Journal of Artificial Intelligence Research 79 (2024) 359-415 Submitted 08/2023; published 01/2024

Human-in-the-Loop Reinforcement Learning: A Survey and Position on Requirements, Challenges, and Opportunities

Carl Orge Retzlaff carl.retzlaff@human-centered.ai
Human-Centered AI Lab, University of Natural Resources and Life Sciences Vienna, Austria

Srijita Das srijita1@ualberta.ca
Department of Computing Science, University of Alberta, Canada

Christabel Wayllace cwayllac@nmsu.edu
Department of Computer Science, New Mexico State University, USA

Payam Mousavi payam.mousavi@amii.ca
Alberta Machine Intelligence Institute (Amii), Canada

Mohammad Afshari mafshari@ualberta.ca
Department of Computing Science, University of Alberta, Canada

Tianpei Yang tianpei.yang@ualberta.ca
Department of Computing Science, University of Alberta, Canada

Anna Saranti anna.saranti@human-centered.ai
Human-Centered AI Lab, University of Natural Resources and Life Sciences Vienna, Austria

Alessa Angerschmid alessa.angerschmid@human-centered.ai
Human-Centered AI Lab, University of Natural Resources and Life Sciences Vienna, Austria

Matthew E. Taylor matthew.e.taylor@ualberta.ca
Department of Computing Science, University of Alberta, Canada &
Alberta Machine Intelligence Institute (Amii), Canada &
AI Redefined, Canada

Andreas Holzinger andreas.holzinger@human-centered.ai
Human-Centered AI Lab, University of Natural Resources and Life Sciences Vienna, Austria &
xAI Lab, Alberta Machine Intelligence Institute, University of Alberta, Canada

Abstract

Artificial intelligence (AI) and especially reinforcement learning (RL) have the potential to enable agents to learn and perform tasks autonomously with superhuman performance. However, we consider RL as fundamentally a Human-in-the-Loop (HITL) paradigm, even when an agent eventually performs its task autonomously.

In cases where the reward function is challenging or impossible to define, HITL approaches are considered particularly advantageous.

The application of Reinforcement Learning from Human Feedback (RLHF) in systems such as ChatGPT demonstrates the effectiveness of optimizing for user experience and integrating their feedback into the training loop. In HITL RL, human input is integrated during the agent's learning process, allowing iterative updates and fine-tuning based on human feedback, thus enhancing the agent's performance. Since the human is an essential part of this process, we argue that human-centric approaches are the key to successful RL, a fact that has not been adequately considered in the existing literature. This paper aims to inform readers about current explainability methods in HITL RL. It also shows how the application of explainable AI (xAI) and specific improvements to existing explainability approaches can enable a better human-agent interaction in HITL RL for all types of users, whether for lay people, domain experts, or machine learning specialists.

Accounting for the workflow in HITL RL and based on software and machine learning methodologies, this article identifies four phases for human involvement for creating HITL RL systems: (1) Agent Development, (2) Agent Learning, (3) Agent Evaluation, and (4) Agent Deployment. We highlight human involvement, explanation requirements, new challenges, and goals for each phase.

We furthermore identify low-risk, high-return opportunities for explainability research in HITL RL and present long-term research goals to advance the field. Finally, we propose a vision of human-robot collaboration that allows both parties to reach their full potential and cooperate effectively.

1. Introduction

Reinforcement learning (Sutton & Barto, 2018) (RL) is a general framework in which an agent can autonomously learn to take actions to best maximize the discounted sum of future rewards, allowing agents to learn to outperform humans, sometimes generating novel and unanticipated strategies. RL agents have had many impressive successes in board games, video games, robotics, natural language processing, and other applications (Li, 2017). RL is also increasingly finding its way into the industry, with it being applied in some of the largest companies in the world, such as in recommender systems in Netflix and Spotify (Akanksha et al., 2021), for video optimization at Meta (Mao et al., 2020), or in robotic automation at Covariant (Liu et al., 2022).

The development and impact of models like ChatGPT and the emergence of reinforcement learning from human feedback (RLHF) exemplify the remarkable success achieved by combining reinforcement learning with Human-in-the-Loop (HITL) approaches. These models have demonstrated the ability to generate high-quality human-like responses and have greatly benefited from human feedback during the training process. Using RLHF, these models are able to learn from human guidance and iterate through multiple rounds of improvement, resulting in impressive performance gains and enhanced capabilities in natural language understanding and generation tasks (Aiyappa et al., 2023). This combination of reinforcement learning and Human-in-the-Loop interaction has proven to be a powerful paradigm for training AI models that can surpass human-level performance in specific domains while incorporating human values and expertise into the learning process (Choi et al., 2023).

Although the strategies and approaches learned by AI and specifically RL systems are effective in solving many specified problems, they often prove to be substantially different from the causal approaches a human would take (Lake et al., 2017) and can lack robustness and generalization (Holzinger & Müller, 2021). As model complexity increases, there is an increased risk of model bias due to the inability to sanitize large amounts of training data, resulting in undesired model behaviour. Data drift further amplifies this issue, decreasing the reproducibility of possibly important decisions (Baniecki et al., 2022). Especially in high-stakes situations where failures can cause direct harm to humans, the safety and responsibility of AI systems are essential, not only due to regulatory but also ethical concerns (Baniecki et al., 2022; Holzinger et al., 2020).

Explainability approaches play a major role in overcoming these problems by ensuring the safety and responsibility of such systems and, consequently, generating acceptance for its application in fields like medicine, finance, and defense (Baniecki et al., 2022; Heuillet et al., 2021). In addition to this, explainability can also help with the practical deployment process by allowing programmers to discover and fix bugs in the development process of complex models, which can speed up implementation (Heuillet et al., 2021).

Learning from human feedback is one of the key techniques leading to the success of ChatGPT as one of the first Large Language Models (LLM) to gain mass adoption. However, this process also resulted in data contamination, casting doubts about its robustness in different domains (Aiyappa et al., 2023). We argue that explainability forms the basis for further improvement of this interactive process, as the underlying algorithms and their decisions must be understood by a variety of different audiences with different goals to allow humans to understand and trust the behavior of the agent (Heuillet et al., 2021).

Even in cases with a low degree of human-agent interaction, such as abstract software development or industrial applications, HITL can be beneficial. In these applications, humans can, for example, improve agent robustness by monitoring the agent's performance and fixing potential errors or biases in its decision-making process and by providing the necessary input to adapt the agent's behavior to accommodate changes in the environment or new requirements (Hussein et al., 2017). A well-known example of incorporating human feedback is ChatGPT, one of the most impactful RL applications of 2022. ChatGPT learns a reward model from human feedback and then optimizes the policy against this reward model, showing how incorporating human feedback can outperform traditional approaches based on supervised learning alone (Stiennon et al., 2020).

We claim that the current framing of RL overlooks the significant human input and biases encoded in the RL problems and argue that:

1. Reinforcement learning is fundamentally a Human-in-the-Loop paradigm.

2. Explainability is critical for the success of real-world RL applications.

Firstly, we argue that RL is fundamentally a HITL paradigm and identify four phases where human involvement is critical to the goal of deploying and using RL agents: Agent Development, Agent Learning, Agent Evaluation and Agent Deployment (see Subsection

1.1). We emphasize that these phases are sequential but cyclical, meaning that individual phases can be repeated throughout the overall deployment process.

Secondly, this article serves as a position paper. We argue that ignoring humans and treating RL as a fundamentally autonomous learning paradigm is short-sighted — we highlight where and how explainability can play a critical role in those four phases of human-agent collaboration.

Thirdly, this article is a survey of explainability in RL. We teach readers about current explainability methods in HITL RL, describe how they can enable a better human-agent interaction in HITL RL applications, and present long-term research goals to advance the field.

While we emphasize that we consider the HITL paradigm critical for ML as a whole, in this paper, we focus on human-agent interaction in an RL setting. We argue that the HITL paradigm is particularly applicable to RL because it allows for integrating human input and oversight into the learning process, which helps to ensure controlled agent behavior (Lee et al., 2021).

We want to clarify that HITL applications in RL also cover topics such as bias, fairness, and personalization. As these issues are complex and require a more in-depth and specialized analysis, we refer to the related work of others, such as the European Parliament (2020), Mehrabi et al. (2021), and Arrieta et al. (2019). Our work focuses on explainable RL and contributes to the field of HITL by shedding light on why explainable RL is important and showing how we can ensure that these systems are transparent, trustworthy, and provide a human-understandable decision-making processes.

1.1 Four Phases of HITL RL Deployment

The following steps contain the identified, distinct phases for RL deployment and how humans are involved in each. Figure 1 gives an overview of the sequence of phases for agent deployment. We argue that explainability is a critical and underdeveloped technology in each of these four phases.

1. Initial Agent Development: An ML specialist lays the technical groundwork of the planned RL model. The team defines the problem to be solved, defines the agent's environment, and makes decisions about hyperparameters. Explainability helps to show how those decisions influence the agent's learning process.

2. Agent Learning: The model is trained in interactively, with the human expert providing feedback and guidance to the agent. The expert can also bias the learning process or disallow certain actions to help the model learn faster. Explainability is used during training to show the current policy and the impact of the human expert's guidance on the model.

3. Agent Evaluation: In the evaluation phase, the model is tested to see if it is ready for deployment. Specifically, domain experts must decide if the model is ready for deployment, if more training is needed, or if the problem definition needs to be changed in this phase. Here, explainability can help with an in-depth inspection of learned policies and emerging behavior.

4. Agent Deployment: The agent is deployed in a working environment and needs explainability to be safe, understandable, and reliable. It also has to provide fluent interactions to be trustworthy and effective at its intended task. In this phase, the developers determine if the agent should continue learning, have a fixed policy, or retrain if there is a change in the environment. The customer and the end-user decide where and how the agent should be used in the real world. Explainability can help users understand the final policy, improve trust, evaluate safety, and understand the stability of the policy.

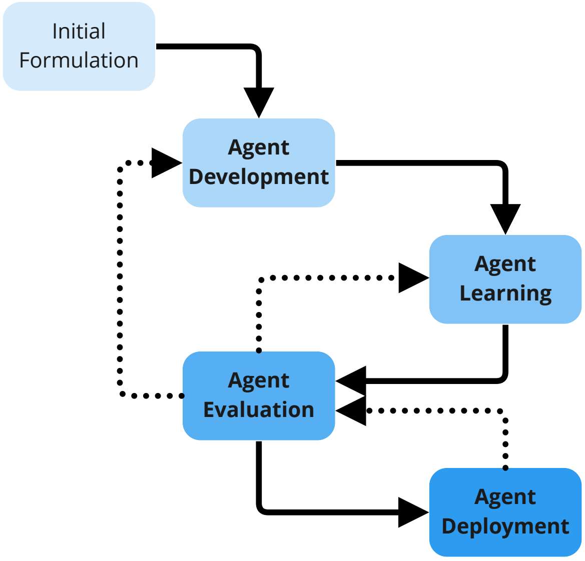

Figure 1: Overview of the sequence of the four phases for deployment. The ideal path from the initial formulation is shown with the bold arrows and goes over initial agent development, agent learning, and agent evaluation to agent deployment. The dotted arrows show the possible paths for revising different aspects of the agent and are centered around the agent evaluation phase for assessing the model flaws.

We base these phases on software and ML models and adapt them to the nature of a HITL RL workflow. The general life cycle model for software development (Ragunath et al., 2010) includes the phases of requirements, design, implementation, and testing. Similarly, Nascimento et al. (2019) use four stages of development, i.e., understanding the problem, data handling, model building, and model monitoring. Since our focus is on the deployment of models rather than the design aspect, we omit the requirements and data handling stages, which leaves us with three phases: implementation or model building, testing, and monitoring. We add a training or learning stage between the implementation and testing stage to represent the ML nature of HITL RL applications based on the phases proposed by Amershi et al. (2019) for an ML workflow (feature engineering, model training, model evaluation, model deployment, and model monitoring). We merge the deployment and monitoring phases into a single deployment phase to simplify the model and reflect the similarities between both stages. This results in the proposed four steps of agent development, agent learning, agent evaluation, and deployment. We also allow reiterations of the agent development phase, as proposed by Amershi et al. (2019), to reflect the cyclic nature of the workflow.

1.2 Overview and Goals

With this paper, we want to shift the discussion about RL to embrace human interaction and cooperation. Furthermore, we provide an entry point into this exciting area of contemporary research at the intersection of explainability in RL, with the goal of also giving non-experts a starting point for HITL RL research. Throughout the paper, we focus on human-agent interaction[1] and, therefore, center our research on topics concerning embodied intelligence (i.e. physical entities such as robots controlled by AI systems). Where appropriate, we also refer to and discuss topics about the superset of human-agent interaction, including software-only applications. We recognize that the field of HITL RL is very young and, in some sense, has significant room to improve. We highlight that not all techniques we list in this paper are ready to use in production but rather aim to steer the discussion and research on approaches we deem most promising.

The paper is structured as follows. After the introduction and motivation, Section 2 is an overview of background work for explainability and interactive learning. In the background, we review the fundamental and current challenges for embodied intelligence. After the background, we explore the four phases for the deployment of HITL RL systems and analyze where to apply explainability in HITL RL:

1. Agent Development (Section 3)
2. Agent Learning (Section 4)
3. Agent Evaluation (Section 5)
4. Agent Deployment (Section 6)

Each of the phases has specific requirements for the success of human participation. Furthermore, we explore different approaches and discuss challenges and possible directions for further research. In order to better illustrate the recommendations we have for these four phases, we added the use-case of robot operation in forestry, which we will discuss throughout the paper.

We choose forestry as our use-case example because of the various possibilities for applying HITL this field provides. Forests are furthermore of great economic value not only because they provide renewable raw materials but also for their contribution to CO_2 sequestration and, with that, the fight against climate change. The use of robots in forestry, therefore, has enormous economic significance (Holzinger et al., 2022b; Holzinger et al., 2022d). Consequently, robot operation in forestry is a use-case that receives more and more attention from the research community (Mikhaylov & Lositskii, 2018; Mowshowitz et al., 2018; Zhang et al., 2019a) and lends itself to the application of RL, as it involves making decisions over time in uncertain and dynamic environments. RL is well-suited to these types of problems because agents can learn from experience and adapt to changing conditions. Successful autonomous navigation in a forest then lays the groundwork for more demanding tasks such as weeding, species identification, and other forest management applications. While robot operation has been thoroughly explored in simulations, it has seen few real-world applications due to robustness and security issues (Surmann et al., 2020).

[1] In the case of embodied intelligence approaches, we refer to this also as human-robot interaction.

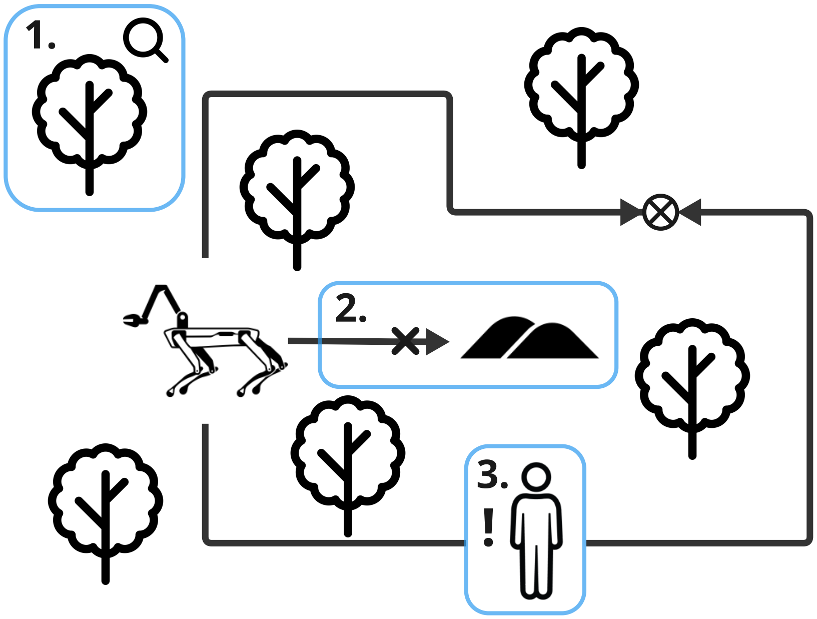

Figure 2: Use-case of robot operations in the forest with navigation to a given target. We highlight three important challenges for a successful application: (1) completing a primary task such as identifying tree species, (2) overcoming difficult terrain or replanning the route, (3) interacting with the operator or bystander.

