Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.072146 Published Online: 21 November 2025

REVIEW

Implementation of Human-AI Interaction in Reinforcement Learning: Literature Review and Case Studies

Shaoping Xiao[1,*], Zhaoan Wang[1], Junchao Li[2], Caden Noeller[1], Jiefeng Jiang[3] and Jun Wang[4]

[1]Department of Mechanical Engineering, Iowa Technology Institute, The University of Iowa, Iowa City, IA 52242, USA
[2]Talus Renewables, Inc., Austin, TX 78754, USA
[3]Department of Psychological and Brain Sciences, Iowa Neuroscience Institute, The University of Iowa, Iowa City, IA 52242, USA
[4]Department of Chemical and Biochemical Engineering, Iowa Technology Institute, The University of Iowa, Iowa City, IA 52242, USA

*Corresponding Author: Shaoping Xiao. Email: shaoping-xiao@uiowa.edu

Received: 20 August 2025; Accepted: 25 October 2025

ABSTRACT: The integration of human factors into artificial intelligence (AI) systems has emerged as a critical research frontier, particularly in reinforcement learning (RL), where human-AI interaction (HAII) presents both opportunities and challenges. As RL continues to demonstrate remarkable success in model-free and partially observable environments, its real-world deployment increasingly requires effective collaboration with human operators and stakeholders. This article systematically examines HAII techniques in RL through both theoretical analysis and practical case studies. We establish a conceptual framework built upon three fundamental pillars of effective human-AI collaboration: computational trust modeling, system usability, and decision understandability. Our comprehensive review organizes HAII methods into five key categories: (1) learning from human feedback, including various shaping approaches; (2) learning from human demonstration through inverse RL and imitation learning; (3) shared autonomy architectures for dynamic control allocation; (4) human-in-the-loop querying strategies for active learning; and (5) explainable RL techniques for interpretable policy generation. Recent state-of-the-art works are critically reviewed, with particular emphasis on advances incorporating large language models in human-AI interaction research. To illustrate some concepts, we present three detailed case studies: an empirical trust model for farmers adopting AI-driven agricultural management systems, the implementation of ethical constraints in robotic motion planning through human-guided RL, and an experimental investigation of human trust dynamics using a multi-armed bandit paradigm. These applications demonstrate how HAII principles can enhance RL systems' practical utility while bridging the gap between theoretical RL and real-world human-centered applications, ultimately contributing to more deployable and socially beneficial intelligent systems.

KEYWORDS: Human-AI interaction; reinforcement learning; partially observable environments; trust model; ethical constraints

1 Introduction

In the era of artificial intelligence (AI), its technologies have been increasingly applied across a wide range of science and engineering domains, including cyber-physical systems (CPS), materials science, hydrology, agriculture, and manufacturing [1–5]. This broad integration is largely driven by the pursuit of greater efficiency in prediction and recommendation, leveraging data-driven insights to solve complex problems. A key research goal within this context is the development of AI-driven autonomous systems

capable of operating independently and making complex decisions without direct human supervision. These systems hold the potential to revolutionize industries by enabling faster responses, reducing human errors, and maintaining continuous operation in dynamic or hazardous environments [6].

Among the various AI approaches, reinforcement learning (RL) stands out as one of the most advanced and influential techniques for enabling autonomy [7]. RL empowers systems to learn complex behavior patterns through trial-and-error interactions with their environments. Unlike supervised learning, which relies on labeled data, RL is driven by reward signals that guide an intelligent agent to take actions aimed at maximizing long-term cumulative rewards. Through this iterative learning process, the agent gradually develops strategies, referred to as policies in the RL framework, to inform which actions to take under various environmental conditions.

These learned policies function as decision-making blueprints, allowing the agent to select optimal actions as circumstances change. As environments evolve or increase in complexity, RL agents continuously refine their policies to improve performance. This makes RL particularly well-suited for real-time, adaptive tasks such as autonomous driving, robotic control, and dynamic resource management [8–11]. Furthermore, RL's capacity to manage sequential decisions under uncertainty positions it as a powerful tool for applications where long-term consequences must be carefully considered.

However, while the capabilities of autonomous systems continue to grow, they raise important ethical and philosophical questions [12]. Can machines be entrusted to make decisions solely with moral implications? Should AI systems act autonomously in scenarios involving human welfare, justice, or environmental impact? These concerns are particularly pressing given that intelligent agents, whether algorithms, robots, or other computational systems, are neither legally nor morally responsible. Unlike humans, they do not possess consciousness, intent, or empathy, and thus cannot be held accountable in traditional ways. Consequently, the responsibility for their behaviors must ultimately rest with the humans who design, develop, and deploy them. Addressing these questions is critical to ensuring that AI technologies are not only technically effective but also ethically grounded and socially responsible.

Furthermore, these ethical concerns become even more pronounced when AI-based systems operate autonomously or semi-autonomously, shifting decision-making authority from humans to machines. This transformation reshapes human-technology interaction and introduces new ethical challenges, including questions of distributive justice, risks of discrimination and exclusion, and transparency issues stemming from the perceived opacity and unpredictability of AI systems [13,14]. Building on these shifts in autonomy and cognitive capabilities, recent research highlights trust, usability, and understandability as three fundamental pillars of human-AI interaction (HAII), offering potential pathways to address these ethical challenges [15]. These aspects are closely tied to the aforementioned ethical considerations and play a pivotal role in shaping how people perceive, adopt, and collaborate with AI systems.

Trust is widely regarded as a cornerstone for effective HAII and the successful adoption of AI systems. It can be broadly defined as the willingness of individuals to accept and rely on AI-generated suggestions or decisions, and to engage in task-sharing or information exchange with these technologies. High or appropriately calibrated trust is essential: when trust is too low, users may avoid using AI systems even when those systems offer substantial benefits. Conversely, excessive or misplaced trust may lead to over-reliance and critical judgment errors. In many high-stakes domains such as healthcare, finance, autonomous driving, and customer service, distrust remains a major barrier to adoption. This often stems from the so-called "black-box" nature of AI models, which makes their internal logic opaque and their behavior difficult to anticipate or validate. As a result, users may perceive AI decisions as arbitrary or unreliable, undermining confidence in the system. On the other hand, establishing and maintaining trust can significantly accelerate AI adoption and integration. When users perceive AI systems as transparent, reliable, and aligned with their

goals, they are more likely to engage with and benefit from these technologies. Ultimately, the full societal value of AI can only be realized when individuals, organizations, and communities trust these systems enough to meaningfully interact with and atop them [16].

The second pillar of HAII is usability. Even the most intelligent AI system may fail to deliver its value if it is not usable or if humans find it difficult to understand, interact with, or incorporate it into their workflow. In traditional human–machine interaction research, usability has often been assessed through interface design and task performance metrics such as efficiency, accuracy, and error rates [17]. However, in collaborative human–AI environments, the concept of usability extends well beyond interface-level interactions. It encompasses the AI system's ability to adapt to the human partner, responding dynamically to the users' needs, preferences, goals, and context. An AI system that can adjust its behavior, level of autonomy, or recommendations in real time contributes more effectively to human performance and decision-making. This perspective represents a shift from viewing AI as a tool to reviewing it as a teammate, a symbiotic partner working toward shared goals. Importantly, an adaptive AI that is trainable or context-aware can better align its assistance with the user's workflow. For instance, by learning from a human's decision-making patterns, the system can provide more timely and context-relevant support. Such alignment has been shown to not only improve team efficiency and effectiveness, but also to enhance the human user's sense of agency and competence [18]. Recent studies confirm that working with an adaptive AI collaborator, as opposed to a static assistant, leads to higher task success and greater user satisfaction [19].

The third crucial facet of HAII is understandability, often referred to as explainability or XAI (explainable AI). This concept is deeply intertwined with the other pillars of trust and usability, as it directly influences a user's ability to comprehend, evaluate, and confidently act on AI-generated outputs. At its core, explainability refers to an AI system's capacity to communicate the rationale behind its decision in ways that are interpretable and meaningful to humans. This capability is particularly important given a persistent trade-off in machine learning (ML): as models become more accurate and powerful, especially in the case of deep learning (DL), they also tend to become more opaque and complex. As a result, it is difficult to discern how specific outcomes are predicted and recommended [20]. When users and stakeholders can understand the underlying reasoning behind an AI's recommendations or actions, they are more likely to trust the system and are better equipped to verify, question, or override its decisions if necessary [21]. Indeed, one of the primary motivations for XAI is the recognition that, without such clarity, users find it difficult (if not impossible) to fully trust AI, especially given its potential for making unpredictable or inexplicable errors. Explainability thus serves as a mechanism for building confidence and ensuring accountability, helping to demystify AI processes and reinforce human oversight.

In the following section, this article will begin by reviewing pioneering works in the domain of HAII, with particular emphasis on the aforementioned three foundational pillars: trust, usability, and understandability. These elements have been widely recognized as critical to fostering effective, ethical, and sustainable collaboration between humans and AI systems. By synthesizing key contributions across disciplines, this review aims to highlight how each of these dimensions has evolved and identify existing gaps and emerging challenges in current literature. Subsequently, the focus will shift to HAII in the context of RL, a growing area of interest due to its capacity to enable adaptive, autonomous behavior. Various approaches, ranging from human-in-the-loop training and reward shaping to imitation learning, will be reviewed to explore how RL-based systems can enhance or hinder the human-AI relationship. This section aims to bridge the conceptual foundations of trust, usability, and understandability with the technical mechanisms of RL. Finally, this article will present several recent contributions from the authors' research group, which seeks to address identified limitations and push the boundaries of current HAII frameworks. These include novel algorithmic developments to incorporate ethical constraints and trust awareness. Additionally, an experiment is designed

to investigate the human trust evolution in AI systems. Collectively, these contributions aim to advance the theoretical and practical understanding of how to build AI systems that are not only intelligent but also aligned with human values, expectations, and needs.

The novel contributions of this paper include:

- Providing a comprehensive overview of foundational works on trust, usability, and understandability in HAII.
- Categorizing and reviewing five major RL-based HAII approaches, including state-of-the-art developments reported in recent publications.
- Presenting new contributions from the authors' research group, focusing on (i) ethical constraint integration in RL, (ii) trust-aware algorithmic design, and (iii) experimental investigation of human trust evolution in AI systems.

2 Human-AI Interaction

In HAII, trust, usability, and understandability form three foundational pillars that collectively determine the effectiveness and acceptance of AI systems, as illustrated in Fig. 1. Understandability enables users to grasp how and why an AI system makes decisions, fostering accurate mental models and informed oversight. Usability ensures that users can interact with the system efficiently and intuitively, minimizing friction and cognitive load. Trust governs whether users feel confident relying on the AI, influencing their willingness to adopt and engage with it. These three pillars are deeply interconnected: improvements in understandability and usability tend to promote trust, while trust can shape how users perceive the system's usability and explanations.

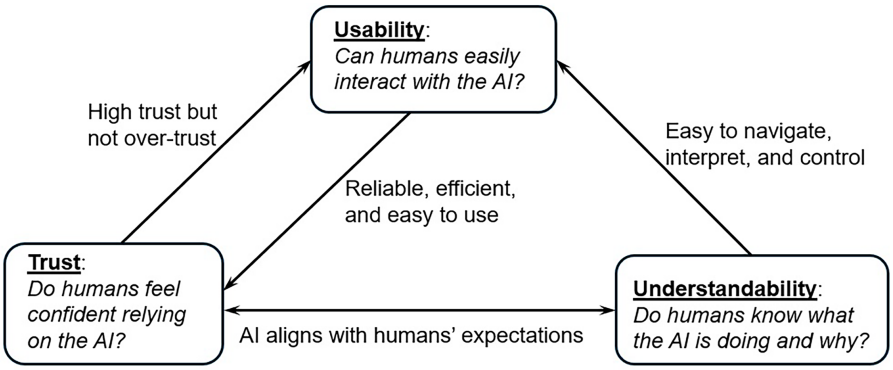


Figure 1: Three foundational pillars in HAII

2.1 Trust

Previous research suggests that trust in AI emerges from a combination of human, contextual, and technology-based factors [16]. Human characteristics (e.g., a user's general propensity to trust technology) and contextual variables (e.g., high-stakes vs. low-stakes scenarios) significantly influence the formation of trust. However, technology-based factors are often unique to AI systems. Key attributes such as the accuracy and reliability of an AI's predictions and recommendations, its predictability and consistency over time, and its transparency and explainability in decision-making are consistently identified as critical determinants of perceived trustworthiness [22]. In particular, because modern AI, especially ML models, can exhibit unexpected or opaque behaviors, unlike knowledge-based systems, providing users with insight into how the AI works and clearly communicating its uncertainties and limitations becomes essential. For instance, users

are more likely to trust an AI agent that can clearly signal its confidence level or justify its recommendations, rather than one that delivers opaque decisions without context [23]. Recent work has therefore focused on integrating features such as confidence scores, explanations, uncertainty estimates, and trustworthiness cues to foster appropriately calibrated user trust.

A central theme in human-AI trust research is the notion of trust calibration, ensuring that users' trust levels are aligned with the AI system's actual capabilities and performance. Miscalibrated trust can manifest as either over-trust (excessive reliance, even when unwarranted) or under-trust (unwarranted skepticism, leading to under-use). Both outcomes are problematic: over-trust may lead users to follow incorrect AI suggestions, while under-trust may cause them to dismiss valuable guidance, undermining the AI's benefits.

One recent behavioral study found that simply labeling advice as AI-generated led participants to over-rely on it, even when it contradicted their correct judgments or the available evidence [24]. This blind trust resulted in suboptimal or harmful decisions, highlighting how over-trust can degrade decision quality and negatively affect others. To mitigate this, researchers advocate for dynamic trust calibration strategies. One promising approach involves designing AI systems to be both interpretable and uncertainty-aware, enabling users to adjust their trust based on the system's likelihood of error. For example, Okamura and Yamada [25] demonstrated a method of adaptive trust calibration, where the system monitored user reliance patterns and delivered targeted alerts or explanations when users appeared to be over- or under-trusting. This interaction helped users recalibrate trust in line with the system's true performance. Overall, interactive transparency features can serve as important safeguards that promote critical engagement, preventing users from uncritically deferring to AI outputs.

2.2 Usability

Usability, a core pillar of Human-AI Interaction, refers to how effectively and efficiently users can interact with an AI system to accomplish their goals with minimal friction or cognitive strain. A growing body of empirical work highlights that adaptive AI systems, which personalize their behavior based on user actions or situational context, can significantly enhance usability by improving interaction quality, reducing user workload, and increasing task efficiency. These systems adjust their responses in real time, tailoring support to user needs, which leads to more intuitive and productive human-AI collaboration. For example, one controlled study found that domain experts using an AI decision aid with explainable visual feedback (such as heatmap explanations) achieved higher accuracy than those using a non-transparent, black-box AI, improving task success rates by approximately 5%–8% [26]. Another study shows that AI teammates capable of learning from user input and dynamically adapting their guidance can reduce decision time and cognitive load, thereby streamlining the interaction process [18]. Furthermore, adaptive interfaces that provide timely, customized feedback or modulate task difficulty help keep users in a performance-optimized zone, contributing to sustained usability and satisfaction [27]. These findings underscore the importance of designing AI systems that are not only intelligent but also easy to use, responsive, and context-aware, ensuring that users can interact with them seamlessly and effectively.

Designing human-centered AI (HCAI) offers a practical solution to addressing usability challenges in HAII by aligning AI system behavior with human needs, preferences, and limitations. HCAI design emphasizes building systems that account for the capabilities and constraints of both humans and AI. This approach directly enhances usability by making AI systems more intuitive, responsive, adaptive, and aligned with human-centered goals. The development of HCAI systems puts people first, supporting human values such as rights, justice, and dignity, while also promoting goals like self-efficacy, creativity, responsibility, and social connections [28,29]. A recent review highlights the breadth of HCAI research, identifying established clusters as well as emerging areas at the intersection of HAII and ethics [30]. In addition to proposing a

new definition and conceptual framework of HCAI, the review outlines key challenges and future research directions. Furthermore, Bingley and colleagues conducted a qualitative study involving AI developers and users, revealing that current HCAI guidelines and frameworks need to place greater emphasis on addressing real human needs in everyday life [31].

2.3 Understandability

Understandability, also referred to as explainability, is another fundamental pillar of HAII that enables users to interpret, evaluate, and make informed decisions based on AI outputs. Broadly, there are two complementary approaches to achieving understandability: (1) building interpretable models, and (2) generating post-hoc explanations for "black-box" models [21,32–35]. Interpretable models are inherently transparent; that is, users can directly inspect their internal logic. Examples include decision trees (where the decision path can be followed step-by-step), linear regression with a limited number of features (where coefficients indicate variable importance), and rule-based or knowledge-based systems [36–39]. These models offer intrinsic explainability, meaning that the model's structure itself explains. However, a trade-off often exists: while these models are easier to interpret, they may lack the capacity to capture complex patterns in high-dimensional data, potentially compromising performance for transparency.

In contrast, many high-performing AI models, particularly deep neural networks (DNNs), are not inherently interpretable. For these "black-box" systems, post-hoc explanation techniques are used to enhance understandability without modifying the model's internal structure. These techniques generate explanations after a decision is made, offering insights into the model's reasoning process. Common post-hoc methods include: (1) generating feature importance scores, which identify which input variables most influenced the output [40]; (2) providing local explanations for individual predictions, such as with LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), which use simplified surrogate models or Shapley values to attribute importance [41,42]; (3) visualizing internal activations, for example with heatmaps that highlight the image regions most relevant to a vision model's decision [43,44]; (4) generating natural language justifications, which explain the model's outputs in human-readable text [45,46]; and (5) constructing counterfactual explanations, which illustrate how minimal changes to input features could have altered the model's decision [47,48]. These techniques play a vital role in making complex AI systems more transparent and comprehensible, especially for non-technical users or in domains requiring accountability.

Moreover, explainability is increasingly recognized as a cornerstone of accountable and transparent AI, particularly in high-stakes domains such as healthcare, finance, and criminal justice. In these settings, users and regulators must be able to interrogate AI decisions, asking not only what decision was made, but why and how. For instance, if an AI system recommends a treatment or a bail decision, stakeholders must receive a satisfactory justification. Without such transparency, trust erodes, and systems may fail to meet ethical or legal requirements. Recent policy developments reflect this urgency. The European Union's General Data Protection Regulation (GDPR) has been interpreted as granting individuals a "right to explanation" for algorithmic decisions, and the EU's Ethics Guidelines for Trustworthy AI explicitly identify transparency and explainability as core principles [49,50]. These regulatory pressures have further accelerated research in XAI, emphasizing that understandability is not merely a desirable feature but is often a prerequisite for legal compliance, ethical alignment, and public acceptance [21].

3 Reinforcement Learning and Its Extensions

Learning from interaction is a foundational principle that underlies many theories of intelligence, both biological and artificial. RL formalizes this concept as a framework for goal-directed learning through

sequential interactions between an agent and its environment [7]. In this framework, the agent learns to make decisions over time by receiving feedback in the form of rewards, aiming to discover policies that maximize cumulative returns.

This section introduces the mathematical foundations of RL, focusing on the formalism of Markov Decision Processes (MDPs). We then explore several key extensions that enhance RL's expressiveness and applicability. In particular, we examine Partially Observable Markov Decision Processes (POMDPs), which address environments where the agent lacks full observability and must instead maintain and reason over belief states. We also discuss approaches for handling temporally complex or logically structured tasks by integrating temporal logic, automata-based task representations, and related formal methods. These extensions allow RL systems to operate effectively in settings characterized by long-term dependencies, safety constraints, and structured objectives beyond scalar reward signals, bringing RL closer to the demands of real-world decision-making under uncertainty.

3.1 Reinforcement Learning

Fig. 2 illustrates the core structure of RL, which consists of five key components: the agent, the environment, the internal state, actions, and the reward function. The environment, whether physical or virtual, represents the external setting with which the agent interacts. Through this interaction, the agent receives observations that provide (possibly partial) information about the environment, which it uses to update its internal state, i.e., an internal representation or belief about the current configuration of the environment. Based on this internal state, the agent selects and executes a goal-oriented action that influences the environment. In response, the agent receives a reward signal, a scalar feedback value that evaluates the desirability of its recent behavior. This trial-and-error learning process, driven by continuous interaction and feedback, forms the foundation of RL.

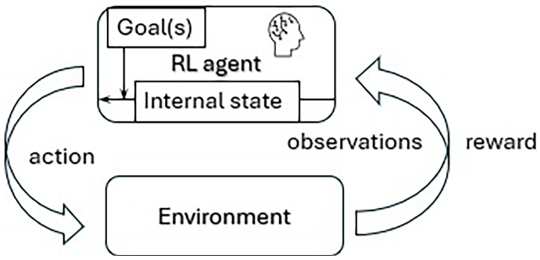


Figure 2: RL agents learn from interaction with the environment

The agent's objective is to learn a policy, a mapping from observations or states to actions, that maximizes the expected cumulative reward over time, often in the presence of uncertainty, partial observability, and delayed action effects. In conventional RL, it is typically assumed that the environment is fully observable, meaning the agent has direct access to the true state of the environment at each time step. Under this assumption, the environment-agent interaction is formally modeled using an MDP, which provides a mathematical framework for defining the states, actions, transition dynamics, and reward structure [51].

An MDP can be mathematically represented by a tuple

$$\mathcal{M} = \{S, A, T, s_0, R, \Pi, L\} \tag{1}$$

where S is a finite state space, A is a finite action space, $T(s, a, s')$ denotes the transition function with $s, s' \in S$, s_0 is the initial state, and $R(s)$ is the reward function. Additionally, Π is a set of atomic propositions, and $L(s)$

is a labeling function. The transition function defines the probability of transition to the next state s' after taking an action a in the current state s, and satisfies the normalization condition $\sum_{s' \in S} T(s, a, s') = 1$. The labeling function $L : S \to 2^{\Pi}$ assigns to each state a subset of atomic propositions, enabling task recognition and specification.

It shall be noted that a set of actions is available in each state, e.g., $A(s) = \{a | a \in A\}$. The output of an RL learning process is an optimal policy π, i.e., the behavior function, which guides the agent's decision-making process by determining how to choose an action at the current state. The objective function is the expected return, also referred to as the accumulated rewards, as shown below.

$$U(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \middle| s_{t=0} = s\right] \tag{2}$$

where $\gamma \in [0, 1]$ is the discount factor that captures the importance of future rewards, and s_t denotes the state at time t. While many policies may be feasible, the goal of an RL problem is to identify an optimal policy ξ^*, as expressed in Eq. (3), which yields a sequence of actions to maximize the expected return over a finite or infinite horizon.

$$\xi^* = \text{argmax}_{\xi} U^{\xi}(s) \tag{3}$$

To numerically derive the optimal policy, two value functions are typically introduced: the state-value function $V(s)$ and the action-value function $Q(s, a)$. A value function estimates the expected accumulated reward that an RL agent can obtain. For example, $Q^{\xi}(s, a)$, also called Q-values, quantifies the expected return when the agent begins in states s, executes action a, and then proceeds according to a given policy ξ. After estimating these Q-values, the agent can derive the optimal policy by selecting the action in each state that yields the highest value: $\xi^*(s) = \arg\max_{a \in A} Q^*(s, a)$. Therefore, a major category of RL approaches, value-based methods, focuses on estimating value functions that correspond to optimal policies.

Q-learning is a value-based RL algorithm, originally introduced to estimate the Q-value function by exploring the state space of an MDP over discrete time steps [52]. It is considered a model-free method because it does not require prior knowledge of the transition probabilities ($T(s, a, s')$) and the reward function ($R(s)$), which makes it well-suited to many real-world scenarios. In Q-learning, after taking action a in the current state s, the corresponding Q-value can be updated using a modified version of Bellman's equation, as shown in Eq. (4) [53].

$$Q_{new}(s, a) = Q(s, a) + \alpha[R(s) + \gamma \max_{a' \in A} Q(s', a') - Q(s, a)] \tag{4}$$

where $\alpha \in (0, 1]$ is the learning rate.

Classical Q-learning assumes that both the state space (S) and action space (A) are small and discrete, enabling the use of a tabular representation to store and update Q-values for each state-action pair individually. However, in many practical applications, the state and/or action spaces are either continuous or high-dimensional, making tabular methods infeasible due to the large number of state-action pairs. To address this challenge, it is common to employ function approximation techniques, particularly DNNs, to generalize Q-value estimates across similar states and actions. A notable advancement in this direction is the deep Q-network (DQN) algorithm, which leverages DNNs to approximate the Q-function and enables RL in complex, high-dimensional environments [54].

DQN introduces the use of two DNNs to stabilize learning: an evaluation Q-network $Q_e(s, a; \theta_e)$ and a target Q-network $Q_t(s, a; \theta_t)$ where θ_e and θ_t denote the parameters (weights) of the respective networks.

DQN is also a model-free RL method in which the agent interacts with the environment over multiple episodes, each consisting of a sequence of state transitions. During each transition, the standard Q-value update in Eq. (4) is adapted using the Q-networks as follows

$$Q_{\text{new}}(s,a) = Q_e(s,a;\theta_e) + \alpha\left[R(s) + \gamma \max_{a' \in A} Q_t(s',a';\theta_t) - Q_e(s,a;\theta_e)\right] \tag{5}$$

Using the above equation, the agent records each transition as an experience tuple and stores it in a memory buffer (also known as experience replay memory) [55]. During training, the evaluation Q-network $Q_e(s,a;\theta_e)$ is updated at each step by randomly sampling a mini-batch of past experiences from the buffer, which helps break temporal correlations and improve learning stability. In contrast, the target Q-network $Q_t(s,a;\theta_t)$ serves as a fixed reference during Q-value updates in Eq. (5). Its parameters remain unchanged for a specified number of steps and are then periodically synchronized with the evaluation network, i.e., $\theta_t \leftarrow \theta_e$) to reduce oscillations and divergence during training.

