity, the learning framework, the design of the human-robot interface, and the supervisor's expertise. To the author's best knowledge, these characteristics have not been well studied in the realm of interactive imitation learning.

Consequently, rather than assigning absolute numerical values, this survey characterizes each feedback paradigm qualitatively using relative categories (e.g., *High*, *Medium*, and *Low*) across the following three dimensions:

- **Cost per Intervention (C):** The effort required from the human to provide a single feedback instance. This includes both the cognitive effort involved in interpreting the robot's behavior and deciding on an appropriate response, as well as any physical effort required to deliver the feedback.
- **Information Gain ($\mathcal{I}$):** The amount of useful learning signal conveyed by a single feedback instance. Different forms of feedback vary in how much information they provide to improve the learned policy or reward function.
- **Total Human Interventions (N):** The total number of feedback instances required from the human during the learning process until the policy converges or achieves the desired task performance. Feedback paradigms differ substantially in the amount of supervision they require throughout training.

[1] Although GAIL employs a discriminator-derived reward during training, its objective is occupancy measure matching rather than recovering an explicit reward function.

Table 8.1: Taxonomy of human feedback paradigms organized by learning objective: reward learning versus direct policy/controller learning.

Category	Feedback Type	Reward Learning	Direct Policy Learning
Passive Feedback			
P1	Demonstrations	Apprenticeship Learning [57], MaxEnt IRL [27], Deep MaxEnt IRL [58], AIRL [59], GCL [60]	BC [61], IBC [62], Diffusion Policy [63], GAIL[1] [31]
Active Robot Queries			
A1	Pairwise Preferences	Deep RLHF [15], PEBBLE [16], SURF [17], BRIDGE [41], FDPP [64]	DPO [45]
A1	Ranking Preferences	DemPref [39], LiRE [65]	–
Human-Initiated Evaluative Feedback			
A2	Evaluative	TAMER [66], Deep TAMER [67], RLIF [51]	COACH [68], Policy Shaping [69]
Human-Initiated Corrective Feedback			
A3	Relative Correction (Single-step)	–	Safe MPC Alignment [70]
A3	Absolute Correction (Single-step)	–	IBC [62]
A3	Absolute Correction (Multi-step)	–	DAgger [71], HG-DAgger [49], CLIC [53], SDP [54]

8.1.1. The Information Bandwidth versus Human Effort Trade-Off

The qualitative characteristics introduced in Table 8.1 reveal a fundamental trade-off between supervision quality and human effort. We approximate the overall supervisory burden as

$$\text{Total Human Effort} \approx C \times N$$

where C denotes the average cost per intervention and N denotes the total number of human interventions required throughout learning.

Paradigms with low intervention cost, such as evaluative feedback (A2) and pairwise preferences (A1), are easy for humans to provide but typically convey relatively coarse supervision. Consequently, many interactions are often required before an accurate reward model can be learned, particularly in high-dimensional continuous control tasks. At the opposite end of the spectrum, demonstrations (P1) provide rich supervision over complete trajectories and largely eliminate the need for online exploration. However, they require humans to generate successful demonstrations, resulting in substantial offline data collection effort and inheriting the limitations discussed in earlier chapters.

This trade-off suggests that human-initiated corrective feedback (A3) occupies an attractive yet largely unexplored region for reward learning. Unlike scalar evaluations or preference comparisons, corrective interventions specify the desired behavior precisely at states where the current policy fails, providing dense, localized supervision while avoiding the need to demonstrate entire trajectories. If such corrections could be converted into reward representations, they may enable more sample-efficient reward learning by reducing the amount of exploration required.

Nevertheless, several challenges remain.

- **Potential advantages.** Corrective actions provide explicit action-level supervision at failure states, tightly constraining the desired behavior. This richer supervision may reduce the number of human interventions required to learn an accurate reward function compared with lower-bandwidth feedback modalities.
- **Challenges.** Providing corrective actions generally requires real-time intervention through teleoperation or other continuous control interfaces, increasing the cost of each intervention. More