This use-case can therefore strongly profit from robustness and explainability benefits of HITL approaches.

In Figure 2, we highlight three important challenges in robot operation in forestry: (1) completing a primary task such as identifying tree species; (2) overcoming difficult terrain; and (3) interacting with the operator or bystander while navigating to a designated target. In the beginning of each phase, we list how these challenges apply to this phase. In the discussion section of each phase, we propose specific methods suited for this phase and how they could be used to overcome the discussed challenges and enhance the agent's performance, robustness, and trustworthiness.

Section 7 discusses the general challenges observed and how future solutions can be shaped to overcome them. Finally, Section 8 outlines the general problems and goals of the field and proposes future work.

Phase	Human Involvement	Explanation Requirements	Explainability	Goals	Metrics
Development	•Define problem •Construct state space, action space, and reward function Design model	•Comparable to other model versions •Fast and simple explanations for shallow inspection •Complex and exhaustive explanations for deep inspection	**Focus:** Interpretable models such as decision trees, causal models, compositional language **Explanations**: Counterfactuals, policy querying, decision rules **Users:** RL experts	•Create thorough and comparable model summaries •Use compositional, representational language •Further integrate causal learning into HITL approaches	**Fidelity** •Correctness (correlation between output on controlled synthetic data or single feature deletion) •Explainable Model Discrepancy (error between model predictions)
Agent Learning	•Give evaluative feedback •Deliver action-advice •Select action preferences •Provide demonstrations	•Understandable by domain experts •Fluent interactions	**Focus:** HITL approaches such as human preferences querying, uncertainty highlighting **Explanations:** Counterfactuals, textual explanation in user language, saliency maps **Users:** Domain experts, RL experts	•Make use of imitation learning and preference-based learning as complementary approaches •Adapt xAI approaches to HITL context •Apply Human-as-Teacher approach •Find hybrid methods of different kinds for human interaction	**Fidelity** **Relevancy** •Human Feedback (evaluate survey with users on relevance to task) **Performance** •Time to Explanation (in milliseconds)
Evaluation	•Understand and evaluate learned policies on micro and macro-level •Test model boundaries and safety •Decide whether model is ready for deployment	•Summarise learned behavior •Scalable to large models •Comparable to untrained models •Understandable by domain experts	**Focus:** Safety evaluation by modeling uncertainty, using shield-based defenses **Explanations**: Policy summarization with natural language, rules or code, graph-based explanations **Users:** Domain Experts, RL Experts	•Ensure understandability for domain expert •Enable thorough and comparable explanations that scale with model size and complexity •Further develop dashboards for policy inspection from different viewpoints	**Fidelity** **Relevancy** **Cognitive Load** •Compactness (absolute size of explanation in number of features, path length, percent reduction to complete data) •Redundancy (overlap between parts of explanations)
Deployment	•Deploy agent •Interact with agents •Define agents' real-world application goal and context	•Fast, clear, and concise to reduce cognitive load •Understandable by end-users •Non-intrusive to prevent detrimental effects on user performance	**Focus:** Building User Trust with intent and uncertainty communication, allowing error corrections **Explanations:** Saliency maps, dendrograms, bounding-boxes, textual explanations, visual and auditory indicators **Users:** End-users	•Develop and apply new approaches beyond image and driving-based explanations •Use simple and fast explanations •Implement cohesive error and uncertainty handling •Communicate agent intent via different modalities	**Fidelity** **Relevancy** **Cognitive Load** **Performance**

Table 1: Overview of the different explanations contexts in the four different phases. *Explanation Requirements* enumerates desirable properties of explanations at this phase, and *Human Involvement* describes how the human is involved in it. The *Explainability* column lists (1) example techniques currently used, (2) directionality of explanations, i.e., agent-to-human, human-to-agent, or both, and (3) the types of users interacting with the agent at this phase. The *Goals* column describes targets that help achieve a comprehensive HITL RL experience, while the *Metrics* column lists metrics as per Milani et al. (2022) we recommend focusing on in each phase and how to quantify them with approaches listed by Nauta et al. (2022). See Section 3 for agent development, 4 for agent learning, 5 for agent evaluation, and 6 for agent deployment.

Table 2: Comparing strengths and weaknesses of major xAI approaches with an example in the forestry robot use-case.

xAI Technique	Example	Strengths	Weaknesses
Counterfactuals	If the robot encounters obstacles, a counterfactual explanation could be provided to show how the robot could change its positioning or path to avoid the obstacle.	•Provide direct insights into how changing inputs affects model outputs, facilitating quick feedback from humans (Karalus & Lindner, 2021). •Easy to understand as humans may prefer contrastive explanations (Miller, 2019a).	•Lower coverage for complex models and high-dimensional data, requiring multiple queries to better understand the overall model behavior (Keane et al., 2021).
Policy Querying	User can query the policy of the navigation system to understand why the robot chose a specific path to reach a particular forest location, considering factors like terrain conditions and efficiency.	•Allows users to interactively explore the model's decision-making process. •Can be combined with counterfactuals (Madumal et al., 2020a).	•Challenging to implement, since the model architecture needs to support the extraction and inspection of human-readable policies (Hayes & Shah, 2017). •May not provide a comprehensive understanding of the entire model's behavior.
Policy Summarization	Provide a summarized explanation to the operator, explaining its overall strategy for reaching the target location while avoiding obstacles, ensuring safety, and optimizing energy consumption.	•Provides an overview of model behavior and the overall set of actions the agent makes. •Facilitates the identification of potential biases or unintended behavior.	•Can oversimplify behavior, leading to an unfaithful representation of the real policy. •Does not explain individual actions and why they were taken (Wells & Bednarz, 2021b).
Decision Rules	When the robot needs to replan its route, decision rules can explain the criteria (such as path length, time to target and energy expenditure) and corresponding thresholds used to decide the next action.	•Easy to interpret and understand. •Can be used to build simple, transparent models that mimic the original black-box model.	•May not capture complex interactions and dependencies present in the data, especially for more complex models. •These substitute models have limited expressiveness compared to more complex models like neural networks (Liu et al., 2018).
Textual Explanations in User Language	Generate natural language explanations to interact with the operator or bystanders, providing context on its actions, intentions, and safety precautions during navigation.	•Provides explanations in a natural language format, making them easily understandable to non-experts (Xu et al., 2023). •Enables communication of complex model behavior without requiring technical expertise (Ben-Younes et al., 2022).	•Generating high-quality textual explanations may require advanced language models (Xu et al., 2023). •May require combination with other approaches (such as regional highlighting) to increase understandability (Xu et al., 2023).
Saliency Maps	Highlight areas in the robot's visual perception where important obstacles or tree species are detected.	•Identify important input features that influence model predictions. •Relatively fast and computationally efficient for feature attribution.	•May not capture global interactions and complex relationships between features. •Can be misleading and create false trust in a model (Evans et al., 2022; Glanois et al., 2021).
Graph-Based Explanations	Illustrate the relationships between different navigation paths and the corresponding terrain conditions, helping the operator visualize the decision-making process.	•Represent complex relationships and dependencies between features and predictions. •Provide a holistic view of the model's decision-making process.	•Construction and visualization of complex graphs can be challenging and time-consuming, growing worse with scale (Wells & Bednarz, 2021b). •Interpretation of graph structures can be subjective and context-dependent. Better suited for subject-matter experts than end-users (Song et al., 2019).

Continued on next page

Dendrograms	Group similar tree species based on their characteristics, helping the robot in the efficient identification process.	•Represent hierarchical relationships between data points or clusters.	•Interpretation and understanding requires domain knowledge and should best be combined with other approaches like textual explanations (Serradilla et al., 2020).
		•Enable the visualization of data clusters and their relationships in a compact manner.	•Can be sensitive to data preprocessing, chosen number of clusters, distance function etc., and may produce very different dendrograms for small changes (Kulkarni & Gkountouna, 2021).
Bounding-Boxes	Highlight the regions of interest and identified objects, ensuring safe interaction and communication.	•Localize important regions within images, providing more granular explanations (Kashyap et al., 2020).	•Selection of bounding-box size and location can influence the interpretation.
		•Useful for object detection and localization tasks in computer vision due to fast computation and being easy to understand (Behl et al., 2017).	•Best applicable to image data and may not generalize well to other data types.
Visual and Auditory Indicators	Use visual indicators and auditory signals to alert bystanders or the operator of its presence and actions, ensuring safety during navigation and interactions in the forest environment.	•Enable low-level, real-time feedback and interaction with the model in multimodal environments.	•Designing effective and informative indicators may require domain expertise (Jain et al., 2021).
		•Improve user engagement and understanding by incorporating multiple sensory inputs.	•Potential risk of information overload, leading to decreased interpretability.
		•Very helpful for communicating an agent's intent (Dragan, 2015).	•Should be combined with other approaches (textual or visual explanations) to communicate complex intents (Ben-Younes et al., 2022).

Table 1 gives the reader a comprehensive overview of the main insights of our paper. It highlights the requirements, challenges, and human context for the four phases discussed extensively in the following pages. The row corresponding to the agent deployment phase is also shown in each section to facilitate reading. Table 2 expands on this by providing examples for each of the explainability approaches mentioned in Table 1 aimed at the forestry robot use-case. The table furthermore provides a concise overview of their respective strengths and weaknesses, offering insights into their applicability, possible synergies with other xAI techniques, and the potential use in HITL RL.

In this paper, we argue that RL greatly benefits from being thought of as a human-centered process and that explainability is required to enable this HITL RL approach. We highlight how current xAI methods can be used to facilitate such HITL approaches and identify research gaps for further explainability research, ultimately enabling a more productive interaction of humans and RL agents.

2. Background

In this section, we establish the technical background for our discussion of the four phases of the deployment of HITL RL agents. We detail how to provide insight into ML models with the help of explainability approaches. We then explain how interactive learning allows integrating the HITL into RL. Finally, we summarize current challenges for reinforcement learning in general and, more specifically, in the HITL context.

2.1 Explainability

Explainable artificial intelligence (xAI) is a framework for helping human users understand the process and output of machine learning models. As ML models are deployed in a growing number of applications that affect human life (e.g., agriculture, health, smart home scenarios, etc.), the need for such xAI frameworks is ever more apparent. Moreover, xAI approaches are essential for many human-AI collaboration scenarios, where understanding and trusting model outputs are prerequisites for their use (Holzinger, 2021).

Explainability has grown from the 1980s and 1990s with researchers aiming to extract rules from knowledge or rule-based systems and neural networks (Buchanan & Shortliffe, 1984; Chandrasekaran et al., 1989; Tickle et al., 1998) towards a dedicated and expanding research community, as seen in the example of the DARPA initiative (Gunning & Aha, 2019). xAI efforts have since led to several successful xAI methods (Holzinger et al., 2022c; Zhou et al., 2021). However, we acknowledge that the terms "explainability" and "interpretability" are not used consistently and sometimes interchangeably in the literature (Miller, 2019a). Therefore, we choose to follow the recommendations of Holzinger et al. (2022c), which define explainability as the collection of methods highlighting the decision-relevant components of machine representations and machine models. Interpretability, on the other hand, is defined by Doran et al. (2017) as a system in which the user not only sees but also understands how inputs are mathematically mapped to outputs. This definition of interpretability implies model transparency and requires an understanding of the technical details of mapping but, as opposed to explainability, does not consider decision-relevant aspects and external factors. This understanding also aligns with the notion of Glanois et al. (2021), which defines interpretability as the methods that passively make the model understandable, whereas explainability actively generates explanations for a model. Therefore, we differentiate interpretability as methods that enable a mechanistic understanding of the model, whereas explainability also encompasses active explanations of decision-relevant factors.

The following examples show how xAI frameworks and human users can interact:

1. Explaining the role of a data source in the final decision, for example, to identify which data samples were used for a specific action or decision. This is, for example, important for assigning credit to (and potentially compensating) the individuals who produced the data (Zanzotto, 2019).

2. Building trust in human users is especially important when safety is a concern. In AI applications in medicine, the human user needs a reliable explanation for the decision made by the AI agent. Therefore, transparency and accountability are essential (Schneeberger et al., 2020; Stoeger et al., 2021).

3. Enabling humans to provide richer feedback through additional counterfactual examples (Del Ser et al., 2024). AI agents can use feedback in the form of explanations provided by humans, leading to more accurate, robust, and transparent models (Karalus & Lindner, 2021; Puiutta & Veith, 2020).

These examples show the various applications of xAI frameworks. Especially in the context of human-machine cooperation, the cognitive ability of the human operator paired with the computational power of a machine has the potential to handle complex tasks (Buchelt et al., 2024; Liang et al., 2017). Here, it is essential for the machine, as well as the human operator, to be able to react to the environment and for the operator to understand and interpret the actions of the machine correctly. Therefore, the underlying algorithm and its decisions must be understandable to different audiences with various goals (Heuillet et al., 2021), which shows the importance of explainability in the context of human-machine cooperation.

We use the categorization presented by Glanois et al. (2021) to classify the interpretability approaches surveyed. We extend their classification to also include explainability approaches, which results in the following three categories:

1. interpretable/explainable inputs of RL models

2. interpretable/explainable transition and reward models for RL

3. interpretable/explainable decision-making processes of RL

The first category focuses on the input to the RL model used to make decisions. It includes not only the agent's state but also other structural information, such as the problem descriptions from human experts (Hasanbeig et al., 2021) and the relational (Battaglia et al., 2018; Martínez et al., 2017) or hierarchical structure (Andreas et al., 2017; Lyu et al., 2019) of the problem. This context information helps to better understand the decisions made by RL models.

An important building block for explaining model inputs is visualizing them as perceived by the model. This visualization is often combined with showing the relevance and importance of a given decision, which helps to evaluate whether the model is focusing on the right aspects of the input, but can also mislead the user if used incorrectly. See Evans et al. (2022) for a further discussion of this set of problems. Saliency maps are one of the most common examples for visualization approaches and work by highlighting important image regions. Liu et al. (2018) show an example where continuous "super-pixels" with large feature influence are highlighted. Bach et al. (2015) developed the technique of "Layer-Wise Relevance Propagation" to iteratively change the model input to find the relative importance of individual (image) parts or features.

The second category of explainable transition and reward models leverages understandable models of the task or environment, e.g., a transition model (Martínez et al., 2016; Zhu et al., 2020) or a preference model (Icarte et al., 2018; Icarte et al., 2019). Such models help explain both the RL agent's reasoning about its decision-making and humans' understanding of the decision-making process.