Another major class of RL methods is policy-based methods, differing from value-based approaches, which directly optimize the policy without explicitly estimating value functions. Rather than learning a value function and then deriving a policy from it, these methods aim to update the policy parameters θ to learn a stochastic policy, $\pi_\theta(a|s)$, which defines the probability of selecting action a given state s. For example, policy gradient methods improve the policy by applying gradient ascent on a performance objective and are known to converge toward an optimal policy using gradient-based or evolutionary strategies [56]. A commonly used objective function is the expected log-likelihood of the section actions, weighted by an estimate of their advantage, computed over a finite batch of trajectories:

$$Loss(s) = \mathbb{E}_t\left[log\pi_\theta(a_t|s_t)\hat{A}_t\right] \tag{6}$$

where $\hat{A}_t$ is an estimator of the advantage function at time t, reflecting how much better action a_t is compared to the average action at s_t. For a comprehensive overview of RL algorithms, readers are referred to recent survey articles such as [57]. In this paper, however, we primarily focus on value-based RL methods.

3.2 Partially Observable Environments

Assuming a fully observable environment, MDPs have been extensively employed in robotics and autonomous systems. However, in many real-world scenarios, agents often operate with incomplete perceptual information, making it difficult to fully determine the underlying system state, i.e., the environment is partially observable. In such cases, a more general mathematical framework, POMDP, must be adopted [58].

To accommodate both static and dynamic events, a probabilistic-labeled POMDP (PL-POMDP, or simply POMDP hereafter) can be formally defined as a tuple $\mathcal{P} = (S, A, T, s_0, R, O, \Omega, \Pi, L, P_L)$. The elements S, A, T, s_0, R, Π and L are defined analogously to those in the MDP formulation (see Eq. (1)). In addition, a POMDP explicitly incorporates observation uncertainty and probabilistic state labeling. $\Omega : S \times A \times O \rightarrow [0,1]$ is the observation probability function, indicating the likelihood of receiving observation o after taking action $a \in A(s)$ and transitioning to state $s' \in S$. It satisfies the normalization condition $\sum_{o \in O(s')} \Omega(s',a,o) = 1$ for all s' and a. On the other hand, $P_L : S \times L \rightarrow [0,1]$ is the labeling probability function, denoting the probability of associating label $l \in L(s)$ associated with state $s \in S$. It satisfies $\sum_{l \in L(s)} P_L(s,l) = 1, \forall s \in S$.

The set of state labels, $L(s) = l_1, l_2, \ldots$, consists of atomic propositions that represent the occurrence of specific events at a given state s. The labeling probability function $P_L(s,l)$ distinguishes between static and dynamic events: if $P_L(s,l) = 1$, the event labeled by l is deterministically associated with state s (static), whereas $P_L(s,l) < 1$ reflects uncertainty or variability (dynamic). These state labels can naturally identify goal

states, enabling straightforward specification of simple go-to-goal tasks. For more complex tasks, a formal language framework can be used to describe desired behaviors using state labels as symbolic inputs. Similarly, various constraints can also be encoded within such a formalism. Further details are provided in Section 3.3.

Over the past decade, researchers have developed a range of model-based algorithms to address POMDP problems. Many state-of-the-art solvers are capable of handling complex domains involving thousands of states [59]. One prominent class of these algorithms is Point-Based Value Iteration (PBVI), which approximates the solution to a POMDP by computing the value function over a selected, finite subset of the belief space [60,61]. Notably, a belief state represents a probability distribution over all possible system states. Following each state transition, the belief must be updated using the system's transition and observation models. In essence, model-based approaches, including those based on value iteration, solve an equivalent MDP defined over the continuous belief space.

A commonly used alternative for handling POMDPs is the model-free approach, which does not require prior knowledge of state transition or observation probabilities. In this framework, the agent relies on a history of observations to determine actions, effectively mapping sequences of observations to decisions. To extend the DQN framework for partially observable environments, the Q function (i.e., Q-networks) is adapted to process a series of past observations rather than a single input. Specifically, let $\mathbf{o}_t = (o_{t-j}, o_{t-j+1}, \ldots, o_t)$ denote the observation sequence over the past $j+1$ time steps. The policy then becomes dependent on this observation history, with the agent aiming to optimize the expected cumulative reward defined in Eq. (2).

To better capture temporal dependencies in such settings, recurrent architectures like Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs), or other recurrent neural network (RNN) variants are often integrated into the DQN architecture, as shown in Fig. 3 [8]. This leads to reformulated Q-networks: the evaluation network $Q_E(\mathbf{o}_t, a_t; \theta_E)$ and the target network $Q_T(\mathbf{o}_t, a_t; \theta_T)$. As a result, the standard DQN update rule (Eq. (5)) is modified to accommodate the sequential input structure as below.

$$Q_{new}(\mathbf{o}_t, a_t) = Q_E(\mathbf{o}_t, a_t; \theta_E) + \alpha \left[r_t + \gamma \max_{a_{t+1} \in A} Q_T(\mathbf{o}_{t+1}, a_{t+1}; \theta_T) - Q_E(\mathbf{o}_t, a_t; \theta_E) \right] \tag{7}$$

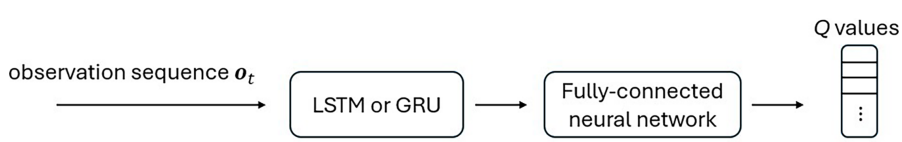


Figure 3: RNN-based Q networks in DQN for POMDP problems

3.3 Temporal Logic and Finite-State Automaton

Linear temporal logic (LTL), a formalism within the domain of formal languages, is widely used to specify linear-time properties by describing how state labels evolve over sequential system executions [62]. LTL builds upon propositional logic, incorporating not only standard Boolean operators but also temporal ones. Two fundamental temporal operators are $\bigcirc$ (read as "next") and $\mathcal{U}$ (read as "until"). In this context, let $\beta \in \Pi$ denote an atomic proposition, and let ϕ, ϕ_1 and ϕ_2 represent LTL formulas. The syntax of valid LTL formulas can be described using the following grammar rule adapted from [63]:

$$\phi := \text{True} \mid \beta \mid \phi_1 \wedge \phi_2 \mid \neg\phi \mid \bigcirc\phi \mid \phi_1 \mathcal{U} \phi_2 \tag{8}$$

here, $\neg$ and $\wedge$ stand for negation and conjunction, respectively. The expression $\bigcirc\phi$ is satisfied at a given time step if ϕ holds at the immediate next time step. The expression $\phi_1 \mathcal{U} \phi_2$ is satisfied if ϕ_2 holds at some point in the future, and ϕ_1 holds at every time step up to (but no including) that point. LTL also includes other commonly used temporal constructs: $\diamond$ ("eventually") asserts that ϕ becomes true at some point in the future, while $\square$ ("always") guarantees that ϕ holds at all time steps from the current one onward.

The meaning of LTL formulas is evaluated over infinite sequence, or words, denoted as $\boldsymbol{w} = w_0 w_1 \ldots$, where w_i is an element of 2^{Π} for $i \geq 0$, and Π is the set of atomic propositions defined in MDP and POMDP (Sections 3.1 and 3.2). Using the satisfaction relation symbol $\models$, the semantics of LTL formulas can be formally defined as

$$
\begin{aligned}
&\boldsymbol{w} \models \text{True} \\
&\boldsymbol{w} \models \beta && \Leftrightarrow && \beta \in L(\boldsymbol{w}[0]) \\
&\boldsymbol{w} \models \phi_1 \wedge \phi_2 && \Leftrightarrow && \boldsymbol{w} \models \phi_1 \text{ and } \boldsymbol{w} \models \phi_2 \\
&\boldsymbol{w} \models \neg\phi && \Leftrightarrow && \boldsymbol{w} \not\models \phi \\
&\boldsymbol{w} \models \bigcirc\phi && \Leftrightarrow && \boldsymbol{w}[1:] \models \phi \\
&\boldsymbol{w} \models \phi_1 \mathcal{U} \phi_2 && \Leftrightarrow && \exists t \text{ s.t. } \boldsymbol{w}[t:] \models \phi_2, \forall t' \in [0, t), \boldsymbol{w}[t':] \models \phi_1
\end{aligned}
\tag{9}
$$

LTL formulas can be systematically transformed into equivalent representations in the form of automata, enabling automated verification techniques such as model checking to assess whether a system satisfies a specified temporal property [62]. One commonly used class of automata for this purpose is the Limit-Deterministic Generalized Büchi Automaton (LDGBA) [64]. LDGBAs are particularly advantageous because they strike a balance between the expressiveness of nondeterministic automata and the efficiency of deterministic ones. By converting an LTL formula into an LDGBA, it becomes possible to encode temporal tasks as automata that accept exactly the sequences (executions) satisfying the formula. This transformation enables checking whether all possible system executions satisfy the temporal specification, or identifying counterexamples when violations occur.

An LDGBA can be defined as a tuple $\mathcal{A} = (Q, \Sigma, \delta, q_0, \mathcal{F})$, where Q is a finite set of states, and $\Sigma = 2^{\Pi}$ is the finite input alphabet from the power set of atomic propositions Π (defined in MDP and POMDP). δ is the transition relation, defined as a function δ: $Q \times (\Sigma \cup \{\epsilon\}) \to 2^{Q}$. $q_0 \in Q$ is the initial state, and $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_f\}$ is a set of accepting components, where each $\mathcal{F}_i$ is a subset of Q. The set of states Q is partitioned into two disjoint subsets: a deterministic part Q_D and a non-deterministic part Q_N, such that $Q_D \cup Q_N = Q$ and $Q_D \cap Q_N = \varnothing$. Specifically, each transition from a state in Q_D under any input symbol $\beta \in \Sigma$ leads to exactly one successor, i.e., $|\delta(q, \beta)| = 1$ for all $q \in Q_D$. Transitions originating from states in Q_D only point to other states in Q_D. In other words, for all $q \in Q_D$ and $\beta \in \Sigma$, there exist $\delta(q, \beta) \subseteq Q_D$. Transitions labeled with ϵ (which do not consume input) are allowed only from nondeterministic to deterministic states, that is, from $q \in Q_N$ to $q' \in Q_D$. Additionally, each accepting set $\mathcal{F}_i$ is entirely contained within Q_D, i.e., $\mathcal{F}_i \subseteq Q_D$ for all i.

Given an infinite input word $\boldsymbol{w} = w_0 w_1 \ldots$, a run of the LDGBA is an infinite sequence of states $\boldsymbol{q} = q_0 q_1 \ldots$, such that each successive state is determined by the transition relation, i.e., $q_{i+1} \in \delta(q_i, w_i)$. Define $\inf(\boldsymbol{q})$ as the set of states that appear infinitely often in the run $\boldsymbol{q}$. The word $\boldsymbol{w}$ is accepted by the LDGBA if, for every accepting set $\mathcal{F}_i$, the intersection $\inf(\boldsymbol{q}) \cap \mathcal{F}_i$ is non-empty. For practical guidance and examples on how such automata are constructed, readers may refer to the Owl tool [65]. In addition, a detailed example and explanation of how LTL encodes ethical constraints and guides the agent's behavior in the clinical setting is provided in the case study in Section 5.2.

3.4 Complex Tasks and Specifications in RL

LDGBAs are frequently applied in RL and motion planning under temporal constraints by constructing the product of an MDP and an LDGBA. This product formulation enables the seamless incorporation of logic specifications into the decision-making processes for complex tasks [66–68]. Similarly, methods have been extended to construct the product of a POMDP and an LDGBA, addressing the challenges in environments with partial observability [8], as illustrated in Fig. 4.

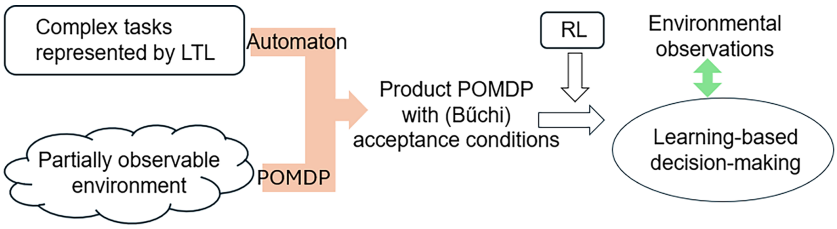


Figure 4: The product of LDGBA and POMDP for RL problems with complex tasks

Combining a POMDP $\mathcal{P} = (S, A, T, s_0, R, O, \Omega, \Pi, L, P_L)$ with an LDGBA $\mathcal{A} = (Q, \Sigma, \delta, q_0, \mathcal{F})$ results a product POMDP, which is defined as $\mathcal{P}^\times = \mathcal{P} \times \mathcal{A} = (S^\times, A^\times, T^\times, s_0^\times, R^\times, O, \Omega^\times, \mathcal{F}^\times)$. The product POMDP augments the environment's state space with automaton states to track satisfaction of the LTL specification. Each component is defined as follows:

- State Space: $S^\times = S \times Q$ with $s^\times = \langle s, q \rangle$.
- Action Space: $A^\times = A \cup \{\epsilon\}$ including both environment actions and ϵ-transitions.
- Initial State: $s_0^\times = \langle s_0, q_0 \rangle$.
- Probabilistic Transition Function: $T^\times = S^\times \times A^\times \times S^\times \to [0,1]$ is defined by:

$$T^\times\left(s^\times, a^\times, s^{\times\prime}\right) = \begin{cases} T(s, a^\times, s\prime) & q' = \delta\left(q, l\right), l \in L(s') \text{ and } a^\times \in A \\ 1 & a^\times \in \{\epsilon\} \text{ and } q' \in \delta(q, \epsilon) \text{ and } s' = s, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

  where $s^{\times\prime} = \langle s', q' \rangle$.
- Reward Function: Rewards are defined over product states:

$$R^\times(s^\times) = \begin{cases} R(s) & l \in L(s), q' = \delta\left(q, l\right) \in \mathcal{F}_i, \text{ and } \mathcal{F}_i \in \mathcal{F} \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

- Observation Function: $\Omega^\times = S^\times \times A^\times \times O \to [0,1]$ is

$$\Omega^\times(s^{\times\prime}, a^\times, o) = \Omega(s', a^\times, o) \tag{12}$$

  if $a^\times \in \{\epsilon\}$, no observation is emitted; the environment remains in state s, while the automation transition via $\delta(q, \epsilon)$.
- Accepting Conditions: The set of accepting sets in the product POMDP is $\mathcal{F}^\times = \left\{\mathcal{F}_1^\times, \mathcal{F}_2^\times, \ldots, \mathcal{F}_f^\times\right\}$ where $\mathcal{F}_i^\times = \{\langle s, q \rangle | s \in S; q \in \mathcal{F}_i\}$.

A run on $\mathcal{P}^\times$ is represented by an infinite sequence $(s_0, q_0)(s_1, q_1)\ldots$, integrating both the environment's path and the automaton's progression. The expected return from an initial state under a policy $\xi^\times$ is:

$$U^{\xi^\times}(s_0^\times) = \mathbb{E}^{\xi^\times}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t^\times)\Big| s_{t=0}^\times = s_0^\times\right] \tag{13}$$

This formulation enables evaluating whether a policy not only performs well in the environment but also satisfies the desired temporal logic objectives.

The optimal policy $\xi^{\times *}(\mathbf{o}_t, \mathbf{q}_t)$ defined over the product POMDP $\mathcal{P}^\times$ corresponds to an optimal policy $\xi^*(\mathbf{o}_t)$ on the original POMDP $\mathcal{P}$, while ensuring compliance with the given LTL specification. Here, $\mathbf{o}_t$ represents the sequence of observations encountered by the agent, and $\mathbf{q}_t$ denotes the associated sequence of automaton states resulting from label transitions. This formulation assumes that the agent receives labels, i.e., automaton input symbols, as part of the environment's feedback. When the agent has full knowledge of the LTL specification and its corresponding automaton, it can reconstruct the automaton state trajectory $\mathbf{q}_t$. However, in scenarios where the agent lacks such prior knowledge, the automaton state transitions cannot be directly inferred. In such cases, the policy must instead rely on the history of observations and perceived label sequences, and is thus expressed as $\xi^{\times *}(\mathbf{o}_t, \mathbf{l}_t)$ where $\mathbf{l}_t = (l_{t-j}, l_{t-j+1}, \ldots, l_t)$.

3.5 Applications of Reinforcement Learning

RL has become a cornerstone of modern AI, with applications spanning a wide range of scientific, industrial, and societal domains. Its core strength lies in solving sequential decision-making problems under uncertainty, where an agent must learn effective strategies through interaction and feedback from the environment.

- **Robotics and autonomous systems:** In robotics, RL enables machines to learn complex motor skills such as grasping, locomotion, and manipulation without explicit programming, often outperforming traditional control methods in unstructured settings [69–71]. For autonomous systems, including self-driving vehicles, unmanned aerial vehicles (UAVs), and mobile robots, RL provides a framework for real-time decision-making, motion planning, and obstacle avoidance under uncertainty [72–74]. By leveraging simulated environments for training and transferring learned policies to the physical world, RL allows these systems to achieve robust, adaptive, and scalable performance in tasks ranging from warehouse automation to autonomous exploration.
- **Energy and sustainability:** RL has shown significant promise in advancing energy efficiency and sustainability by enabling intelligent decision-making in complex, resource-constrained systems. In power grid management, RL has been applied to optimize demand response, balance renewable energy integration, and improve fault detection and recovery, thereby enhancing grid stability and resilience [75–77]. In building energy systems, RL supports dynamic control of heating, cooling, and lighting to minimize energy consumption while maintaining occupant comfort [78,79]. Similarly, in renewable energy operations, RL can optimize wind farm coordination, solar panel tracking, and battery storage management to maximize output and reduce waste [80,81]. Beyond energy systems, RL has been applied to water resource management, smart irrigation, and waste reduction strategies, contributing to broader sustainability goals [82–84]. By continuously adapting policies to variable conditions such as weather patterns, consumption behavior, and resource availability, RL enables sustainable solutions that are both cost-effective and environmentally responsible.
- **Healthcare and medicine:** RL is increasingly applied in healthcare and medicine, where adaptive decision-making is critical for improving patient outcomes. In personalized medicine, RL has been used to optimize treatment strategies for chronic diseases such as diabetes, cancer, and sepsis by continuously adjusting drug dosages or therapy schedules based on patient responses [85–87]. In medical imaging, RL assists in automating diagnostic tasks, improving image acquisition protocols, and

guiding interventions with greater precision [88,89]. Robotic surgery and rehabilitation also benefit from RL, where agents learn to perform delicate maneuvers or adapt therapy regimens to individual patient needs [90,91]. Beyond clinical care, RL contributes to hospital operations by optimizing resource allocation, scheduling, and patient flow management [92,93]. By learning from complex, uncertain, and high-dimensional health data, RL systems can support clinicians in making evidence-based, real-time decisions, thereby enhancing both efficiency and quality of care.

- **Finance and economics:** RL has emerged as a powerful tool in finance and economics, where decision-making under uncertainty and dynamic environments is fundamental [94]. In financial markets, RL is applied to algorithmic trading, enabling agents to learn adaptive strategies for portfolio optimization, risk management, and high-frequency trading [95]. In economic policy and macroeconomic modeling, RL has been used to simulate and evaluate interventions such as monetary policy adjustments or carbon pricing schemes, allowing exploration of long-term impacts in complex systems [96,97]. Beyond markets, RL supports operations in banking and insurance by optimizing credit scoring, fraud detection, and dynamic pricing [98,99]. Its ability to balance risk and reward in uncertain and evolving contexts makes RL particularly suited to financial and economic domains, where small improvements in decision-making can translate into significant economic value.
- **Natural sciences and engineering:** RL is increasingly being applied across the natural sciences and engineering to address problems characterized by complex dynamics, uncertainty, and the need for adaptive control. In environmental and climate sciences, RL has been used to optimize flood mitigation strategies and water management in agricultural systems, supporting efficient resource allocation and risk reduction under variable conditions [100,101]. In materials science and chemistry, RL guides simulations and experiments to discover novel compounds, catalysts, and metamaterials [102,103]. In engineering, RL enables intelligent control of systems ranging from adaptive traffic networks and structural health monitoring to advanced manufacturing processes [3,104]. Together, these applications demonstrate RL's capacity to operate in high-dimensional, data-rich environments, accelerating scientific discovery while enhancing the resilience, efficiency, and sustainability of natural and engineered systems.

4 Human-AI Interaction in Reinforcement Learning

This study classifies the implementation of the HAII mechanism into RL with the following different approaches:

- *Learning from human feedback (LHF)*: In this approach, RL agents learn by using human-provided feedback, such as scalar reward signals, preference comparisons, or evaluative judgments, to guide policy improvement and optimize behavior [105]. This learning is typically passive, as the human evaluates agent behavior after it is produced, rather than being actively queried by the agent.
- *Learning from human demonstrations (LHD)*: This approach uses human demonstrations of desired behavior to kickstart or guide RL. Methods range from inverse RL (inferring reward functions from expert trajectories) to interactive imitation learning (where the agent may iteratively request corrective demonstrations from a human).
- *Shared autonomy (SA)*: In this approach, human and AI agents collaboratively share control of a system. The agent assists the human by inferring goals, predicting intent, or adjusting suboptimal inputs—often using RL to learn when and how to intervene or assist.
- *Human-in-the-Loop querying (HLQ)*: This approach allows agents to actively query a human for information, such as demonstrations, preferences, or advice, to improve learning efficiency. By selecting informative queries or asking targeted questions, the agent can reduce the total amount of human effort needed.

- *Explainable RL (XRL)*: In this approach, the RL agent generates human-understandable explanations about its decision-making process, including why specific actions are chosen. These explanations enhance transparency and trust, enabling humans to provide better-informed feedback, corrections, or constraints.

Although this study reviewed each approach individually, these approaches form a spectrum of HAII mechanisms that are often used together in the same system, as shown in Table 1. They complement each other to enable more effective, efficient, and trustworthy collaboration between humans and RL agents. Additionally, their relationships with three HAII pillars (discussed in Section 2) are illustrated in Table 2.

Table 1: The intertwining of five approaches to implementing HAII mechanisms in RL

	LHF	**LHD**	**SA**	**HLQ**	**XRL**
LHF	–	Feedback can complement demos to refine policies.	SA can use feedback for better collaboration.	Queries often request feedback as information.	Explainability helps humans understand feedback context.
LHD	Demonstrations can be combined with feedback for guidance.	–	Shared control may rely on demonstrations to bootstrap the agent.	Queries may ask for corrective demonstrations.	Explanations may clarify what the agent learned from demonstrations.
SA	Feedback informs when/how the agent intervenes.	Demonstrations help agents learn collaborative behaviors.	–	Queries can help negotiate control or clarify intent.	Explanations may build trust in agent interventions.
HLQ	Queries actively request feedback.	Queries actively request demonstrations or corrections.	Queries negotiate shared control.	–	Explanations justify queries, making responses more effective.
XRL	Explains decisions to improve feedback quality.	Explains demo-based learning outcomes.	Explains autonomous actions to human collaborators.	Explains why queries are made.	–

Table 2: Relationships between five HAII approaches and the three HAII pillars

	Trust	**Usability**	**Understandability**
LHF	High: Aligns AI with human values.	Moderate: Adapts to human workflows.	Low: Limited transparency of reasoning.

(Continued)

Table 2 (continued)

	Trust	Usability	Understandability
LHD	Moderate: Learns reliable behaviors from demonstrations.	High: Easy for humans to provide examples.	Moderate: Shows behavior patterns but internal reasoning may remain opaque.
SA	High: Human-guided AI actions increase reliability.	High: AI assists without taking full control.	Moderate: AI behavior is partially transparent through shared control.
HLQ	High: Human feedback ensures reliable AI decisions.	High: Humans can guide AI interactively.	Moderate: Queries reveal reasoning partially.
XRL	High: Explanations make AI decisions more reliable.	Moderate: Users can act on insights but need some expertise.	High: Clear rationale for AI actions.

4.1 Learning from Human Feedback

Human feedback can often be combined with environmental feedback and integrated into the reward function, enabling RL agents to perform tasks that better reflect human preferences and needs. This approach is known as reward shaping [105]. In addition, recent developments have employed generative AI models, such as large language models (LLMs), to incorporate human feedback into the RL learning process [106]. Importantly, there is a distinction between direct human feedback, which reflects authentic user values and preferences, and synthetic feedback generated by LLMs, which can simulate or approximate human input at scale. While LLMs offer clear practical advantages, such as reducing annotation costs, accelerating prototyping, and enabling richer language-based interaction, they also raise ethical and methodological concerns. Synthetic feedback may unintentionally embed model biases, misrepresent user intent, or obscure accountability if used as a replacement for real human input. Thus, a critical open question is not whether LLMs can generate feedback, but how they should be used: as a complement to, rather than a replacement for, genuine human oversight. While human feedback can be used in various ways, including value shaping, policy shaping, and decision biasing, this section focuses specifically on reward shaping and the integration of feedback via LLMs, as illustrated in Fig. 5.

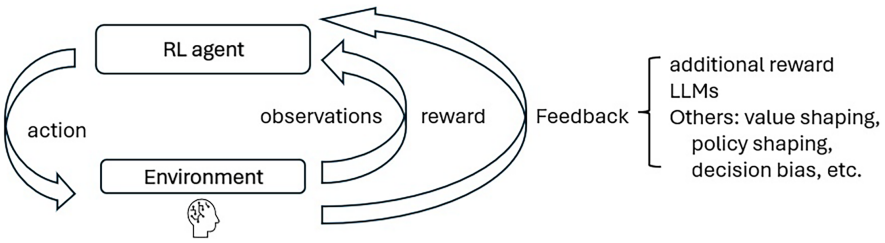


Figure 5: The implementation of LHF into RL

LHF directly implements the three pillars of HAII by aligning AI systems more closely with human expectations and needs. It strongly enhances trust, as human feedback allows AI to adopt behaviors that reflect users' values, preferences, and priorities, making decisions more reliable and socially acceptable. It

moderately improves usability by enabling AI systems to adapt to human workflows and practical constraints, reducing friction in interactions and facilitating adoption in real-world settings. LHF also provides partial support for understandability, as analyzing feedback signals can shed light on the factors influencing AI behavior, though full transparency of internal decision-making may remain limited. By integrating human guidance directly into learning processes, LHF helps bridge the gap between theoretically optimal AI policies and solutions that are actionable, interpretable, and widely accepted by users, demonstrating its pivotal role in human-centered AI design.