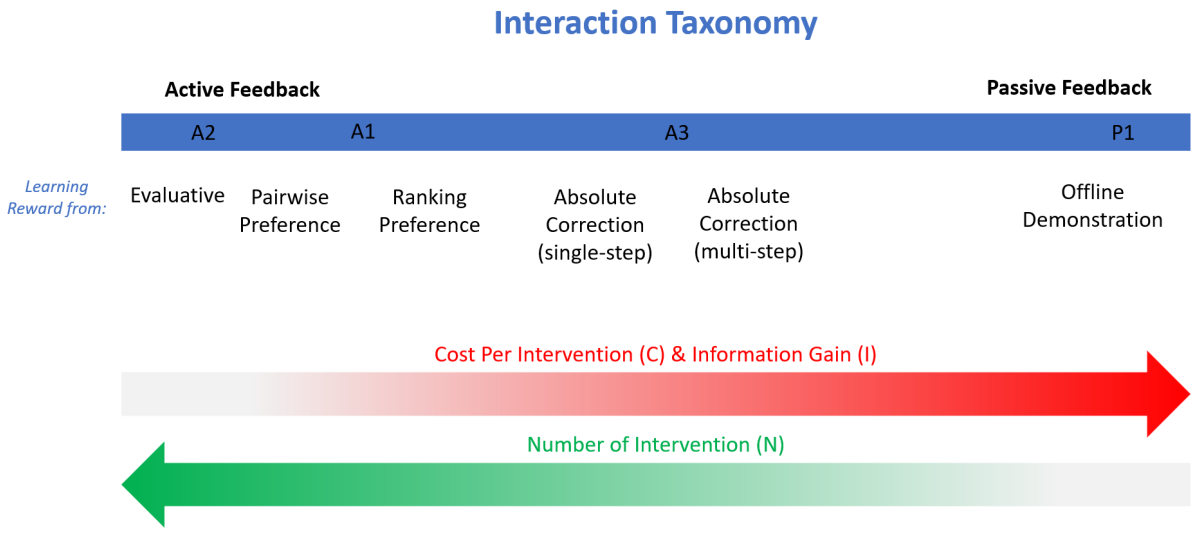

Figure 8.1: Spectrum of human feedback modalities, illustrating the relationship between per-intervention cost/information gain and the total number of required interventions. Appendix A.1 explains the rationale behind the qualitative taxonomy.

importantly, converting corrective actions into transferable reward representations remains an open algorithmic problem. Existing corrective feedback methods predominantly optimize policies or controller parameters directly, while learning reward functions that are robust to execution noise, delayed human reactions, and supervisor variability remains largely unexplored.

8.1.2. Reward Learning versus Direct Policy Learning

The type of human feedback alone does not determine the underlying learning objective. As summarized in Table 8.1, the same feedback modality can be used either to infer a reward function or to directly optimize a policy. Distinguishing these objectives is important because they differ fundamentally in the learned representation, its reusability, and the role of human supervision.

Reward learning seeks to infer an explicit or implicit reward representation from human feedback, after which a planning or reinforcement learning algorithm optimizes the resulting policy. By separating reward inference from policy optimization, the learned reward can potentially be reused across different policies, environments, or planning algorithms. In contrast, direct policy learning bypasses reward modeling and directly updates the policy or controller from human supervision, treating demonstrations, preferences, or corrective interventions as supervisory signals for action generation.

This distinction is reflected throughout the literature. Demonstrations have been used both for inverse reinforcement learning and behavior cloning, while pairwise preferences support either reward learning (e.g., Deep RLHF [15] and PEBBLE [16]) or direct policy optimization (e.g., DPO [45]). In contrast, human-initiated corrective feedback has been employed almost exclusively for direct policy or controller adaptation, with methods such as DAgger [36], HG-DAgger [49], and Safe MPC Alignment [70] directly optimizing policies from corrective actions rather than inferring transferable reward representations.

The limited adoption of corrective feedback for reward learning does not arise from a lack of informative supervision. On the contrary, corrective interventions provide rich action-level information precisely at policy failures. The challenge lies in converting these corrections into generalized reward representations that capture the supervisor's intent while remaining robust to execution-specific artifacts, such as delayed reactions, teleoperation noise, and imperfect recovery motions. Consequently, reward learning from corrective feedback remains a largely unexplored direction with the potential to combine the rich supervision of corrective interventions with the flexibility and generalization afforded by learned reward functions.

8.1.3. Open Question 1: Relative versus Anchored Preference Supervision

Standard preference-based reward learning predominantly relies on robot-vs-robot comparisons, where a human supervisor selects the better of two segments generated by the current policy. While this pairwise feedback (A1) is cognitively lightweight, it provides a purely relative supervisory signal. Because the human only indicates the gradient direction of the reward without providing an absolute reference point, the learning algorithm must exhaustively query the supervisor to map the contours of the reward landscape.

This structural limitation raises a critical open question regarding sample complexity. Theoretical arguments suggest that purely relative preference feedback requires an inherently large number of samples due to the lack of an absolute kinematic reference. It remains an open area of investigation whether introducing an "anchored" supervisory signal such as robot-vs-human comparisons, where policy rollouts are judged against an optimal expert demonstration, can significantly reduce the exploration space. Understanding the mathematical and empirical trade-offs between standard relative queries and anchored preference supervision is essential for improving the data efficiency of interactive learning.