The third category is interpretable/explainable decision-making of RL agents. It consists of approaches to represent decision policies in an intuitively understandable manner. Some approaches learn such interpretable policies in the form of decision trees (Likmeta et al., 2020; Silva et al., 2020; Topin et al., 2021), formulas (Hein et al., 2018, 2019), fuzzy rules (Akrour et al., 2019; Hein et al., 2017; Zhang et al., 2021), logic rules (Jiang & Luo, 2019), or programs (Sun et al., 2019; Verma et al., 2019).

Generally, it is challenging to reliably assess the quality and efficacy of xAI solutions, as user cost (that is, cognitive load and other user requirements) is often difficult to objectively measure (Bruneau et al., 2002). To better assess user cost, Milani et al. (2022) name four key metrics for evaluating xAI solutions, which we will also use for evaluating different classes of xAI solutions: *Fidelity*, the truthfulness of the explanation with respect to the model itself; *Performance*, the default metric used to evaluate the success of the AI solution to be explained; *Relevancy*, the relevance of the explanations provided to the task at hand; and *Cognitive load*, the mental effort required to understand the explanations provided. In the following paragraphs, we present different approaches to how the metrics of Milani et al. (2022) could be quantified. A related paper is presented by Nauta et al. (2022), where the authors conduct an exhaustive survey of quantitative measurements for different properties of xAI approaches.

Model fidelity can be measured with correctness as a key element to ensure that the model performs correctly and produces accurate outputs (Nauta et al., 2022). Correctness can be quantified by generating synthetic data and then testing the model on data known to be correct to determine how well the model is able to replicate the correct output. Another way to measure fidelity is to test the model's performance on data with single deletions of features, which allows determining the correlation between change in output and importance score of the feature. For explainable models, one can also compare the decisions made by the model and the explanation, and quantify correctness with error measurements such as the mean squared error (MSE) between the model's predictions and the explanation. A lower MSE indicates that the model and explanation are in closer agreement (Nauta et al., 2022). Intuitively, the computational performance of explanations can be quantified by measuring (in milliseconds or seconds, if appropriate) how quickly an explanation can be generated.

Both the relevancy metric and cognitive load are difficult to quantify, as they are highly dependent on the user's expertise and other environmental factors (Milani et al., 2022). We, therefore, propose to rely on the qualitative measurement of relevancy by evaluating human feedback for the given explanations as the most promising and comprehensive approach to determine whether the explanation is relevant to the human.

The cognitive load can also be evaluated by measuring the compactness of the explanation as surveyed by Nauta et al. (2022). A compact explanation is one that is small in size (i.e., bytes of information) or sparse, meaning it contains only the most important information. Compactness can be measured in different ways depending on the underlying model and modality, for example, with the number of features in the explanation, the path length in a decision tree, or the reduction in size compared to the complete data. Furthermore, the redundancy of explanations can be computed to evaluate and minimize the overlap between parts of the explanation.

In addition to quantifying explanations with these metrics, they can be classified as unidirectional or bidirectional. In the unidirectional case, the system simply provides an explanation to the user. In the bidirectional case, the user can give feedback or additional questions to the system, which can improve model accuracy and allow the system to generate updated explanations (Smith-Renner et al., 2020). Although this provides a more in-depth and interactive explanation process, the interactive approach is still in its infancy, with challenges in technical implementation and application scale that still need to be overcome (Smith-Renner et al., 2020; Sreedharan et al., 2022). Since we speak from the perspective of a position paper, we recommend using a bidirectional explanation in the stages from agent learning onward. However, implementing bidirectional explanations will not always be possible or viable due to the added technical complexity.

After the high-level categorization of explainability by Glanois et al. (2021), we focus in the two following paragraphs on approaches that explain model decisions as parameterized by their learned policies. As the policies determine the decision-making of a model, they belong to the "explainable decision-making" category of the categorization by Glanois et al. (2021). In policy summarization, the general aim is to make the underlying model and its policy tangible. This can be achieved by codifying its decision process as rules, as seen in the linear model U-trees by Liu et al. (2018) or by representing the learned model with generated code blocks (Verma et al., 2018).

The second approach to explain model decisions by learned policies is policy querying. In policy querying, the decision process that leads to a given result is explained. This can be general ("when do you do X") or specific to a given action. An example of a specific explanation is a natural language explanation for a classification in the ML space (Alonso et al., 2018) or the generation of a summary of "when do you do X?" type questions in natural language to explain the actions of an agent (Hayes & Shah, 2017).

As in the last subsection of explainability approaches, causal models provide a fundamentally different approach to explainability. Methods available within this framework generally fall into the categories of explainable transition and reward models and decision-making — both provide causal explanations for the task, environment, or the policies themselves.

One causal approach is graph neural networks (GNN) (Vu & Thai, 2020), which can generate explanations for a prediction via a probabilistic graphical model (PGM) that identifies crucial graph elements (e.g., nodes and edges) causally responsible for that prediction. Along similar lines, Madumal et al. (2020b) encode causal models using action influence graphs to generate explanations using causal chains. Adding these causal explanations yields better explanations as well as improved prediction performance compared with baseline explanation models.

Causal imitation learning, on the other hand, allows one to learn a structural causal model (SCM) (Pearl et al., 2000) from policies performed by humans. This is the case even if the actual reward is not specified and the environment is not perceived as the same by the learner and the human expert demonstrator. Dynamic SCMs are incorporated to formalize the partially observable Markov decision process (POMDP) (Sutton & Barto, 2018) as perceived by the agent and to take into account human intervention and its implications. The so-called counterfactual agent does not blindly take the human's advice and execute it, but rather compares it with other possible actions and decides correspondingly. In cases where the reward and transition functions are the same, human feedback is beneficial, even if the instructions are suboptimal. Counterfactual explanations are an especially powerful explanation approach since they leverage the fact that humans prefer contrastive explanations (see Miller (2019a) for a more detailed discussion).

2.2 Interactive Learning in Reinforcement Learning

A fundamental way of learning in nature is parents interactively teaching their offspring. Similar learning dynamics exist between a teacher and a student, where the teacher tries to guide the student with their experience and knowledge. Following the same rationale, interactive learning (Arzate Cruz & Igarashi, 2020a) in RL aims to include the human as teacher to guide the RL agent by using domain knowledge and rich human experience. Even without HITL, RL has been successfully applied to solve various real-world problems, ranging from drug discovery (Popova et al., 2018), navigating high-pressure balloons in the stratosphere (Bellemare et al., 2020) to robot manipulation tasks (Nguyen & La, 2019).

Although these are exciting research directions, recent deep RL systems still face many challenges, including sample inefficiency, sim-to-real transfer issues, generalization, exploration exploitation trade-off, etc., to name a few (see Subsection 2.3 for further discussion) (Ibarz et al., 2021). In response to these challenges, interactive RL aims to overcome them by involving a human prior to training (Guo et al., 2022), during training (Knox & Stone, 2008), or in the deployment phase of the RL system (Guo et al., 2021). Interactions could either be teacher-initiated (Torrey & Taylor, 2013), student-initiated (Da Silva et al., 2020; Mandel et al., 2017), or jointly initiated by both parties (Amir et al., 2016).

In interactive learning, the human is often characterized as a teacher, and the teaching loop can contain different types of critique, advice modalities, and guidance that can be fed back to the RL algorithm. There can be different modalities of human advice, such as binary evaluative feedback [+1/-1] (Knox & Stone, 2008), action-advice (Torrey & Taylor, 2013), preference-based feedback (Christiano et al., 2017; Lee et al., 2021), and sub-goal specification (Le et al., 2018). A comprehensive survey of various types of human guidance in Deep RL can be found in Zhang et al. (2019b). Thomaz and Breazeal (2006) proposed one of the earliest works on Interactive RL, which allowed human trainers to give binary feedback for the agent's behavior and specific objects associated with the task.

A common approach to modifying the reward function is called reward shaping. In reward shaping, the teacher provides useful information to shape the reward function to encourage favourable parts of the state space and penalize unfavourable parts (Ng et al., 1999). Reward shaping is useful in sparse reward environments and facilitates reward specification in complex domains. TAMER (Knox & Stone, 2008; Knox et al., 2013) is a well-known reward shaping framework where a human expert provides evaluative reinforcement (positive or negative feedback) signal by observing an agent in action, and the agent maximizes the human's feedback with classification. Subsequent variants of this method (Knox & Stone, 2010, 2012) optimized the human reinforcement with the environment reward function to learn a reward model.

Methods that consider modifying the agent's policy are called policy shaping (Cederborg et al., 2015; Griffith et al., 2013; Wu et al., 2021a). These methods augment an agent's policy directly using human knowledge. This technique does not require a well-formulated reward function but assumes that the trainer knows a near-optimal policy to guide the agent. Human advice can also be useful in guiding the agent in its exploration phase so that the agent can identify highly rewarding states or trajectories in fewer environmental interactions (Amir et al., 2016). Action pruning is another way HITL RL can guide exploration and improve learning (Abel et al., 2017).

Lastly, human-advice-based value functions can be combined with agent value functions to effectively guide the agent (Jiang et al., 2021; Kartoun et al., 2010; Taylor et al., 2011; Wu et al., 2021a). Demonstrations from humans can also increase the value function by biasing it according to the actions taken by the expert (Hester et al., 2018; Nair et al., 2018;

Vecerik et al., 2017). These approaches have been particularly successful in complex robotic tasks such as pushing, sliding, etc., which humans can easily demonstrate. Demonstrations, by default, may contain human biases that can, in turn, be removed by experts (Wang et al., 2022).

The first indicator of the need for human intervention is poor model performance, since a bad model is more likely to produce suboptimal policies. However, good models can also lead to suboptimal policies in deployment scenarios due to the sim-to-real gap (the task of translating behavior learned in a simulation to reality (Zagal et al., 2004)). In many cases, human intervention is simulated from pseudo-agents in development. This approach can help to evaluate the potential benefits and imperfections of the model before deployment. The designer of the interactive framework must also consider that human interventions might not always be perfect or beneficial; the user might need special training and an informative user interface (UI) to effectively improve the RL algorithm.

In turn, the type of UI (hardware-driven or natural interaction) determines the degree of expertise required and can affect the quality of the feedback (Lin et al., 2020). Keyboard keys, mouse clicks with sliders, and game controllers are examples of UIs in hardware-delivered interactions, and experts or knowledgeable trainers generally use these UIs. On the other hand, sound interfaces that use the techniques of audification and sonification[2] (Hermann et al., 2011; Kartoun et al., 2010; Saranti et al., 2009; Scurto et al., 2021), cameras to capture facial expressions (Arakawa et al., 2018), etc. are examples of UIs for natural interaction that non-expert users prefer.

2.3 Challenges for Reinforcement Learning and HITL Applications

To conclude the background section, we discuss the underlying challenges for reinforcement learning in general and those more specific to HITL approaches to better understand fundamental and current challenges in the field.

The first fundamental challenge in RL is the exploration/exploitation trade-off, which is defined by the decision of when to continue exploiting a current option and when to explore further for new options. The decision maker must balance between exploring a set of unknown options to find the best one (exploration) and exploiting the best option already discovered (exploitation). A more in-depth description of this problem can be found in Audibert et al. (2009).

The second important challenge is the “sim-to-real gap,” which results from the fact that simulations always under-model the target system (i.e. the real-world), which means that various aspects of reality are missing. This presents agents with unforeseen challenges and sometimes even prevents a policy learned in a simulation from being transferable to the real world. However, real-world samples are very expensive regarding cost, complexity, and time, making modeling much more appealing despite its challenges (Kober et al., 2013).

The third challenge concerns the difficulties in pixel-based learning. In real-world applications, vision is often the central modality for agents, which makes learning from images and videos essential. Current solutions, however, often rely on weak assumptions and, consequently, do not generalize well (Tomar et al., 2021). Refer to Ibarz et al. (2021) for a

[2] Audification involves visualizing an existing sound, while sonification involves creating a sound to represent data or information.

further discussion of the main challenges in RL as identified by them, namely sample efficiency, sim-to-real-gap, exploration challenges, generalization challenges, goal and reward shaping, and safety issues.

Following the general challenges for RL, Roy et al. (2021) provide a comprehensive overview of challenges specific to reinforcement learning for embodied intelligence. We highlight five particular constraints:

- Interaction with the real world involves safety risks for exploration and hard limits on resources like energy.
- Poor alignment of learned models with the real world.
- Require stronger generalizations and adaptation than regular Deep Learning approaches since specifications, goals, and rewards might change.
- Observed data are drawn from a local distribution, but generalization requires the agent to learn a reasonable world model beyond what is currently observed.
- Agent morphology defines what can be learned from the environment and has to be considered when designing agents.

Finally, there are challenges specific to HITL approaches. The first challenge is to what extent RL should imitate a good human player, and when and how RL can be used to surpass the performance of the human (Abel et al., 2017). Additionally, surpassing human performance often requires the HITL RL agent to discover new action sequences, which is hindered by engineering overly specific and complex reward functions (Liu & Abbeel, 2020). This challenge of striking a balance between human intervention and the agent's ability of exploration is another extension of the exploration-exploitation trade-off described above.

Glanois et al. (2021) discuss challenges with regard to explainability. Due to the close connection between HITL and explainability, we also include challenges for explainability in our discussion of HITL challenges. Glanois et al. (2021) state that explanations are not reliable and do not make sense when the neural network is not yet fully trained. Therefore, the network does not exhibit a cohesive behavior that could be explained. This can be due to lack of training, poor overall performance, lack of generalization capability, misclassified examples, and other underlying errors. Therefore, ML specialists will need to consider at what point in development the application of different explanation methods is appropriate and capable of providing insights into the model. Glanois et al. (2021) further name as open challenges the problem complexes of scalability, performance, and achieving full interpretability in general with RL xAI methods.

Another challenge in explainability is that most of the current xAI methods invented for deep neural networks are not created with RL principles in mind. They are driven by the mathematical principles of neural networks and typically developed with the intent of uncovering a simpler, interpretable model or pinpointing the important elements of a potential input. In the example of the convolutional neural network (CNN) that was used to process the Atari images (Mnih et al., 2013), layer-wise relevance propagation (LRP) could be used (Alber et al., 2019; Bach et al., 2015). However, this would only provide the user with a heatmap about what is relevant for positive or negative prediction, meaning that it would only characterize (in RL terms) one input state. Those heatmaps are not juxtaposed or combined with the possible actions from that state or their expected reward as a whole — the human would not know why the RL algorithm decided for the selected next action. Reconstructing the complete strategy of a model, its rules, and the underlying purposes of all (or at least the representative) state-action pairs out of those heatmaps would be a very cumbersome task.

We argue that the challenges discussed show that a generalizable and performant RL is a fundamentally challenging problem. We claim that many of these challenges can be overcome with the application of HITL approaches, supported by Mathewson and Pilarski (2022), who argue that human-centered interactive approaches are essential for designing and deploying machine learning systems. While Mathewson and Pilarski (2022) take the bird's-eye perspective on machine learning systems and formulate a high-level guideline for human-centered design of ML systems, we focus specifically on designing, evaluating and deploying interactive HITL RL systems and associated short-term and long-term challenges.

As the second focus of our paper, we argue that xAI approaches are fundamental to the success of HITL approaches, a notion supported by other researchers (Heuillet et al., 2021; Milani et al., 2022). Therefore, in each of the four deployment phases we outlined above, xAI is a central component. In the following sections, we aim to show where xAI can be applied in the deployment of HITL RL agents, which solutions exist, and how they might be adapted to allow for a productive HITL interaction. We furthermore discuss how explainability can impact the safety considerations of RL applications in each phase and how this can, in turn, help to build trust and accountability.