4.1.1 Reward Shaping

Reward shaping converts human feedback into a numerical value as a supplemental reward (either incentive or penalty) to RL agents during HAII. This additional reward is often added to the reward provided by the environment [107–109]. Consequently, the overall reward function is revised as $R = R^e + R^h$, where R^e represents the environmental reward and R^h is the human feedback reward. For example, in the TAMER (Training an Agent Manually via Evaluative Reinforcement) framework, the human feedback reward, also known as the human reinforcement function, is expressed as $R^h(s, a) = \beta\hat{H}(s, a)$. Here, β is a decaying weight factor, and $\hat{H}$ is the agent's internal model that predicts the expected human feedback for a given state-action pair [110–112].

Ibrahim et al. comprehensively reviewed reward engineering and shaping in RL [113]. Their work highlights reward shaping as a critical technique for accelerating learning and guiding agents toward desired behaviors, particularly in environments with sparse or delayed rewards. Among various methods, potential-based reward shaping (PBRS) stands out for its strong theoretical guarantees, as it modifies the reward using a potential function to provide additional guidance without altering the optimal policy [108,114]. The paper also emphasizes the growing importance of incorporating human feedback into reward shaping, where human evaluations, preferences, or demonstrations help construct internal reward models that better align agent behavior with human intentions [115–117].

In a recent work done by Golpayegani et al., the authors applied RL to advancing sustainable manufacturing for job shop scheduling (JSS) [118]. They proposed an ontology-based adaptive reward machine to address unforeseen events in dynamic and partially observable environments. In their paper, the generation of additional rewards does not rely on explicit real-time human feedback. Instead, it is achieved automatically through an ontology-based mechanism that embeds domain expert knowledge. Particularly, the ontology encodes concepts and properties relevant to the JSS environment and labels them with positive or negative beliefs that reflect desirable or undesirable outcomes from a human perspective, such as high machine utilization or low waiting time [119]. When the RL agent observes its environment, it maps observations to high-level ontological concepts, extracts new propositional symbols, and dynamically updates the reward machine. This process creates new reward functions by maximizing properties with positive beliefs and minimizing those with negative beliefs, allowing the agent to adapt its behavior to unforeseen events.

In another study, Jeon et al. proposed a robot teaching framework using RL in which robots learn cooperative tasks by integrating real-time sentiment feedback from human trainers [120]. Using a speech-based sentiment analysis module, the system detects humans' emotional tone, such as encouragement or frustration, and adaptively transforms this into an additional reward signal that augments the robot's standard task-based rewards [121,122]. This human-derived reward dynamically shapes the learning process, helping robots adjust their behavior to better align with human expectations and cooperation goals. Experiments show that robots guided by sentiment-based rewards achieve faster convergence, more stable learning, and improved cooperative performance compared to purely environment-driven RL.

4.1.2 Reward Shaping via LLMs

Recent research explores leveraging LLMs to automate or enhance reward shaping in RL. By interpreting natural language feedback, demonstrations, or preferences, LLMs can generate auxiliary rewards that guide RL agents more efficiently than manually designed reward functions [123–125]. For instance, LLMs can translate human instructions (e.g., "Avoid unsafe actions") into quantitative shaping rewards or infer implicit preferences from textual critiques. This approach is particularly valuable in complex, real-world tasks where sparse or ambiguous environmental rewards hinder learning. However, challenges remain, such as mitigating biases in LLM-generated rewards and ensuring alignment with true objectives [126,127].

Chaudhari et al. provided a comprehensive critique of LHF as applied to LLMs, dissecting its components, challenges, and practical implications [128]. The paper systematically analyzes the LHF pipeline, from human preference data collection and reward modeling to policy optimization, highlighting key bottlenecks like reward misgeneralization, scalability limitations, and annotator biases. It contrasts with some other alternative approaches, such as direct preference optimization, adversarial training, and underscores trade-offs between alignment, diversity, and computational cost. The authors also discussed ethical concerns (e.g., amplification of harmful preferences) and proposed future directions, including hybrid methods and better human-AI collaboration frameworks. By synthesizing empirical findings and theoretical insights, this survey serves as a roadmap for improving LHF's efficacy and reliability in LLM deployment.

A recent study introduced Text2Reward, a framework that integrates LLMs to automatically generate dense, interpretable reward functions for RL tasks from natural language task descriptions [129]. By prompting an LLM to decompose high-level goals into structured reward components (e.g., "move forward" to linear velocity reward) and synthesize them into executable code, Text2Reward eliminates the need for manual reward engineering and iteratively incorporates human feedback from users to improve reward functions. Experiments in robotic locomotion and manipulation tasks showed that policies trained with LLM-generated rewards achieve performance comparable to those trained with human-designed rewards, while also enabling real-time adjustments via natural language. The method bridges the gap between high-level intent and low-level RL implementation, making RL more accessible to non-experts.

It is worth mentioning that, in many state-of-the-art achievements, LLMs in RL act as proxies for human feedback, even without real-time or runtime HAII. For example, Sun et al. introduced an LLM-as-reward-designer paradigm where the LLM automatically shapes rewards through iterative dynamic feedback, without requiring direct human input during training [130]. The LLM evaluates agent behavior, identifies suboptimal actions, and adjusts dense/sparse rewards to guide learning, mimicking human-like reward engineering but in a scalable and automated manner. In another study, Qu et al. use LLMs to interpret environment states and agent behaviors, generating intermediate rewards that guide learning more effectively [131]. By integrating LLM-derived latent rewards, the framework improves sample efficiency and policy optimization in complex, long-horizon tasks.

Another approach included LLM-guided pretraining for RL agents, where an LLM (e.g., GPT) provides high-level task guidance to accelerate learning [132]. This approach simulates human-like feedback by critiquing agent behaviors during pretraining and provides reward-shaped advice to bootstrap RL policies. Additionally, Guo et al. presented an LLM-based framework for automated reward shaping in robotic RL, where LLMs interpret task descriptions and generate dense, semantically aligned reward functions [133]. The key innovation is a hierarchical reward decomposition: the LLM first breaks tasks into sub-skills, then designs weighted sub-rewards for each skill, dynamically adjusting their importance based on real-time agent performance. This approach also maintains interpretability by mapping LLM outputs to modular reward components.

4.1.3 Others

The other techniques of LHF in RL include value shaping, policy shaping, and preference modeling. Value shaping addresses the challenge of sparse or delayed rewards in RL by incorporating additional immediate feedback signals [134,135]. These signals, including human feedback, provide intermediate guidance about the relative value of actions or states, helping to bridge the temporal credit assignment problem [136]. Policy shaping incorporates human feedback by directly modifying the policy update rule in RL [137]. Model-free policy shaping is typically implemented by adding a human guidance term to the policy gradient, which biases the agent's actions toward those preferred by the human demonstrator or critic [138]. Several model-based policy shaping can be found in [110–112,139–142]. Additionally, preference modeling is a framework for learning reward functions or policies directly from human judgments (e.g., rankings or comparisons of trajectories). Instead of requiring explicit rewards or demonstrations, the agent infers objectives from relative human preferences over pairs of state-action sequences [143,144]. Table 3 presents an overview of the articles discussed in detail subsequently.

Table 3: Reviewed papers using LHF with value shaping, policy shaping, and preference modeling

Paper	LHF techniques
Tarakli et al. convert natural language feedback into reward signals via Q-augmentation, improving agent training through human guidance [145].	Value shaping
Faulkner et al. propose a human-guided policy shaping method that accelerates RL agent learning by focusing exploration on supervisor-attended states [146].	Policy shaping
Juan et al. introduce a human-guided RL framework where evaluative feedback shapes policy transfer across tasks by softly weighting gradient updates [147].	Policy shaping
Lee et al. present a benchmark for preference-based RL that simulates human feedback imperfections, enabling systematic evaluation of algorithmic robustness and efficiency in learning from preferences [148].	Preference modeling
Knox et al. investigate how agents in preference-based RL can misinterpret the advantage function that measures relative action quality as the reward function [149].	Preference modeling

Tarakli et al. presented a novel RL framework, ECLAIR (Evaluative Corrective Guidance Language as Reinforcement), in which an agent learns from natural language feedback instead of predefined numeric rewards [145]. The authors employed value shaping, especially Q-augmentation methods, to convert qualitative human feedback into structured reward signals, dynamically adjusting the agent's value function to align with the trainer's guidance [150]. By integrating LLMs with RL, the method enables more intuitive human-robot collaboration, as demonstrated in simulated environments where it improves learning efficiency and task performance over traditional RL approaches in advanced robotics.

Faulkner et al. introduced Supervisory Attention Driven Exploration (SADE), a policy shaping method where a human supervisor guides an RL agent's exploration by directing attention to key states (e.g., via gaze or clicks) and providing shaping rewards [146]. The approach dynamically scales exploration noise using an attention mask and converts feedback into potential-based rewards, reducing sample complexity in robotic tasks (e.g., block stacking) while preserving convergence guarantees. Unlike traditional policy

shaping, SADE focuses on biasing exploration rather than directly modifying actions, balancing human guidance with autonomous learning.

In another work, Juan et al. proposed a framework for policy transfer in RL using human evaluative feedback to shape exploration [147]. The method integrates human-provided evaluations into a progressive neural network, where feedback dynamically scales the policy's gradient updates, biasing learning toward preferred behaviors without hard action overrides. By combining human guidance with multi-task RL, the approach achieves efficient policy adaptation across related tasks, demonstrating faster convergence and higher success rates compared to standard RL baselines. While not traditional policy shaping, the human feedback implicitly shapes the policy through loss function modifications, offering a flexible alternative to direct action corrections.

Lee et al. presented a comprehensive benchmark designed to evaluate and advance preference-based RL methods, which learn from human feedback rather than pre-defined rewards [148]. The authors addressed the lack of standardized evaluation in general preference-based RL approaches by introducing simulated human teachers that exhibit realistic irrationalities, including stochastic preferences, myopic decision-making, occasional mistakes, and query skipping, to better reflect real-world human input. The benchmark covers diverse tasks from locomotion to robotic manipulation, enabling systematic comparison of algorithms under varying conditions. Using this benchmark, the authors evaluated state-of-the-art methods and found that while these algorithms perform well with perfectly rational teachers, their effectiveness degrades significantly when faced with noisy or biased feedback [115]. They also pointed out that preference-based RL may have potential drawbacks if malicious users teach the bad behaviors or functionality, which raises safety and ethical issues.

Furthermore, Knox et al. investigated how RL agents, particularly those trained with preference-based LHF, can misinterpret the advantage function (which measures relative action quality) as the reward function (which measures absolute desirability) [149]. The authors showed that when agents are trained using preference comparisons, where humans rank trajectories, they often conflate advantage with reward, leading to suboptimal or unintended behaviors. Through theoretical analysis and experiments, they demonstrated that this confusion can cause agents to prioritize actions that merely outperform alternatives rather than those that are objectively best. The paper highlights implications for LHF in LLMs, where such misalignment could propagate biases or reward hacking. Proposed solutions include careful reward shaping and explicit separation of advantage and reward during training.

4.2 Learning from Human Demonstration

While reward shaping through human feedback provides an efficient means of steering agent behavior, it often captures only local preferences or corrective signals. Demonstrations complement this limitation by offering richer, trajectory-level guidance that illustrates not only what outcomes are preferred but also how to achieve them. In this way, learning from demonstration extends the strengths of feedback, providing structured examples that can anchor RL in more complex, real-world tasks.

LHD, also known as imitation learning (IL), is a subfield of ML and robotics where an agent learns to perform tasks by observing how humans (or sometimes other agents) perform them, rather than being explicitly programmed or learning solely through trial and error [151,152]. LHD supports the HAII pillars by allowing AI systems to observe and imitate human behavior. It moderately enhances trust as AI's actions reflect demonstrated strategies and appear more reliable to users. LHD strongly improves usability, since providing demonstrations is a highly intuitive form of teaching that lowers the barrier for human-AI collaboration, especially for non-experts. It also moderately supports understandability, as observing the AI's learned behaviors can reveal patterns aligned with human actions, though the internal decision-making

process may still be difficult to fully interpret. By leveraging human demonstrations, LHD enables AI systems to adopt human-aligned strategies efficiently while maintaining practical relevance and user acceptance.

There are three common methods to incorporate LHD in RL: (1) Behavioral cloning (BC) that directly mapping states to action by supervised learning [153], (2) inverse RL (IRL) to infer the reward function based on the demonstration and use it to train the agent [154], and (3) interactive imitation learning (IIL) where the agent actively interacts with the human during learning, instead of only passively copying static demonstrations. The overview of the three approaches is listed in Table 4.

Table 4: Three approaches of LHD in RL

	BC	**IRL**	**IIL**
Goal	Directly copy expert behavior.	Infer the reward function that explains expert behavior.	Improve imitation by involving the expert interactively during learning.
Learning paradigm	Supervised learning.	Reward learning and RL.	Supervised imitation, but with interactive expert corrections.
Input	Offline dataset of expert state-action pairs.	Expert trajectories (states and actions).	Expert demonstrations plus on-demand corrections during training.
Output	Policy that mimics expert.	Reward function and derived policy.	Policy that mimics experts more robustly.
Main motivation	Simple and fast imitation.	Understand underlying preferences; generalize better.	Reduce compounding errors; improve generalization.
Advances	Implicit BC [155], active exploration BC [156], and counterfactual BC [157]	Addressing computational and memory limitations [158], sequential tasks [159], and application to LLMs [160]	Exploration efficiency improvement [161], LLM-based IIL [162], and generative adversarial IL [163].

It should be noted that Zare et al. provided a comprehensive overview of IL, focusing on its evolution, methodologies, and open challenges [164]. The authors discussed the approaches of IL, such as BC, IRL, and adversarial IL, as well as emerging areas like imitation from observation [165,166]. Key challenges discussed include handling imperfect demonstrations and domain discrepancies like dynamics or viewpoint mismatches [167–169]. The paper concludes by outlining future directions, emphasizing the need for robust, scalable methods to enhance real-world applicability in domains such as autonomous driving and robotics.

4.2.1 Behavior Cloning

BC is essentially supervised learning. Given a set of demonstrations, $\mathcal{D} = \{(s_1, a_1), (s_2, a_2) \ldots (s_n, a_n)\}$ where s_i, $i = 1 \ldots n$ are observed states, and a_i are the corresponding actions taken by the human demonstrator. The agent's goal is to learn a policy $\pi : S \rightarrow A$ that maps states to actions.

$$\pi^* = argmin_\pi \sum_{(s,a)\in\mathcal{D}} ||\pi(s) - a||^2 \tag{14}$$

That is, BC directly learn to imitate the demonstrated behavior by minimizing the difference between the agent's chosen action and the demonstrator's action.

Jia and Manocha presented a hybrid approach that leverages BC to initialize painting policies from human demonstrations before refining them via RL in a simulated environment [170]. This BC framework bridges the sim-to-real gap by pre-training the agent on stroke sequences, enabling efficient transfer to a real UltraArm robot equipped with a RealSense camera. By combining BC with Gaussian stroke modeling and vision-based contact force estimation (eliminating the need for physical sensors), the system achieves precise, adaptive brush control in real-world sketching, outperforming RL-only baselines in both stroke quality and policy convergence. This work highlights BC as a critical enabler for scalable and realistic robotic artistry.

Zhang et al. proposed Implicit BC (IBC) combined with Dynamic Movement Primitives (DMPs) to enhance RL for robot motion planning [155]. The authors address the inefficiency of traditional RL by leveraging human demonstrations through IBC, which avoids overfitting by penalizing deviations via an energy function rather than explicit action matching, unlike explicit BC (EBC). The framework integrates a multi-DoF DMP model to simplify the action space and a dual-buffer system (demonstration and replay) to bootstrap training. Experiments on a 6-DoF Kinova arm demonstrated that IBC-DMP achieves faster convergence, higher success rates, and better generalization in pick-and-place tasks compared to baselines such as the Deep Deterministic Policy Gradient (DDPG) method. The authors also discussed IBC's energy-based loss and demonstration-augmented RL, bridging human expertise with RL robustness for complex motion planning.

In another study, Xu and Zhao introduced Active Exploration Behavior Cloning (AEBC), an innovative IL algorithm designed to overcome the sample inefficiency limitations of traditional BC [156]. Unlike conventional BC, which passively replicates expert demonstrations, AEBC actively enhances exploration by integrating a Gaussian Process (GP)-based dynamics model that predicts future state uncertainties. The algorithm's novel entropy-driven exploration mechanism prioritizes high-uncertainty states during training, optimizing sample utilization through a dual objective, consisting of the standard BC's action prediction loss and an entropy term that maximizes the log determinant of the covariance matrices of the predicted states. In other words, GP-guided active exploration avoids expensive online expert relabeling, and a hybrid loss function dynamically balances imitation and exploration. Consequently, AEBC addresses BC' core weaknesses, compounding errors and poor sample efficiency, making it particularly suitable for real-world applications like robotics, where interactions are expensive. The authors conducted benchmark problems and found that AEBC achieved expert-level performance with much fewer interactions than traditional BC and outperformed Advantage-Weighted Regression (AWR) in sample efficiency [171].

Recently, Sagheb and Losey proposed Counterfactual Behavior Cloning (Counter-BC), a novel offline IL method that addresses the limitations of traditional BC by inferring the human's intended policy from noisy demonstrations [157]. Unlike standard BC, which naively clones all actions, including errors, Counter-BC introduces counterfactual actions that are plausible alternatives the human might have intended but did not demonstrate perfectly. The algorithm's core innovation is a generalized loss function that minimizes policy

entropy over these counterfactuals, favoring simple, confident explanations while remaining close to the demonstrated data, all without requiring additional human labels or environment interactions. Evaluated on simulated tasks and real-world robotics, Counter-BC outperforms BC and baselines in sample efficiency and task success, particularly with noisy or heterogeneous demonstrations. Counter-BC advances BC's practicality by interpreting what the human meant rather than what they did, bridging noisy data and robust policy learning.

Echeverria et al. introduced another novel approach, in which an offline RL method is tailored for combinatorial optimization problems like job-shop scheduling [172]. The authors addressed limitations of traditional deep RL (slow learning) and BC (poor generalization) by proposing a hybrid approach that combines reward maximization with imitation of expert solutions. The innovative contributions include modeling scheduling problems as MDPs with heterogeneous graph representations, encoding actions in edge attributes, and a new loss function balancing RL and imitation terms. The method outperforms state-of-the-art techniques on several benchmarks, demonstrating superior performance in minimizing makespan, particularly for large-scale instances. The work highlights offline RL's potential for real-time scheduling with complex constraints and suggests future extensions to other combinatorial problems.

4.2.2 Inverse Reinforcement Learning

Although BC is computationally efficient, it faces another major limitation: the covariate shift problem [173,174]. When training, the agent learns from expert-generated state-action pairs. However, during deployment, it encounters states influenced by its actions, potentially deviating from the training distribution [175,176]. Since the agent isn't trained to recover from these unfamiliar states, errors compound over time [177]. This makes BC particularly unreliable in safety-critical applications, such as autonomous driving, where unseen scenarios can lead to catastrophic failures [178]. One of the solutions to address covariate shifting is IRL.

A recent survey comprehensively explored IRL as an approach to infer an agent's reward function from observed behavior rather than manually designing it for handling RL problems [179]. The paper outlines key challenges, including ambiguity in reward inference, generalization to unseen states, sensitivity to prior knowledge, and computational complexity [180,181]. It categorizes foundational methods, including maximum margin optimization, maximum entropy approaches, Bayesian learning techniques, and classification or regression-based solutions [182–185]. The survey also discusses extensions addressing imperfect or partial observations, multi-task learning, incomplete models, and nonlinear reward functions [186–188]. The paper concludes by emphasizing future directions, such as enhancing efficiency for high-dimensional spaces and adapting IRL to dynamic real-world applications like robotics and autonomous systems. Another survey can be found in [189].

Lin et al. proposed a computationally efficient IRL approach to recover reward functions from expert demonstrations while addressing the computational and memory limitations of existing methods [158]. The authors introduced three key contributions, including a simplified gradient algorithm for the reward network, a loss function based on feature expectations rather than state-visiting frequency, and an automated featurization network to process trajectories without manual intervention. The method eliminates repetitive gradient calculations and excessive storage by using streamlined feature expectations for both expert and learner trajectories. Validated on benchmarks, such as CartPole, this approach demonstrated superior performance in terms of average return, required steps, and robustness to noise, while reducing training time and memory demands compared to state-of-the-art methods [190]. The approach is particularly effective for high-dimensional, continuous-state tasks with variable-length demonstrations.

Melo and Lopes addressed the challenge of teaching sequential tasks to multiple heterogeneous IRL agents [159]. They identified conditions under which a single demonstration can effectively teach all learners (e.g., when agents share the same optimal policy) and demonstrated that significant differences in transition probabilities, discount factors, or reward features render class teaching impossible. After formalizing the problem, they proposed two novel algorithms: SplitTeach and JointTeach. SplitTeach combines group and individualized teaching to ensure all learners master the task optimally, while JointTeach minimizes teaching effort by selecting a single demonstration, albeit with potential suboptimal learning outcomes. Simulations across multiple scenarios validated the theoretical findings, demonstrating that SplitTeach guarantees perfect teaching with reduced effort compared to individualized teaching, whereas JointTeach achieves minimal effort at the cost of performance. This work bridges gaps in ML for sequential tasks, offering practical solutions for heterogeneous multi-agent settings.

Wulfmeier et al. explored the application of IRL to improve IL in LLMs [160]. Unlike traditional maximum likelihood estimation (MLE), which optimizes individual token predictions, the authors proposed an IRL-based approach that optimizes entire sequences by extracting implicit rewards from expert demonstrations. They reformulated inverse soft-Q-learning as a temporal difference-regularized extension of MLE, bridging the gap between supervised learning and sequential decision-making. Experiments on T5 and PaLM2 models demonstrated that the proposed IRL methods outperform MLE in both task performance (e.g., accuracy) and generation diversity. This paper also discusses improved robustness to dataset size, better trade-offs between diversity and performance, and potential for more aligned reward functions in LLM fine-tuning. The work positions IRL as a scalable and effective alternative to MLE for IL in language models.

In addition, recent efforts include generative adversarial imitation learning (GAIL) that uses ideas from generative adversarial networks (GANs), so the agent tries to produce behavior indistinguishable from the demonstrator, judged by a discriminator network. Ho and Ermon introduced an early work of the GAIL framework that directly learns a policy from expert demonstrations without requiring interaction with the expert or access to a reinforcement signal [191]. Unlike traditional BC or IRL, which can be inefficient or indirect, GAIL draws an analogy to GANs. It trains a policy to match the expert's occupancy measure by optimizing a minimax objective between a discriminator (which distinguishes expert from learner actions) and a policy (which aims to fool the discriminator). The authors demonstrated that GAIL outperforms existing model-free methods, achieving near-expert performance across various high-dimensional control tasks, while being more robust to limited expert data. Another recent study by Zuo et al. proposed a GIRL approach to handle imperfect demonstrations from multiple sources [192]. The method introduces confidence scores to weight demonstrations, enabling the agent to prioritize higher-quality data. Additionally, it leverages maximum entropy RL and a reshaped reward function to enhance exploration and avoid local optima.

4.2.3 Interactive Imitation Learning

IIL is a framework where an agent learns to perform tasks by iteratively querying a human expert for corrective feedback during execution, refining its policy through interactions. Unlike BC, which passively mimics expert demonstrations without feedback, IIL actively corrects errors in real-time, improving robustness to distributional shift. In contrast to IRL, which infers a reward function from demonstrations to guide RL, IIL directly learns a policy through supervised corrections, avoiding the need for reward estimation. Thus, IIL combines the sample efficiency of supervised learning with interactive refinement, bridging the gap between BC's simplicity and IRL's adaptability.

Celemin et al. conducted a survey paper on IIL in Robotics, providing a comprehensive overview of methods where human feedback is intermittently provided during robot execution to improve behavior

online [193]. The authors categorized IIL approaches based on feedback modalities (evaluative vs. transition space, absolute vs. relative), the models learned (policies, state transitions, or reward functions), and auxiliary models like task features or human feedback interpretation [117,194–196]. They also discuss algorithmic properties (on/off-policy learning), interfaces for human-robot interaction, and applications in real-world robotics. The paper highlighted IIL's advantages over traditional IL and RL, such as data efficiency and robustness, while addressing challenges like human feedback inconsistency and covariate shift. Other similar surveys can be found in [164,197]. Table 5 summarizes several IIL papers reviewed in this section.

Table 5: Reviewed papers related to IIL

Paper	Feedback modality	Model learned	Auxiliary models
Human motion imitation with interactive RL [161].	Transition space.	Policy via Proximal Policy Optimization.	Bayesian neural network and priority state initialization algorithm.
LLM-based IIL [162].	Evaluative space.	Stochastic policy via negative log-likelihood optimization.	Similarity-checking mechanism.
Social robot with generative adversarial IL [163].	Evaluative space.	Generator via Proximal Policy Optimization and a discriminator using adversarial training.	Human reward network.
IIL with human expert [198].	Transition space.	Novice policy as an ensemble of neural networks.	A risk metric and a safety threshold.

Cai et al. proposed an interactive RL framework to accelerate human motion imitation for robots by improving exploration efficiency and reducing training time [161]. The method involves human experts labeling critical states in sampled trajectories, which are then used to train a Bayesian Neural Network (BNN) to identify and prioritize effective samples automatically. Additionally, a priority state initialization (PSI) algorithm allocates training resources based on the difficulty of initial states. Experiments demonstrate that the approach, combining human labeling termination (HLT) and PSI, significantly reduces the number of samples and training time while achieving higher imitation accuracy, compared to baseline methods.

Werner et al. introduced an IIL framework that leverages LLMs as cost-effective, automated teachers for robotic manipulation tasks [162]. The method employs a hierarchical prompting strategy to generate a task-specific policy and uses a similarity-based feedback mechanism to provide evaluative or corrective feedback during training. Evaluations on benchmark tasks demonstrated that this LLM-based IIL outperforms BC in success rates and matches the performance of a state-of-the-art IIL method using human teachers, while reducing reliance on human supervision [195]. While this approach offers a scalable, efficient alternative to human-in-the-loop training in robotics, the authors discussed the limitations, including LLMs' restricted environmental awareness and dependence on ground-truth data, and suggested future improvements with vision-language models and real-world sensory inputs.