8.1.4. Open Question 2: Can Corrective Actions Improve Reward Learning Efficiency?

As discussed in Section 8.1.2, human-initiated corrective actions (A3) provide dense, step-wise supervision exactly where the policy fails. By supplying explicit kinematic targets, corrections theoretically bypass the prolonged exploration bottleneck associated with evaluative and preference-based feedback.

Despite this theoretical advantage, the interactive learning literature lacks comprehensive benchmarks comparing the sample efficiency of reward extraction from corrective actions against standard preference learning. An open research question is whether the high information bandwidth of corrective actions translates into empirically faster convergence speeds during reward learning. Rigorously evaluating if, and to what extent, incorporating corrective feedback reduces the total number of human interventions required to learn an optimal policy remains a vital step toward deploying scalable, human-in-the-loop robotic systems.

8.1.5. Open Question 3: Robustness under Sub-Optimal Human Feedback

Interactive feedback is fundamentally shaped by the physical and cognitive limitations of the human supervisor. During real-time continuous control, human corrections are rarely optimal; they are frequently corrupted by teleoperation jitter and cognitive reaction delays.

When extracting rewards from corrective actions, a major open question is how this sub-optimal human feedback impacts the learned reward landscape. Treating noisy corrections as flawless supervised targets often leads to severe kinematic overfitting, where the policy replicates the supervisor's physical execution errors rather than their underlying semantic intent. It remains to be investigated how sensitive modern reward learning algorithms are to varying degrees of human sub-optimality. Furthermore, identifying whether injecting structural priors, specifically geometric constraints that enforce smoothness, symmetry, or known spatial boundaries, can robustly isolate the true human intent from physical execution noise is a critical frontier for ensuring safe and reliable interactive learning.

Key Insights

Chapter 8: Discussion and Open Questions

- **Information vs. Effort Trade-off:** Human supervision navigates a fundamental tension between information gain ($\mathcal{I}$) and total human effort ($C \times N$). High-bandwidth modalities drastically reduce the required intervention volume (N) but impose higher transient per-intervention costs (C).
- **Reward vs Direct Policy Learning:** While demonstrations and preferences are widely utilized for reward learning, corrective feedback is predominantly restricted to direct policy optimization (e.g., DAgger, CLIC). This bypasses the potential of corrections to learn robust, transferable reward representations.
- **Relative vs. Anchored Feedback:** Purely relative preferences structurally lack absolute kinematic references, increasing sample complexity. It remains an open question whether introducing anchored supervisory signals can effectively constrain the exploration space.
- **Corrective Reward Efficiency:** Despite theoretical advantages, benchmarking the empirical sample efficiency of extracting rewards from dense corrective actions against standard relative preference learning remains a critical gap in the literature.
- **Robustness to Sub-Optimality:** Translating noisy, non-Markovian corrections into robust reward signals without kinematically overfitting to human execution errors requires novel algorithmic solutions, highlighting the need to investigate geometric constraints and structural priors.

9

Thesis Formulation

9.1. Problem Statement

Standard Preference-based reward learning methods rely on relative comparisons (i.e., $a_i \succ a_j$) without any notion of potential optimal value. As a result, each query constrains the reward function weakly, leading to a large hypothesis space of consistent reward functions and low sample efficiency. Additionally, this is exacerbated when the human feedback is noisy.

Therefore, this research aims to bridge this gap by developing a structured preference learning framework that integrates corrective actions to improve sample efficiency, while utilizing geometric constraints to ensure algorithmic robustness against human sub-optimality.

9.2. Research Questions

To systematically evaluate the efficacy and robustness of the proposed framework, this thesis investigates the following research questions:

- **RQ1:** Can corrective-action-based feedback improve sample efficiency compared to standard preference learning?
- **RQ2:** How does sub-optimal human feedback impact reward learning?
- **RQ3:** Do geometric constraints improve robustness to noisy corrections?

9.3. Hypotheses

Corresponding to the formulated research questions, this work proposes and empirically evaluates the following core hypotheses:

- **H1:** Relative preference feedback requires more samples due to the lack of an absolute reference.
- **H2:** Incorporating corrective actions provides stronger supervision than standard preference queries from a robot policy, thus improving convergence speed.
- **H3:** Geometric constraints improve robustness to sub-optimal feedback.

Figure 9.1 illustrates this proposed methodology.

9.4. Broader Impact of the Research

By investigating whether corrective actions accelerate convergence relative to standard preference queries (RQ1, H2), this work mitigates both sample inefficiency and supervisory workload. High-bandwidth corrective anchors minimize required human queries while rapidly establishing safe directional gradients, thereby preventing compounding execution errors and dangerous excursions into out-of-distribution state spaces.