This section summarizes the background on explainability, interactive learning, and challenges in HITL RL. In the next sections, we will dive into the four phases of deployment, starting in Section 3, discussing how xAI techniques could be applied to the initial phase of HITL RL development. In Section 4, we focus on the agent learning phase, followed by the discussion of the subsequent evaluation and deployment of these systems in Sections 5 and 6, respectively.

3. Initial Agent Development

The first steps of the HITL RL model deployment entail underlying model development, problem formulation, and pre-learning considerations. These steps make the model understandable to machine learning (ML) specialists who seek detailed insight into their model. The development phase is characterized by the ML specialist laying the technical groundwork. In this phase, the focus is on the RL model itself.

In all phases of RL agent development, we distinguish three user types: machine learning specialist, domain expert, and end-user. The ML specialist, also called a developer, has expertise in developing and interpreting ML solutions. The domain expert has experience and authority in the field where the AI solution is applied but does not have a technical background to fully understand the AI model. Finally, end-users represent customers who buy and use a product available on the market. Although the end-user has a background in the respective field, they do not have the expertise to decide whether a given policy is correct and appropriate.

The ML specialist defines the problem to be solved by the agent and selects the agent's environment accordingly. For that, they construct a Markov decision process (MDP), defining the state space, action space, and reward function. All three have a critical impact on the speed of learning, the agent's final performance, and what policy is learned. The ML specialists also set the agent's algorithm and hyperparameters and decide whether to incorporate prior knowledge by adding hand-crafted features or transferring knowledge from an existing model. Another consideration is whether the agent should pre-train on existing data.

In the initial agent development phase, explainability can help show the impact of algorithm or hyperparameter selection, how prior knowledge biases the agent, and how pretraining changes the agent's learning process. Considering explainability in this stage allows ML specialists to start in the right direction instead of being forced to make costly (post hoc) changes later on in the learning lifecycle. Many decisions and considerations made in this phase have broad implications for the overall life cycle of RL and are likely to be difficult to change later on. Table 4 gives an overview of key aspects of the initial agent development phase.

Phase	**Human Involvement**	**Explanation Requirements**	**Explainability**	**Goals**	**Metrics**
Development	•Define problem •Construct state space, action space, and reward function •Design model	•Comparable to other model versions •Fast and simple explanations for shallow inspection •Complex and exhaustive explanations for deep inspection	**Focus:** Interpretable models such as decision trees, causal models, compositional language **Explanations**: Counterfactuals, policy querying, decision rules **Users:** RL experts	•Create thorough and comparable model summaries •Use compositional, representational language •Further integrate causal learning into HITL approaches	**Fidelity** •Correctness (correlation between output on controlled synthetic data or single feature deletion) •Explainable Model Discrepancy (error between model predictions)

Table 4: Overview of the types of human involvement, the specific requirements for explanations, the focus, and types of explainability approaches for the initial agent development phase. Furthermore, we list different goals in this phase as well as metrics and options for their quantification.

We highlight the importance of making a model understandable for developers, even if they implemented the model and have a basic understanding of how it works. First, explainability can help researchers understand the underlying mechanisms of the model and how it makes decisions, which can be useful for debugging the model and identifying potential issues or improvements. Additionally, explainability can facilitate the development of new models and methods. By understanding how a model works, researchers can better iterate on it and create new models that improve or extend its capabilities.

In the agent development phase, working with an understandable model helps ensure that the model is based on reasonable assumptions and is able to reach consistent conclusions. Numerous errors in training data, model initialization, and the initial learning process can be monitored and limited in this phase. Designing the proper function and architecture of the model in this phase also ensures a suitable baseline for comparison with the trained model.

We regard the agent development phase as the only phase where interactivity is not required. Bidirectional, interactive explanations are essential when involving experts and end-users in the development process, but at this initial stage of development, the low overhead of simpler xAI methods is preferable.

For this phase, we primarily consider the exploration-exploitation trade-off and sim-to-real gap as relevant challenges (see Subsection 2.3). These constrain the learning algorithm and parameter selection and subsequently define which type of explanation is applicable to the model.

In reference to our forest operation use-case, we highlight the risk of introducing unnoticed, fundamental errors to the model in this phase. This could, for example, be an incorrectly specified reward function, which has the potential to impede the agent's ability to perform basic navigation tasks. Since the reward function is the central building block of an agent's behavior, this can hamper the overall progress of the agent's development. In the discussion subsection of this phase (see 3.5), we discuss how explainability and HITL approaches can be used to minimize those risks.

3.1 Requirements

We propose primary considerations for the development phase and identify two approaches for an explainable development process. The first is a faster and more superficial evaluation of the general model behavior, which allows for a high-level inspection of the model behavior. The second is an in-depth assessment that provides a more detailed and complex view of the model.

In the first case, explanations must be computed quickly to enable a feedback loop during training. Xin et al. (2018) explore the implications of a faster feedback loop and highlight aspects such as introspection, the ability to rapidly analyze and compare the impact of changes to reuse intermediate results. They find that a faster feedback loop also enables an easier end-to-end optimization by the ML specialist. With the metrics of Milani et al. (2022), we therefore propose high demands for computational performance and reduced cognitive load to enable more explanation breadth.

In the second case of an in-depth assessment, explanations require more time to compute, understand, and interpret correctly. Therefore, the ML specialist can use these explanations to analyze a small number of snapshots of the model in depth, allowing a better understanding of the behavior of complex models. Both approaches should complement each other, since a thorough assessment of the model behavior requires both depth and breadth.

In the beginning of the initial agent development phase, the generated explanations should furthermore be comparable to those of other model versions to track the progress of development by contrasting different behaviors and their explanations. This is essential to enable researchers to monitor how the model behavior evolves and assess if the development progresses in the desired direction.

Complex and detailed explanations are best suited for ML specialists, as they require detailed insights into the model and can afford the required cognitive load. In addition, computational resources are the least constrained in this phase, and explanations do not yet need to scale to large model sizes. In the metrics of Milani et al. (2022), these requirements result in lower demands for computational performance and cognitive load but high fidelity.

In the following section, we identify different approaches that can help during this phase. First, pre-training can help with the groundwork of intelligent behavior and enable sensible debugging. Second, interpretability approaches should be considered in this phase. They allow for an inherent understandability, which can benefit all subsequent phases if implemented here. Third, we argue which types of explainability approaches are suitable for this phase.

3.2 Approach: Pre-Training

Pre-training models benefit the RL training workflow since they prepare the groundwork for the Human-Robot interaction and benefit HITL in several ways. Pre-training models helps to develop useful priors, diverse behaviors, generalized policies/feedback, and efficient initial feedback from human input (Daniel et al., 2016; Eysenbach et al., 2018; Florensa et al., 2017; Hazan et al., 2019; Lee et al., 2021). Recently, Parisi et al. (2022) demonstrated the effectiveness of pretraining with out-of-domain computer vision data sets for downstream robotics control tasks, which shows that pre-training can learn useful representations even from out-of-domain datasets. Pre-training involves training a model on a large dataset in an unsupervised manner to learn general language representations. These pre-trained models can then be fine-tuned on specific tasks or domains with smaller datasets with transfer learning (Taylor & Stone, 2009). The pre-trained model acts as a knowledge base, capturing general language understanding, and the fine-tuning process adapts this knowledge to the target task, making it more efficient and effective.

Preference-based learning is a HITL technique that strongly benefits from pre-training, in which an embodied agent allows a human to decide which is the preferred option of two or more possibilities of behavior (for example, two different movement policies). This choice simplifies the difficult reward-selection process, i.e., sidesteps the need to define the reward function explicitly. It is advantageous for this approach if the robot already exhibits two "meaningful" movement policies rather than the normal frenetic behavior found in newly instantiated models. Judging the policies of an already trained model is generally easier because, after the initial noise of random initialization, it shows meaningful behavior (Akrour et al., 2011). Lee et al. (2021) demonstrated the importance of unsupervised pre-training using intrinsic reward to make the agent learn diverse skills, which further helps generate informative queries for receiving human preference. Generally, approaches such as transfer learning and lifelong learning for RL agents are promising, as they also alleviate the problem of the noisy warm-up phase of RL (Taylor & Stone, 2009; Yang et al., 2021).

The application of pre-training in ChatGPT has demonstrated the effectiveness of leveraging large-scale datasets to initialize Large Language Models (LLMs), leading to a reasonable parameter initialization and, with that, improved performance and more efficient learning in downstream tasks (Zhou et al., 2023). This success showcases the paradigm shift that pre-training can bring to RL, as it enables models to acquire useful knowledge from vast amounts of unlabeled data, providing a strong foundation for subsequent fine-tuning and learning from specific tasks.

Pre-training holds relevance not only during the development stage but also impacts the other stages due to the superior downstream performance. During the learning stage, pre-training helps initialize the RL agent, facilitating faster convergence and enabling better sample efficiency. In the evaluation stage, pre-training can help improve the model's performance and generalization capabilities on diverse tasks. Lastly, in the deployment stage, pre-training ensures that the RL agent is well-equipped with foundational knowledge, enhancing its ability to adapt to real-world scenarios and perform effectively (Yang et al., 2023).

3.3 Approach: Explainability

Explanations can provide various benefits for building RL models during the initial agent development phase. Contextual information should be taken into account when defining what constitutes a "good" explanation for an RL model. This can be background knowledge, different levels of expertise, as well as the needs and expectations of the addressee of this explanation. There are various types of explanations, like visual (Atrey et al., 2019; Gupta et al., 2019), textual (Fukuchi et al., 2017b; Hayes & Shah, 2017), causal (Madumal et al., 2020a, 2020b), or decision tree explanations (Bastani et al., 2018). The approach of Liu et al. (2018) sets out the decision process as rules to make the influence and learning of the network more transparent. Another approach is to represent the learned model with generated code blocks, as presented by Verma et al. (2018). A policy network is codified by learning a neural policy network and searching for the optimal policy, which results in human-readable policies and improves generalization, but it also incurs a performance penalty during training. A third approach is to represent network policies in natural language, such as Alonso et al. (2018), who showed an example of justifying classifications with a textual explanation of the choice made by a decision tree.

Furthermore, the subset of policy querying approaches that allow one to look at questions like "When do you do X?" can be used for this phase. Hayes and Shah (2017) provide an example of such questions and generate a summary of a "When do you do X?" type question in natural language to explain the actions of an agent. An important addition to this approach is the use of counterfactuals (see Evans et al. (2022) for a more thorough assessment of the importance of counterfactuals). Madumal et al. (2020b) generate a structural causal model for RL agents, which allows one to generate explanations of taken actions (see Subsection 2.2). This approach also allows one to respond to counterfactual queries, like: "why did you not do Y?". Providing such counterfactuals is shown by the authors to produce satisfactory explanations and increase user trust.

Finally, causal learning strategies can assist in understanding the underlying model and move from just interpretability to full explainability. This shift is accomplished by not only making the model understandable, but also by actively explaining decisions and their context. An example of such xAI methods is probabilistic graphical models (PGM), which help to construct causal models and are often applied to graph neural networks (GNN) (Saranti et al., 2019). Graph neural networks are especially suitable for explainability methods in the initial agent development phase, as they allow for an intuitive and more direct visualization of critical components. Vu and Thai (2020), for example, support the informed creation of a causal model by identifying essential graph components and then generating PGMs that approximate that prediction. These essential components can help to identify cause and effect in neural networks and determine cause-and-effect relations.

3.4 Focus: Interpretability

A central approach to adding explainability in the HITL (in the broader sense) is to make the model interpretable, i.e., to provide an inherently understandable AI solution. Interpretable models, in combination with pre-trained or otherwise initialized systems, facilitate the judgment of the model and its decision-making by the ML specialist.

Roy et al. (2021) enumerate several approaches that would translate to more interpretable models. For example, embedding core knowledge into models (such as physical constraints) could provide agents with innate reasoning capabilities. This ability to reason can, in turn, allow checking and debugging the entire reasoning process (Ha & Schmidhuber, 2018). As second aspect, Roy et al. (2021) highlight the use of compositional language, which can facilitate a high-level understanding of the concepts the model learned. Koditschek (2021) suggests that the use of model composition and, with it, compositional language are key elements of embodied intelligence.

In addition to the approaches listed above, using a representative language could allow abstractive reasoning with rigorous generalization (Roy et al., 2021). This can be achieved by graph neural networks, natural language, or attention mechanisms in combination with sys1/sys2 separation and further the inherent understandability of the learned model. The advantage of a combination of innate reasoning with understandable language could allow an intuitive understanding of the model. Intuitive understanding of model behavior can be one building block for tackling the challenge of adversarial attacks, for example with image recognition approaches. Many image recognition models focus on very different aspects than what humans do and “understand” images on a fundamentally different level. This makes image recognition susceptible to changes that are unnoticeable for a human observer, but drastically alter the classification of an image (Chakraborty et al., 2021). By ensuring the model “understands” images in the same way that humans do, this attack surface can be reduced.

3.5 Discussion, Outlook, and Use-Case

Regarding the current challenges in the development phase, we find many xAI approaches intended for this setting. However, we encounter a lack of interactivity and comparability of the explanations, which hinders a thorough evaluation of the model at this phase. Furthermore, we find that current explainability approaches often lack the causality and intuitive understandability required for a thorough introspection of a newly developed model.

Firstly, we suggest using compositional and representational language for explainability to enhance a model’s intuitive understandability, as mentioned by Roy et al. (2021). Representing the model as a hierarchical, graphical, or topological structure (Battaglia et al., 2018; Lyu et al., 2019) is more understandable to humans than a traditional neural network model. Such structural models are however not as powerful as neural network models, since their expression and computation ability are limited. An optimal approach would therefore be to integrate the structural models with the neural network model itself without losing explainability.

Secondly, we think that causal learning approaches should be adapted and integrated much more deeply into HITL approaches. In addition to generating better explanations, they could also yield improved prediction performance, as exemplified by Madumal et al. (2020b).

Thirdly, policy querying approaches such as those presented by Hayes and Shah (2017) could be adapted to allow specific inquiries into the model structure and be expanded with counterfactual structures. We ultimately envision interactive, thorough, and comparable model summaries. Individual components of such a solution can already be found, but the simplicity of a comprehensive solution could greatly benefit the development process. Roy et al. (2021) encourage us to think about the opportunities that other forms of sensors, sensor fusion, and new components enable, such as novel interaction approaches, application areas and ways of learning (see Subsection 2.3 for a perspective on the respective challenges). Another direction involves human teaching (Kulick et al., 2013) or programs (Penkov & Ramamoorthy, 2019; Sun et al., 2019) to guide the agent in learning the symbolic structure or representations of the task, which greatly reduces the complexity of the task.

In the example of robot operation in the forest, core knowledge of physical properties helps develop a more robust model of the agent and the environment (Ha & Schmidhuber, 2018). The model policy can, for example, be represented and tested with code blocks (Verma et al., 2018). The representation with code blocks is chosen since this phase focuses on the model developers and, therefore, speaks the user's language. To further enhance the model debugging process, it can be helpful to allow developers to query the model policies and allow inspection of why the robot stopped and what object it considered an obstacle (Hayes & Shah, 2017). Finally, the environment could be represented as a graph with points of interest to help developers better understand how the agent perceives the landscape (Lyu et al., 2019) and ensure that this understanding is in line with the real-world conditions.

We emphasize that explainability is an essential tool for uncovering model errors at the initial agent development phase. However, explainability alone will not be enough to discover all possible errors. We recommend aligning subject-matter experts and ML specialists as well as considering performance metrics and other indicators to ensure that the reward function aligns with the desired objectives. Explainability serves as a valuable tool in this process, facilitating the collaboration between experts and the agent in a HITL process, highlighting where the agent takes incorrect decisions, and providing insights to enhance the understanding of the implications of the reward function (see Section 5 for further discussion).