Recently, a method named GAILHF (Generative Adversarial Imitation LHF) was introduced to enable a social robot to learn and imitate human game strategies in real-time two-player games by combining human demonstrations and evaluative feedback [163]. This framework employs a generator-discriminator architecture, where the generator learns from human trajectories. At the same time, the discriminator distinguishes

between human and robot actions, supplemented by a human reward network that predicts feedback and reduces human burden. Experiments showed the robot successfully imitated strategies and outperformed conventional IL and RL in game scores. Overall, this study demonstrates the efficacy of interactive learning for social robots and highlights the role of evaluative feedback in surpassing demonstration performance.

Kelly et al. developed HG-DAgger, an improved variant of the DAgger (Dataset Aggregation) algorithm for IIL with human experts [198]. Unlike standard DAgger, which stochastically switches control between the novice policy and the expert (potentially causing instability and degraded human performance), HG-DAgger lets the human expert take full control whenever they perceive the novice entering unsafe states, collecting corrective demonstrations only during these recovery phases. The method also learns a risk metric from the novice's uncertainty using neural network ensembles and a safety threshold derived from human interventions. Evaluated on autonomous driving tasks, HG-DAgger outperforms DAgger and BC in sample efficiency, safety, and human-like behavior, while providing interpretable risk estimates for test-time deployment.

4.3 Shared Autonomy

Feedback and demonstrations together lay the groundwork for aligning agents with human intentions: feedback supplies evaluative cues, while demonstrations encode procedural knowledge. However, both approaches are primarily offline and cannot always anticipate the uncertainties and novelties of dynamic environments. SA complements these methods by enabling real-time collaboration, allowing humans to intervene when learned strategies fall short and letting agents assist when human control alone becomes burdensome. This balance ensures adaptability beyond what either feedback or demonstrations can provide in isolation.

SA is an emerging paradigm in human-robot interaction and HAII that blends autonomous control with real-time human input to achieve better performance, safety, and user satisfaction than either fully manual or fully autonomous systems alone [199]. In this framework, the autonomous system and the human operator collaboratively share control of the task, often dynamically adjusting their respective levels of influence based on context, uncertainty, or user preference. This approach leverages the complementary strengths of humans, such as intuition, high-level reasoning, and adaptability, and machines like precision, speed, and the ability to process large volumes of sensory data. Applications of SA span a wide range of domains, including assistive robotics for people with disabilities, teleoperation in hazardous environments, and intelligent driving systems [200–202]. By facilitating seamless human-machine collaboration, SA aims to enhance task efficiency, reduce cognitive load, and improve overall user experience.

SA supports the HAII pillars by combining human input with AI assistance, creating a collaborative partnership where each agent leverages its strengths. This strongly enhances trust, as humans retain ultimate control and accountability while benefiting from AI support, leading to more reliable and acceptable outcomes. SA also strongly improves usability, since the system assists users without requiring them to relinquish control entirely, fitting naturally into human workflows. It moderately supports understandability, as users can observe and influence AI actions, gaining partial insight into its reasoning, though full transparency may still be limited. By effectively balancing human guidance with AI autonomy, SA enables efficient, user-centered decision-making that leverages the strengths of both humans and machines.

SA approaches can be broadly categorized into classical methods and learning-based methods, particularly those leveraging RL. Classical SA typically relies on manually designed control blending strategies, where the system assists the user according to predefined rules or heuristics, for instance, by combining user preferences on autonomous vehicle service attributes [203]. While these methods offer predictability and

are often easier to verify for safety-critical applications, they lack flexibility and personalization and cannot easily adapt to new tasks, user preferences, or changes in the environment without manual retuning.

In contrast, RL-based SA frameworks aim to learn optimal assistance policies directly from data, often through human demonstrations, real-time feedback, or preference learning. This enables the system to dynamically infer user intent, personalize assistance based on individual skill or context, and continuously improve through interaction [204–206]. Although RL-based methods face challenges related to data efficiency, safety, and interpretability, they hold significant promise for creating more adaptive, user-centered SA systems that generalize beyond the scenarios anticipated by classical designs [207]. The comparison is summarized in Table 6.

Table 6: The comparison between classical and RL-based SA

	Classical SA	**RL-based SA**
Control blending strategy	Rule-based, predefined mixing (e.g., linear blending).	Learned from data or interaction (adaptive blending).
Human model	Handcrafted assumptions (e.g., fixed goals, behavior models).	Learned from demonstrations, corrections, or preferences.
Adaptability	Static and task-specific.	Dynamically adapts to user skill, task difficulty, or context.
Policy design	Manually designed or based on classical control theory.	Learned via trial and error or human-guided reinforcement.
Goal inference	Often assumed known or manually specified.	Inferred from partial input using learned latent representations.
Personalization	Rare or limited.	Personalization through user-specific reward models or policies.
Uncertainty	Deterministic or basic probabilistic models.	Bayesian or ensemble-based exploration and uncertainty estimation.
Performance over time	Performance remains fixed unless re-tuned.	Improves with interaction and user feedback.
Application examples	Shared control of wheelchairs, surgical robots (with fixed blending).	Preference-based RL, interactive RL with goal inference, assist-as-needed exoskeletons.
Limitations	Hard to generalize across users/tasks, inflexible.	May require more data and training time.
Advances	Intelligent driving systems [200], assistive robotics [201], teleoperation [202], and autonomous vechile [203].	Agent without prior knowledge of environments [208,209] and safety constraints [210,211].

In one of their early works, Reddy et al. presented a model-free deep RL framework for SA, where a semi-autonomous agent assists human users in real-time control tasks without prior knowledge of the environment dynamics, user policy, or goal representation [208]. The method combines human input with RL by embedding user actions into the agent's observations and using DQN to balance task optimization

with adherence to user commands. Specifically, a behavior policy selects actions close to user inputs while ensuring they meet a performance threshold. In their work, experiments in simulated and real-world tasks demonstrated that the approach improved task success rates and reduced crashes compared to unassisted human or autonomous control. In addition, the framework was adapted to diverse user behaviors and could leverage additional structure (e.g., known goals) when available, offering a flexible solution for human-robot collaboration in complex, uncertain environments.

In another work, Schaff and Walter introduced a residual policy learning framework for SA in continuous control tasks, enabling collaboration between humans and robots without restrictive assumptions of prior knowledge of the environment dynamics, goal space, or human intent [209]. The method leverages model-free deep RL to train an agent that minimally adjusts human actions to satisfy goal-agnostic constraints, while preserving the human's control authority. By framing the problem as a constrained MDP and using residual policy learning, the agent learns corrective actions added to human inputs, ensuring safety and efficiency. This approach generalizes across tasks and users, highlighting its potential for real-world applications where human-robot collaboration is essential.

Furthermore, Li and co-workers proposed a human-in-the-loop RL method with policy constraints to enhance shared autonomous systems where humans and agents collaborate on tasks [210]. The approach addresses key challenges in RL, including the high cost and safety risks associated with exploration during training, as well as the potential for human errors during execution. By incorporating human prior knowledge as policy constraints, the method accelerates training by guiding agents' exploration and ensures safety during testing through an arbitration mechanism that overrides unsafe human inputs. Experiments in a Lunar Lander simulation demonstrated that the proposed approach improved convergence speed, reduced risky behaviors, and increased task success rates compared to standard RL methods, highlighting its effectiveness in balancing efficiency and safety in human-machine systems.

For a specific application, Backman et al. presented a SA approach to assist novice and expert pilots in safely landing unmanned aerial vehicles (UAVs) under challenging conditions, such as limited depth perception and ambiguous landing zones [211]. The system consists of a perception module that encodes stereo RGB-D camera data into a compressed latent representation and a policy module trained using RL with simulated users to infer pilot intent and provide control inputs. The assistant, trained solely in simulation, was validated in a physical user study (n = 28), significantly improving task success rates without prior knowledge of the environment or pilot goals. The study demonstrated that the assistant enabled participants of all skill levels to perform at or above the proficiency of the most experienced unassisted pilots, while reducing workload and improving efficiency.

4.4 Human-in-the-Loop Querying

By combining the strengths of feedback and demonstrations, SA enables humans and AI to coordinate effectively in real time. Yet, smooth collaboration also requires agents to recognize when human input is most valuable. HLQ complements SA by shifting the interaction from reactive intervention to proactive engagement: instead of waiting for humans to take control, the agent can strategically request guidance at critical moments. Together, these approaches reduce cognitive burden, enhance trust, and ensure that human expertise is applied precisely where it has the greatest impact.

HLQ is an approach within HAII where human users actively engage with AI systems by issuing queries, providing feedback, or validating outputs. Rather than operating purely autonomously, the AI system incorporates human expertise and preferences to guide or refine its decision-making process [212]. This interactive paradigm is particularly useful in complex or high-stakes applications, such as medical diagnosis, scientific discovery, or content moderation, in which purely automated models might struggle with

ambiguous cases or domain-specific nuances [213]. By enabling iterative querying and feedback, human-in-the-loop systems can achieve higher accuracy, improve transparency, and align outcomes more closely with human values and expectations [214].

HLQ supports the HAII pillars by integrating human feedback directly into AI decision-making through interactive queries. This strongly enhances trust, as humans can validate, correct, or refine AI actions, ensuring outcomes align with expectations. HLQ also strongly improves usability, allowing humans to guide AI behavior dynamically and intuitively within their natural workflow. It moderately supports understandability; the process of giving feedback and seeing the AI's response helps humans build a mental model of its reasoning, though full transparency into its internal logic may not be achieved. By incorporating humans in the loop, HLQ ensures that AI systems remain accountable, adaptive, and aligned with user intent.

In the context of RL, HLQ has emerged as a powerful method to augment traditional reward-driven learning [215]. Instead of relying solely on environmental rewards or predefined objectives, RL agents can actively query humans to receive additional feedback about specific actions or states. This can take the form of explicit scalar rewards, natural language explanations, or preference comparisons, effectively shaping the agent's behavior toward more desirable policies [216,217]. Such integration not only speeds up convergence in environments where rewards are sparse or delayed but also helps embed human norms and safety considerations into the learning process. As a result, human-in-the-loop querying bridges the gap between purely autonomous agents and HCAI systems, fostering more trustworthy and adaptable decision-making [218].

While HLQ and LHF both aim to integrate human knowledge into AI systems, they differ in how and when human input is collected and applied. LHF typically involves passively receiving evaluations or preferences from humans after the AI generates outputs, such as preference ranking, demonstration, or critique, which are then used to adjust the model. In contrast, HLQ is a more proactive process, where the AI agent initiates interaction by selectively querying humans to resolve uncertainty or gather targeted information during training or inference. This active querying enables the AI to focus human effort on the most ambiguous or impactful decisions, potentially improving sample efficiency and alignment with human intent. Both approaches enhance human-AI collaboration, but HLQ is distinguished by its dynamic, interactive nature, where the AI system itself drives the request for human guidance.

In RL, several techniques implement HLQ to incorporate human knowledge efficiently. One common method is active preference learning, where the agent presents pairs of trajectories or actions and queries the human to indicate which is preferred, allowing the agent to learn a reward model that better reflects human values [219,220]. Another approach involves reward shaping via queries, where the agent asks for additional scalar rewards or corrections when encountering states with high uncertainty or conflicting signals from the environment [221]. Critique-based learning allows humans to provide natural language or symbolic feedback on specific behaviors, which the agent can translate into policy adjustments [222]. Additionally, interactive policy adjustment methods let agents query humans during exploration to avoid dangerous or unethical actions, supporting safer learning in real-world tasks [223]. Together, these techniques help RL agents efficiently allocate limited human input to the most informative parts of the state or action space, improving learning speed, safety, and alignment with human intent. We will focus on reviewing active preference learning and reward shaping via queries, which have recently benefited from LLMs. The comparison between the two techniques is listed in Table 7.

Table 7: The comparison between active preference learning and reward shaping via querying

	Active preference learning	**Reward shaping via querying**
Querying type	Requests human to compare pairs of trajectories or behaviors (relative preference).	Requests human to provide scalar feedback or corrections on specific states/actions (evaluative feedback).
Human input	Relative preference judgments (“Which behavior do you prefer?”).	Scalar rewards or corrections (“This action is good/bad”).
Goal	Learn or refine a reward model representing human values.	Directly shape or augment the reward signal to guide learning.
Feedback timing	Typically after observing pairs of behaviors.	Typically at uncertain or critical decision points.
Strengths	Good for complex, subjective, or hard-to-specify rewards; humans better at comparisons.	Simple and intuitive feedback; speeds up learning in sparse or noisy reward environments.
Limitations	Can be computationally expensive; requires generating meaningful behavior pairs.	Depends on consistent and reliable human scalar feedback; may miss nuanced preferences.
Use of LLMs	LLMs generate informative query pairs, predict human preferences, and provide natural language explanations or justifications.	LLMs parse natural language feedback, ask clarifying questions, and generalize shaping rewards to unseen states.

4.4.1 Active Preference Learning

Active preference learning, also known as learning from pairwise comparisons, is an HLQ approach where the agent actively requests humans to compare short trajectories, state-action sequences, or outcomes by asking questions such as “Which do you prefer?” Rather than requiring humans to explicitly design a complete reward function, this method leverages relative human judgments to fit a reward model that more accurately captures human values and preferences. This approach is highly sample-efficient since it focuses human effort on simple comparison tasks instead of complex reward engineering. Additionally, active preference learning scales well to complex domains, including robotics, video games like Atari, and other simulated environments, making it a popular and effective technique for aligning RL agents with human intentions.

Ge et al. presented a method for designing reward functions in RL for autonomous driving navigation using active preference learning to align rewards with human intentions [224]. The method addresses challenges in reward design and enhances policy learning by better capturing human preferences. Traditional reward functions often fail in complex, continuous state spaces, resulting in suboptimal policies. The proposed method leverages human preferences to iteratively refine reward weights, using mutual information to generate informative queries and the No-U-Turn Sampler (NUTS) for efficient belief model updates. Experiments in the CARLA simulation environment demonstrated that this approach significantly improves the success rate of navigation tasks (up to 60%) compared to baseline methods, enabling autonomous vehicles to navigate between random start and end points using only forward vision, without reliance on high-precision maps or predefined routes.

When combined with LLMs, active preference learning becomes even more powerful and efficient. LLMs can help design smarter queries by generating trajectory pairs that are more informative, diverse, or challenging, rather than relying on random sampling; this ensures each human comparison provides maximal learning value. Beyond query generation, LLMs can also act as preference predictors: once some human preference data is collected, the LLM can generalize and predict likely human choices for new trajectory pairs, reducing the need for frequent human input and improving sample efficiency. Additionally, LLMs can produce natural language justifications explaining why certain behaviors might be preferred, which not only aids human understanding but also encourages humans to provide more consistent and thoughtful feedback. Together, these enhancements make the querying process more strategic, scalable, and aligned with human values

Muldrew and co-workers employed active preference learning to enhance the efficiency of fine-tuning LLMs using human or AI feedback, specifically with Direct Preference Optimization (DPO) [225]. Unlike traditional approaches that collect preference labels upfront, this approach iteratively selects the most informative prompt-completion pairs for labeling, based on predictive entropy and preference certainty, to optimize the use of limited labeling resources. On the other hand, this method avoids the complexity of RL with human feedback and leverages DPO's implicit reward model, demonstrating practical advantages in alignment and efficiency.

In another study, Mehta et al. introduced an approach of sample-efficient preference alignment in LLMs using active exploration, optimizing human feedback collection to align LLMs with user preferences [226]. The authors formalized the problem as an active contextual dueling bandit task, where an algorithm actively selects prompts and completions for human feedback to maximize learning efficiency. Their approach leverages a contextual Borda function to reduce preference-based feedback to a single-action optimization problem, with theoretical guarantees on polynomial worst-case regret. The method is extended to LLMs via DPO, with online and offline variants that use dropout-based uncertainty estimation to prioritize informative data points.

Mahmud et al. introduced MAPLE (Model-guided Active Preference Learning), a framework that integrates LLMs with Bayesian active learning to efficiently infer human preferences from natural language feedback and pairwise trajectory rankings [227]. MAPLE leverages LLMs to model preference distributions and employs a language-conditioned active query selection mechanism to prioritize informative and easy-to-answer queries, reducing human effort. The framework combines linguistic feedback (e.g., explanations) with conventional preference data (e.g., trajectory rankings) within a Bayesian framework, enhancing interpretability and sample efficiency. Additionally, MAPLE includes LLM-guided weight sampling, oracle-guided active query selection, and modular preference functions for adaptability.

Similarly, Wang et al. presented a novel framework that enhances preference-based RL by leveraging crowdsourced LLMs as synthetic teachers [228]. Unlike traditional preference-based RL that relies on extensive human feedback and is resource-intensive, this framework uses multiple LLM agents to generate diverse evaluation functions for robot trajectories, aggregating their preferences via Dempster-Shafer Theory for nuanced and adaptive feedback [229,230]. A human-in-the-loop module further refines these evaluations based on user inputs, enabling personalized robot behaviors in human-robot interaction scenarios.

4.4.2 Reward Shaping via Querying

Reward shaping via querying is a human-in-the-loop approach where an RL agent actively requests scalar feedback or corrective signals from humans during the learning process. Instead of relying exclusively on predefined or sparse environmental rewards, the agent strategically queries humans at uncertain, risky,

or high-impact decision points—asking questions. The human-provided feedback is then used to directly augment or reshape the reward signal, effectively steering exploration and policy updates in a way that aligns better with human goals and safety constraints. This technique is particularly valuable in complex or real-world environments where designing a fully accurate reward function is difficult, and timely human input can prevent unsafe or undesirable behaviors. By focusing human effort where it is most informative, reward shaping via querying helps improve both learning efficiency and alignment with human intent.

When combined with LLMs, reward shaping via querying becomes even more expressive and efficient. Rather than limiting human input to scalar values alone, agents can actively solicit free-form natural language feedback, which LLMs parse and translate into actionable shaping signals or policy updates. LLMs also enable agents to ask clarifying follow-up questions when human feedback is ambiguous or incomplete, ensuring that the guidance is correctly interpreted. Additionally, by leveraging their extensive pretrained knowledge, LLMs can help generalize shaping rewards to similar but unseen situations, thereby reducing the number of direct human queries required. Together, these capabilities make reward shaping via querying not only more data-efficient but also better equipped to capture subtle preferences, safety considerations, and domain-specific nuances that are difficult to express through numeric rewards alone.

Chan et al. introduced Attention-Based Credit (ABC), a reward shaping method that implicitly queries the transformer architecture of a reward model from LLMs to densify sparse human feedback [231]. Instead of relying only on a final scalar reward, ABC extracts the reward model's attention maps, effectively requesting which tokens most influenced its judgment, and redistributes the reward accordingly. This process acts as an automated query mechanism, where the attention weights serve as fine-grained, model-generated feedback, similar to potential-based shaping. By leveraging the LLMs' built-in attention dynamics, ABC provides token-level credit assignment without additional human queries or computational overhead.

Additionally, a fair and stable RL was introduced for optimizing LLMs by dynamically balancing diverse rewards, such as factuality, completeness, and harmlessness, according to human or AI-generated feedback [232]. This method implicitly queries the reward landscape through mirror descent-based weight updates, treating the composite reward as a dynamic weighted sum to minimize disparity and maximize stability across conflicting objectives. By framing the problem as a max-min optimization inspired by distributionally robust optimization [233], the method automatically adjusts reward weights without gradients, ensuring balanced policy improvements. Experiments on dialogue, question/answer, and safety tasks showed that this RL approach outperforms fixed-weight baselines and avoids overfitting to dominant rewards. While not requiring explicit queries, the approach functionally mimics internal querying by iteratively probing and rebalancing reward signals, offering a scalable alternative to manual reward shaping.

Wu et al. presented an RL framework that enhances LLM training by leveraging fine-grained human feedback to create dense, multi-category reward signals [234]. This approach is closely related to reward shaping via querying, as it relies on querying external sources (e.g., humans or auxiliary models like Perspective API) to provide intermediate rewards during generation. Moreover, RL with fine-grained human feedback extends traditional reward shaping by structuring feedback into specific error types (including factual inaccuracies, irrelevance, and toxicity) and assigning rewards at granular levels (per sentence or sub-sentence), rather than relying on a single holistic score. While offering automated feedback, the fine-grained human feedback mechanism emphasizes human-annotated feedback for training specialized reward models. This framework addresses limitations of sparse feedback in RL, providing a more precise way to align LLMs with human preferences, albeit at a higher computational cost.

4.5 Explainable Reinforcement Learning

LHF, LHD, SA, and HLQ collectively form a powerful toolkit for aligning and adapting AI behavior with human needs. Still, their effectiveness hinges on human understanding: users must be able to interpret why an agent requests input, follows certain strategies, or overrides prior preferences. XRL complements all prior methods by providing transparency, making the agent's reasoning process accessible and actionable. In doing so, XRL strengthens the feedback loop, empowering humans to deliver more precise feedback, demonstrations, or answers to queries, and fostering mutual trust in SA.

XRL is a complementary approach within human-AI interaction that focuses on making an agent's decision-making process transparent and interpretable to human users [235,236]. Rather than actively querying humans for feedback, XRL helps humans understand why an agent chose certain actions or followed specific policies by generating explanations, visualizations, or natural language rationales. This transparency enables humans to build trust in the system, identify potential flaws or unsafe behaviors, and provide more informed guidance or corrections when necessary. While XRL does not directly shape the learning process through feedback, it plays a vital role in fostering collaboration, accountability, and effective oversight in human-in-the-loop systems.

XRL supports the HAII pillars by providing interpretable explanations for decisions made by RL agents. This strongly enhances trust, as users can verify that AI actions align with expectations and goals, making the system more reliable and accountable. XRL provides foundational support for usability; the insights from explanations can help users interact with the AI more effectively, though this depends on the explanations being clear and concise to avoid adding cognitive load. It strongly supports understandability, as explanations reveal the rationale behind AI actions, highlighting the factors and strategies driving behavior. By making RL agents' decision-making transparent, XRL facilitates safe, informed, and user-aligned human-AI collaboration.

In RL, HAII through XRL is typically implemented by designing agents that can produce interpretable representations of their internal policies, value functions, or decision pathways [237,238]. Common techniques include visual explanations such as saliency maps that highlight influential states or features, counterfactual explanations showing how behavior would change under different circumstances, and natural language rationales that describe why specific actions were chosen. Other methods extract symbolic rules or simplified models summarizing agent behavior, making it easier for humans to inspect and understand complex policies. These mechanisms transform opaque, black-box learning processes into transparent models that support human users in debugging, supervising, and refining agent behavior.

While both querying-based methods and XRL are part of the broader landscape of HAII, they serve distinct yet complementary roles. Query-based methods, including active preference learning and reward shaping via querying, involve the agent actively requesting human feedback during learning, that is, using it to directly adjust the reward model or policy. In contrast, XRL focuses on the agent explaining its reasoning to the human, enabling the human to interpret and potentially correct or refine the agent's behavior. On the other hand, querying-based methods directly shape learning by incorporating human judgments, whereas XRL indirectly supports alignment by equipping humans with insight into why the agent acts as it does. Together, these approaches form a synergistic loop: querying leverages human input to refine agent behavior, and XRL helps humans provide better, more context-aware feedback by making the agent's internal processes transparent. The comparison is shown in Table 8.

Table 8: The comparison between querying-based methods and XRL

	Querying-based methods	XRL
Agent role	Agent actively queries humans for feedback to guide learning.	Agent explains its decisions, policies, or reasoning to humans.
Human role	Provide direct feedback (comparisons, scalar rewards, corrections).	Interpret explanations; supervise, debug, or refine the agent.
Primary goal	Improve or shape learning directly by incorporating human feedback.	Increase transparency, trust, and human understanding.
Interaction flow	Human feedback → modifies reward model or policy.	Agent's internal reasoning → presented to human.
Effect on learning	Direct: changes the reward or policy updates.	Indirect: helps humans give better or safer feedback.
Complementary	Uses human input to align behavior.	Helps humans understand behavior to offer better input.

A recent survey highlighted the critical role of XRL in HAII by emphasizing how explanations can bridge the gap between opaque RL agents and human users [239]. The authors argued that understanding an agent's decisions, such as why it took a specific action or how it learned its policy, is essential for building trust, enabling human oversight, and facilitating real-world deployment. Their taxonomy categorizes XRL methods based on how they support human interpretability: feature importance provides immediate, action-level insights, the learning process and MDP reveal training dynamics and reward structures, and policy-level offers long-term behavioral summaries [240–243]. All of these help humans diagnose, validate, and collaborate with AI systems. The paper also critiques current evaluation practices, noting that many methods rely on visualizations rather than human-centered metrics like comprehensibility, actionability, or cognitive load. By framing XRL as a tool for enhancing human-AI collaboration, the survey underscores its potential in high-stakes domains like healthcare, autonomous driving, and robotics, where human trust and intervention are paramount.

In another study, Raees et al. systematically reviewed the evolution of HAII, shifting the focus from XAI to more active and collaborative forms of Interactive AI [244]. It highlights the limitations of current research, which often prioritizes system performance and passive explanations over empowering users with agency, such as the ability to contest, adapt, or co-design AI systems [245]. The analysis identifies gaps in practical implementations, particularly in high-stakes domains like healthcare, and underscores the dominance of low-risk applications where interactivity is more tolerated [246–248]. Key findings reveal that while transparency and user experience are central to HCAI, few studies enable meaningful user control over AI mechanics, with Interactive Machine Learning (IML) emerging as a promising approach for fostering collaboration [249]. The paper advocates for balancing AI autonomy with human agency, emphasizing the need for participatory design, multi-modal interfaces, and real-world experimentation to advance HAII beyond theoretical proposals.

A recent study used deep RL to design a resource allocation mechanism that promotes sustainable cooperation in a multiplayer trust game, where participants must reciprocate investments to maintain a shared pool of resources [250]. In this study, the RL agent learns a dynamic policy that balances fairness and efficiency, temporarily sanctioning defectors while redistributing resources more equally when the pool is

abundant. By analyzing the agent's behavior, the authors derived a simpler, explainable heuristic that mimics the RL policy and is preferred by human players for its transparency. The results demonstrated that AI-driven mechanisms can outperform traditional baselines in fostering long-term cooperation, with implications for real-world applications like welfare systems or environmental management.