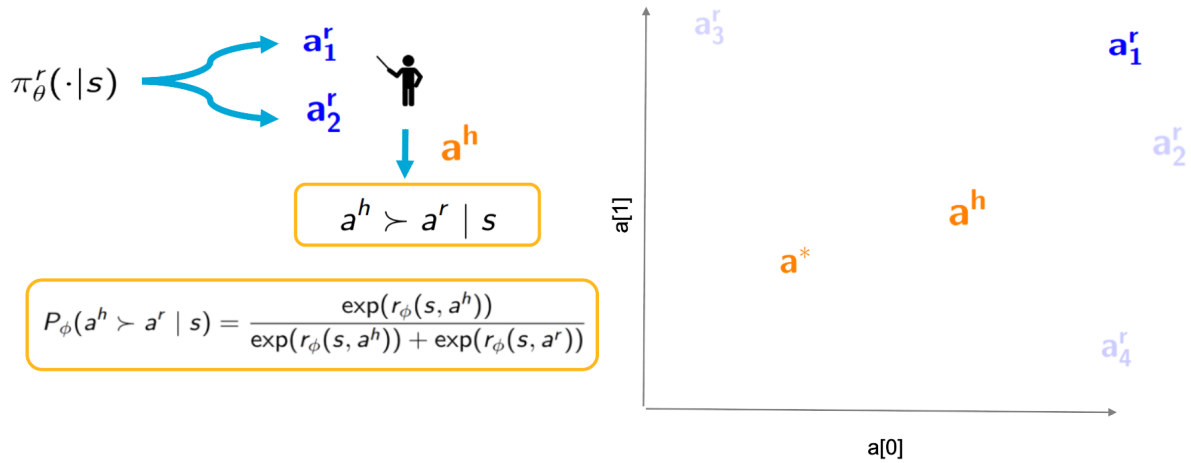

Figure 9.1: Proposed methodology: modeling human corrective actions (a^h) as anchored preferences over sub-optimal robot actions (a^r) to provide absolute spatial guidance toward the optimal action (a^*).

Furthermore, analyzing the impact of sub-optimal feedback (RQ2) and introducing geometric constraints to enhance algorithmic robustness (RQ3, H3) confronts the inevitability of human cognitive fatigue. By relaxing the assumption of a flawless supervisor, encoding task-relevant invariances isolates human intent from noise in physical execution. Establishing this structural robustness is critical to ensure that learned reward functions generalize reliably under distributional shifts without kinematically overfitting to flawed human guidance.

10

Conclusion

This report has systematically reviewed the foundational challenges of aligning reinforcement learning agents with complex human intent. As highlighted throughout the literature, the manual specification of reward functions is inherently brittle, often leading to objective misalignment and specification gaming. While offline imitation learning techniques mitigate these vulnerabilities by extracting rewards from expert demonstrations, their reliance on static, high-cost datasets necessitates a transition toward dynamic, human-in-the-loop interactive paradigms.

The central tension identified in this survey lies in the fundamental trade-off between information bandwidth and human supervisory effort. Active evaluative methods, particularly preference-based reward learning, impose low cognitive burdens but suffer from severe sample inefficiency due to their reliance on purely relative, low-bandwidth signals. Conversely, human-initiated corrective interventions supply dense, localized kinematic targets that effectively bypass the algorithmic exploration bottleneck. However, the prevailing literature has predominantly restricted corrective feedback to direct policy optimization. This methodological divide leaves policies highly susceptible to compounding execution errors, forcing algorithms to kinematically overfit to sub-optimal human execution noise rather than acquiring a generalized, transferable reward representation.

To safely scale interactive learning, the computational paradigm must shift from direct behavioral cloning to robust reward extraction. Synthesizing the high-bandwidth supervision of corrective actions with the structural regularity of geometric priors offers a mathematically principled pathway to isolate a supervisor's underlying semantic intent from the transient artifacts of physical teleoperation.

Ultimately, the open questions identified in this review establish the foundation for the proposed thesis. By rigorously evaluating the sample efficiency of anchored corrective feedback relative to standard preferences, and by investigating the capacity of geometric constraints to regularize learning against sub-optimal human interventions, this research aims to address a critical gap in the interactive learning literature. Resolving these challenges will directly contribute to the development of structured, sample-efficient reward learning frameworks, thereby facilitating the deployment of adaptable, safe, and aligned autonomous systems in complex physical environments.

References

[1] Volodymyr Mnih et al. “Playing Atari with Deep Reinforcement Learning”. In: *arXiv preprint arXiv:1312.5602* (2013).