3.6 Initial Agent Development Summary

To summarize the development phase, we refer to Table 1. The model explanations should be comparable to each other, potentially allowing the user to understand the differences between the tested architectures. Explanations should be carefully designed to be broad and shallow or detailed and complex, depending on the cognitive resources and goals of the audience. Humans are involved in this phase for defining the problem, the overall model design, as well as the specification of state space, action space and reward function, and evaluation metrics of the agent.

We recommend using pre-training approaches such transfer learning in combination with preference-based learning to leverage human knowledge effectively. Possible xAI approaches for HITL in this phase are focused on making the underlying model more understandable, for example, by enabling interpretable model policies by representing them as trees, causal models, or with compositional language. Finally, post hoc explanations, such as counterfactuals or extracted rules, can be useful to gain insights even at this early stage. Causal approaches should also be considered at this stage to help identify cause and effect in this phase and those further downstream.

In this section, we covered the development phase, the first of the four phases in creating HITL RL systems. In the next section, we discuss the subsequent phase of interactively training the agent.

4. Agent Learning

Once developed, the agent is trained either autonomously or interactively with the help of human input. Section 3 discussed the current explanation techniques used mainly during the development phase under the close supervision of the developer. This section focuses on approaches where humans act as a trainer or teacher (as domain expert), guiding the agent learning process.

In this phase, the model is trained interactively with a domain expert, under the close supervision of ML specialists. This phase focuses on understanding the model perceptions and fundamentals, and assesses the agent's cooperation capabilities with the end-user. The human can decide to disallow the agent from selecting invalid or suboptimal actions. The agent's action selection can also be expanded with an expert bias to enable faster learning rates. Furthermore, developers can decide to include an interactive paradigm with human demonstration, feedback, advice, or other types of cooperation.

Eventually, the domain expert judges whether the learning process is successful or if the MDP (or other agent components) should be revised. This can be either because the agent is learning too slowly or because the policy being learned is otherwise unsuited for the problem. Finding and determining the appropriate role of HITL should be the first step in developing an optimal mechanism for interaction. We propose that the focus should be on providing interpretable information and clear explanations to the human during this teaching process to enable transparency between the parties. As a consequence, human trainers can understand how agents perceive the world and provide better feedback. Explainability can show the impact of bias through an existing controller or human advice, how learning is progressing, or how the current policy functions.

Among the challenges identified in Subsection 2.3, the following are relevant to this phase: (1) determining to what extent agents should imitate or follow human advice, (2) optimizing reward shaping, (3) choosing explanation methods and complexity according to the background knowledge and expertise of teachers, and (4) reducing the sim-to-real gap. Table 5 gives an overview of key aspects of the agent learning phase.

In the forest operation use-case, we are now beginning to engage with non-technical experts. During this phase, we will be communicating and interacting with inexperienced operators, which may present unexpected scenarios that stress the agent's ability to generalize. Additionally, it is important to consider the potential risks associated with the agent's interactions with the environment, including navigating unknown obstacles or unstable ground. This emphasizes the importance of thorough testing and validation during this phase, to ensure the agent's ability to operate effectively in the field.

Phase	Human Involvement	Explanation Requirements	Explainability	Goals	Metrics
Agent Learning	●Give evaluative feedback ●Deliver action-advice ●Select action preferences ●Provide demonstrations	●Understandable by domain experts ●Fluent interactions	**Focus:** HITL approaches such as human preferences querying, uncertainty highlighting **Explanations:** Counterfactuals, textual explanation in user language, saliency maps **Users:** Domain experts, RL experts	●Make use of imitation learning and preference-based learning as complementary approaches ●Adapt xAI approaches to HITL context ●Apply Human-as-Teacher approach ●Find hybrid methods of different kinds for human interaction	**Fidelity** **Relevancy** ●Human Feedback (evaluate survey with users on relevance to task) **Performance** ●Time to Explanation (in milliseconds)

Table 5: Overview of the types of human involvement, the specific requirements for explanations, the focus and types of explainability approaches, the goals as well as metrics for the agent learning phase.

4.1 Requirements

This phase is the first where a novice user may interact with the RL agent, which imposes certain requirements for a reduced complexity of the explanation. This also coincides with the requirement for rapid explanations for the sake of fluency, enabling a truly interactive training process. Interpretable inputs, such as symbolic representations of the problem structure or visualizations of the agent's perception, are well suited for these requirements since they give rapid introspection into what the agent perceives and bases its decisions on (Glanois et al., 2021). This rapidity only allows for a more shallow introspection, which is alleviated to a certain degree by the thorough testing in the development and evaluation phases. Fundamental errors in the model should be taken into account during the development phase, while hidden biases introduced during training are in focus during the evaluation phase.

We emphasize that even suboptimal explanations by human teachers can be better than none. Current literature focuses on agents using human advice during the learning phase, taking advantage of humans' a priori knowledge. However, even though human decisions could be less accurate, Zhang and Bareinboim (2020) demonstrate that agents are more likely to learn suboptimal policies if they ignore human advice. To refer to the metrics of Milani et al. (2022), we propose the requirements of high fidelity, relevancy, and performance to ensure that the model is able to learn and interact with the user effectively.

4.2 Approach: Explainability

As stated in Subsection 2.2, interactive RL uses human feedback to reduce problems in areas such as sample efficiency, sim-to-real transfer problems, and generalization. Like any interaction, interactive RL requires a level of agent-human understanding, and one effective way to improve communication involves explaining one's and others' behavior (De Graaf & Malle, 2017). Therefore, more and more recent approaches augment interactive RL using explainability techniques. For instance, researchers found that most people training an AI agent assume that their behavior reveals their knowledge (Habibian et al., 2021).

Fukuchi et al. (2017b) proposed a method to explain an agent's future behavior to its trainer using the same expressions used by the trainer. The agent selects the phrases assuming that a higher reward means that the agent correctly followed the advice. In later work, they apply the approach to agents that dynamically change policies (Fukuchi et al., 2017a).

Furthermore, simple visualization techniques, like saliency maps or layer-wise relevance propagation, could be used to explain the agent's perception of the world at a glance. These techniques each have their own set of benefits and drawbacks more explicitly discussed in Liu et al. (2018) and Bach et al. (2015)

4.3 Approach: Interactive Learning

While explanations usually assume a human explainee, interactive RL can also allow the human to give feedback or explanations to the agent. One common way of providing feedback is to evaluate agent actions as positive or negative (Arakawa et al., 2018; Knox & Stone, 2008; Knox et al., 2013; MacGlashan et al., 2017). However, this limited feedback could improve if the trainer explains why specific actions are wrong. Guan et al. (2020) augment the binary evaluative feedback with visual explanations using saliency maps from humans. In addition to improving the agent's sample efficiency, the approach also reduces the human input required. Likewise, Karalus and Lindner (2021) enhance evaluative feedback but this time using counterfactual explanations, yielding significant improvements in convergence speed. In case of negative feedback, humans can communicate to the agent that the feedback would have been positive *if* action a had been performed in a different state s'. The authors limit counterfactual feedback only to negative reward cases, where it has the largest impact.

Another example of counterfactual application is provided by Pearl (2009), which uses dynamical structural causal models (DSCM) to explicitly model the differences in the capabilities of the agent and the human operator as the world states evolve. In this framework, the agent views human feedback as the intended action and adjusts it (using counterfactual reasoning) if the action is suboptimal. A trade-off between autonomy and optimality is demonstrated, meaning that fully autonomous agents are likely to be suboptimal and could only achieve optimality if they receive critical feedback from their human operators. The counterfactual approach proposed by the authors improves on standard methods even when human advice is imperfect.

Further, human trainers tend to give more positive feedback and the learning agent should be able to inherently accommodate this feedback bias. Making the agent's assumptions transparent to the trainer can improve the overall process. Additionally, RL agents who learn from human demonstration, imitation, or querying the trainer's preferences can inherit biased human behavior. In that case, xAI can also shed light on the biases of the model before deployment. The robot then selects different behaviors and asks people about their preferences. Habibian et al. (2021) study the influence of robot questions on how their trainers perceive them. In their approach, the robot chooses informative questions that simultaneously reveal its learning. Compared to other approaches that do not account for human perception, Habibian et al. (2021) found that people prefer revealing and informative questions since they find these questions easier to answer, more revealing about the robot's learning, and better focused on uncertain aspects of the task.

4.4 Focus: Interaction Design

The agent learning step is the first phase where interaction with a non-technical user, i.e., the domain expert, takes place. Therefore, we recommend using this phase to focus on the interaction between user and agent, enabling a fluent and cooperative workflow. Wu et al. (2021b) propose that a human can play different roles for interaction with RL agents, such as Supervisor, Controller, Assistant, Collaborateur, or Impactfactor. This encourages us to take into account how collaboration is framed and what it entails in developing efficient teaching approaches. A learning agent will possibly interact with the designer, trainer, and final user. Therefore, it is important to consider experts to naïve collaborators. For example, tools for visual explanations that use typical data visualization techniques, such as bars, may be useful for people with a scientific background; however, they add mental load to others (Anderson et al., 2020).

Wu et al. (2021b) state that the ideal interaction for HITL would be fluent, performant, and reliable. For systems geared at performance, the interaction is usually framed as collaboration, while a focus on reliability favors the role of supervisor for the human. For fluency, however, new roles of interaction are proposed and discussed, which would also require new kinds of interface. An agent could integrate implicit empathic feedback from a human in the form of gestures, vocalizations, and facial expressions as shown by Cui et al. (2020), allowing a more intuitive interaction and richer feedback from the human to the learning agent. To effectively leverage such feedback, appropriate user interfaces should be developed, and the underlying model should be able to process multimodal data such as speech or image. Another form of implicitly improving feedback is the approach presented by Peng et al. (2016), which makes the agent move slower if it is uncertain, enabling the human to provide feedback where it is most useful with an intuitive cue. Identifying and leveraging useful implicit feedback from humans would facilitate agent learning by moving beyond what the human explicitly mentions as a teacher.

Ultimately, we propose that Agent Learning approaches should consider different approaches in their HITL framework rather than forcing one specific technique, which allows one to better determine the mentioned sweet spot of interaction.

4.5 Discussion, Outlook, and Use-Case

For the agent learning phase, we find that many approaches support the Human-as-Teacher paradigm itself, but identify a lack of explainability approaches that facilitate this interaction. More generally, xAI approaches in this phase need to evolve to support interactivity in the HITL setting, which is currently underdeveloped.

To proceed, the various current approaches should be further adapted for the HITL and RL contexts. An important adaptation is the development of a suite of tools to facilitate the systematic deployment and comparison of the different approaches to discover sweet-spot mixes. An example of an effective combination could be to build fundamental behavior via imitation learning, followed by fine-turning actions by preference-based learning, and finally identifying and solving weak spots using querying approaches. We propose to strive for a solution that allows efficient HITL training in the field along with domain experts, enabling users to rapidly bootstrap agent behavior and further support this process with replanning and corrections. A combination of such approaches could be very efficient and fast in bringing about robust agents suited for real-world applications.

With regard to our use-case, we now move to real-world tests with a domain expert. Since at this stage we aim to improve the model's robustness and performance, the focus should be on thorough explainability. We recommend providing visual explanations of robot perception at this point (Glanois et al., 2021), augmented, for example, by easily understandable saliency maps that highlight important image regions (Liu et al., 2018). In this phase, we also recommend starting to integrate human feedback, which can be used to classify elements in the environment, for example, to determine whether a way is passable or not (Guan et al., 2020). If uncertain, the robot can move slowly and highlight the problem in its visual output to communicate with the user (Peng et al., 2016). Additionally, imitation learning could be integrated to allow the robot to simply follow the user path and, with that, learn more about the environment in a safe way (Pearl et al., 2000).

4.6 Agent Learning Summary

To summarize the agent learning phase according to Table 1, explanations should provide the user with interpretable inputs and allow for a rich interaction to better understand the behavior of the model. This also requires explanations to be in the language of the domain expert, facilitating a productive "Human-as-Teacher" interaction. We recommend focusing on designing and testing the optimal interaction between human and agent, balancing feedback, explainability, and nonintrusiveness. Humans are involved in this phase to provide evaluative feedback, give advice and preferences to the agent, and provide demonstrations for complicated tasks. The main explainability techniques are those explaining the agent's perceptions and evaluating behavior, for example, with counterfactuals or textual explanations.

Furthermore, techniques which enable the agent to query human preferences or give interactive feedback to instructions from the human teacher are recommendable, while visualization approaches like saliency maps could be used for the sake of their low cognitive and computational load. Explainability is interactive at this phase, and RL experts as well as domain experts are involved. These requirements differ from the agent development phase in that a domain expert is involved in the process, and explainability should be geared towards interactivity and efficient agent learning, whereas the development phase was more focused on generating insights into the model architecture.

In this section, we covered the second phase of development, discussing the training of HITL RL agents. In the next section, we discuss the subsequent phase of thoroughly evaluating the learned policies and the emerging agent behavior.

5. Agent Evaluation

In this section, we detail the requirements of the agent evaluation phase and how explainability approaches and safety considerations can help build trust in the learned model. For the success of human-robot teamwork, safety is essential. The robot must meet the innate expectations of the human to be predictable and safe and communicate its intentions (Eder et al., 2014).

The third phase is characterized by testing the behavior of the trained model, which requires tools that enable comparability, quantify the learning progress, and can furthermore scale up to thoroughly evaluate larger-scale models. ML specialists and domain experts exhaustively test the learned model. Developers need to ensure that no erroneous behavior or glitches have emerged during training, focusing mainly on the syntactical level. The domain expert will need to understand and evaluate the learned policies for sensible micro- and macro-behavior, therefore focusing more on an evaluation of the semantic behavior.

The judgment for either moving forward with the deployment to market or engaging in another development-learning-evaluation cycle lies in this phase and is made by the developers and project owners. The domain expert must decide if the learned policy is ready for deployment, if more training is needed, or if the problem definition needs to be changed. In the evaluation phase, explainability can help with an in-depth inspection of learned policies and emerging behavior.

To ensure this, the trained system has to be tested extensively. It is important to make a distinction between errors in the underlying model and errors learned during training. In the agent deployment phase described in Section 6, underlying errors in the model architecture should be discovered and fixed. Therefore, the agent evaluation phase can focus on discovering errors acquired during training and, ultimately, ensuring a safe decision-making process.

This type of acquired error becomes apparent in various forms. One is shortcut learning, where a model finds undesired shortcuts in the training data instead of learning the desired concept. This often prevents a generalization since the shortcut is, in most cases, not present in the application context. For example, an algorithm that learns the hospital token embedded in an image rather than the targeted signs of pneumonia in X-ray images represents a case of acquired error (Geirhos et al., 2020).

Another symptom of errors acquired during training is adversarial attacks, which show that the model did not learn the desired concept, but rather invisible patterns in the image (Goodfellow et al., 2014). There are several approaches to reduce the attack surface for adversarial attacks with optimizations in the training process, but we will focus on the underlying issue of models failing to learn concepts. This problem is also part of the challenge of alignment of the learned model and the real world, identified in Subsection 2.3 and discussed by Roy et al. (2021).

Both issues show why it is important to examine the behavior of the trained model. We propose that at this phase of development, the focus should be on the evaluation of safety aspects, such as represented by the decision-making process of the model, which reflects the learned behavior. This decision-making process can be made tangible with approaches such as policy summarization, graph-based explanations, and causal models. Furthermore, we consider different safety approaches, which ensure that the agent takes only allowed actions and is transparent with regard to its level of uncertainty. Table 6 gives an overview of key aspects of the agent evaluation phase.