Moreover, Gebellí et al. proposed a framework for Personalized Explainable Robots Using LLMs, focusing on enhancing human-robot interaction (HRI) through XAI grounded in Theory of Mind (ToM) [251]. The system reconciles the robot's internal model with its estimate of the human's mental model to generate tailored explanations for users. Notably, this study uses LHF to fine-tune LLMs for deeper personalization, which aligns with XAI principles in RL by adapting explanations based on iterative user interactions. The framework demonstrates practical applications in a hospital patrolling robot, showcasing adaptive explanations for diverse user types (e.g., non-active, active, or returning users), emphasizing the role of XAI in dynamic, real-world RL environments where transparency and user trust are critical.

5 Case Studies

This section presents three case studies demonstrating the practical application of newly developed HAII algorithms in RL. The first study addresses the critical gap between theoretically optimal AI policies and their practical adoption, focusing on farmers who rely on experiential knowledge. While RL can generate "trust-agnostic" policies that outperform expert strategies in simulation, these solutions often prioritize numerical efficiency over practical feasibility, leading to low user acceptance. To bridge this gap, we introduce a novel, quantitative trust model that translates subjective farmer confidence into a computable score integrated directly into the RL objective. By framing the problem as multi-objective optimization, we pioneer a method for deriving a "trust-aware" policy that Pareto-dominates other strategies, successfully balancing high agricultural productivity with user trust to enhance real-world deployability. This case directly verifies the theoretical pillar of trust, demonstrating that explicit modeling of human confidence is necessary for practical acceptance of AI policies. It also expands the usability pillar, showing that usability is not limited to interface design but also encompasses the alignment of AI optimization with human decision-making practices. Finally, it reveals limits to understandability, since while farmers could observe how trust scores influenced outcomes, the internal policy trade-offs of the RL system remained difficult to interpret, highlighting the need for improved transparency in multi-objective AI.

The second case study integrates ethical reasoning into robotic motion planning for a medical clinic, motivated by the need to ensure autonomous systems operate safely in high-stakes environments. The novel contribution is a hierarchical framework that formalizes ethical norms within a POMDP as distinct hard and soft constraints. Deontological obligations (hard constraints) are mapped to LTL for guaranteed compliance, while utilitarian permissions (soft constraints) are encoded into the reward function to incentivize adaptive behavior. This hybrid approach ensures rigorous adherence to critical safety rules while permitting flexibility on secondary guidelines, providing a robust methodology for embedding ethical decision-making into RL under uncertainty. The case strongly verifies the trust pillar, showing that explicit ethical guarantees are essential for human reliance on AI in safety-critical contexts. It expands the usability pillar, as the integration of ethical reasoning demonstrates that usability in medical environments depends not only on efficiency and performance but also on compliance with human moral expectations. Finally, it enhances understandability, since the separation of hard and soft constraints provides a transparent logic for why certain motions are strictly prohibited while others remain adaptable, making the AI's decision-making process more interpretable to clinicians.

The third study quantitatively investigates the evolution of human trust in AI through a multi-armed bandit experiment. The motivation is to move beyond subjective surveys by developing a quantitative,

behavioral model of how trust evolves through interaction, which is key to designing AI systems that can calibrate user trust effectively. The novelty of this case study is the application of an RL paradigm to model the human's internal decision-making process itself. We frame human trust as a value function over states of "following" or "ignoring" AI advice, which is updated based on received rewards. This computational model, fitted to experimental data, provides a novel lens to objectively track and predict trust dynamics over time. The case verifies the trust pillar by demonstrating that trust can be rigorously quantified as a dynamic, learnable function, offering predictive power beyond subjective measures. It also expands the usability pillar, since modeling trust dynamics allows for adaptive AI designs that calibrate assistance in line with user confidence, thereby improving sustained interaction quality. Finally, it contributes to understandability, as the trust value function makes the evolution of human-AI reliance explicit and interpretable, providing researchers and practitioners with a structured framework to anticipate and manage user behavior over time.

5.1 Farmer Trust Model in Optimal Agricultural Management

Precision agriculture increasingly integrates advanced technologies, such as remote sensing, UAVs, and ML, to address global food shortages [252]. RL, in particular, has been applied alongside crop simulators to optimize agricultural management practices, including irrigation and fertilization strategies [253–255]. These RL frameworks typically aim to maximize crop yield while minimizing resource use (e.g., water and nitrogen) and environmental impacts, such as nitrate leaching.

In this case study, we focused on maize crops in Iowa, where conventional farming practices rely primarily on rainfall rather than irrigation. To optimize fertilization strategies, we employed Gym-DSSAT, a crop simulator serving as a virtual agricultural environment, to approximate system dynamics and predict crop yield [256]. The simulator models 28 internal state variables, capturing diverse crop growth stages and environmental conditions. Building on our prior work [4,255,257], we framed this problem as a POMDP to account for real-world agricultural complexity and variability. From these, we selected 10 key variables (Table 9) as primary observations. Consequently, we implemented a GRU-based DQN approach, as detailed in Section 3.2, to derive optimal agricultural strategies (policies).

Table 9: State variables of the agricultural environment used in this study as observations

Variables	Definitions
cumsumfert	Cumulative nitrogen fertilizer applications (kg/ha)
dap	Days after planting
istage	DSSAT maize growing stage
pltpop	Plant population density (plant/m^2)
rain	Rainfall for the current day (mm/d)
sw	Volumetric soil water content in soil layers (cm^3 [water]/cm^3 [soil])
tmax	Maximum temperature for the current day (°C)
tmin	Minimum temperature for the current day (°C)
vstage	Vegetative growth stage (number of leaves)
xlai	Plant population leaf area index

The action space consists of discrete nitrogen application rates N_t on day t, ranging from 0 to 200 kg/ha in 10 kg/ha increments. The simulator utilizes soil data from an experimental field in Iowa and 1999 weather conditions to predict daily nitrate leaching L_t and end-of-season corn yield Y. Accordingly, we adopted the following reward function [4].

$$R_t = \begin{cases} w_1 Y - w_2 N_t - w_3 L_t & \text{at harvest} \\ -w_2 N_t - w_3 L_t & \text{otherwise} \end{cases} \tag{15}$$

where the weight coefficients include $w_1 = 0.07087$, $w_2 = 0.39$, $w_3 = 1.95$.

While RL-generated policies outperformed Gym-DSSAT's expert policy in agricultural outcomes, they often prioritized theoretical agricultural optimization over farmers' practical needs and experiential knowledge. To bridge this gap, we developed a trust model that quantifies farmers' confidence in AI-driven strategies. By integrating farmers' behavioral patterns with publicly available agricultural data, our model enables the adaptation of management solutions to better align with real-world expectations. In this case study, we implemented a three-dimensional trust framework evaluating: (1) technical ability, (2) system benevolence, and (3) operational integrity [258]. This approach yields a comprehensive trust score (T_s) calculated as:

$$T_s = T_{\text{ability}} \times T_{\text{benevolence}} \times T_{\text{integrity}} \tag{16}$$

where $T_{\text{integrity}} = 1.0$, assuming that the RL agent developed by a research group with outstanding expertise in the field delivers AI-generated policies with equivalent transparency and fairness.

Within trust theory, ability is defined as the collection of skills, competencies, and traits that empower an entity to influence outcomes in a specific domain, underscoring the critical role of demonstrated proficiency [258]. The ability of the RL agent's recommended fertilization strategy was assessed by comparing its performance-in terms of yield output, total nitrogen application ($\sum N_t$), and application frequency (FF)-against 1999 U.S. agricultural benchmarks in Eq. (17): 8649 kg/ha average corn yield, 213 kg/ha average nitrogen application, and the typical 1–3 applications per growing season [259,260].

$$T_{\text{ability}} = \frac{Y}{8649} \times \frac{213+1}{\sum N_t + 1} \times \frac{2}{(|FF-2|+2)} \tag{17}$$

Benevolence in trust theory reflects the trustee's perceived altruistic intentions [258]. While the original reward function in Eq. (15) penalizes nitrate leaching, AI policies may over-prioritize environmental benefits, potentially raising farmer concerns about whether AI's decision-making aligns with their interests. Using a benchmark (0.26 kg/ha total leaching) from the expert policy provided by Gym-DSSAT, we assess deviations' impact on perceived benevolence in Eq. (18), as both higher and lower values may affect trust in the RL agent's fertilization strategy.

$$T_{\text{benevolence}} = 1 - \left| \frac{\sum L_t - 0.26}{0.26} \right| \tag{18}$$

Consequently, this case study addresses the dual challenge of optimizing both expected returns (Eq. (15)) and decision-making trustworthiness (Eq. (16)). Recognizing the inherent trade-off between agricultural productivity and trust, we employ multi-objective RL combining POMDP and GRU-based DQN to identify the Pareto front and determine the optimal policy [261]. We distinguish between two optimal policies: (1) the trust-aware policy that balances both objectives, and (2) the trust-agnostic policy that solely maximizes expected returns. Table 10 compares their agricultural outcomes against the Gym-DSSAT expert policy.

Table 10: Agricultural outcomes following different nitrogen fertilization strategies

	Trust-aware optimal policy	**Trust-agnostic optimal policy**	**Expert policy**
Total reward	583	585	568
Corn yield (Kg/ha)	9165	9248	9248
Nitrogen fertilizer amount (Kg/ha)	170	180	224
Nitrate leaching (Kg/ha)	0.14	0.09	0.26
Number of fertlizer applications	2	7	1
Trust score	0.715	0.012	0.678

The table shows that the trust-agnostic optimal policy achieves the highest reward among the three strategies. However, it records the lowest trust score because it does not consider farmers' trust. This low trust mainly results from its frequent, small-dose fertilization schedule aimed at minimizing nitrate leaching, which substantially increases the farmers' workload and labor hours. In comparison, the expert policy adopts a different approach. By limiting the number of fertilizer applications and maintaining nitrate leaching at baseline levels, it attains a high trust score in the simulation. Yet, its greater fertilizer use and higher nitrate leaching reduce overall performance, yielding the lowest reward among the three.

The trust-aware optimal policy offers a more balanced solution. It reaches a reward similar to that of the trust-agnostic optimal policy while also achieving a higher trust score than the expert policy. This outcome arises from careful regulation of both the timing and amount of fertilizer applied. Fig. 6 provides a detailed view of the fertilization schedule in conjunction with precipitation patterns.

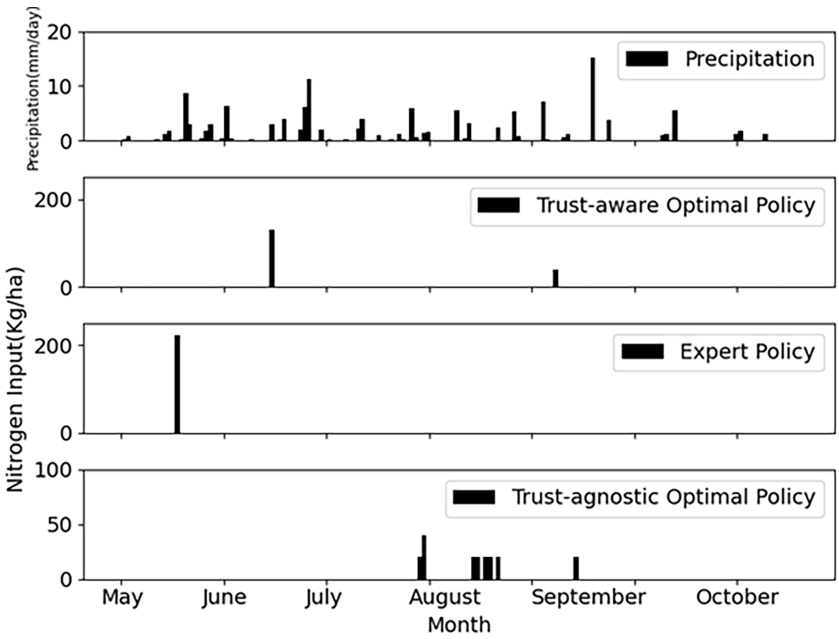


Figure 6: Precipitation during the simulation and fertilization under different policies

Fig. 6 highlights the differences in fertilization strategies. The trust-agnostic policy relies on frequent but small applications, concentrating most fertilizer in August, the month with the lowest rainfall in 1999. By avoiding applications during wetter periods, it effectively minimizes nitrate leaching. In contrast, the trust-aware policy takes a simpler approach, applying a large dose at planting followed by a second application three months later. This design reduces farmers' workload, though it is less focused on avoiding rainy days. Despite these differences, the results demonstrate that the trust-aware optimal policy successfully balances high rewards with farmers' trust. Its strategy manages to meet both objectives effectively, offering a practical and efficient approach to fertilization management.

5.2 Ethical Constraints in Robotics Motion Planning

AI ethics involves the study of moral principles and methodologies designed to promote the responsible creation and deployment of AI, especially as AI systems grow increasingly sophisticated. A central challenge lies in developing ethical AI agents that adhere to human values, preferences, and limitations while making morally sound decisions—identifying ethical dilemmas, assessing alternatives, and acting in accordance with established moral standards [262]. The ethical framework often incorporates normative concepts such as obligation, prohibition, and permission, which are deeply rooted in legal theory and formalized ethical systems [263]. These principles are further shaped by foundational ethical theories, including utilitarianism and deontology, each offering distinct perspectives on morality, justice, and human rights. Utilitarianism evaluates actions based on their outcomes, prioritizing the greatest good for the greatest number [264]. In contrast, deontology assesses the morality of actions by their alignment with predefined rules and duties, independent of their consequences [265].

Obligations represent positive normative imperatives, typically articulated as statements about what ought to be done, whereas prohibitions function as their inverse by specifying what must not occur. In contrast, the notion of permission is more nuanced and holds significant importance across various ethical and normative frameworks. A key distinction in classifying permissions lies in differentiating between weak and strong permissions [266]. Weak permission implies that an action is allowed unless explicitly forbidden, operating on a default assumption of permissibility. Conversely, strong permission applies only when a normative system explicitly approves an action [266]. Furthermore, strong permissions may also define exceptions that supersede existing obligations or prohibitions, introducing flexibility into rigid normative structures [267].

We reclassified and reformulated the core ethical norms-obligation, permission, and prohibition-into hard and soft ethical constraints [268]. Hard constraints encapsulate the normative demands of obligations and prohibitions, including their conditional variants, and can be formally specified using LTL alongside complex task requirements. In contrast, soft ethical constraints incorporate both strong and weak permissions, modeled as ethical guidelines with varying degrees of reinforcement or deterrence through a redesigned reward function within a POMDP framework, which accounts for partial observability in the environment.

This framework transforms the original problem into an optimization task seeking the optimal policy over the product of the POMDP and an LTL-induced automaton. Here, the fulfillment of complex tasks and adherence to hard ethical constraints are guaranteed through automaton-based model checking, while soft ethical constraints are promoted by maximizing the accumulated reward. To tackle this problem, we adapt and extend the RL methodology outlined in Section 3, with the updated model representation provided in Fig. 7.

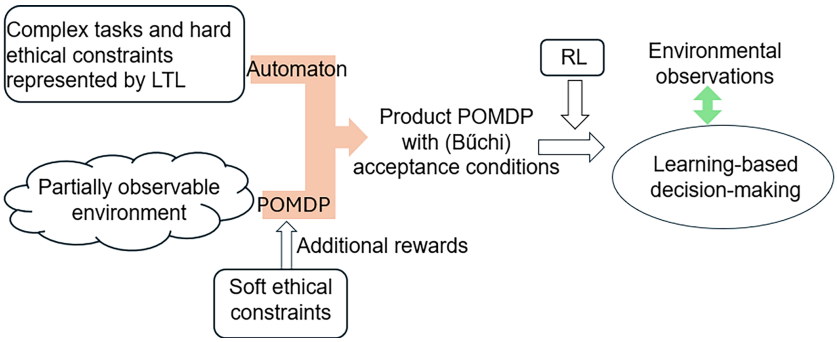


Figure 7: The product of LDGBA and POMDP for RL problems with complex tasks under ethical constraints

This case study examines a medical clinic (see Fig. 8) comprising several specialized areas: four treatment rooms (a–d), a medical storage room (S), a surgical equipment room (Equip), a central front desk (FD), and a charging station (Chrg) for the robotic assistant. Rooms 'a' and 'd' include large windows, while multiple doors ensure accessibility. The clinic's layout is modeled as a 4 × 4 grid world. The robot perceives its surroundings in four fixed directions (North, West, South, East), detecting environmental features such as walls, hallways, doors, and windows. Since these observations are non-unique, that is, multiple states may yield identical sensor readings, the robot operates under partial observability, framing the task as a POMDP.

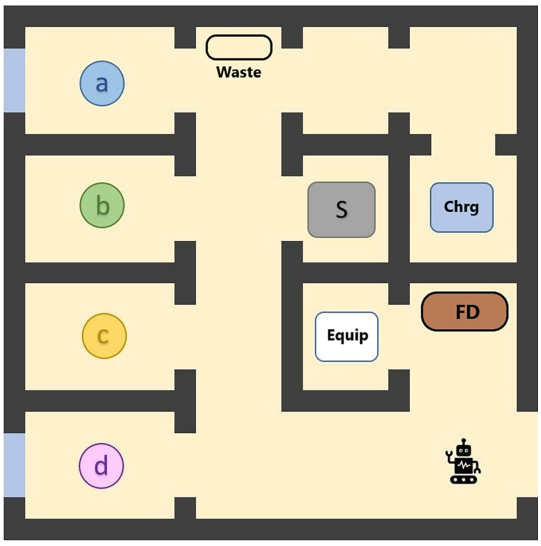


Figure 8: The clinic environment

The robot's motion is restricted to four primary navigation directions: up, left, down, and right, representing simplified discrete actions for path planning. To account for real-world execution uncertainty, we model the robot's movement as stochastic: with a 0.9 probability of successfully moving in the intended direction and a 0.1 probability of deviating uniformly to adjacent directions. Collisions with walls result in the robot remaining stationary. Each movement incurs a penalty (cost = −0.1), while reaching a designated acceptance state yields a reward (+1.0). The clinic's workflow requires the robot to assist doctors in treatment rooms 'a' and 'c' by frequently transporting medical supplies from the equipment room (Equip). The robot's

operational cycle consists of: Retrieving supplies from the Equipment room, delivering them to 'a' or 'c', returning to the charging station for replenishment. This cycle repeats indefinitely, with the critical constraint that the robot must never enter the storage room. The task specifications are formally encoded as an LTL formula, denoted as φ_{task}, in Eq. (19).

$$\varphi_{task} = \Box\Diamond(Equip \wedge \Diamond((a \vee c) \wedge \Diamond Chrg)) \wedge \Box(\neg S) \tag{19}$$

where $Equip$, a, c, $Chrg$, and S denote atomic propositions indicating that the robot is at the surgical equipment room, treatment room 'a', treatment room 'c', the charging station, and the medical storage room, respectively.

As shown in Fig. 8, a waste bin (Waste) is positioned at the end of the hallway. Although primarily designated for general trash, it is sometimes used by patients to discard medical waste (e.g., used wipes, specimen containers), which introduces biohazard risks to the clinic. To uphold hygiene and safety standards, an ethical protocol is implemented: when the bin's sensor detects medical waste, the robot must immediately transport it to a secure disposal zone near the front desk (FD). The robot is equipped with a sealed, biohazard-safe compartment to prevent contamination of the clinical environment or cross-contact with medical supplies during transport. The waste bin has a 10% probability of containing biohazardous waste at any given time, as determined by historical clinic data. This stochastic occurrence is modeled as a Bernoulli-distributed event in the system's state transitions.

In the first scenario, our architecture enforces ethical compliance through a hard constraint that the robotic platform must execute without exception. Grounded in the clinic's deontological operational policies, the robot initiates immediate countermeasure protocols upon detecting biohazardous materials, prioritizing waste transportation to the authorized decontamination zone (FD) above all other functions. This hard ethical constraint is mathematically formalized as an invariant property in LTL as

$$\varphi_{norm} = \Box(Waste \rightarrow \bigcirc FD) \tag{20}$$

The LTL formula components are formally defined as: (1) Proposition '$Waste$' holds when biohazardous material is present in the disposal receptacle, and (2) Proposition 'FD' becomes true when the robotic agent reaches the front desk disposal station. The complete LTL specification for this case is expressed as where '$Waste$' means the waste bin has the medical waste, and 'FD' indicates visiting the front desk to dispose of the waste. The complete LTL formula for this case is defined as Eq. (21), and the result can be found in [268].

$$\varphi_{hard} = \varphi_{task} \wedge \varphi_{norm} \tag{21}$$

To illustrate how the LTL formulas guide behavior, consider the following scenario: A nurse requests medical supplies in treatment room 'c'. The robot begins its routine by navigating from the equipment room toward 'c', following its planned task cycle. On its way, the robot's waste sensor detects that the hallway bin contains biohazardous material. The hard ethical constraint encoded in Eq. (20) immediately overrides its delivery task: instead of continuing to room 'c', the robot diverts to the front desk 'FD' to dispose of the hazardous waste. Only after fulfilling this obligation does it resume its original route, retrieving a fresh set of supplies before completing the delivery. In this way, the LTL specification acts as a "moral compass," ensuring that urgent ethical requirements always take precedence over efficiency goals, while still guaranteeing eventual task completion.

In another scenario with the utilitarian implementation, the clinical framework prioritizes both efficient patient care in treatment rooms and responsible waste management by treating ethical norms as permission rather than obligation. This soft-constraint approach is encoded in the POMDP model through an additional

dynamic reward function that imposes a minor penalty (−0.01 per movement) when the robot detects medical waste but postpones disposal, creating an opportunity cost that gently incentivizes timely waste management without strictly prohibiting other actions. The penalty resets to zero upon successful waste disposal at the designated front desk location, allowing the robot to automatically restore standard operational priorities while maintaining compliance with ethical requirements. This design effectively balances urgent patient needs with environmental safety protocols through quantifiable, reversible trade-offs in the decision-making framework.

Under the soft ethical constraint, the robot behaves differently. Suppose it is on the way to the charging station after delivering supplies to room 'a' when it detects biohazardous waste in the bin. Instead of immediately diverting to the front desk, the robot evaluates its priorities: because the waste penalty is small compared to the risk of depleting its battery, the robot chooses to recharge first. Only afterward does it return to dispose of the waste, thereby balancing clinical efficiency with environmental safety. This demonstrates how treating ethical requirements as permissions rather than obligations allows the system to adaptively trade off between competing goals, while still promoting desirable ethical outcomes.

Our RL framework operates under the assumption that the robotic agent has no explicit knowledge of the predefined task specification (Eq. (19)). To compensate for this task-agnostic condition, the Q-network processes both raw sensor observations and perceived state label sequences as dual input streams, as detailed in Section 3.4. This architecture enables implicit task learning through the reward structure rather than explicit task awareness. After achieving convergence (Fig. 9), the resulting optimal policy generates the navigation trajectories shown in Fig. 10, which demonstrate three key characteristics under soft ethical constraints: (1) maintenance of primary mission objectives as the dominant path determinant, (2) conditional incorporation of the soft ethical constraint, waste disposal when detected, and (3) robust accommodation of the action uncertainty through stable yet adaptable path variations that preserve overall mission integrity while responding to environmental contingencies.

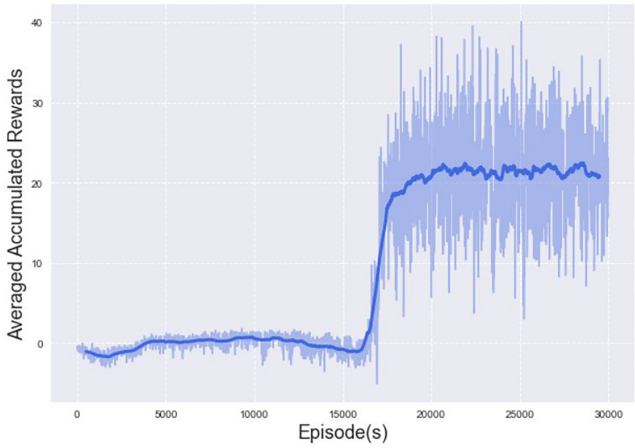


Figure 9: The evolution of the accumulated rewards in the scenario with a soft constraint

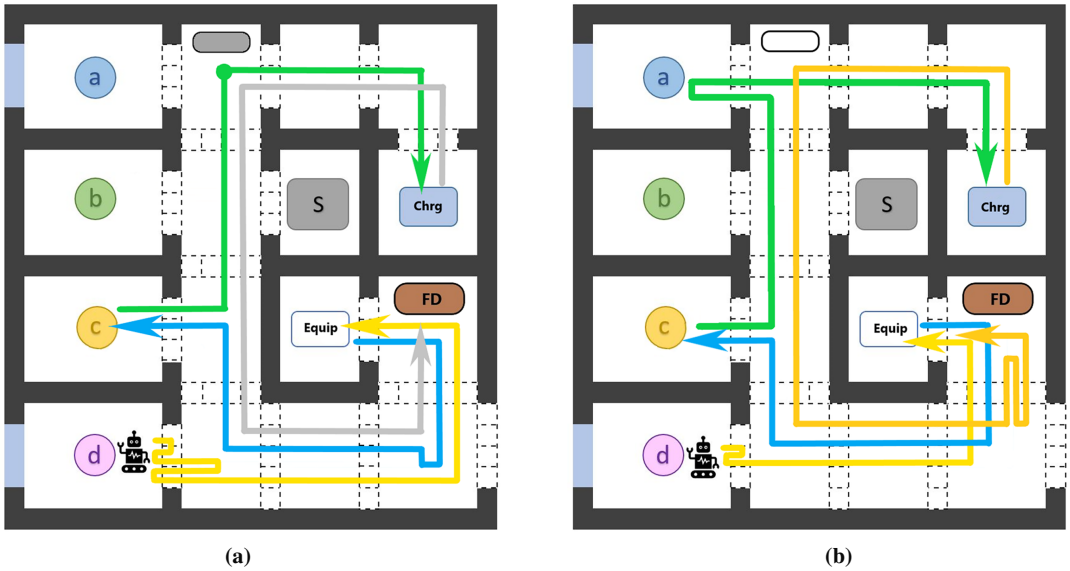


Figure 10: The generated trajectory from the optimal policy demonstrates the robot's behavior when operating under the soft ethical constraint without explicit task awareness. Key observations include **(a)** while en route to the charging station (Chrg), the robot passes the waste bin containing medical waste, and **(b)** the robot accomplishes the task when there is no medical waste

The trajectories depicted in 10 demonstrate structurally similar navigation patterns to accomplish the task under different operational conditions, highlighting how the agent's decision-making adapts to environmental changes while maintaining consistent movement logic. In both scenarios (a) and (b), the agent follows a comparable route from treatment room 'd' to equipment room 'Equip', then to treatment room 'c', and finally toward charging station 'Chrg', with the primary divergence occurring when medical waste is detected (visualized by the gray rounded rectangle in (a)). When biohazardous waste is identified during the charging phase in (a), the same fundamental trajectory is temporarily maintained before the robot picks and transports the bio-hazardous waste to the front desk 'FD', showcasing how the soft constraint implementation modifies but doesn't fundamentally alter the baseline navigation strategy. It should be noted that the robot's trajectory exhibits characteristic oscillatory movements at specific segments, a direct consequence of action uncertainties.