[2] Jens Kober and Jan R. Peters. “Policy Search for Motor Primitives in Robotics”. In: *Advances in Neural Information Processing Systems 21 (NeurIPS 2008)*. Ed. by Daphne Koller et al. Curran Associates, Inc., 2008, pp. 849–856. URL: https://proceedings.neurips.cc/paper/2008/file/7647966b7343c29048673252e490f736-Paper.pdf.

[3] Jianlan Luo et al. “Precise and Dexterous Robotic Manipulation via Human-in-the-Loop Reinforcement Learning”. In: *arXiv preprint arXiv:2410.21845* (2024). arXiv: 2410.21845 [cs.RO]. URL: https://hil-serl.github.io/static/hil-serl-paper.pdf.

[4] Sinan Ibrahim et al. *Comprehensive Overview of Reward Engineering and Shaping in Advancing Reinforcement Learning Applications*. 2024. arXiv: 2408.10215 [cs.LG]. URL: https://arxiv.org/abs/2408.10215.

[5] Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. “Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping”. In: *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*. Morgan Kaufmann, 1999, pp. 278–287.

[6] Andrew Y. Ng and Stuart J. Russell. “Algorithms for Inverse Reinforcement Learning”. In: *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*. Morgan Kaufmann, 2000, pp. 663–670. ISBN: 1-55860-707-2.

[7] Baohe Zhang et al. *On the Importance of Hyperparameter Optimization for Model-based Reinforcement Learning*. 2021. arXiv: 2102.13651 [cs.LG]. URL: https://arxiv.org/abs/2102.13651.

[8] Carlos Celemin et al. “Interactive Imitation Learning in Robotics: A Survey”. In: *arXiv preprint arXiv:2211.00600* (2022).

[9] Viola Zhou and Kinling Lo. *In Chinese data factories, workers teach humanoid robots boring tasks*. Rest of World. Jan. 2026. URL: https://restofworld.org/2026/china-robots-training-centers-workers/ (visited on 07/06/2026).

[10] Stephen Casper et al. “Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback”. In: *Transactions on Machine Learning Research* (2023).

[11] Yuke Zhu et al. “RoboSuite: A Modular Simulation Framework and Benchmark for Robot Learning”. In: *Proceedings of the 4th Conference on Robot Learning (CoRL)*. 2020.

[12] Ajay Mandlekar et al. “RoboMimic: A Benchmark for Robot Learning from Demonstration”. In: *Proceedings of the 5th Conference on Robot Learning (CoRL)*. 2021.

[13] Tianhe Yu et al. “Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning”. In: *Proceedings of the Conference on Robot Learning (CoRL)*. 2020.

[14] Long Ouyang et al. *Training language models to follow instructions with human feedback*. 2022. arXiv: 2203.02155 [cs.CL]. URL: https://arxiv.org/abs/2203.02155.

[15] Paul Christiano et al. *Deep reinforcement learning from human preferences*. 2023. arXiv: 1706.03741 [stat.ML]. URL: https://arxiv.org/abs/1706.03741.

[16] Kimin Lee, Laura Smith, and Pieter Abbeel. “PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training”. In: *arXiv preprint arXiv:2106.05091* (2021).

[17] Jongjin Park et al. *SURF: Semi-supervised Reward Learning with Data Augmentation for Feedback-efficient Preference-based Reinforcement Learning*. 2022. arXiv: 2203.10050 [cs.LG]. URL: https://arxiv.org/abs/2203.10050.

[18] Zhaoting Li et al. *From Action Labels to Sets: Rethinking Action Supervision for Imitation Learning from Corrective Feedback*. 2026. arXiv: 2502.07645 [cs.RO]. URL: https://arxiv.org/abs/2502.07645.

[19] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Second. The MIT Press, 2018. URL: http://incompleteideas.net/book/the-book-2nd.html.

[20] Victoria Krakovna. *Specification Gaming Examples in AI*. Google Sheets. Accessed: 2026-06-21. 2018. URL: https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml.

[21] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. *The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models*. 2022. arXiv: 2201.03544 [cs.LG]. URL: https://arxiv.org/abs/2201.03544.

[22] Wikipedia contributors. *Goodhart's law — Wikipedia, The Free Encyclopedia*. Accessed: 2026-06-10. 2026. URL: https://en.wikipedia.org/wiki/Goodhart%27s_law.

[23] Embodiment Collaboration et al. *Open X-Embodiment: Robotic Learning Datasets and RT-X Models*. 2025. arXiv: 2310.08864 [cs.RO]. URL: https://arxiv.org/abs/2310.08864.

[24] Alexander Khazatsky et al. *DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset*. 2025. arXiv: 2403.12945 [cs.RO]. URL: https://arxiv.org/abs/2403.12945.

[25] Homer Walke et al. *BridgeData V2: A Dataset for Robot Learning at Scale*. 2024. arXiv: 2308.12952 [cs.RO]. URL: https://arxiv.org/abs/2308.12952.