The agent evaluation phase of our forest operations use-case is crucial to discover potential pitfalls that could arise. One pitfall is the failure to discover emerging errors in the model, which impedes performance or prevents application in the field. Failure to consider

Phase	Human Involvement	Explanation Requirements	Explainability	Goals	Metrics
Evaluation	•Understand and evaluate learned policies on micro and macro-level •Test model boundaries and safety •Decide whether model is ready for deployment	•Summarise learned behavior •Scalable to large models •Comparable to untrained models •Understandable by domain experts	**Focus:** Safety evaluation by modeling uncertainty, using shield-based defenses **Explanations**: Policy summarization with natural language, rules or code, graph-based explanations **Users:** Domain Experts, RL Experts	•Ensure understandability for domain expert •Enable thorough and comparable explanations that scale with model size and complexity •Further develop dashboards for policy inspection from different viewpoints	**Fidelity** **Relevancy** **Cognitive Load** •Compactness (absolute size of explanation in number of features, path length, percent reduction to complete data) •Redundancy (overlap between parts of explanations)

Table 6: Overview of the types of human involvement, the specific requirements for explanations, the focus and types of explainability approaches, the goals as well as metrics for the agent evaluation phase.

how the agent's performance scales with increasing complexity and size of the tasks can lead to great results in theory and less useful results in practice. This could manifest in, for example, an agent that can only successfully complete ten consecutive tasks for which it was trained, but then starts to fail or misbehave after that.

Another key consideration is ensuring that the agent is thoroughly tested across a wide range of edge cases and scenarios to ensure that it can handle a diverse range of operating conditions. An agent that is not able to handle a wide range of tasks and operate effectively under different conditions will be restricted to a very narrow field of applications, while potentially failing under unfavorable conditions. This could, for example, be a forest with a different biome, terrain roughness or weather conditions than trained on. This shows why it is important to ensure that the agent's reliability and safety are properly evaluated. This includes both ensuring that the agent is able to detect and avoid potential hazards, such as obstacles or unstable ground, and that it can operate in a generalizable and reliable manner that minimizes the risk of injury to both the end-user and other personnel.

5.1 Requirements

During the evaluation phase, xAI approaches have to work with large models and complex decision-making processes. This requirement, for example, makes the use of text- or rule-based approaches (Hayes & Shah, 2017; Tabrez & Hayes, 2019) more challenging, since they might be helpful when producing one page of output, while parsing and understanding many pages of model policy explanations will become prohibitive. In a similar vein, visual approaches such as trees or DAGs, in general, should not exceed a certain size to still be useful.

This requirement is supported by Wells and Bednarz (2021a), who find that the authors of several xAI approaches identify scaling up their approaches as a major challenge, which also shows why many xAI approaches are only applied to toy examples. This is especially relevant to text-or graph-based explanations, which can rapidly become unwieldy.

Furthermore, models should provide explanations that are understandable to domain experts. It will often be the case that only the domain expert, instead of the developer, can judge whether a learned policy is consistent, which makes it a requirement that the domain expert can evaluate it.

A final consideration is that the provided explanations should be able to highlight differences in the learned behavior with regard to a newly initialized or only pre-trained model, in order to gain an understanding of what the agent actually learned during the different phases of the training process.

To refer back to the metrics of (Milani et al., 2022), we propose adding cognitive load to the demands on fidelity and relevancy in the previous phases. Cognitive load is added to ensure explanations scale to the full-size model, which constitutes a major challenge with approaches relying on visual or textual summarization (Vu & Thai, 2020).

5.2 Approach: Explainability

Policy summarization approaches focus on showing and explaining model policies to the user. We refer to examples that codify the decision process of the model as rules (Liu et al., 2018), as code blocks (Verma et al., 2018), or through natural language. Alonso et al. (2018) shows an example of justifying classifications with a textual explanation of the choice made by a decision tree, which could in turn be transferred to RL applications. Other examples for providing an introspection into LLMs with textual explanations are Zini and Awad (2022) and Xu et al. (2023).

Policy summarization is well suited for assessing a trained model and checking its policies for unexpected and undesired behavior. Depending on whether and which domain experts are included in the process, different summarization approaches are advisable. Summarizing model policies as code blocks can be intuitive for computer science and adjacent fields, but are likely inadvisable for domain experts with a non-technical background. Here, special care should be given to assessing how a model could best be summarized to be intuitively understandable for the explanation target, since the additional cognitive load for understanding the explanation modality should be kept to a minimum. A second aspect is that the scale of the model should be considered. Ten blocks of model policy code can easily be evaluated, but a hundred blocks of code will be very difficult to understand and thoroughly inspect. Here, the graph-based explanation approach may be a useful addition.

Graph-based explanations can furthermore be very helpful in providing a quick and intuitive overview of model behavior. Holzinger et al. (2021) recommend the use of graph-based explanations for HITL systems, as they can be used to intuitively compare expert domain knowledge with learned model behavior. Song et al. (2019) show how graph-based explanations can be applied in recommender systems, a field where knowledge graphs are often used. With their presented system, the user is shown a meaningful path within that graph on how a recommendation was formed, which helps provide effective recommendations and good explanations. The use of graph-based explanations can, however, become overwhelming for the user if the model behavior or explained decision becomes too complex. Approaches such as PGExplainer (Vu & Thai, 2020) alleviate this by focusing the explanation graph on relevant parts of the decision graph.

5.3 Focus: Safety Evaluation

RL systems can be highly complex and non-transparent, making it difficult to understand their decision-making processes and identify potential safety issues. We emphasize the need to proactively evaluate the safety of the developed application to ensure that they are designed and deployed in a way that minimizes risks to the agent itself and others. The mentioned approaches help to evaluate the agent's behavior and test it for unexpected or undesirable actions.

However, while testing the model is an essential part of the training loop, it should not be the only component to ensure safe operation. An example of how safety can be ensured is presented by Xiong et al. (2020), who propose using shield-based defenses, where agents learn to stay within predefined, safe boundaries during training and application, and with that increase robustness.

In addition, approaches that estimate model uncertainty in different scenarios can be useful. Lutjens et al. (2018) present a collision avoidance policy to provide computationally tractable and parallelizable uncertainty estimations in navigation tasks. This can be used to ensure that the model is sufficiently confident in the test scenarios employed by the developers, but also able to discover blind spots. These blind spots can then be used to test how the model behaves when encountering them.

5.4 Discussion, Outlook, and Use-Case

In the agent evaluation phase, we identify a major challenge in the scalability of xAI approaches, which are often suited to smaller models but fail to provide explainability for large trained models and their learned policies. We also find that few approaches allow for identifying variations between different versions of a trained model, which we consider a central feature for this phase in order to conduct an exhaustive evaluation of the learned policy. Visual approaches such as saliency maps can be applied in this phase to evaluate the behavior of the model, but should be used with caution due to providing only superficial insights about the "where", but not the "how", and being susceptible to confirmation bias (Evans et al., 2022).

We highlighted that many explainability approaches that suffice in the development and learning phase need adaptation to the evaluation phase due to model size and complexity. Approaches such as code block summarization as provided by Verma et al. (2018) could be extended by focusing only on relevant parts of the explanations as illustrated by Vu and Thai (2020). Furthermore, expert readability must be ensured to allow domain experts to help test and assess whether the learned policies are sensible, while post hoc explanations such as visual (Atrey et al., 2019; Gupta et al., 2019) or textual (Fukuchi et al., 2017b; Hayes & Shah, 2017) explanations would increase overall explainability.

We recommend integrating various tools to help evaluate and scrutinize a model from many different points of view. We propose that detailed explanations of the decision of the model are essential for this process. Furthermore, the focus should be on evaluating model safety by providing explanations in such a way that they are understandable to the domain experts, allowing one not only to debug superficial model behavior, but also to check the learned routines for semantically sensible behavior, which is a task that also calls on the knowledge of the domain expert.

We suggest that a combination of the explainability methods described can help significantly by ensuring that the model has only learned the desired behavior. The use of graph-based explanations is recommended as complementary to the policy summarization approach since the summarization provides a broad overview of model policy, which can then be further inspected by querying specific explanations. The rapidity and intuitiveness of graph-based explanations allow inspection of the learned model policies together with domain experts, and together with safety measures and uncertainty estimation build a trustworthy model which informs the user of its limits. With this in mind, developers, domain experts and project owners should decide whether the model is ready for deployment or needs further development.

To refer back to the use-case of forest operation, we ensure that the agent is based on a reliable model capable of autonomously navigating in different environments and safely responding to user errors. To ensure reliability, the model should be tested in different environments. Furthermore, intentional user errors (such as trying to navigate to an unreachable location or entering malformed data) can be used to test the agent and see how it behaves in such edge cases (Xiong et al., 2020). Moreover, the model policy should be summarized and could be presented in a form of a policy graph, which will allow developers and domain experts to easily assess the behavior of the robot and see how the trained policy works (Vu & Thai, 2020). This could also be used to compare different model versions against each other to see what the agent has learned. Finally, the model should provide an understanding of the uncertainties and blind spots that it may encounter when navigating in the forest, for example, by using uncertainty estimates (Lutjens et al., 2018).

5.5 Agent Evaluation Summary

To summarize the agent evaluation phase as per Table 1, the explanations in this phase should scale to large trained models and be comparable to previous versions of the agent. They must be understandable by the domain experts to allow an exhaustive comparison of the learned model behavior. We also recommend focusing on evaluating the overall safety of the system and the robustness of its behavior. Humans are involved in understanding and evaluating these policies and the resulting behavior at the micro level (sensible individual decision) and at the macro level (cohesive overall behavior) and deciding if to deploy the model in the real world.

Useful explainability techniques include policy summarization, graph-based explanations, and approaches to interpretable decision-making, such as extracting decision trees and logic rules, involving domain experts and RL experts in a bidirectional fashion. In this phase, the benefit of utilizing causal models can lead to intrinsic explanation of the model. The model should also be evaluated with regard to safety aspects to ensure lasting user trust. In comparison to the other phases, this phase has many similarities in the xAI techniques used with the initial agent development phase, but requires in addition the understandability for the domain expert and the scaling of xAI methods to more complex policies.

In this section, considerations regarding the evaluation of HITL RL systems were discussed. In the next section, we discuss the final step of deploying the agent in a real-world context.

Phase	Human Involvement	Explanation Requirements	Explainability	Goals	Metrics
Deployment	•Deploy agent •Interact with agents •Define agents' real-world application goal and context	•Fast, clear, and concise to reduce cognitive load •Understandable by end-users •Non-intrusive to prevent detrimental effects on user performance	**Focus:** Building User Trust with intent and uncertainty communication, allowing error corrections **Explanations:** Saliency maps, dendrograms, bounding-boxes, textual explanations, visual and auditory indicators **Users:** End-users	•Develop and apply new approaches beyond image and driving-based explanations •Use simple and fast explanations •Implement cohesive error and uncertainty handling •Communicate agent intent via different modalities	**Fidelity** **Relevancy** **Cognitive Load** **Performance**

Table 7: Overview of the types of human involvement, the specific requirements for explanations, the focus and types of explainability approaches, the goals as well as metrics for the agent deployment phase.

6. Agent Deployment

The agent is then deployed in the real world and interacts with the end-user. It now has to provide an efficient interaction for different users, while also building trust by providing explanations for its behavior. A human typically needs to argue why the agent is safe and should be deployed in the real world from a vendor perspective. The developers will need to determine if the agent should continually learn, if its policy should be frozen, or if it should retrain if an environment change makes this necessary. The customer and the end-user then make the final deployment decision, determining the usage context and specific application of the agent in the field. Here, explainability can help users understand the final policy, improve trust, assess safety, and understand the stability of the policy.

In this section, we focus on approaches that facilitate an efficient interaction between the trained agent and its human user. This frequent and repeated interaction requires finding a balance between appropriately showing explanations and not hindering the task at hand — a trade-off also highlighted by Anderson and Bischof (2013), who state that while initially guides can be helpful, they can also be detrimental to long-term performance and learning. Additionally, challenges in terms of safety and generalization to overcome unforeseen situations come into play (see Roy et al. (2021) and Subsection 2.3). Table 7 gives an overview of key aspects of the agent deployment phase.

We propose that the use of HITL agents can lead to significant performance gains during the four phases of development, learning, evaluation, and deployment. In this section, we focus on approaches ensuring that those benefits actually reach the end-user, factoring in issues like mental overload and distrust.

When it comes to the agent deployment phase for the forest operations use-case, there are several challenges that must be considered. One significant challenge is communication. If the agent's actions and intentions are not effectively communicated to the operator or if the operator is unable to provide commands and feedback to the agent, it may not be able to function effectively in the field. Additionally, it is important to ensure that the agent can mostly function autonomously and ask for help from the human only when necessary, as too much dependence on the HITL may reduce the agent's autonomy and practicality for real-world use. Furthermore, safety should always be of highest priority, as the agent can only be successful if it operates without causing injury to personnel or otherwise damaging itself and its ability to operate.

6.1 Requirements

The xAI systems used by end-users can draw on the vast pool of research on HCI usability. Therefore, we adapt considerations and requirements from the well-known "golden rules of interface design" (Shneiderman et al., 2016). It must be taken into account that designing interactive RL systems has special design requirements (Arzate Cruz & Igarashi, 2020b), although the underlying rules of HCI design still apply.

To reduce the memory load, the explanations must be simple and quickly understandable. We aim to facilitate difficult tasks and should avoid complicating the human-robot interaction with overly complex explanations. Since operators are likely to work with rapidly changing perspectives and environments, explanations should be computed in real time to ensure they correspond to the current situation. An example of this is an autonomous car. If explanations are provided with a delay of several seconds, actions that could require intervention will already have occurred.

Regarding the guideline to allow experienced users to take shortcuts, agents should provide explanations on demand and should be deactivated if desired, as discussed in Anderson and Bischof (2013). This option is essential to prevent information fatigue and to allow natural and efficient human-robot collaboration.

Our third consideration is to simplify error handling and give the user the feeling of being in control. We propose that a HITL model should provide some means to show whether it is uncertain about a given situation or decision, for example, in the form of a warning light as seen in cars. Such a mechanism can give the user a notification that something is wrong or uncertain, and allow us to investigate what causes this. Along similar lines, we propose some kind of startup check sequence, again based on the warning light startup sequence of a car, where users can ensure that the system is in order and correctly understands the situational context.

In contrast to the other three phases of HITL RL deployment, we do not propose that the major requirements can be lifted at this phase. We rather suggest that this phase is the most demanding of the four enumerated, as it combines constraints on computational and cognitive capacities, requiring high fidelity, relevancy, performance, and low cognitive load (Milani et al., 2022) to ensure that the model performs well in real world situations and can handle a variety of inputs.

6.2 Approach: Explainability

Several researchers provide examples of what real-time explanations for different use-cases could look like. Rodriguez et al. (2021) provide feature-based explanations for COVID-19 case predictions, while Kulkarni and Gkountouna (2021) developed a classroom dashboard that gives an overview of student performance with dendrograms and text-based explanations. Another example is the utilization of dendrograms for estimating the remaining useful lifetime of industrial machinery by machine learning combined with domain knowledge (Serradilla et al., 2020). The majority of those real-time explainability systems are data/software-based, while for the area of explanations for robotic systems, there are much fewer examples. Most autonomous driving systems provide explanations in the form of bounding-boxes and labels for recognized objects, which is an appropriate option for explaining model perception and helps with localizing important image regions (Behl et al., 2017; Kashyap et al., 2020).