5.3 Evolution of Human Trust in AI

The multi-armed bandit problem is a one-state RL task where, in each trial, an agent selects an action, receives a reward, and resets to the initial state—enabling iterative learning under consistent dynamics. In this case study, we designed an experiment in which a human and an AI agent collaboratively solve such a problem, investigating how human decision-making interacts with AI guidance. Specifically, we quantify the evolution of human trust in the AI throughout the task over the course of 200 trials.

Here, we describe the roles of the AI agent and the human participant, as well as their interaction protocol, in each trial.

- **Initialization:** Five slot machines (labeled 0–4) are generated, each with a random reward value (2–10) and a success probability (0%–100%). If the selected machine pays out (determined probabilistically), the participant receives the assigned reward; otherwise, they receive a fallback reward of 1.

- **AI Recommendation:** The AI agent identifies the optimal machine with the highest expected reward and recommends this machine to the participant. With 20% probability, the AI agent provides an explanation comparing the top two machines, for example, "Machine 3: expected reward 6.58 (70% chance of 9, 30% chance of 1). Machine 1: 6.48 (78% chance of 8, 22% chance of 1). Recommendation: Machine 3."
- **Participant Decision:** The participant selects a machine to play, either following or ignoring the AI's recommendation.
- **Outcome Observation:** The participant receives their reward. At the same time, the AI also observes which machine was selected and the reward obtained.
- **Trust Assessment:** Every fifth trial, the participant completes a brief survey rating their likelihood (in percentage) of following the AI recommendation.

We recruited several undergraduate students to participate in this experiment and collected preliminary data that was analyzed individually. We hypothesize that human trust in AI is based on the human's internal values of following and not following AI recommendations. The internal values are modeled using RL. Unlike conventional one-state multi-armed bandit problems, we introduce two distinct states to track participants' behavioral dynamics: S_0 (representing independent decision-making, where participants either disregard the AI's recommendation or randomly select a non-recommended machine) and S_1 (representing compliance with the AI's suggestion). By monitoring state transitions and updating state values ($V(S_0)$ and $V(S_1)$) throughout the experiment, we can quantitatively model the evolution of human trust in AI based on the collected behavioral data [269].

At the start of each experiment, state values $V(S_0)$ and $V(S_1)$ are initialized to zero. The initial active state is determined by the participant's first survey response: S_1 (AI-following) is assigned if the self-reported likelihood of compliance is ≥50%, otherwise S_0 (independent decision-making) is assigned. During each trial, the current state A (S_0 or S_1) may either persist ($B = A$) or transition to the alternate state ($B \neq A$), contingent on whether the participant follows the AI's recommendation. Following reward R observation, $V(A)$ is updated via Eq. (22), regardless of whether B equals A or not.

$$V(A) = V(A) + \alpha(R + \gamma(V(B) - V(A))) \tag{22}$$

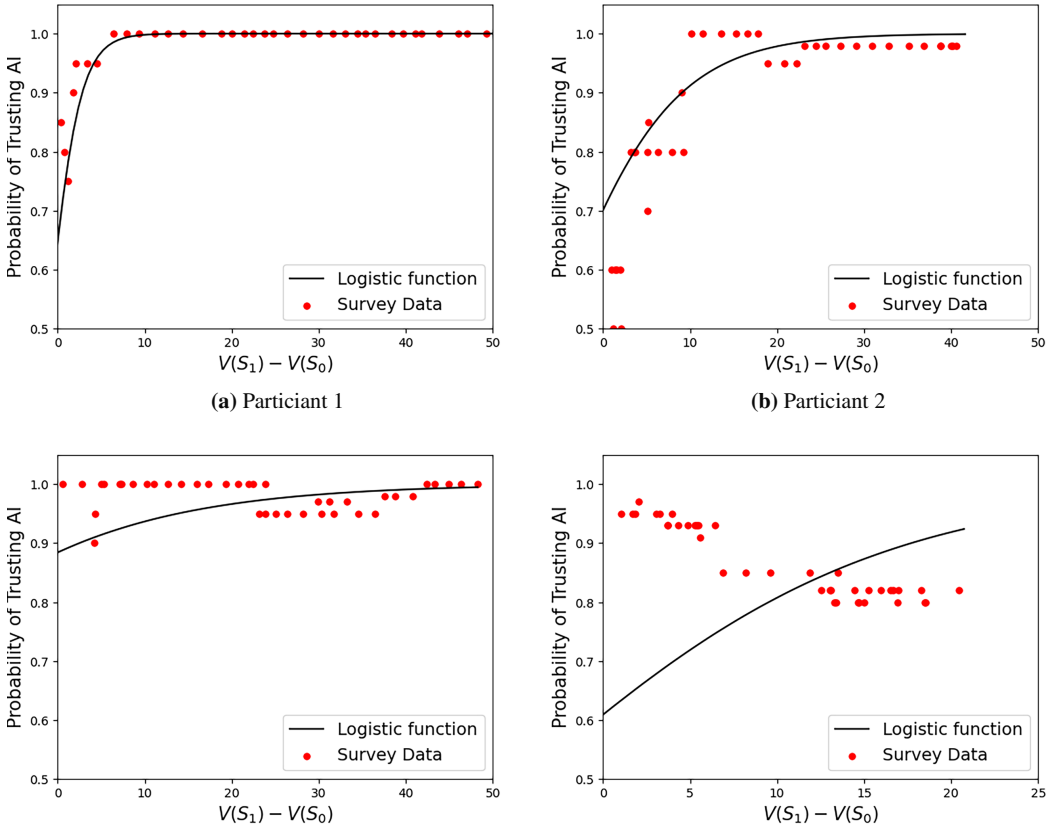
where α (learning rate) and γ (discount factor) are hyperparameters governing the value update process.

Consequently, each experiment generates 200 samples, and an individual data sample contains the state value difference $X = V(S_1) - V(S_0)$ as input and a binary label (follow AI = 1 / not follow AI = 0) as output. This constitutes a binary classification problem, for which we model trust probability via the sigmoid function in Eq. (23).

$$P_{AI} = \frac{1}{1 + e^{aX+b}} \tag{23}$$

where parameters a and b are estimated through logistic regression applied to the experimental dataset.

The dataset is partitioned into training and validation sets, with hyperparameters α and γ optimized via grid search ($\alpha \in [0.05, 0.5]$, $\gamma \in [0.75, 0.95]$). Survey responses are held out as an independent test set to evaluate the performance of the trust probability model (Fig. 11). If the model successfully captures the interaction between human trust and AI, its predictions should align with the participants' subjective survey reports collected throughout the experiment. Consistent with the R^2 values, visual inspection of the plots shows clear variation in how well the logistic function captures individual trust dynamics.



(a) Particiant 1 **(b)** Particiant 2

(c) Particiant 3 **(d)** Particiant 4

Figure 11: Approximated evolution of human trust in AI using a logistic function, derived from experimental data and compared to survey responses

For Participants 1 and 2, the model predictions track the survey data reasonably well (R^2 = 0.640 and 0.623, respectively). In both cases, the logistic curves capture the overall upward trend in trust probability across trials, suggesting that the two-state RL framework provides a suitable abstraction for modeling human trust when participants display moderate variability in compliance. Notably, according to the collected data, we also observed that Participant 2 exhibited increased reliance on AI recommendations when explanations were provided compared to when they were absent.

For Participant 3, the logistic model also appears to align visually with the survey data in Fig. 11. However, the near-exclusive following of AI recommendations yields a strongly negative R^2 (−2.43). This paradox reflects the lack of behavioral variance: although the curve matches the reported trust levels, the triviality of the participant's responses undermines statistical fit and limits interpretability. In contrast, Participant 4 presents a more challenging case. Both the negative R^2 (−6.48) and the divergence seen in Fig. 11 indicate that the logistic function fails to capture their highly fluctuating trust behavior. The participant's survey data suggest dynamic, nonlinear shifts in reliance on AI that cannot be adequately represented by the binary state framework of compliance vs. independence. This underscores the influence of unmodeled factors, such as changing risk preferences, attention fluctuations, or evolving interpretations of AI explanations.

Taken together, the quantitative metrics and Fig. 11 highlight the heterogeneity of trust dynamics in human-AI collaboration. While the proposed two-state RL model is effective for participants with stable yet flexible trust trajectories, it struggles when behavior is either overly uniform (Participant 3) or highly volatile (Participant 4). These limitations suggest that future extensions should incorporate richer state representations, potentially capturing gradations of trust, contextual variability, or cognitive-affective signals, to more faithfully represent human decision-making in collaborative settings.

6 Conclusions and Future Directions

6.1 Conclusions

The integration of HAII into RL represents a transformative approach to developing intelligent systems that are not only technically proficient but also aligned with human values, trust, and usability. This review has systematically explored the foundational pillars of HAII, including trust, usability, and understandability, and HAII implementation through five key approaches: LHF, LHD, SA, HLQ, and XRL. Each of these methods offers unique advantages and challenges, and their synergistic applications can significantly enhance the effectiveness and acceptance of RL systems in real-world scenarios. In real-world deployments, these approaches rarely function in isolation. Instead, effective human-AI interaction emerges from their integration. For example, systems may combine demonstrations for initialization, feedback for refinement, SA for safety, querying for uncertainty resolution, and explanations for transparency. Together, these methods form a holistic framework that enables RL systems to align with human needs, values, and trust requirements.

The case studies presented in this review illustrate the practical applications of HAII principles across diverse domains, from agriculture to robotics and human-AI trust dynamics. These examples highlight the importance of balancing technical performance with human-centric considerations, such as ethical constraints, trust calibration, and interpretability.

In the first case study, the trust-aware RL framework produced fertilization strategies that achieved comparable yields to trust-agnostic policies while significantly increasing farmer acceptance in user studies. This demonstrates that trust can be quantified and optimized jointly with productivity, highlighting the feasibility of incorporating human-centered objectives directly into the learning process. Beyond agricultural fertilization, the methodology exemplifies a generalizable approach: embedding human trust as an explicit optimization objective enables RL systems to move beyond narrow performance metrics and balance technical efficiency with social acceptability. Future work will extend validation to diverse crops, soil conditions, and regional contexts, ensuring robustness under varying ecological and socioeconomic conditions, while also exploring the transfer of this framework to other high-stakes domains such as healthcare decision support, energy management, and autonomous vehicle control. By adapting the trust-modeling component to domain-specific user populations, the approach holds promise as a blueprint for designing AI systems that achieve dual goals of technical performance and sustained human adoption.

In the second case study, the developed hierarchical POMDP framework enabled robots to flexibly adapt to patient scheduling and workflow needs while ensuring that hard safety and ethical constraints were never violated. Simulation results demonstrated that this approach improved safety compliance without sacrificing efficiency compared to baseline planners, showing that formal methods such as LTL can be effectively embedded within decision-making frameworks to guarantee ethical alignment. Beyond the clinical context, this methodology illustrates a generalizable strategy for integrating symbolic constraints with probabilistic planning, allowing autonomous agents to operate in dynamic environments without compromising critical human values. Future directions include pilot testing in real clinical settings, expanding and refining

the library of ethical rules to capture diverse stakeholder perspectives, and exploring scalability to other high-stakes domains such as eldercare, manufacturing, and public service robotics.

In the third case study, the bandit-based trust model successfully reproduced observed patterns of how users increase or decrease reliance on AI advice over time, providing a quantitative predictor of trust evolution. This marks a significant step beyond traditional survey-based approaches by offering a computational mechanism for tracking trust dynamically and continuously. The methodology further demonstrates generalizability: by framing trust adaptation as a sequential decision-making process, it can be applied across domains where human reliance on AI evolves through repeated interactions, such as medical diagnostics, financial decision support, and autonomous driving. Future research will focus on integrating this trust model into adaptive AI systems, enabling real-time trust calibration and proactive adjustment of autonomy levels to prevent both under-reliance and over-reliance.

6.2 Future Directions

As AI systems grow more capable and autonomous, the principles of HAII will be critical for bridging the gap between theoretical advancements and safe, effective real-world deployment. While recent progress in RL has shown great promise, substantial research gaps remain. Future work, as detailed below, must move beyond isolated technical solutions and instead develop integrated frameworks that explicitly address the core pillars of HAII (trust, usability, and understandability) through the unique capabilities of RL.

- **Advancing dynamic trust calibration.** A critical gap exists in the development of RL agents that can perceive, model, and actively calibrate user trust in real time. Current systems often assume trust as static, but in practice, trust fluctuates depending on context, prior interactions, and perceived system performance. Future research should integrate trust modeling into the RL framework itself. For instance, by representing the human's trust state as part of a POMDP belief state. This would allow agents to dynamically adjust their behavior, becoming more conservative or offering richer explanations when trust is low, while exercising greater autonomy when trust is high. Such dynamic trust calibration prevents both under-reliance and over-reliance, ensuring balanced human-AI partnerships.
- **Enhancing usability through adaptive collaboration.** Existing systems often rely on static interaction paradigms that do not flexibly adapt to human needs. RL is uniquely suited to support adaptive collaboration, where the level of autonomy, assistance, and communication is personalized based on user expertise, cognitive load, and contextual demands. Future RL algorithms should be designed to switch seamlessly between different HAII modes. For example, transitioning from SA to proactive querying or from high-level suggestions to direct control. This adaptability would maximize team performance while minimizing user burden, embodying the vision of AI as an intelligent, responsive teammate rather than a rigid tool.
- **Achieving causal understandability with XRL.** XRL has made progress in offering post-hoc justifications, but a significant gap remains in generating explanations that are not only interpretable but also causal and actionable. Future work should integrate causal inference with RL, enabling agents to answer "what-if" questions, explain the necessity of actions in the context of long-term goals, and provide counterfactual insights. Moreover, bidirectional explainability should be pursued: human feedback on explanations should directly inform and refine the agent's policy, creating a closed-loop system where both human and AI learn together.
- **Integrating emerging technologies for HAII.** Another frontier lies in the integration of RL with emerging technologies such as LLMs. LLMs can serve as explanation mediators, translating complex RL strategies into natural language narratives, or as feedback channels, capturing nuanced human input beyond numeric rewards. However, relying on LLM-generated synthetic feedback introduces ethical and

methodological concerns: it may embed preexisting model biases, misrepresent user intent, or obscure accountability if used as a substitute for authentic human oversight. Future research should therefore focus on unified RL–LLM architectures that not only enhance adaptivity and communication but also establish safeguards for transparency, bias mitigation, and human-in-the-loop validation, ensuring alignment with HAII principles.

By prioritizing these human-centered challenges and leveraging RL's capacity for learning and adaptation, we can unlock the full potential of this synergy. The next generation of intelligent systems will not merely optimize performance but will be trustworthy through continuous calibration, usable through adaptive collaboration, and understandable through causal transparency. In doing so, they can become not just powerful technologies but reliable and ethical partners that are deeply aligned with human values and genuinely beneficial to society.

In conclusion, the intersection of HAII and RL holds immense promise for advancing autonomous systems that collaborate seamlessly with humans, adapt to complex environments, and operate responsibly. This review serves as a roadmap for researchers and practitioners, offering insights and frameworks to guide the development of next-generation AI systems that are both intelligent and human-aligned.

Acknowledgement: Shaoping Xiao, Zhaoan Wang and Jun Wang acknowledge the support from the University of Iowa OVPR Interdisciplinary Scholars Program for this study. Shaoping Xiao, Caden Noeller and Jiefeng Jiang acknowledge the support from the University of Iowa OUR research for undergraduate program. Shaoping Xiao acknowledges the support from the University of Iowa OVPR professional development award. We also acknowledge the project support from the private-public-partnership (P3) program at the University of Iowa.

Funding Statement: This research was funded by the U.S. Department of Education under Grant Number ED#P116S210005 and the National Science Foundation under Grant Numbers 2226936 and 2420405. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Department of Education and the National Science Foundation.

Author Contributions: Conceptualization, Shaoping Xiao, Zhaoan Wang and Jiefeng Jiang; methodology, Zhaoan Wang, Junchao Li and Caden Noeller; software, Zhaoan Wang, Junchao Li and Caden Noeller; validation, Zhaoan Wang and Junchao Li; formal analysis, Shaoping Xiao, Zhaoan Wang, Junchao Li and Caden Noeller; investigation, Zhaoan Wang, Junchao Li and Caden Noeller; resources, Zhaoan Wang, Junchao Li, Caden Noeller, Jiefeng Jiang and Jun Wang; data curation, Zhaoan Wang, Junchao Li and Caden Noeller; writing—original draft preparation, Shaoping Xiao, Zhaoan Wang, Junchao Li and Caden Noeller; writing—review and editing, Jiefeng Jiang and Jun Wang; visualization, Shaoping Xiao, Junchao Li and Caden Noeller; supervision, Shaoping Xiao, Jiefeng Jiang and Jun Wang; project administration, Shaoping Xiao and Jun Wang; funding acquisition, Shaoping Xiao, Jiefeng Jiang and Jun Wang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data supporting the findings of Section 5 are available from the corresponding author, Shaoping Xiao, upon reasonable request.

Ethics Approval: The study reported in Section 5.3 was approved by the University of Iowa Institutional Review Board (Approval No.: 202001345) and informed consent was obtained from all participants.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Radanliev P, De Roure D, Van Kleek M, Santos O, Ani U. Artificial intelligence in cyber physical systems. AI Soc. 2021;36(3):783–96. doi:10.1007/s00146-020-01049-0.

2. Chen Y, Deierling P, Xiao SP. Exploring active learning strategies for predictive models in mechanics of materials. Appl Phys A. 2024;130(8):588. doi:10.1007/s00339-024-07728-9.
3. Xiao SP, Li J, Wang Z, Chen Y, Tofighi S. Advancing additive manufacturing through machine learning techniques: a state-of-the-art review. Future Int. 2024;16(11):419. doi:10.3390/fi16110419.
4. Wang Z, Jha K, Xiao SP. Continual reinforcement learning for intelligent agricultural management under climate changes. Comput Mater Contin. 2024;81(1):1319–36. doi:10.32604/cmc.2024.055809.
5. Gurbuz F, Mudireddy A, Mantilla R, Xiao SP. Using a physics-based hydrological model and storm transposition to investigate machine-learning algorithms for streamflow prediction. J Hydrol. 2024;628:130504. doi:10.1016/j.jhydrol.2023.130504.
6. Andreoni M, Lunardi WT, Lawton G, Thakkar S. Enhancing autonomous system security and resilience with generative AI: a comprehensive survey. IEEE Access. 2024;12(1):109470–93. doi:10.1109/ACCESS.2024.3439363.
7. Sutton RS, Barto AG. Reinforcement learning: an introduction. London, UK: The MIT Press; 2018.
8. Li J, Cai M, Kan Z, Xiao SP. Model-free reinforcement learning for motion planning of autonomous agents with complex tasks in partially observable environments. Auton Agents Multi-Agent Syst. 2024;38(1):14. doi:10.1007/s10458-024-09641-0.
9. Aradi S. Survey of deep reinforcement learning for motion planning of autonomous vehicles. IEEE Trans Intell Transp Syst. 2020;23(2):740–59. doi:10.1109/TITS.2020.3024655.
10. Cai M, Kan Z, Li Z, Xiao SP. Optimal probabilistic motion planning with potential infeasible LTL constraints. IEEE Trans Autom Control. 2023;68(1):301–16. doi:10.1109/TAC.2021.3138704.
11. Chowdhury S, Budhwar P, Wood G. Generative artificial intelligence in business: towards a strategic human resource management framework. Br J Manag. 2024;35(4):1680–91. doi:10.1111/1467-8551.12824.
12. Heyder T, Passlack N, Posegga O. Ethical management of human-AI interaction: theory development review. J Strateg Inf Syst. 2023;35(3):101772. doi:10.1016/j.jsis.2023.101772.
13. Boada JP, Maestre BR, Genís CT. The ethical issues of social assistive robotics: a critical literature review. Technol Soc. 2021;67:101726. doi:10.1016/j.techsoc.2021.101726.
14. Giermindl LM, Strich F, Christ O, Leicht-Deobald U, Redzepi A. The dark sides of people analytics: reviewing the perils for organisations and employees. Eur J Inf Syst. 2022;31(3):410–35. doi:10.1080/0960085X.2021.1927213.
15. Naveed S, Stevens G, Robin-Kern D. An overview of the empirical evaluation of explainable AI (XAI): a comprehensive guideline for user-centered evaluation in XAI. Appl Sci. 2024;14(23):11288. doi:10.3390/app142311288.
16. Afroogh S, Akbari A, Malone E, Kargar M, Alambeigi H. Trust in AI: progress, challenges, and future directions. Humanit Soc Sci Commun. 2024;11(1):1568. doi:10.1057/s41599-024-04044-8.
17. Fragiadakis G, Diou C, Kousiouris G, Nikolaidou M. Evaluating human-AI collaboration: a review and methodological framework. arXiv:2407.19098. 2025. doi:10.48550/arXiv.2407.19098.
18. Zhao M, Simmons R, Admoni H. The role of adaptation in collective human-AI teaming. Top Cogn Sci. 2025;17(2):291–323. doi:10.1111/tops.12633.
19. Du T, Li X, Jiang N, Xu Y, Zhou Y. Adaptive AI as collaborator: examining the impact of an AI's adaptability and social role on individual professional efficacy and credit attribution in human-AI collaboration. Int J Hum-Comput Interact. 2025;41(19):12422–33. doi:10.1080/10447318.2025.2462120.
20. Xu F, Uszkoreit H, Du Y, Fan W, Zhao D, Zhu J. Explainable AI: a brief survey on history, research areas, approaches and challenges. In: Proceedings of Natural Language Processing and Chinese Computing; 2019 Oct 9–14; Dunhuang, China. Cham, Switzerland: Springer; 2019. p. 563–74.
21. Yang W, Wei Y, Wei H, Chen Y, Huang G, Li X, et al. Survey on explainable AI: from approaches, limitations and applications aspects. Hum-Cent Intell Syst. 2023;3(3):161–88. doi:10.1007/s44230-023-00038-y.
22. Li Y, Wu B, Huang Y, Luan S. Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust. Front Psychol. 2024;15:1382693. doi:10.3389/fpsyg.2024.1382693.
23. Mehrotra S, Degachi C, Vereschak O, Jonker CM, Tielman ML. A systematic review on fostering appropriate trust in human-AI interaction: trends, opportunities and challenges. ACM J Responsib Comput. 2024;1(4):1–45. doi:10.1145/3696449.

24. Klingbeil A, Grützner C, Schreck P. Trust and reliance on AI—an experimental study on the extent and costs of overreliance on AI. Comput Hum Behav. 2024;160(1):108352. doi:10.1016/j.chb.2024.108352.
25. Okamura K, Yamada S. Adaptive trust calibration for human-AI collaboration. PLoS One. 2020;15(2):e0229132. doi:10.1371/journal.pone.0229132.
26. Senoner J, Schallmoser S, Kratzwald B, Feuerriegel S, Netland T. Explainable AI improves task performance in human-AI collaboration. Sci Rep. 2024;14(1):31150. doi:10.1038/s41598-024-82501-9.
27. Halkiopoulos C, Gkintoni E. Leveraging AI in e-learning: personalized learning and adaptive assessment through cognitive neuropsychology—a systematic analysis. Electronics. 2024;13(18):3762. doi:10.3390/electronics13183762.
28. Braunschweig B, Ghallab M. Reflections on artificial intelligence for humanity. New York, NY, USA: Springer; 2021.
29. Dubber MD, Pasquale F, Das S. The oxford handbook of ethics of AI. New York, NY, USA: Oxford University Press; 2020.
30. Capel T, Brereton M. What is human-centered about human-centered AI? A map of the research landscape. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2023. doi:10.1145/3544548.3580959.
31. Bingley WJ, Curtis C, Lockey S, Bialkowski A, Gillespie N, Haslam SA, et al. Where is the human in human-centered AI? Insights from developer priorities and user experiences. Comput Hum Behav. 2023;141(1):107617. doi:10.1016/j.chb.2022.107617.
32. Saranya A, Subhashini R. A systematic review of explainable artificial intelligence models and applications: recent developments and future trends. Decis Anal J. 2023;7(5):100230. doi:10.1016/j.dajour.2023.100230.
33. Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Inf Fusion. 2023;99(3):101805. doi:10.1016/j.inffus.2023.101805.
34. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting black-box models: a review on explainable artificial intelligence. Gogn Comput. 2024;16(1):45–74. doi:10.1007/s12559-023-10179-8.
35. Kostopoulos G, Davrazos G, Kotsiantis S. Explainable artificial intelligence-based decision support systems: a recent review. Electronics. 2024;13(14):2842. doi:10.3390/electronics13142842.
36. Blockeel H, Devos L, Frénay B, Nanfack G, Nijssen S. Decision trees: from efficient prediction to responsible AI. Front Artif Intell. 2023;6:1124553. doi:10.3389/frai.2023.1124553.
37. Salih AM, Wang Y. Are linear regression models white box and interpretable? Authorea. 2024. doi:10.22541/au.172122905.57636646/v1.
38. Mendel JM, Bonissone PP. Critical thinking about explainable AI (XAI) for rule-based fuzzy systems. IEEE Trans Fuzzy Syst. 2021;29(12):3579–93. doi:10.1109/tfuzz.2021.3079503.
39. Mekonnen ET, Longo L, Dondio P. A global model-agnostic rule-based XAI method based on parameterized event primitives for time series classifiers. Front Artif Intell. 2024;7:1381921. doi:10.3389/frai.2024.1381921.
40. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. doi:10.1023/A:1010933404324.
41. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2016. p. 1135–44. doi:10.1145/2939672.2939778.
42. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, NY, USA: Association for Computing Machinery; 2017. p. 4768–77.
43. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision. Piscataway, NJ, USA: IEEE; 2017. p. 618–26. doi:10.1109/ICCV.2017.74.
44. Hendricks LA, Akata Z, Rohrbach M, Donahue J, Schiele B, Darrell T. Generating visual explanations. In: Leibe B, Matas J, Sebe N, Welling M, editors. Lecture notes in computer science. Computer vision–ECCV 2016. Cham, Switzerland: Springer; 2016. p. 3–19. doi:10.1007/978-3-319-46493-0_1.
45. Cambria E, Malandri L, Mercorio F, Mezzanzanica M, Nobani N. A survey on XAI and natural language explanations. Inf Process Manag. 2001;60(1):10311. doi:10.1016/j.ipm.2022.103111.