[26] Pieter Abbeel and Andrew Y. Ng. "Apprenticeship learning via inverse reinforcement learning". In: *Proceedings of the Twenty-First International Conference on Machine Learning*. ICML '04. Banff, Alberta, Canada: Association for Computing Machinery, 2004, p. 1. ISBN: 1581138385. DOI: 10.1145/1015330.1015430. URL: https://doi.org/10.1145/1015330.1015430.

[27] Brian D. Ziebart et al. "Maximum Entropy Inverse Reinforcement Learning". In: *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI-08)*. AAAI Press, 2008, pp. 1433–1438.

[28] Faraz Torabi, Garrett Warnell, and Peter Stone. *Behavioral Cloning from Observation*. 2018. arXiv: 1805.01954 [cs.AI]. URL: https://arxiv.org/abs/1805.01954.

[29] Haozhuo Li, Yuchen Cui, and Dorsa Sadigh. "How to Train Your Robots? The Impact of Demonstration Modality on Imitation Learning". In: *CoRR* abs/2503.07017 (2025). DOI: 10.48550/arXiv.2503.07017. URL: https://arxiv.org/abs/2503.07017.

[30] Saurabh Arora and Prashant Doshi. "A Survey of Inverse Reinforcement Learning: Challenges, Methods and Progress". In: *Artificial Intelligence* 297 (2021), p. 103500. ISSN: 0004-3702. DOI: 10.1016/j.artint.2021.103500.

[31] Jonathan Ho and Stefano Ermon. *Generative Adversarial Imitation Learning*. 2016. arXiv: 1606.03476 [cs.LG]. URL: https://arxiv.org/abs/1606.03476.

[32] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML]. URL: https://arxiv.org/abs/1406.2661.

[33] Justin Fu, Katie Luo, and Sergey Levine. *Learning Robust Rewards with Adversarial Inverse Reinforcement Learning*. 2018. arXiv: 1710.11248 [cs.LG]. URL: https://arxiv.org/abs/1710.11248.

[34] Moo Jin Kim, Jiajun Wu, and Chelsea Finn. "Giving Robots a Hand: Learning Generalizable Manipulation with Eye-in-Hand Human Video Demonstrations". In: *arXiv preprint arXiv:2307.05959* (2023). DOI: 10.48550/arXiv.2307.05959. arXiv: 2307.05959 [cs.RO]. URL: https://arxiv.org/abs/2307.05959.

[35] Aakash Yadav et al. "Enhancing Trust Examinations with Neural Measures during Human–Robot Collaboration under Cognitive Fatigue". In: *J. Hum.-Robot Interact.* 15.2 (Dec. 2025). DOI: 10.1145/3773905. URL: https://doi.org/10.1145/3773905.

[36] Stephane Ross, Geoffrey Gordon, and Drew Bagnell. “A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning”. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 15. Proceedings of Machine Learning Research. PMLR, 2011, pp. 627–635. URL: https://proceedings.mlr.press/v15/ross11a.html.

[37] Sergey Levine. *Supervised Learning of Behaviors*. Lecture 2 slides, CS 185/285: Deep Reinforcement Learning, University of California, Berkeley. Accessed: 2026-07-06. 2024. URL: https://rail.eecs.berkeley.edu/deeprlcourse/static/slides/lec-2.pdf.

[38] Marvin Minsky. “Steps Toward Artificial Intelligence”. In: *Proceedings of the IRE* 49.1 (Jan. 1961), pp. 8–30. DOI: 10.1109/JRPROC.1961.287775. URL: https://courses.csail.mit.edu/6.803/pdf/steps.pdf.

[39] Erdem Bıyık et al. *Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences*. 2021. arXiv: 2006.14091 [cs.RO]. URL: https://arxiv.org/abs/2006.14091.

[40] Joey Hejna and Dorsa Sadigh. *Few-Shot Preference Learning for Human-in-the-Loop RL*. 2022. arXiv: 2212.03363 [cs.RO]. URL: https://arxiv.org/abs/2212.03363.

[41] Maël Macuglia, Paul Friedrich, and Giorgia Ramponi. *Fine-tuning Behavioral Cloning Policies with Preference-Based Reinforcement Learning*. 2025. arXiv: 2509.26605 [cs.AI]. URL: https://arxiv.org/abs/2509.26605.

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

[43] John Schulman et al. *Proximal Policy Optimization Algorithms*. 2017. arXiv: 1707.06347 [cs.LG]. URL: https://arxiv.org/abs/1707.06347.

[44] Tuomas Haarnoja et al. *Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor*. 2018. arXiv: 1801.01290 [cs.LG]. URL: https://arxiv.org/abs/1801.01290.