The next step is decision explanations. Here, Ben-Younes et al. (2022) present a method where object saliency is combined with a textual explanation of an action. For example, the observed traffic light is highlighted, in combination with the textual explanation of a "stop" action. Such an approach is already helpful and quick to evaluate by the end-user. Still, it could, for example, be even further refined when using known symbols and signs instead of text along with regional highlighting. Xu et al. (2023) provides an example of how textual explanations could be used to create textual explanations aligned with human judgement, increasing their understandability.

A major component for trusting an agent is the predictability of the agents' actions. Therefore, we suggest that xAI approaches used in the real-world focus on making the agents' decisions and planning transparent for the user by showing intended actions. Strictly, since actions do not have to be explained, just announced, this approach belongs to the area of interpretability, which constitutes a subset of explainability (Dragan, 2015). Requiring only interpretability and not explainability simplifies the requirements for such an indication, though, of course, HCI principles still have to be taken into account to avoid incurring too much mental load. Caltagirone et al. (2017), for example, shows a predicted trajectory for autonomous driving applications, which could easily be translated into other movement-based domains. An open challenge is how these predictions can be communicated in other contexts than autonomous cars and with other modalities. Here, items such as smartwatches, headphones, or simple visual indicators could provide familiar and flexible interfaces.

6.3 Focus: Building User Trust

Trust is essential for AI development because it enables effective deployment and adoption of AI systems. We established that without trust, people may be hesitant to use or rely on AI systems, which can limit their potential benefits (see Subsection 2.1). In the agent deployment phase, we therefore recommend to focus on building trust with the end-user and highlighted how different explainability approaches can be used for that.

Furthermore, we suggest that the use of "warning light" alerts could be beneficial, recognizing when the agent is uncertain about a decision and advising the user. This could, on the one hand, increase the general robustness of the agents' decisions and also foster human trust in the agents' decisions, since the user can now estimate better if the agent is certain about a decision.

Such a warning light could be based on uncertainty estimation and be activated when the uncertainty rises above a given threshold. Jain et al. (2021) give an example of how epistemic uncertainty can be estimated to a certain degree. The introduction of such an approach could help the user focus on the given task and interaction with the robot, while still being in control and able to intervene when required.

Such an intervention approach is demonstrated by Wu et al. (2021a). They allow the HITL operator to intervene when the agent makes erroneous decisions and, furthermore, allow the model to learn from those interventions.

The startup sequence approach could complement this concept of error handling. Liu et al. (2021) for example proposes an error detection framework, where the HITL operator is presented with a list of the most relevant and explainable features to detect unusual or nonsensical behavior. This list is evaluated during startup with a quick glance and provides considerable trust benefits.

6.4 Discussion, Outlook, and Use-Case

In the agent deployment phase, we find only a few suitable xAI approaches. The application of current approaches is most often hindered by the failure to speak the user's language, limited available computation, or too much complexity to be usable in a real-time context. Additionally, many HITL RL approaches fail to gain (and deserve) the user's trust, in addition to failing to communicate uncertainty when warranted.

We propose that the currently available approaches look beyond the use-cases of autonomous driving focused on visual aspects and consider other modalities. An example is the context of credit or policy computations, which require explaining textual facts. Additionally, modalities beyond graphic dashboards must be considered to ensure, such as auditory and tactic perspectives. Furthermore, visual perspectives should be explored in different form factors such as smartwatches, LED indicators, and image projections.

We emphasize the need for simple and fast explanations, as there are only a few such current approaches. We furthermore recommend considering explaining agent actions via different modalities, such as visual indicators, but also haptic or auditory signals, aspects which are largely unexplored as of now. Finally, we envision a suite of tools equipped with warning indicators that show when the agent encounters difficult situations, allowing the user to trust the agent when it is within its generalization capabilities, and inform the user if that is not the case. Ultimately, a start-up sequence with different checks would allow the user to ensure that the agent is properly initialized and trustworthy.

With regard to our use-case, we now similarly recommend relying on simple, robust, and nonintrusive indicators to communicate the agent state to the user. For example, the robot may use visual cues to show its operating intentions accompanied by a warning light or an audible alarm that communicates uncertainty (Jain et al., 2021). Additionally, a display that can be connected to the machine could be used to allow the user to view and understand the models' perception (Glanois et al., 2021). This will provide easy and quick explanations for the user, helping them identify and avoid obstacles along the way. Furthermore, its actions could be justified with textual explanations (Ben-Younes et al., 2022), which ideally also allow the user to provide feedback and correct mistakes (Wu et al., 2021a).

6.5 Agent Deployment Summary

To summarize, the agent deployment phase according to Table 1, the main requirement is that the explanations are understandable by the end-user. Additionally, they will be used in the field and therefore should not incur significant overhead, either with too much cognitive load or long and costly computation times. We highlight the need to focus on how to best handle errors and communicate model uncertainty to the end-user. Humans are involved as end-users in deploying the agent, defining the usage context, and specific agent tasks.

Explainability can be provided in the form of a combined explainability dashboard, communicating the agent's intents, actions, and ways of handling uncertainty and errors in a bidirectional fashion. Explanations can visualize important aspects of the agents' perception, or provide textual or other explanations for the agent's decision and intent. For visualization, currently used techniques are saliency maps, dendrograms, and bounding-boxes, which, in combination with textual explanations and visual or auditory indicators, could be used to communicate the agent's intent. This phase is the first entirely focused on the end-user and on generating user trust in the system, allowing one to accurately assess its strengths and weaknesses and enabling efficient real-world cooperation. This does not require providing insights into the model architecture as in the agent development and deployment phase, but rather highlighting relevant aspects of its perception and decision-making as in the agent learning phase.

We have now covered the four phases that lead to the deployment of HITL RL systems. In the next section, we open up the discussion on propositions for general research directions that emerged in this paper.

7. Research Directions

We refer to Subsection 2.3 to emphasize that we consider RL to be a challenging problem setting that could greatly benefit from HITL approaches. We highlight that there is no one-size-fits-all solution for explainability and that the requirements for suitable HITL xAI differ between each phase. We do not propose a strict separation of explainability techniques in different phases, but rather recognize the suitability of certain types of xAI for each phase depending on the nature of the human involvement, the aspect of the agent in focus (e.g., model architecture, its decision-making process, or its perception of the world), and the task to be tackled. The types of xAI approaches we recommend are informed by the explanation requirements listed in Table 1, the strengths and weaknesses of xAI methods as per Table 2, as well as the nature of human participation, the types of users, and the directionality of the interaction between the human and the agent. Explainability plays a central role in each phase of the agent deployment process, and discussed how it influences safety and trust considerations at each step.

Our broader vision is that the HITL RL approaches depicted could, in the future, enhance the productivity of a human-robot team. Khatib et al. (1999) stated that the HITL contributes experience, domain knowledge, and is able to ensure the correct execution of tasks. The robot, on the other hand, can increase human capabilities in terms of force, speed, and precision. Moreover, the robot should reduce human exposure to harmful and hazardous conditions. However, it is essential to allow both a machine and a human operator to react to the environment and human beings to correctly understand and interpret the machine's actions, as the underlying algorithms and their decisions must be understandable to a wide variety of different audiences with different goals (Heuillet et al., 2021). Therefore, we highlight the importance of explainability in HITL RL.

An exemplary instance of applying human feedback can be seen with RLHF in ChatGPT, where the integration of human feedback for fine-tuning the model plays a pivotal role in its success (Li et al., 2023). RLHF allows models to learn from human interactions and feedback, resulting in improved performance and more natural conversation generation, exhibiting impressive capabilities in generating coherent and contextually relevant responses. RLHF can furthermore help to enhance sample efficiency, relying on only moderately large labelled sample sets ($\tilde{5}$0,000 samples), thereby also possibly reducing training time (Lambert et al., 2022). This advantage is particularly valuable considering the extensive training durations needed for renowned RL models like AlphaGo and OpenAI Five, which require one month and ten months of training time, respectively (OpenAI et al., 2019; Silver et al., 2017). These instances highlight the ample scope for improvement in terms of both training time and sample efficiency.

However, it is essential to consider the limitations associated with incorporating human feedback. Biases present in training data and introduced through human feedback can lead to incorrect or biased responses, and ChatGPT models have been known to generate hallucinations and provide factually inaccurate information, posing significant drawbacks to their applicability (Peng et al., 2023).

Explainability techniques can alleviate these limitations by shedding light on the decision-making process, enabling users to identify and address issues. Incorporation of HITL and explainability mitigates the risks of biased or incorrect outputs by involving human oversight and intervention (Peng et al., 2023), underscoring the importance of explainability in comprehending the decision-making process of LLMs. Additionally, lack of transparency, coupled with biased training data, may result in the dissemination of inaccurate or harmful content by users (Ray, 2023).

Various explanation approaches, including classic attention mechanisms in NLP (Glanois et al., 2021), text-based explanations by LLMs (Xu et al., 2023; Zini & Awad, 2022), and symbolic representation (Saba, 2023), can enable users to understand the model's reasoning and identify potential flaws. In particular, the use of LLMs to generate explanations has garnered attention recently. For example, Zini and Awad (2022) employ the model itself to provide transparent explanations for its decision-making. Similarly, Xu et al. (2023) introduce an explainable metric that combines human instructions and the implicit knowledge of GPT, offering explanations aligned with human judgment for given outputs.

Another caveat to explainability is that most of the work on xAI is heavily biased by what researchers assumed to be good explanations for a given task or domain (Miller, 2019b), not taking into account the preferences and expertise of human end-users. Based on the feedback of the operators, models and approaches must adapt their language and modalities to be effective, which requires a more human-centered development (Puiutta & Veith, 2020). To ensure that explainability methods actually align with users' expectations, we call for a comprehensive set of guidelines and requirements for developing xAI systems.

But human-robot collaboration comes with challenges beyond explainability. When the agent is deployed, the trust requirement is essential; otherwise, the agent will not be used. According to De Santis et al. (2008), only trustworthy robots can work in this team. Humans tend to anthropomorphize robots (Damiano & Dumouchel, 2018), thus overestimating their cognitive capabilities. De Santis et al. (2008) argue that a user's mental model might result in a fake robot dependability, which exacerbates the problem of safety in human-robot collaboration. This collaboration relies on the predictability of the robot's actions, testability, explainability of the policies, as well as performance increases, which leads us to emphasize a research focus on trust and safety issues in HITL RL in the long term.

We finally propose future research on the use of explainable AI in safety. One area of interest is the development of safety-critical systems that incorporate explainable AI methods to ensure that the agent's decision-making process is transparent, interpretable, and can be easily understood by human decision-makers. This could include the use of techniques such as counterfactuals, which allows for the examination of the factors that led to a particular decision, and the use of natural language explanations to communicate the agent's decision-making process to human users. Another area of research could be on the development of methods for evaluating and testing the safety and robustness of AI agents, such as testing for robustness to distributional shift and adversarial attacks, and methods for detecting and mitigating bias in the agent's decision-making process. We also want to highlight that both trust and user experience play critical roles in determining the overall success of an AI application. While safety and trust are paramount in many applications, especially fields such as healthcare, education, and entertainment rely on user experience for their effectiveness and acceptance. Users must feel comfortable, confident, and satisfied with the AI system to fully engage and benefit from it (Holzinger, 2021; Holzinger et al., 2022a). Building trust involves ensuring transparency, explainability, and accountability in the decision-making process, enhancing user confidence. Simultaneously, providing a positive and seamless user experience through intuitive interfaces, personalized interactions, and effective problem-solving capabilities is essential for user satisfaction and adoption (Arzate Cruz & Igarashi, 2020b). Striking a balance by establishing both trust and an exceptional user experience is crucial in driving the success and widespread acceptance of AI applications in various domains. Further research is needed to explore how to effectively combine both goals in order to avoid the fact that trust-building measures hinder (or are sacrificed for the sake of) a fluid user experience. In this section, we have proposed different research directions. In the next section, we summarize the content and central insights of this paper and conclude with the vision we have for the future of HITL RL.

8. Conclusion and Future Outlook

In summary, we emphasize that RL is a fundamentally difficult problem setting and could benefit greatly from HITL interactions. A human expert can contribute to the conceptual understanding gained through many years of experience in the task at hand, thus significantly improving robustness and explainability (Holzinger, 2021). Numerous approaches have shown that RL benefits from human-centered approaches (Li et al., 2019; Mathewson & Pilarski, 2022).

We further argue that HITL RL in particular benefits greatly from xAI approaches. These are, after all, fundamentally human approaches, which, in turn, ensure successful interactions, acceptance, and trust, as well as conceptual knowledge about the agents' limitations (Heuillet et al., 2021; Milani et al., 2022).

We identify the following phases for deploying HITL RL solutions: (1) initial agent development, (2) agent learning, (3) agent evaluation, and (4) agent deployment. In our work, we discuss how xAI can support each of these phases and what are some considerations for a successful deployment. Thus, the HITL combination enables better human-robot collaboration and ultimately increases on-task productivity and efficiency.

In the deployment phase, interactive, thorough, and coherent model summaries can enable an agile and transparent workflow. During the agent's interactive learning, xAI approaches can enable more efficient training in the field through interactive replanning and imitation learning. In the evaluation phase, comprehensive explanations of model decisions can provide detailed insight into the trained model and lead to informed decisions about whether to proceed with the deployment or start another development cycle. In the deployment phase, simple and quick explanations of actions and different explanations for error handling could significantly increase user confidence in the agent and lead to more efficient collaboration.

Last but not least, we propose a vision of an interactive human-robot collaboration that enables new use-cases for RL applications and allows both humans and robots to realize their full potential and respective strengths. Such a collaboration requires strong interaction and trust between both parties, which can only be achieved through comprehensive explanation and a deep and intuitive understanding of the mental model generated by the agent.

Acknowledgments

Parts of this work have been funded by the Austrian Science Fund (FWF), Project: P-32554 "explainable Artificial Intelligence". Part of this work has taken place in the Intelligent Robot Learning (IRL) Lab at the University of Alberta, which is supported in part by research grants from the Alberta Machine Intelligence Institute (Amii); a Canada CIFAR AI Chair, Amii; Compute Canada; Huawei; Mitacs; and NSERC. We also thank Dr. Antonie Bodley for her editorial and proofreading assistance.

References

Abel, D., Salvatier, J., Stuhlmüller, A., & Evans, O. (2017). Agent-agnostic human-in-the-loop reinforcement learning. *arXiv preprint arXiv:1701.04079.*

Aiyappa, R., An, J., Kwak, H., & Ahn, Y.-Y. (2023, March). Can we trust the evaluation on ChatGPT? [arXiv:2303.12767 [cs]]. https://doi.org/10.48550/arXiv.2303.12767

Akanksha, E., Jyoti, Sharma, N., & Gulati, K. Review on Reinforcement Learning, Research Evolution and Scope of Application. In: In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*. 2021, April, 1416–1423. https://doi.org/10.1109/ICCMC51019.2021.9418283.

Akrour, R., Schoenauer, M., & Sebag, M. Preference-based policy learning (D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis, Eds.). In: *Machine learning and knowledge discovery in databases* (D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis, Eds.). Ed. by Gunopulos, D., Hofmann, T., Malerba, D., & Vazirgiannis, M. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, 12–27. https://doi.org/978-3-642-23780-5.

Akrour, R., Tateo, D., & Peters, J. Towards reinforcement learning of human readable policies. In: In *Workshop on deep continuous-discrete machine learning*. 2019.