46. Park DH, Hendricks LA, Akata Z, Rohrbach A, Schiele B, Darrell T, et al. Multimodal explanations: justifying decisions and pointing to the evidence. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2018. p. 8779–88. doi:10.1109/CVPR.2018.00915.
47. Karimi AH, Barthe G, Balle B, Valera I. Model-agnostic counterfactual explanations for consequential decisions. In: Proceedings of the Twenty-Third International Conference on Artificial Intelligence and Statistics, AISTATS 2020; 2020 Aug 26–28; Online. PMLR; 2020. p. 108:895–905.
48. Liu J, Wu X, Liu S, Gong S. Model-agnostic counterfactual explanation: a feature weights-based comprehensive causal multi-objective counterfactual framework. Expert Syst Appl. 2025;266(5):126063. doi:10.1016/j.eswa.2024.126063.
49. European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). O J Eur Union. 2016;L119:1–88.
50. High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI. Brussels, Belgium: European Commission; 2019 [cited 2025 Oct 24]. Available from: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
51. Howard RA. Dynamic programming and Markov processes. Cambridge, MA, USA: MIT Press; 1960.
52. Watkins CJC, Dayan P. Q-learning. Mach Learn. 1992;8(3):279–92. doi:10.1007/BF00992698.
53. Andrew AM. Reinforcement learning: an introduction. Kybernetes. 1998;27(6):1093–6. doi:10.1108/k.1998.27.9.1093.3.
54. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature. 2015;518(7540):529–33. doi:10.1038/nature14236.
55. Lin LJ. Self-improving reactive agents based on reinforcement learning, planning and teaching. Mach Learn. 1992;8(3):293–321. doi:10.1007/BF00992699.
56. Kakade S, Langford J. Approximately optimal approximate reinforcement learning. In: Proceedings of the Nineteenth International Conference on Machine Learning. New York, NY, USA: Association for Computing Machinery; 2002. p. 267–74.
57. Shakya AK, Pillai G, Chakrabarty S. Reinforcement learning algorithms: a brief survey. Expert Syst Appl. 2023;231(7):120495. doi:10.1016/j.eswa.2023.120495.
58. Kurniawati H. Partially observable markov decision processes and robotics. Annu Rev Control Robot Auton Syst. 2022;5(5):253–77. doi:10.1146/annurev-control-042920-092451.
59. Shani G, Pineau J, Kaplow R. A survey of point-based POMDP solvers. Auton Agents Multi-Agent Syst. 2013;27(1):1–51. doi:10.1007/S10458-012-9200-2.
60. Pineau J, Gordon G, Thrun S. Point-based value iteration: an anytime algorithm for POMDPs. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence. New York, NY, USA: Association for Computing Machinery; 2003. p. 1025–30.
61. Kurniawati H, Hsu D, Lee WS. SARSOP: efficient point-based POMDP planning by approximating optimally reachable belief spaces. In: Brock O, Trinkle J, Ramos F, editors. Robotics: science and systems IV. Cambridge, MA, USA: MIT Press; 2009. p. 65–72. doi:10.7551/mitpress/8344.003.0013.
62. Baier C, Katoen JP. Principles of model checking. Cambridge, MA, USA: MIT Press; 2008.
63. Bozkurt AK, Wang Y, Zavlanos MM, Pajic M. Control synthesis from linear temporal logic specifications using model-free reinforcement learning. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). Piscataway, NJ, USA: IEEE; 2020. p. 10349–55. doi:10.1109/ICRA40945.2020.9196796.
64. Sickert S, Esparza J, Jaax S, Křetínský J. Limit-deterministic Büchi automata for linear temporal logic. In: Chaudhuri S, Farzan A, editors. Computer aided verification. CAV 2016. Cham, Switzerland: Springer; 2016. p. 312–32. doi:10.1007/978-3-319-41540-6_17.
65. Křetínský J, Meggendorfer T, Sickert S. Owl: a library for ω-words, automata, and LTL. Lect Notes Comput Sci. 2018;11138:543–50. doi:10.1007/978-3-030-01090-4_34.
66. Cai M, Xiao SP, Li J, Kan Z. Safe reinforcement learning under temporal logic with reward design and quantum action selection. Sci Rep. 2023;13(1):1–20. doi:10.1038/s41598-023-28582-4.

67. Cai M, Hasanbeig M, Xiao S, Abate A, Kan Z. Modular deep reinforcement learning for continuous motion planning with temporal logic. IEEE Robot Autom Lett. 2021;6(4):7973–80. doi:10.1109/LRA.2021.3101544.
68. Cai M, Zhou Z, Li L, Xiao SP, Kan Z. Reinforcement learning with soft temporal logic constraints using limit-deterministic generalized Büchi automaton. J Autom Intell. 2025;4(1):39–51. doi:10.1016/j.jai.2024.12.005.
69. Han D, Mulyana B, Stankovic V, Cheng S. A survey on deep reinforcement learning algorithms for robotic manipulation. Sensors. 2023;23(7):3762. doi:10.3390/s23073762.
70. Elguea-Aguinaco I, Serrano-Muñoz A, Chrysostomou D, Inziarte-Hidalgo I, Bøgh S, Arana-Arexolaleiba N. A review on reinforcement learning for contact-rich robotic manipulation tasks. Robot Comput-Integr Manuf. 2023;81(4):102517. doi:10.1016/j.rcim.2022.102517.
71. Wang Y, Sagawa R, Yoshiyasu Y. Learning advanced locomotion for quadrupedal robots: a distributed multi-agent reinforcement learning framework with riemannian motion policies. Robotics. 2024;13(6):86. doi:10.3390/robotics13060086.
72. Li J, Cai M, Wang Z, Xiao S. Model-based motion planning in POMDPs with temporal logic specifications. Adv Robot. 2023;37(14):871–86. doi:10.1080/01691864.2023.2226191.
73. Li J, Cai M, Xiao S, Kan Z. Online motion planning with soft metric interval temporal logic in unknown dynamic environment. IEEE Control Syst Lett. 2022;6:2293–8. doi:10.1109/LCSYS.2022.3145058.
74. Cai M, Xiao S, Li B, Li Z, Kan Z. Reinforcement learning based temporal logic control with maximum probabilistic satisfaction. In: Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA). Piscataway, NJ, USA: IEEE; 2021. p. 806–12. doi:10.1109/ICRA48506.2021.9561903.
75. Bui VH, Mohammadi S, Das S, Hussain A, Hollweg GV, Su W. A critical review of safe reinforcement learning strategies in power and energy systems. Eng Appl Artif Intell. 2025;143(1):110091. doi:10.1016/j.engappai.2025.110091.
76. Abdul Shakir W. Reinforcement learning challenges in power grid management: a case study with city learn simulator. In: Service-Oriented Computing – ICSOC 2024 Workshops. ICSOC 2024. Singapore: Springer; 2024. p. 301–13. doi:10.1007/978-981-96-7238-7_24.
77. Erick AO, Folly KA. Reinforcement learning approaches to power management in grid-tied microgrids: a review. In: Proceedings of 2020 Clemson University Power Systems Conference (PSC); 2020 Mar 10–13; Clemson, SC, USA. p. 1–6. doi:10.1109/PSC50246.2020.9131138.
78. Michailidis P, Michailidis I, Kosmatopoulos E. Reinforcement learning for optimizing renewable energy utilization in buildings. A Rev Appl Innov Energies. 2025;18(7):1724. doi:10.3390/en18071724.
79. Ding X, Cerpa A, Du W. Exploring deep reinforcement learning for holistic smart building control. ACM Trans Sen Netw. 2024;20(3):70. doi:10.1145/3656043.
80. Soler D, Marino O, Huergo D, Frutos M, Ferrer E. Reinforcement learning to maximize wind turbine energy generation. Expert Syst Appl. 2024;249(PartA):123502. doi:10.1016/j.eswa.2024.123502.
81. Stavrev S, Ginchev D. Reinforcement learning techniques in optimizing energy systems. Electronics. 2024;13(8):1459. doi:10.3390/electronics13081459.
82. Agyeman BT, Decardi-Nelson B, Liu J, Shah SL. Semi-centralized multi-agent RL framework for efficient irrigation scheduling. Control Eng Pract. 2025;155(6):106183. doi:10.1016/j.conengprac.2024.106183.
83. Ding X, Du W. Optimizing irrigation efficiency using deep reinforcement learning in the field. ACM Trans Sen Netw. 2024;20(4):99. doi:10.1145/3662182.
84. Alkaff M, Basuhail A, Sari Y. Optimizing water use in maize irrigation with reinforcement learning. Mathematics. 2025;13(4):595. doi:10.3390/math13040595.
85. Banumathi K, Venkatesan L, Benjamin LS, Vijayalakshmi K, Satchi NS. Reinforcement learning in personalized medicine: a comprehensive review of treatment optimization strategies. Cureus. 2025;17(4):e82756. doi:10.7759/cureus.82756.
86. Liu M, Shen X, Pan W. Deep reinforcement learning for personalized treatment recommendation. Stat Med. 2022;41(20):4034–56. doi:10.1002/sim.9491.

87. Frommeyer TC, Gilbert MM, Fursmidt RM, Park Y, Khouzam JP, Brittain GV, et al. Reinforcement learning and its clinical applications within healthcare: a systematic review of precision medicine and dynamic treatment regimes. Healthcare. 2025;13(14):1752. doi:10.3390/healthcare13141752.
88. Elmekki H, Islam S, Alagha A, Sami H, Spilkin A, Zakeri E, et al. Comprehensive review of reinforcement learning for medical ultrasound imaging. Artif Intell Rev. 2025;58(9):284. doi:10.1007/s10462-025-11268-w.
89. Hu M, Zhang J, Matkovic L, Liu T, Yang X. Reinforcement learning in medical image analysis: concepts, applications, challenges, and future directions. J Appl Clin Med Phys. 2023;24(2):e13898. doi:10.1002/acm2.13898.
90. Kumar A. Reinforcement learning for robotic-assisted surgeries: optimizing procedural outcomes and minimizing post-operative complications. Int J Res Publ Rev. 2025;6(31):5669–84. doi:10.55248/gengpi.6.0325.11166.
91. Martínez-Pascual D, Catalán JM, Lledó LD, Blanco-Ivorra A, Nicolás García-Aracil N. A deep learning model for assistive decision-making during robot-aided rehabilitation therapies based on therapists' demonstrations. J NeuroEng Rehabil. 2025;22(1):18. doi:10.1186/s12984-024-01517-4.
92. Wu Q, Han J, Yan Y, Kuo Y, Shen Z. Reinforcement learning for healthcare operations management: methodological framework, recent developments, and future research directions. Health Care Manag Sci. 2025;28(2):298–333. doi:10.1007/s10729-025-09699-6.
93. Lazebnik T. Data-driven hospitals staff and resources allocation using agent-based simulation and deep reinforcement learning. Eng Appl Artif Intell. 2023;126(Part A):106783. doi:10.1016/j.engappai.2023.106783.
94. Mienye E, Jere N, Obaido G, Mienye ID, Aruleba K. Deep learning in finance: a survey of applications and techniques. AI. 2024;5(4):2066–91. doi:10.3390/ai5040101.
95. Cao G, Zhang Y, Lou Q, Wang G. Optimization of high-frequency trading strategies using deep reinforcement learning. J Artif Intell Gen Sci. 2024;6(1):230–57. doi:10.60087/jaigs.v6i1.247.
96. Baffo I, Leonardi M, D'Alberti V, Petrillo A. Optimizing public investments: a sustainable economic, environmental, and social investment multi-criteria decision model (SEESIM). Reg Sci Policy Pract. 2024;16(11):100140. doi:10.1016/j.rspp.2024.100140.
97. Uddin MA, Chang BH, Aldawsari SH, Li R. The interplay between green finance, policy uncertainty and carbon market volatility: a time frequency approach. Sustainability. 2025;17(3):1198. doi:10.3390/su17031198.
98. Alfonso-Sánchez S, Solano J, Correa-Bahnsen A, Sendova KP, Bravo C. Optimizing credit limit adjustments under adversarial goals using reinforcement learning. Eur J Oper Res. 2024;315(2):802–17. doi:10.1016/j.ejor.2023.12.025.
99. Cui Y, Han X, Chen J, Zhang X, Yang J, Zhang X. FraudGNN-RL: a graph neural network with reinforcement learning for adaptive financial fraud detection. IEEE Open J Comput Soc. 2025;6:426–37. doi:10.1109/OJCS.2025.3543450.
100. Wang L, Ren W. Drought in agriculture and climate-smart mitigation strategies. Cell Rep Sustain. 2025;2(6):100386. doi:10.1016/j.crsus.2025.100386.
101. Feng K, Lin N, Kopp NE, Xian S, Oppenheimer M. Reinforcement learning-based adaptive strategies for climate change adaptation: an application for coastal flood risk management. Proc Natl Acad Sci U S A. 2025;122(12):e2402826122. doi:10.1073/pnas.2402826122.
102. Ahmed N, Farooq MU, Chen F. Materials discovery through reinforcement learning: a comprehensive review. AI Mater. 2025;1(2):0010. doi:10.55092/aimat20250010.
103. Kim H, Choi H, Kang D, Lee WB, Na J. Materials discovery with extreme properties via reinforcement learning-guided combinatorial chemistry. Chem Sci. 2024;15:7908–25. doi:10.1039/D3SC05281H.
104. Zhu Y, Cai M, Schwarz C, Li J, Xiao S. Intelligent traffic light via policy-based deep reinforcement learning. Int J Intell Transp Syst Res. 2022;20(3):734–44. doi:10.1007/s13177-022-00321-5.
105. Najar A, Chetouan M. Reinforcement learning with human advice: a survey. Front Robot AI. 2021;8:584075. doi:10.3389/frobt.2021.584075.
106. González Barman K, Lohse S, de Regt HW. Reinforcement learning from human feedback in LLMs: whose culture, whose values, whose perspectives? Philos Technol. 2025;38(2):35. doi:10.1007/s13347-025-00861-0.
107. Gullapalli V, Barto AG. Shaping as a method for accelerating reinforcement learning. In: Proceedings of the 1992 IEEE International Symposium on Intelligent Control. Piscataway, NJ, USA: IEEE; 1992. p. 554–9. doi:10.1109/ISIC.1992.225046.

108. Ng AY, Harada D, Russell SJ. Policy invariance under reward transformations: theory and application to reward shaping. In: Proceedings of the Sixteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1999. p. 278–87.
109. Tenorio-Gonzalez AC, Morales EF, Villaseñor-Pineda L. Dynamic reward shaping: training a robot by voice. In: Advances in artificial intelligence–IBERAMIA 2010. Berlin/Heidelberg, Germany: Springer; 2010. p. 483–92. doi:10.1007/978-3-642-16952-6_49.
110. Knox WB, Stone P. Combining manual feedback with subsequent MDP reward signals for reinforcement learning. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems; 2010 May 10–14; Toronto, ON, Canada. p. 5–12.
111. Knox WB, Stone P. Augmenting reinforcement learning with human feedback. In: Proceedings of the ICML Workshop on New Developments in Imitation Learning; 2011 Jun 28–Jul 2; Bellevue, WA, USA.
112. Knox WB, Stone P. Understanding human teaching modalities in reinforcement learning environments: a preliminary report. In: IJCAI 2011 Workshop on Agents Learning Interactively from Human Teachers (ALIHT); 2011 Jul 16–17; Barcelona, Spain.
113. Ibrahim S, Mostafa M, Jnadi A, Salloum H, Osinenko P. Comprehensive overview of reward engineering and shaping in advancing reinforcement learning applications. IEEE Access. 2024;12:175473–500. doi:10.1109/ACCESS.2024.3504735.
114. Badnava B, Esmaeili M, Mozayani N, Zarkesh-Ha P. A new potential-based reward shaping for reinforcement learning agent. In: Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC). Piscataway, NJ, USA: IEEE; 2023. p. 1–6. doi:10.1109/CCWC57344.2023.10099211.
115. Lee K, Smith LM, Abbeel P. PEBBLE: feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. arXiv.2106.05091. 2021. doi:10.48550/arXiv.2106.05091.
116. Park J, Seo Y, Shin J, Lee H, Abbeel P, Lee K. SURF: semisupervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning. arXiv:2203.10050. 2022.
117. Ibarz B, Leike J, Pohlen T, Irving G, Legg S, Amodei D. Reward learning from human preferences and demonstrations in Atari. arXiv:1811.06521. 2018.
118. Golpayegani F, Ghanadbashi S, Zarchini A. Advancing sustainable manufacturing: reinforcement learning with adaptive reward machine using an ontology-based approach. Sustainability. 2024;16(14):5873. doi:10.3390/su16145873.
119. Zouaq A, Nkambou R. A survey of domain ontology engineering: methods and tools. In: Nkambou R, Bourdeau J, Mizoguchi R, editors. Advances in intelligent tutoring systems. Studies in computational intelligence. Berlin/Heidelberg, Germany: Springer; 2010. doi:10.1007/978-3-642-14363-2_6.
120. Jeon H, Kim DW, Kang BY. Deep reinforcement learning for cooperative robots based on adaptive sentiment feedback. Expert Syst Appl. 2023;243(1):121198. doi:10.1016/j.eswa.2023.121198.
121. Hattie J, Tipeley H. The power of feedback. Rev Educ Res. 2007;77(1):81–112. doi:10.3102/003465430298487.
122. Hu M, Liu B. Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2004 Aug 22–25; Seattle, WA, USA. p. 168–77.
123. Cao M, Shu L, Yu L, Zhu Y, Wichers N, Liu Y, et al. Enhancing reinforcement learning with dense rewards from language model critic. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language. Stroudsburg, PA, USA: Association for Computational Linguistics; 2024. p. 9119–38. doi:10.18653/v1/2024.emnlp-main.515.
124. Masadome S, Harada T. Reward design using large language models for natural language explanation of reinforcement learning agent actions. IEEJ Trans Electr Electron Eng. 2025;20(8):1203–11. doi:10.1002/tee.70005.
125. Wang K, Zhao Y, He Y, Dai S, Zhang N, Yang M. Guiding reinforcement learning with shaping rewards provided by the vision-language model. Eng Appl Artif Intell. 2025;155(3–4):111004. doi:10.1016/j.engappai.2025.111004.
126. Haider Z, Rahman MH, Devabhaktuni V, Moeykens S, Chakraborty P. A framework for mitigating malicious RLHF feedback in LLM training using consensus based reward. Sci Rep. 2024;15(1):9177. doi:10.1038/s41598-025-92889-7.

127. Fu J, Zhao X, Yao C, Wang H, Han Q, Xiao Y. Reward shaping to mitigate reward hacking in RLHF. arXiv.2502.18770. 2025. doi:10.48550/arXiv.2502.18770.
128. Chaudhari S, Aggarwal P, Murahari V, Rajpurohit T, Kalyan A, Narasimhan K, et al. RLHF deciphered: a critical analysis of reinforcement learning from human feedback for LLMs. ACM Comput. 2025;58(2):53. doi:10.1145/3743127.
129. Xie T, Zhao S, Wu CH, Liu Y, Luo Q, Zhong V, et al. Text2reward: reward shaping with language models for reinforcement learning. arXiv:2309.11489. 2024.
130. Sun S, Liu R, Lyu J, Yang JW, Zhang L, Li X. A large language model-driven reward design framework via dynamic feedback for reinforcement learning. arXiv.2410.14660. 2024. doi:10.48550/arXiv.2410.14660.
131. Qu Y, Jiang Y, Wang B, Mao Y, Wang C, Liu C, et al. Latent reward: LLM-empowered credit assignment in episodic reinforcement learning. Proc AAAI Conf Arti Intell. 2025;39(19):20095–103. doi:10.1609/aaai.v39i19.34213.
132. Du Y, Watkins O, Wang Z, Colas C, Darrell T, Abbeel P, et al. Guiding pretraining in reinforcement learning with large language models. arXiv.2302.06692. 2023. doi:10.48550/arXiv.2302.06692.
133. Guo Q, Liu X, Hui J, Liu Z, Huang P. Utilizing large language models for robot skill reward shaping in reinforcement learning. In: Lan X, Mei X, Jiang C, Zhao F, Tian Z, editors. Intelligent robotics and applications. Singapore: Springer; 2025. doi:10.1007/978-981-96-0783-9_1.
134. Dorigo M, Colombetti M. Robot shaping: developing autonomous agents through learning. Artif Intell. 1994;71(2):321–70. doi:10.1016/0004-3702(94)90047-7.
135. Colombetti M, Dorigo M, Borghi G. Behavior analysis and training-a methodology for behavior engineering. IEEE Trans Syst Man Cybernet B. 1996;26(3):365–80. doi:10.1109/3477.499789.
136. Ho MK, MacGlashan J, Littman ML, Cushman F. Social is special: a normative framework for teaching with and learning from evaluative feedback. Cognition. 2017;167(3):91–106. doi:10.1016/j.cognition.2017.03.006.
137. Najar A, Sigaud O, Chetouani M. Interactively shaping robot behaviour with unlabeled human instructions. Auton Agents Multiagent Syst. 2020;34(2):35. doi:10.1007/s10458-020-09459-6.
138. Rosenstein MT, Barto AG, Si J, Barto A, Powell W, Wunsch D. Supervised actor-critic reinforcement learning. In: Si J, Barto A, Powell W, Wunsch D, editors. Handbook of learning and approximate dynamic programming. Hoboken, NJ, USA: John Wiley & Sons; 2004. p. 359–80.
139. Lopes M, Cederbourg T, Oudeyer PY. Simultaneous acquisition of task and feedback models. In: Proceedings of the 2011 IEEE International Conference on Development and Learning; 2011 Aug 24–27; Frankfurt, Germany. p. 1–7. doi:10.1109/DEVLRN.2011.6037359.
140. Griffith S, Subramanian K, Scholz J, Isbell CL, Thomaz A. Policy shaping: integrating human feedback with reinforcement learning. In: Proceedings of the 26th International Conference on Neural Information Processing Systems; 2013 Dec 5–8; Lake Tahoe, NV, USA. p. 2625–33.
141. Loftin R, Peng B, Macglashan J, Littman ML, Taylor ME, Huang J, et al. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. Auton Agents Multiagent Syst. 2016;30(1):30–59. doi:10.1007/s10458-015-9283-7.
142. Pradyot KVN, Manimaran SS, Ravindran B, Natarajan S. Integrating human instructions and reinforcement learners: an SRL approach. In: Proceedings of the 2012 UAI workshop on Statistical Relational AI; 2012 Aug 18; Catalina Island, CA, USA.
143. Knox WB, Stone P. Interactively shaping agents via human reinforcement: the TAMER framework. In: Proceedings of the Fifth International Conference on Knowledge Capture; 2009 Sep 1–4; Redondo Beach, CA, USA. p. 9–16. doi:10.1145/1597735.1597738.
144. Wirth C, Akrour R, Neumann G, Fürnkranz J. A survey of preference-based reinforcement learning methods. J Mach Learn Res. 2017;18(136):1–46.
145. Tarakli I, Vinanzi S, Nuovo AD. Interactive reinforcement learning from natural language feedback. In: Proceedings of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2024 Oct 14–18; Abu Dhabi, United Arab Emirates. p. 11478–84. doi:10.1109/IROS58592.2024.10803055.