[45] Rafael Rafailov et al. *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*. 2024. arXiv: 2305.18290 [cs.LG]. URL: https://arxiv.org/abs/2305.18290.

[46] Dean A. Pomerleau. “ALVINN: an autonomous land vehicle in a neural network”. In: *Advances in Neural Information Processing Systems 1*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1989, pp. 305–313. ISBN: 1558600159.

[47] Sham M. Kakade and John Langford. “Approximately Optimal Approximate Reinforcement Learning”. In: *Proceedings of the Nineteenth International Conference on Machine Learning (ICML 2002)*. San Francisco, CA, USA: Morgan Kaufmann, 2002, pp. 267–274. ISBN: 1-55860-873-7.

[48] Kazuoki Azuma. “Weighted Sums of Certain Dependent Random Variables”. In: *Tohoku Mathematical Journal* 19.3 (1967), pp. 357–367. DOI: 10.2748/tmj/1178243286.

[49] Michael Kelly et al. *HG-DAgger: Interactive Imitation Learning with Human Experts*. 2019. arXiv: 1810.02890 [cs.RO]. URL: https://arxiv.org/abs/1810.02890.

[50] Volkan Cevher. *Lecture 6: Imitation Learning*. Lecture slides, EE-568: Reinforcement Learning. Spring 2024, accessed 2026-07-06. Laboratory for Information and Inference Systems (LIONS), 2024. URL: https://www.epfl.ch/labs/lions/wp-content/uploads/2024/08/Lecture-6_-Imitation-Learning.pdf.

[51] Jianlan Luo et al. “RLIF: Interactive Imitation Learning as Reinforcement Learning”. In: *The Twelfth International Conference on Learning Representations (ICLR)*. 2024. URL: https://arxiv.org/abs/2311.12996.

[52] Ajay Mandlekar et al. “Leveraging Sub-Optimal Data for Human-in-the-Loop Reinforcement Learning”. In: *The Twelfth International Conference on Learning Representations (ICLR)*. 2024. URL: https://openreview.net/forum?id=0z9B4f2s9K.

[53] Zhaoting Li et al. “From Action Labels to Sets: Rethinking Action Supervision for Imitation Learning from Corrective Feedback”. In: *arXiv preprint arXiv:2502.07645* (2025). arXiv: 2502.07645 [cs.RO]. URL: https://arxiv.org/abs/2502.07645.

[54] Zhaoting Li et al. *Set-Supervised Diffusion Policy: Learning Action-Chunking Diffusion through Corrections*. 2026. arXiv: 2606.01865 [cs.RO]. URL: https://arxiv.org/abs/2606.01865.

[55] Michael A. Goodrich and Alan C. Schultz. *Human-Robot Interaction: A Survey*. Hanover, MA, USA: Now Publishers Inc., 2008. ISBN: 1601980922.

[56] Thomas B Sheridan. “Human–robot interaction: status and challenges”. In: *Human Factors* 58.4 (2016), pp. 525–532.

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

[58] Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. *Maximum Entropy Deep Inverse Reinforcement Learning*. 2016. arXiv: 1507.04888 [cs.LG]. URL: https://arxiv.org/abs/1507.04888.

[59] Justin Fu, Katie Luo, and Sergey Levine. “Learning Robust Rewards with Adversarial Inverse Reinforcement Learning”. In: *International Conference on Learning Representations (ICLR)*. 2018. URL: https://arxiv.org/abs/1710.11248.

[60] Chelsea Finn, Sergey Levine, and Pieter Abbeel. *Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization*. 2016. arXiv: 1603.00448 [cs.LG]. URL: https://arxiv.org/abs/1603.00448.

[61] Michael Bain and Claude Sammut. “A framework for behavioural cloning”. In: *Machine Intelligence 15*. Oxford University Press, 1999, pp. 103–129.

[62] Pete Florence et al. *Implicit Behavioral Cloning*. 2021. arXiv: 2109.00137 [cs.RO]. URL: https://arxiv.org/abs/2109.00137.

[63] Cheng Chi et al. *Diffusion Policy: Visuomotor Policy Learning via Action Diffusion*. 2024. arXiv: 2303.04137 [cs.RO]. URL: https://arxiv.org/abs/2303.04137.

[64] Yuxin Chen et al. *FDPP: Fine-tune Diffusion Policy with Human Preference*. 2025. arXiv: 2501.08259 [cs.RO]. URL: https://arxiv.org/abs/2501.08259.

[65] Heewoong Choi et al. *Listwise Reward Estimation for Offline Preference-based Reinforcement Learning*. 2024. arXiv: 2408.04190 [cs.LG]. URL: https://arxiv.org/abs/2408.04190.

[66] W. Bradley Knox and Peter Stone. “Interactively Shaping Agents via Human Reinforcement: The TAMER Framework”. In: *Proceedings of the Fifth International Conference on Knowledge Capture (K-CAP)*. Redondo Beach, CA, USA, 2009, pp. 9–16.

[67] Garrett Warnell et al. “Deep TAMER: Interactive Agent Shaping in High-Dimensional State Spaces”. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. 2018, pp. 109–116. URL: https://aaai.org/papers/11485-deep-tamer-interactive-agent-shaping-in-high-dimensional-state-spaces/.

[68] James MacGlashan et al. *Interactive Learning from Policy-Dependent Human Feedback*. 2023. arXiv: 1701.06049 [cs.AI]. URL: https://arxiv.org/abs/1701.06049.

[69] Shane Griffith et al. “Policy Shaping: Integrating Human Feedback with Reinforcement Learning”. In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/e034fb6b66aacc1d48f445ddfb08da98-Paper.pdf.

[70] Zhixian Xie et al. *Safe MPC Alignment with Human Directional Feedback*. 2025. arXiv: 2407.04216 [cs.RO]. URL: https://arxiv.org/abs/2407.04216.

[71] Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. *A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning*. 2011. arXiv: 1011.0686 [cs.LG]. URL: https://arxiv.org/abs/1011.0686.

A

Appendix

A.1. Rationale Behind the Qualitative Taxonomy

While the taxonomy presented in Table 8.1 categorizes human feedback paradigms by their learning objective and interaction modality, evaluating their underlying structural trade-offs requires an analysis across three qualitative dimensions: cost per intervention, information gain, and total human interventions. Since no standardized mathematical framework currently exists for comparing these quantities universally across diverse human feedback paradigms, this appendix defines the qualitative rationale used to assess these dimensions.

A.1.1. Cost per Intervention

As summarized in Table A.1, cost per intervention reflects the amount of human effort required to provide a single feedback instance. Rather than attempting to quantify cognitive workload directly, this comparison considers the structural volume of supervision that the human operator must generate for each interaction.

Table A.1: Relative supervision required for a single feedback instance.

Feedback	Human Input
Evaluative feedback	One scalar label
Pairwise preference	One binary comparison
Ranking preference	Ordering of multiple candidates
Single-step correction	One action (possibly multi-DoF)
Multi-step correction	Multiple actions (possibly multi-DoF)
Demonstration	Full trajectory

Accordingly, paradigms requiring supervision over longer temporal horizons or higher-dimensional control (e.g., multi-step corrections and demonstrations) impose a significantly higher intervention cost compared to simple scalar evaluations.

A.1.2. Information Gain

As outlined in Table A.2, information gain is interpreted as the granularity of supervision conveyed by a single feedback instance. Rather than measuring strict information-theoretic quantities (e.g., mutual information), this comparison evaluates how precisely each feedback modality geometrically constrains the desired behavior.

Feedback modalities that specify desired actions directly over longer horizons (such as corrective interventions) generally provide much richer, absolute spatial supervision than those expressing only relative or evaluative judgments.

Table A.2: Relative granularity of supervision conveyed by one feedback instance.

Feedback	Supervision Granularity
Evaluative feedback	Coarse scalar assessment
Pairwise preference	Relative ordering constraint
Ranking preference	Multiple ordering constraints
Single-step correction	Desired action at one state
Multi-step correction	Desired action at multiple states
Demonstration	Desired actions over an entire trajectory

A.1.3. Total Human Interventions

Total human interventions represent the overall volume of human participation required throughout the learning process to achieve policy convergence. Unlike the previous two static characteristics, this dynamic quantity depends heavily on the specific learning algorithm, task complexity, exploration strategy, and the supervisor's sub-optimality.

Consequently, no universal numerical ordering exists in the literature. However, an inherent inverse relationship dictates this dimension: modalities with lower information gain (Table A.2) structurally require a vastly higher frequency of total interventions to constrain the reward landscape. In contrast, passive demonstrations (P1) are collected entirely offline prior to learning, fundamentally separating their collection burden from the active, online intervention frequency required by evaluative (A1, A2) and corrective (A3) paradigms.

A.1.4. Limitations

This qualitative rationale is intended as an organizational framework for comparing structural trade-offs in human feedback paradigms rather than a formal, computable evaluation metric. The relative characteristics summarize recurring trends reported across the interactive learning literature and should not be interpreted as universally absolute for every robotic task or algorithmic implementation. Future work establishing standardized, quantitative benchmarking metrics for human supervisory effort and feedback informativeness will be necessary to enable more rigorous comparisons across human-in-the-loop learning paradigms.