146. Faulkner TK, Short ES, Thomaz AL. Policy shaping with supervisory attention driven exploration. In: Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2018 Oct 1–5; Madrid, Spain. p. 842–7. doi:10.1109/IROS.2018.8594312.
147. Juan R, Huang J, Gomez R, Nakamura K, Sha Q, He B, et al. Shaping progressive net of reinforcement learning for policy transfer with human evaluative feedback. In: Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2021 Sep 27–Oct 1; Prague, Czech Republic. p. 1281–8. doi:10.1109/IROS51168.2021.9636061.
148. Lee K, Smith L, Dragan A, Abbeel P. B-Pref: benchmarking preference-based reinforcement learning. arXiv:2111.03026. 2021.
149. Knox WB, Hatgis-Kessell S, Adalgeirsson SO, Booth S, Dragan A, Stone P, et al. Learning optimal advantage from preferences and mistaking it for reward. Proc AAAI Conf Artif Intell. 2024;38(9):10066–73. doi:10.1609/aaai.v38i9.28870.
150. Knox WB, Stone P. Reinforcement learning from simultaneous human and MDP reward. In: Proceedings of AAMAS '12: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems; 2012 Jun 4–8; Valencia, Spain. Vol. 1, p. 475–82.
151. Fang B, Jia S, Guo D, Xu M, Wen S, Sun F. Survey of imitation learning for robotic manipulation. Int J Intell Robot Appl. 2019;3(4):362–9. doi:10.1007/s41315-019-00103-5.
152. Ozalp R, Ucar A, Guzelis C. Advancements in deep reinforcement learning and inverse reinforcement learning for robotic manipulation: toward trustworthy, interpretable, and explainable artificial intelligence. IEEE Access. 2024;12:51840–58. doi:10.1109/ACCESS.2024.3385426.
153. Hester T, Vecerik M, Pietquin O, Lanctot M, Schaul T, Piot B, et al. Deep Q-learning from demonstrations. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. Palo Alto, CA, USA: AAAI Press; 2018. Vol. 32, p. 3223–30. doi:10.1609/aaai.v32i1.11757.
154. Abbeel P, Ng AY. Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the Twenty-First International Conference on Machine Learning; 2004 Jul 4–8; Banff, AB, Canada. doi:10.1145/1015330.1015430.
155. Zhang Z, Hong J, Enayati AMS, Najjaran H. Using implicit behavior cloning and dynamic movement primitive to facilitate reinforcement learning for robot motion planning. IEEE Trans Robot. 2024;40(2):4733–49. doi:10.1109/TRO.2024.3468770.
156. Xu H, Zhao D. Forward-prediction-based active exploration behavior cloning. In: Proceedings of the 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT); 2024 Mar 22–24; Nanjing, China. p. 2112–7. doi:10.1109/AINIT61980.2024.10581680.
157. Sagheb S, Losey DP. Counterfactual behavior cloning: offline imitation learning from imperfect human demonstrations. arXiv.2505.10760. 2025. doi:10.48550/arXiv.2505.10760.
158. Lin Y, Ni Z, Zhong X. Learning from demonstrations: a computationally efficient inverse reinforcement learning approach with simplified implementation. IEEE Trans Emerg Topics Comput Intell. 2025;9(5):3413–25. doi:10.1109/TETCI.2025.3526502.
159. Melo FS, Lopes M. Teaching multiple inverse reinforcement learners. Front Artif Intell. 2021;4:625183. doi:10.3389/frai.2021.625183.
160. Wulfmeier M, Bloesch M, Vieillard N, Ahuja A, Bornschein J, Huang S, et al. Imitating language via scalable inverse reinforcement learning. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2025.
161. Cai Z, Liu Y, Liang Y. Accelerating human motion imitation with interactive reinforcement learning. In: Proceedings of 2023 IEEE International Conference on Real-time Computing and Robotics (RCAR); 2023 Jul 17–20; Datong, China. p. 719–24. doi:10.1109/RCAR58764.2023.10249889.
162. Werner J, Chu K, Weber C, Wermter S. LLM-based interactive imitation learning for robotic manipulation. arXiv:2504.21769. 2025. doi:10.48550/arXiv.2504.21769.

163. Zheng C, Zhang L, Wang H, Gomez R, Nichols E, Li G. Shaping social robot to play games with human demonstrations and evaluative feedback. In: Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024 May 13–17; Yokohama, Japan. p. 9517–23. doi:10.1109/ICRA57147.2024.10611605.
164. Zare M, Kebria PM, Khosravi A, Nahavandi S. A survey of imitation learning: algorithms, recent developments, and challenges. IEEE Trans Cybern. 2024;54(12):7173–86. doi:10.1109/TCYB.2024.3395626.
165. Liu Y, Gupta A, Abbeel P, Levine S. Imitation from observation: learning to imitate behaviors from raw video via context translation. In: Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA); 2018 May 21–25; Brisbane, QLD, Australia. p. 1118–25. doi:10.1109/ICRA.2018.8462901.
166. Li A, Boots B, Cheng CA. MAHALO: unifying offline reinforcement learning and imitation learning from observations. In: Proceedings of the 40th International Conference on Machine Learning; 2023 Jul 23–29; Honolulu, HI, USA. Vol. 202, p. 19360–84.
167. Yang M, Levine S, Nachum O. TRAIL: near-optimal imitation learning with suboptimal data. In: Proceedings of The Tenth International Conference on Learning Representations; 2022 Apr 25–29; Online. p. 1–20.
168. Xiang G, Su J. Task-oriented deep reinforcement learning for robotic skill acquisition and control. IEEE Trans Cybern. 2021;51(2):1056–69. doi:10.1109/TCYB.2019.2949596.
169. Fickinger A, Cohen S, Russell S, Amos B. Cross-domain imitation learning via optimal transport. In: Proceedings of The Tenth International Conference on Learning Representations; 2022 Apr 25–29; Online. p. 1–15.
170. Jia B, Manocha D. Sim-to-real robotic sketching using behavior cloning and reinforcement learning. In: Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024 May 13–17; Yokohama, Japan. p. 18272–78. doi:10.1109/ICRA57147.2024.10610286.
171. Todorov E, Erez T, Tassa Y. Mujoco: a physics engine for model-based control. In: Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. New York, NY, USA: IEEE; 2012. p. 5026–33. doi:10.1109/IROS.2012.6386109.
172. Echeverria I, Murua M, Santana R. Offline reinforcement learning for job-shop scheduling problems. arXiv:2410.15714. 2024. doi:10.48550/arXiv.2410.15714.
173. Pomerleau DA. ALVINN: an autonomous land vehicle in a neural network. In: Advances in neural information processing systems. Vol. 1. Palo Alto, CA, USA: AAAI Press; 1989. p. 305–13.
174. Belkhale S, Cui Y, Sadigh D. Data quality in imitation learning. In: Advances in neural information processing systems. Vol. 36. Palo Alto, CA, USA: AAAI Press; 2024. p. 1–18.
175. Zhou K, Liu Z, Qiao Y, Xiang T, Loy CC. Domain generalization: a survey. IEEE Trans Pattern Anal Mach Intell. 2023;45(4):4396–415. doi:10.1109/TPAMI.2022.3195549.
176. Belkhale S, Cui Y, Sadigh D. HYDRA: hybrid robot actions for imitation learning. In: Proceedings of The 7th Conference on Robot Learning; 2023 Nov 6–9; Atlanta, GA, USA. p. 2113–33.
177. Reddy S, Dragan AD, Levine S. SQIL: Imitation learning via reinforcement learning with sparse rewards. In: Proceedings of The Eighth International Conference on Learning Representations; 2020 Apr 26–30; Addis Ababa, Ethiopia. p. 1–14.
178. Roche J, De-Silva V, Kondoz A. A multimodal perception-driven self evolving autonomous ground vehicle. IEEE Trans Cybern. 2022;51(9):9279–89. doi:10.1109/TCYB.2021.3113804.
179. Arora S, Doshi P. A survey of inverse reinforcement learning: challenges, methods and progress. Artif Intell. 2021;297(7):103500. doi:10.1016/j.artint.2021.103500.
180. Choi J, Kim KE. Inverse reinforcement learning in partially observable environments. J Mach Learn Res. 2021;12(21):691–730.
181. Neu G, Szepesvári C. Training parsers by inverse reinforcement learning. Mach Learn. 2009;77(2–3):303–37. doi:10.1007/s10994-009-5110-1.
182. Ratliff ND, Bagnell JA, Zinkevich NA. Maximum margin planning. In: Proceedings of the 23rd International Conference on Machine Learning; 2006 Jun 25–29; Pittsburgh, PA, USA. p. 729–36. doi:10.1145/1143844.1143936.
183. Theodorou E, Buchli J, Schaal S. A generalized path integral control approach to reinforcement learning. J Mach Learn Res. 2010;11:3137–81.

184. Ramachandran D, Amir E. Bayesian inverse reinforcement learning. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence; 2007 Jan 6–12; Hyderabad, India. p. 2586–91.
185. Brown D, Goo W, Nagarajan P, Niekum S. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In: Proceedings of the 36th International Conference on Machine Learning; 2019 Jun 9–15; Long Beach, CA, USA. Vol. 97. p. 783–92.
186. Bogert K, Lin JFS, Doshi P, Kulic D. Expectation-maximization for inverse reinforcement learning with hidden data. In: Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems; 2016 May 9–13; Singapore. p. 1034–42.
187. Dimitrakakis C, Rothkopf CA. Bayesian multitask inverse reinforcement learning. In: Proceedings of the 9th European Conference on Recent Advances in Reinforcement Learning; 2011 Sep 9–11; Athens, Greece. p. 273–84.
188. Jain V, Doshi P, Banerjee B. Model-free IRL using maximum likelihood estimation. Proc AAAI Conf Artif Intell. 2019;19:3951–8. doi:10.1609/aaai.v33i01.33013951.
189. Adams S, Cody T, Beling PA. A survey of inverse reinforcement learning. Artif Intell Rev. 2022;55(6):4307–46. doi:10.1007/s10462-021-10108-x.
190. Zhang YF, Luo FM, Yu Y. Improve generated adversarial imitation learning with reward variance regularization. Mach Learn. 2022;111(3):977–95. doi:10.1007/s10994-021-06083-7.
191. Ho J, Ermon S. Generative adversarial imitation learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2016. p. 4572–80.
192. Zuo G, Zhao Q, Huang S, Li J, Gong D. Adversarial imitation learning with mixed demonstrations from multiple demonstrators. Neurocomput. 2021;457(7540):365–76. doi:10.1016/j.neucom.2021.06.053.
193. Celemin C, Pérez-Dattari R, Chisari E, Franzese G, de Souza Rosa L, Prakash R, et al. Interactive imitation learning in robotics: a survey. Found Trends Robot. 2022;10(1–2):1–197. doi:10.1561/2300000072.
194. Chernova S, Thomaz AL. Robot learning from human teachers. Cham, Switzerland: Springer; 2014.
195. Chisari E, Welschehold T, Boedecker J, Burgard W, Valada A. Correct me if I am wrong: interactive learning for robotic manipulation. IEEE Robot Autom Lett. 2022;7(2):3695–702. doi:10.1109/LRA.2022.3145516.
196. Bobu A, Wiggert M, Tomlin C, Dragan AD. Inducing structure in reward learning by learning features. Int J Robot Res. 2022;41(5):497–518. doi:10.1177/02783649221078031.
197. Welte E, Rayyes R. Interactive imitation learning for dexterous robotic manipulation: challenges and perspectives–a survey. arXiv.2506.00098. 2025. doi:10.48550/arXiv.2506.00098.
198. Kelly M, Sidrane C, Driggs-Campbell K, Kochenderfer MJ. HG-DAgger: interactive imitation learning with human experts. In: Proceedings of 2019 International Conference on Robotics and Automation (ICRA); 2019 May 20–24; Montreal, QC, Canada. p. 8077–83. doi:10.1109/ICRA.2019.8793698.
199. Nikolaidis S, Zhu YX, Hsu D, Srinivasa S. Human-robot mutual adaptation in shared autonomy. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. New York, NY, USA: Association for Computing Machinery; 2017. p. 294–302. doi:10.1145/2909824.3020252.
200. Golbabaei F, Yigitcanlar T, Bunker J. The role of shared autonomous vehicle systems in delivering smart urban mobility: a systematic review of the literature. Int J Sustain Transp. 2021;15(10):731–48. doi:10.1080/15568318.2020.1798571.
201. Udupa S, Kamat VR, Menassa CC. Shared autonomy in assistive mobile robots: a review. Disabil Rehabil Assist Technol. 2021;18(6):827–48. doi:10.1080/17483107.2021.1928778.
202. Javdani S, Admoni H, Pellegrinelli S, Srinivasa SS, Bagnell JA. Shared autonomy via hindsight optimization for teleoperation and teaming. Int J Robot Res. 2018;37(7):717–42. doi:10.1177/0278364918776060.
203. Krueger R, Rashidi TH, Rose JM. Preferences for shared autonomous vehicles. Transp Res Part C Emerg Technol. 2016;69:343–55. doi:10.1016/j.trc.2016.06.015.
204. Li Y, Tee KP, Yan R, Ge SS. Reinforcement learning for human-robot shared control. Assem Autom. 2020;40(1):105–17. doi:10.1108/AA-10-2018-0153.
205. Rubagotti M, Sangiovanni B, Nurbayeva A, Incremona GP, Ferrara A, Shintemirov A. Shared control of robot manipulators with obstacle avoidance: a deep reinforcement learning approach. IEEE Control Syst Mag. 2023;43(1):44–63. doi:10.1109/MCS.2022.3216653.

206. Huang W, Liu H, Huang Z, Lv C. Safety-aware human-in-the-loop reinforcement learning with shared control for autonomous driving. IEEE Trans Intell Transp Syst. 2024;25(11):16181–92. doi:10.1109/TITS.2024.3420959.
207. Valiente R, Toghi B, Pedarsani R, Fallah YP. Robustness and adaptability of reinforcement learning-based cooperative autonomous driving in mixed-autonomy traffic. IEEE Open J Intell Transp Syst. 2022;3:397–410. doi:10.1109/OJITS.2022.3172981.
208. Reddy S, Dragan AD, Levine S. Shared autonomy via deep reinforcement learning. arXiv:1802.01744. 2018. doi:10.48550/arXiv.1802.01744.
209. Schaff C, Walter M. Residual policy learning for shared autonomy. In: Proceedings of Robotics: Science and Systems; 2020 Jul 12–16; Online. doi:10.15607/RSS.2020.XVI.072.
210. Li M, Kang Y, Zhao YB, Zhu J, You S. Shared autonomy based on human-in-the-loop reinforcement learning with policy constraints. In: Proceedings of 2022 41st Chinese Control Conference (CCC); 2022 Jul 25–27; Hefei, China. p. 7349–54. doi:10.23919/CCC55666.2022.9902295.
211. Backman K, Kulić D, Chung H. Reinforcement learning for shared autonomy drone landings. Auton Robots. 2023;47(8):1419–38. doi:10.1007/s10514-023-10143-3.
212. Wang S, Li K, Zhang T, Zhang Z, Hu Z. HF-DQN: human in the loop reinforcement learning methods with human feedback for autonomous robot navigation. In: Proceedings of 2024 China Automation Congress (CAC); 2024 Nov 1–3; Qingdao, China. p. 552–7. doi:10.1109/CAC63892.2024.10864643.
213. Yuan H, Kang L, Li Y, Fan Z. Human-in-the-loop machine learning for healthcare: current progress and future opportunities in electronic health records. Med Adv. 2024;2(3):318–22. doi:10.1002/med4.70.
214. Gómez-Carmona O, Casado-Mansilla D, López-de-Ipiña D, García-Zubia J. Human-in-the-loop machine learning: reconceptualizing the role of the user in interactive approaches. Internet Things. 2024;25(4):101048. doi:10.1016/j.iot.2023.101048.
215. Retzlaff CO, Das S, Wayllace C, Mousavi P, Afshari M, Yang T, et al. Human-in-the-loop reinforcement learning: a survey and position on requirements, challenges, and opportunities. J Artif Intell Res. 2024;79:359–415. doi:10.1613/jair.1.15348.
216. Chen X, Zhong H, Yang Z, Wang Z, Wang L. Human-in-the-loop: provably efficient preference-based reinforcement learning with general function approximation. In: Proceedings of the 39th International Conference on Machine Learning; 2022 Jul 17–23; Baltimore, MD, USA.
217. Mandel T, Liu YE, Brunskill E, Popović Z. Where to add actions in human-in-the-loop reinforcement learning. Proc AAAI Conf Artif Intell. 2017;31(1):10945. doi:10.1609/aaai.v31i1.10945.
218. Balamurugan M, Shanmugasamy K, Balaguru S. Humans in the loop, lives on the line: AI in high-risk decision making. Eur J Comput Sci Inf Technol. 2025;13(51):27–31. doi:10.37745/ejcsit.2013/vol13n512731.
219. Christiano PF, Leike J, Brown TB, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. p. 4302–10.
220. Cao Z, Wong K, Lin CT. Weak human preference supervision for deep reinforcement learning. IEEE Trans Neural Netw Learn Syst. 2021;32(12):5369–78. doi:10.1109/TNNLS.2021.3084198.
221. Knox WB, Stone P. TAMER: training an agent manually via evaluative reinforcement. In: Proceedings of the 7th IEEE International Conference on Development and Learning; 2008 Aug 9–12; Monterey, CA, USA. p. 292–7. doi:10.1109/DEVLRN.2008.4640845.
222. MacGlashan J, Ho MK, Loftin R, Peng B, Wang G, Roberts DL, et al. Interactive learning from policy-dependent human feedback. In: Proceedings of the 34th International Conference on Machine Learning. Westminster, UK: PMLR; 2017. Vol. 70, p. 2285–94.
223. Saunders W, Sastry G, Stuhlmüller A, Evans O. Trial without error: towards safe reinforcement learning via human intervention. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems; 2018 Jul 10–15; Stockholm, Sweden. p. 2067–9.
224. Ge L, Zhou X, Li Y. Designing reward functions using active preference learning for reinforcement learning in autonomous driving navigation. Appl Sci. 2024;14(11):4845. doi:10.3390/app14114845.

225. Muldrew W, Hayes P, Zhang M, Barber D. Active preference learning for large language models. In: Proceedings of the 41st International Conference on Machine Learning. Westminster, UK: PMLR; 2024. Vol. 235, p. 36577–90.
226. Mehta V, Belakaria S, Das V, Neopane O, Dai Y, Bogunovic I, et al. Sample efficient preference alignment in LLMs via active exploration. arXiv:2312.00267. 2025. doi:10.48550/arXiv.2312.00267.
227. Mahmud S, Nakamura M, Zilberstein S. MAPLE: a framework for active preference learning guided by large language models. Proc AAAI Conf Artif Intell. 2025;39(26):27518–28. doi:10.1609/aaai.v39i26.34964.
228. Wang R, Zhao D, Yuan Z, Obi I, Min BC. PrefCLM: enhancing preference-based reinforcement learning with crowdsourced large language models. IEEE Robot Autom Lett. 2025;10(3):2486–93. doi:10.1109/LRA.2025.3528663.
229. Dempster AP. Upper and lower probabilities induced by a multivalued mapping. In: Yager RR, Liu L, editors. Classic works of the dempster-shafer theory of belief functions. Studies in fuzziness and soft computing. Vol. 219. Berlin/Heidelberg, Germany: Springer; 2008. doi:10.1007/978-3-540-44792-4_3.
230. Shafer G. A mathematical theory of evidence. Princeton, NJ, USA: Princeton University Press; 1976.
231. Chan AJ, Sun H, Holt S, Schaar MVD. Dense reward for free in reinforcement learning from human feedback. In: Proceedings of the 41st International Conference on Machine. Westminster, UK: PMLR; 2024. Vol. 235, p. 6136–54.
232. Li J, Zhang H, Zhang F, Chang T, Kuang K, Chen L, et al. Optimizing language models with fair and stable reward composition in reinforcement learning. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Allentown, PA, USA: Association for Computational Linguistics; 2024. p. 10122–40. doi:10.18653/v1/2024.emnlp-main.565.
233. Duchi J, Namkoong H. Variance-based regularization with convex objectives. J Mach Learn Res. 2019;20(68):1–55.
234. Wu Z, Hu Y, Shi W, Dziri N, Suhr A, Ammanabrolu P, et al. Fine-grained human feedback gives better rewards for language model training. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2023. p. 59008–33.
235. Puiutta E, Veith EMSP. Explainable reinforcement learning: a survey. In: Proceedings of Machine Learning and Knowledge Extraction. Cham, Switzerland: Springer International Publishing; 2020. p. 77–95. doi:10.1007/978-3-030-57321-8_5.
236. Madumal P, Miller T, Sonenberg L, Vetere F. Explainable reinforcement learning through a causal lens. Proc AAAI Conf Artif Intell. 2020;34(3):2493–500. doi:10.1609/aaai.v34i03.5631.
237. Hayes B, Shah JA. Improving robot controller transparency through autonomous policy explanation. In: Proceedings of 12th ACM/IEEE International Conference on Human-Robot Interaction; 2017 Mar 6–9; Vienna, Austria. p. 303–12.
238. Heuillet A, Couthouis F, Díaz-Rodríguez N. Explainability in deep reinforcement learning. Knowl-Based Syst. 2021;214(7540):106685. doi:10.1016/j.knosys.2020.106685.
239. Milani S, Topin N, Veloso M, Fang F. Explainable reinforcement learning: a survey and comparative review. ACM Comput Surv. 2024;56(7):1–36. doi:10.1145/3616864.
240. Zhang H, Zhou A, Lin X. Interpretable policy derivation for reinforcement learning based on evolutionary feature synthesis. Complex Intell Syst. 2020;6(3):741–53. doi:10.1007/s40747-020-00175-y.
241. Zhang L, Li X, Wang M, Tian A. Off-policy differentiable logic reinforcement learning. In: Oliver N, Pérez-Cruz F, Kramer S, Read J, Lozano JA, editors. Machine learning and knowledge discovery in databases. Rsearch track. ECML PKDD 2021. Lecture notes in computer science. Cham, Switzerland: Springer; 2021. p. 617–32. doi:10.1007/978-3-030-86520-7_38.
242. Zhang Q, Ma X, Yang Y, Li C, Yang J, Liu Y, et al. Learning to discover task-relevant features for interpretable reinforcement learning. IEEE Robot Autom Lett. 2021;6(4):6601–7. doi:10.1109/LRA.2021.3091885.
243. Yau H, Russell C, Hadfield S. What did you think would happen? Explaining agent behaviour through intended outcomes. In: Advances in neural information processing systems (NeurIPS 2020). Vancouver, BC, Canada: Curran Associates, Inc.; 2020. Vol. 33, p. 18375–86.
244. Raees M, Meijerink I, Lykourentzou I, Khan VJ, Papangelis K. From explainable to interactive AI: a literature review on current trends in human-AI interaction. Int J Hum-Comput Stud. 2024;189(4):10331. doi:10.1016/j.ijhcs.2024.103301.

245. Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, et al. Explainable AI (XAI): core ideas, techniques, and solutions. ACM Comput Surv. 2023;55(9):194. doi:10.1145/3561048.
246. Belic M, Bobic V, Badza M, Solaja N, Djuric-Jovicic M, Kostic VS. Artificial intelligence for assisting diagnostics and assessment of Parkinson's disease—a review. Clin Neurol Neurosurg. 2019;184:105442. doi:10.1016/j.clineuro.2019.105442.
247. Raaijmakers S. Artificial intelligence for law enforcement: challenges and opportunities. IEEE Secur Priv. 2019;17(5):74–7. doi:10.1109/MSEC.2019.2925649.
248. Roll I, Wylie R. Evolution and revolution in artificial intelligence in education. Int J Artif Intell Educ. 2016;26(2):582–99. doi:10.1007/s40593-016-0110-3.
249. Feng KK, McDonald DW. Addressing UX practitioners' challenges in designing ML applications: an interactive machine learning approach. In: Proceedings of the 28th International Conference on Intelligent User Interfaces; 2023 Mar 27–31; Sydney, NSW, Australia. p. 37–352. doi:10.1145/3581641.3584064.
250. Koster R, Pîslar M, Tacchetti A, Balaguer J, Liu L, Elie R, et al. Deep reinforcement learning can promote sustainable human behaviour in a common-pool resource problem. Nat Commun. 2025;16(1):2824. doi:10.1038/s41467-025-58043-7.
251. Gebellí F, Hriscu L, Ros R, Lemaignan S, Sanfeliu A, Garrell A. Personalised explainable robots using LLMs. In: Proceedings of 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI); 2025 Mar 4–6; Melbourne, VIC, Australia. p. 1304–8. doi:10.1109/HRI61500.2025.10974125.
252. Zhang N, Wang M, Wang N. Precision agriculture—a worldwide overview. Comput Electron Agric. 2002; 36(2–3):113–32. doi:10.1016/S0168-1699(02)00096-0.
253. Wu J, Tao R, Zhao P, Martin N, Hovakimyan N. Optimizing nitrogen management with deep reinforcement learning and crop simulations. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 1712–20.
254. Sun L, Yang Y, Hu J, Porter D, Marek T, Hillyer C. Reinforcement learning control for water-efficient agricultural irrigation. In: Proceedings of 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC); 2017 Dec 12–15; Guangzhou, China. p. 1334–41.
255. Wang Z, Xiao S, Wang J, Parab A, Patel S. Reinforcement learning-based agricultural fertilization and irrigation considering N_2O emissions and uncertain climate variability. AgriEngineering. 2025;7(8):252. doi:10.3390/agriengineering7080252.
256. Gautron R, Padrón EJ, Preux P, Bigot J, Maillard OA, Emukpere D. gym-DSSAT: a crop model turned into a Reinforcement Learning environment. arXiv:2207.03270. 2022. doi:10.48550/arXiv.2207.03270.
257. Wang Z, Xiao S, Li J, Wang J. Learning-based agricultural management in partially observable environments subject to climate variability. arXiv:2402.08832. 2024. doi:10.48550/arXiv.2402.08832.
258. Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. Acad Manage Rev. 1995;20(3):709–34. doi:10.2307/258792.
259. USDA. Charts and maps-corn: yield by year, US [Internet]. [cited 2025 Oct 12]. Available from: https://ipad.fas.usda.gov/countrysummary/.
260. Sawyer J. Nitrogen use: it's not your grandfather's corn. 2018. Crop advantage series. No. 23. [cited 2025 Oct 12]. Available from: https://www.agronext.iastate.edu/soilfertility/info/N-HyrbidsPage_AEP200.pdf.
261. Moffaert VK, Nowé A. Multi-objective reinforcement learning using sets of pareto dominating policies. J Mach Learn Res. 2014;15(1):3483–512.
262. Ram J. Moral decision-making in AI: a comprehensive review and recommendations. Technol Forecast Soc Change. 2025;217:124150. doi:10.1016/j.techfore.2025.124150.
263. Von Wright GH. Deontic logic. Mind. 1951;60(237):1–15.
264. Mill JS. Utilitarianism. Cambridge, MA, USA: Cambridge University Press; 2015.
265. Davis N. Contemporary deontology. In: Singer P, editor. A companion to ethics. Malden, MA, USA: Blackwell Publishing; 1993. p. 205–18.

266. Governatori G, Olivieri F, Rotolo A, Scannapieco S. Computing strong and weak permissions in defeasible logic. J Philos Logic. 2013;42(6):799–829. doi:10.1007/s10992-013-9295-1.
267. Boella G, Van Der Torre L. Permissions and obligations in hierarchical normative systems. In: Proceedings of the 9th International Conference on Artificial Intelligence and Law; 2003 Jun 24–28; Edinburgh, UK. p. 109–18. doi:10.1145/1047788.1047818.
268. Li J, Cai M, Xiao S. Reinforcement learning-based motion planning in partially observable environments under ethical constraints. AI Ethics. 2025;5(2):1047–67. doi:10.1007/s43681-024-00441-6.
269. Behrens TE, Woolrich MW, Walton ME, Rushworth MF. Learning the value of information in an uncertain world. Nat Neurosci. 2007;10(9):1214–21. doi:10.1038/nn1954.