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., & Kindermans, P.-J. (2019). Innvestigate neural networks! *Journal of Machine Learning Research*, *20*(93), 1–8.

Alonso, J. M., Ramos-Soto, A., Castiello, C., & Mencar, C. Explainable ai beer style classifier. In: In *Sicsa realx*. 2018. https://ceur-ws.org/Vol-2151/Paper_S1.pdf

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. Software Engineering for Machine Learning: A Case Study. en. In: In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. Montreal, QC, Canada: IEEE, 2019, May, 291–300. isbn: 978-1-72811-760-7. https://doi.org/10.1109/ICSE-SEIP.2019.00042.

Amir, O., Kamar, E., Kolobov, A., & Grosz, B. Interactive teaching strategies for agent training. In: In *In proceedings of ijcai 2016*. 2016.

Anderson, A., Dodge, J., Sadarangani, A., Juozapaitis, Z., Newman, E., Irvine, J., Chattopadhyay, S., Olson, M., Fern, A., & Burnett, M. (2020). Mental models of mere mortals with explanations of reinforcement learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *10*(2), 1–37. https://doi.org/10.1145/3366485

Anderson, F., & Bischof, W. F. Learning and performance with gesture guides. In: In *Proceedings of the sigchi conference on human factors in computing systems*. CHI '13. Paris, France: Association for Computing Machinery, 2013, 1109—1118. isbn: 9781450318990. https://doi.org/10.1145/2470654.2466143.

Andreas, J., Klein, D., & Levine, S. Modular multitask reinforcement learning with policy sketches. In: In *Proceedings of the 34th international conference on machine learning. 70*. Proceedings of Machine Learning Research. PMLR. 2017, 166–175.

Arakawa, R., Kobayashi, S., Unno, Y., Tsuboi, Y., & Maeda, S.-i. (2018). Dqn-tamer: Human-in-the-loop reinforcement learning with intractable feedback. *arXiv preprint arXiv:1810.11748*.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI [arXiv: 1910.10045]. *arXiv:1910.10045 [cs]*. Retrieved January 5, 2022, from http://arxiv.org/abs/1910.10045

Arzate Cruz, C., & Igarashi, T. A survey on interactive reinforcement learning: Design principles and open challenges. In: In *Proceedings of the 2020 acm designing interactive systems conference*. Association for Computing Machinery, 2020, 1195–1209. https://doi.org/10.1145/3357236.3395525.

Arzate Cruz, C., & Igarashi, T. A Survey on Interactive Reinforcement Learning: Design Principles and Open Challenges. In: In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. DIS '20. New York, NY, USA: Association for Computing Machinery, 2020, July, 1195–1209. isbn: 978-1-4503-6974-9. https://doi.org/10.1145/3357236.3395525.

Atrey, A., Clary, K., & Jensen, D. D. (2019). Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning. *arXiv preprint arXiv:1912.05743*. http://arxiv.org/abs/1912.05743

Audibert, J.-Y., Munos, R., & Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, *410*(19), 1876–1902. https://doi.org/10.1016/j.tcs.2009.01.016

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Baniecki, H., Kretowicz, W., Piatyszek, P., Wisniewski, J., & Biecek, P. (2022). Dalex: Responsible machine learning with interactive explainability and fairness in Python. *The Journal of Machine Learning Research*, *22*(1), 214:9759–214:9765. https://doi.org/10.5555/3546258.3546472

Bastani, O., Pu, Y., & Solar-Lezama, A. Verifiable reinforcement learning via policy extraction (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett, Eds.). In: *Advances in neural information processing systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett, Eds.). Ed. by Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., & Garnett, R. *31*. Curran Associates, Inc., 2018.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., … Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Behl, A., Hosseini Jafari, O., Karthik Mustikovela, S., Abu Alhaija, H., Rother, C., & Geiger, A. Bounding Boxes, Segmentations and Object Coordinates: How Important Is Recognition for 3D Scene Flow Estimation in Autonomous Driving Scenarios? In: In *Proceedings of the IEEE International Conference on Computer Vision*. 2017, 2574–2583. Retrieved July 24, 2023, from https://openaccess.thecvf.com/content_iccv_2017/html/Behl_Bounding_Boxes_Segmentations_ICCV_2017_paper.html

Bellemare, M. G., Candido, S., Castro, P. S., Gong, J., Machado, M. C., Moitra, S., Ponda, S. S., & Wang, Z. (2020). Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, *588*(7836), 77–82. https://doi.org/10.1038/s41586-020-2939-8

Ben-Younes, H., Zablocki, É., Pérez, P., & Cord, M. (2022). Driving behavior explanation with multi-level fusion. *Pattern Recognition*, *123*, 108421. https://doi.org/10.1016/j.patcog.2021.108421

Bruneau, D., Sasse, M. A., & McCarthy, J. The eyes never lie: The use of eye tracking data in hci research. In: In *Chi 2002: Booktitle=CHI 2002: Conference on Human Factors in Computing Systems conference on human factors in computing systems*. University College London, 2002.

Buchanan, B. G., & Shortliffe, E. H. (1984). Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project. *Artificial Intelligence*. https://doi.org/10.1016/0004-3702(85)90067-0

Buchelt, A., Adrowitzer, A., Kieseberg, P., Gollob, C., Nothdurft, A., Eresheim, S., Tschiatschek, S., Stampfer, K., & Holzinger, A. (2024). Exploring artificial intelligence for applications of drones in forest ecology and management. *Forest Ecology and Management*, *551*, 121530. https://doi.org/10.1016/j.foreco.2023.121530

Caltagirone, L., Bellone, M., Svensson, L., & Wahde, M. Lidar-based driving path generation using fully convolutional neural networks. In: In *2017 ieee 20th international conference on intelligent transportation systems (itsc)*. IEEE. 2017, 1–6. https://doi.org/10.1109/ITSC.2017.8317618.

Cederborg, T., Grover, I., Isbell, C. L., & Thomaz, A. L. Policy shaping with human teachers. In: In *Twenty-fourth international joint conference on artificial intelligence*. 2015, 3366–3372.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2021). A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, *6*(1), 25–45. https://doi.org/10.1049/cit2.12028

Chandrasekaran, B., Tanner, M. C., & Josephson, J. R. (1989). Explaining Control Strategies in Problem Solving [Publisher: IEEE Computer Society]. *IEEE Intelligent Systems*, *4*(01), 9–15, 19. https://doi.org/10.1109/64.21896

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. (2023, January). ChatGPT Goes to Law School. https://doi.org/10.2139/ssrn.4335905

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*.

Cui, Y., Zhang, Q., Allievi, A., Stone, P., Niekum, S., & Knox, W. B. (2020). The empathic framework for task learning from implicit human feedback. *arXiv preprint arXiv:2009.13649*.

Da Silva, F. L., Hernandez-Leal, P., Kartal, B., & Taylor, M. E. Uncertainty-aware action advising for deep reinforcement learning agents. In: In *Proceedings of the aaai conference on artificial intelligence*. *34*. 2020, 5792–5799. https://doi.org/10.1609/aaai.v34i04.6036.

Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in human–robot co-evolution. *Frontiers in psychology*, *9*, 468. https://doi.org/10.3389/fpsyg.2018.00468

Daniel, C., Neumann, G., Kroemer, O., & Peters, J. (2016). Hierarchical relative entropy policy search. *Journal of Machine Learning Research*, *17*, 1–50.

De Graaf, M. M., & Malle, B. F. How people explain action (and autonomous intelligent systems should too). In: In *2017 aaai fall symposium series*. 2017.

De Santis, A., Siciliano, B., De Luca, A., & Bicchi, A. (2008). An atlas of physical human–robot interaction. *Mechanism and Machine Theory*, *43*(3), 253–270. https://doi.org/10.1016/j.mechmachtheory.2007.03.003

Del Ser, J., Barredo-Arrieta, A., Díaz-Rodríguez, N., Herrera, F., Saranti, A., & Holzinger, A. (2024). On generating trustworthy counterfactual explanations. *Information Sciences*, *655*, 119898. https://doi.org/10.1016/j.ins.2023.119898

Doran, D., Schulz, S., & Besold, T. R. (2017, October). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives [arXiv:1710.00794 [cs]]. https://doi.org/10.48550/arXiv.1710.00794

Dragan, A. D. (2015, July). *Legible Robot Motion Planning* (Doctoral dissertation). Carnegie Mellon University. https://doi.org/10.1184/R1/6720419.v1

Eder, K., Harper, C., & Leonards, U. Towards the safety of human-in-the-loop robotics: Challenges and opportunities for safety assurance of robotic co-workers. In: In *The*

23rd ieee international symposium on robot and human interactive communication. IEEE. 2014, 660–665. https://doi.org/10.1109/ROMAN.2014.6926328.

European Parliament. (2020). European Parliament resolution of 20 October 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020IP0275

Evans, T., Retzlaff, C. O., Geißler, C., Kargl, M., Plass, M., Müller, H., Kiehl, T.-R., Zerbe, N., & Holzinger, A. (2022). The explainability paradox: Challenges for xai in digital pathology. *Future Generation Computer Systems*, *133*(8), 281–296. https://doi.org/10.1016/j.future.2022.03.009

Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S. (2018). Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070.*

Florensa, C., Duan, Y., & Abbeel, P. (2017). Stochastic neural networks for hierarchical reinforcement learning. *arXiv preprint arXiv:1704.03012.*

Fukuchi, Y., Osawa, M., Yamakawa, H., & Imai, M. Application of instruction-based behavior explanation to a reinforcement learning agent with changing policy (D. Liu, S. Xie, Y. Li, D. Zhao, & E.-S. M. El-Alfy, Eds.). In: *International conference on neural information processing* (D. Liu, S. Xie, Y. Li, D. Zhao, & E.-S. M. El-Alfy, Eds.). Ed. by Liu, D., Xie, S., Li, Y., Zhao, D., & El-Alfy, E.-S. M. Springer. 2017, 100–108. https://doi.org/10.1007/978-3-319-70087-8_11.

Fukuchi, Y., Osawa, M., Yamakawa, H., & Imai, M. Autonomous self-explanation of behavior for interactive reinforcement learning agents. In: In *Proceedings of the 5th international conference on human agent interaction.* New York, NY, USA: Association for Computing Machinery, 2017, 97–101. https://doi.org/10.1145/3125739.3125746.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Glanois, C., Weng, P., Zimmer, M., Li, D., Yang, T., Hao, J., & Liu, W. (2021). A survey on interpretable reinforcement learning. *arXiv preprint arXiv:2112.13112.*

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572.*

Griffith, S., Subramanian, K., Scholz, J., Isbell, C. L., & Thomaz, A. L. Policy shaping: Integrating human feedback with reinforcement learning (C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger, Eds.). In: *Advances in neural information processing systems* (C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger, Eds.). Ed. by Burges, C., Bottou, L., Welling, M., Ghahramani, Z., & Weinberger, K. *26.* 2013.

Guan, L., Verma, M., Guo, S., Zhang, R., & Kambhampati, S. (2020). Explanation augmented feedback in human-in-the-loop reinforcement learning. *arXiv preprint arXiv:2006.14804v3.*

Gunning, D., & Aha, D. W. (2019). Darpa's explainable artificial intelligence program. *AI Magazine*, *40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Guo, W., Wu, X., Khan, U., & Xing, X. Edge: Explaining deep reinforcement learning policies (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan, Eds.). In: *Advances in neural information processing systems* (M. Ranzato, A. Beygelzimer,

Y. Dauphin, P. Liang, & J. W. Vaughan, Eds.). Ed. by Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., & Vaughan, J. W. *34*. 2021, 12222–12236.

Guo, Z., Yao, C., Feng, Y., & Xu, Y. (2022). Survey of reinforcement learning based on human prior knowledge. *Journal of Uncertain Systems*, *15*(01), 2230001. https://doi.org/10.1142/S1752890922300011

Gupta, P., Puri, N., Verma, S., Kayastha, D., Deshmukh, S., Krishnamurthy, B., & Singh, S. (2019). Explain your move: Understanding agent actions using focused feature saliency. *arXiv preprint arXiv:1912.12191v2*.

Ha, D., & Schmidhuber, J. (2018). World models. *arXiv preprint arXiv:1803.10122*. https://doi.org/10.5281/ZENODO.1207631

Habibian, S., Jonnavittula, A., & Losey, D. P. (2021). Here's what i've learned: Asking questions that reveal reward learning. *arXiv preprint arXiv:2107.01995*.

Hasanbeig, M., Jeppu, N. Y., Abate, A., Melham, T., & Kroening, D. Deepsynth: Automata synthesis for automatic task segmentation in deep reinforcement learning. In: In *The thirty-fifth {aaai} conference on artificial intelligence,{aaai}*. *2*. 2021, 36.

Hayes, B., & Shah, J. A. Improving robot controller transparency through autonomous policy explanation. In: In *2017 12th acm/ieee international conference on human-robot interaction (hri*. IEEE. 2017, 303–312. https://doi.org/10.1145/2909824.3020233.

Hazan, E., Kakade, S., Singh, K., & Van Soest, A. Provably efficient maximum entropy exploration (K. Chaudhuri & R. Salakhutdinov, Eds.). In: *Proceedings of the 36th international conference on machine learning* (K. Chaudhuri & R. Salakhutdinov, Eds.). Ed. by Chaudhuri, K., & Salakhutdinov, R. *97*. Proceedings of Machine Learning Research. PMLR. 2019, 2681–2691.

Hein, D., Hentschel, A., Runkler, T., & Udluft, S. (2017). Particle swarm optimization for generating interpretable fuzzy reinforcement learning policies. *Engineering Applications of Artificial Intelligence*, *65*, 87–98. https://doi.org/10.1016/j.engappai.2017.07.005

Hein, D., Udluft, S., & Runkler, T. A. (2018). Interpretable policies for reinforcement learning by genetic programming. *Engineering Applications of Artificial Intelligence*, *76*, 158–169. https://doi.org/10.1016/j.engappai.2018.09.007

Hein, D., Udluft, S., & Runkler, T. A. Generating interpretable reinforcement learning policies using genetic programming. In: In *Proceedings of the genetic and evolutionary computation conference companion*. Association for Computing Machinery, 2019, 23–24. https://doi.org/10.1145/3319619.3326755.

Hermann, T., Hunt, A., & Neuhoff, J. G. (2011). *The sonification handbook*. Logos Verlag Berlin.

Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Osband, I., Dulac-Arnold, G., Agapiou, J., Leibo, J., & Gruslys, A. Deep q-learning from demonstrations. In: In *Proceedings of the aaai conference on artificial intelligence*. *32*. 2018. https://doi.org/10.1609/aaai.v32i1.11757.

Heuillet, A., Couthouis, F., & Díaz-Rodríguez, N. (2021). Explainability in deep reinforcement learning. *Knowledge-Based Systems*, *214*, 106685. https://doi.org/10.1016/j.knosys.2020.106685

Holzinger, A. The next frontier: Ai we can really trust (M. Kamp, Ed.). In: *Proceedings of the ecml pkdd 2021, ccis 1524* (M. Kamp, Ed.). Ed. by Kamp, M. Cham: Springer Nature, 2021, pp. 427–440. https://doi.org/10.1007/978-3-030-93736-2_33.

Holzinger, A., Carrington, A., & Mueller, H. (2020). Measuring the quality of explanations: The system causability scale (scs). comparing human and machine explanations. *KI - Künstliche Intelligenz (German Journal of Artificial intelligence), Special Issue on Interactive Machine Learning, Edited by Kristian Kersting, TU Darmstadt*, *34*(2), 193–198. https://doi.org/10.1007/s13218-020-00636-z

Holzinger, A., Dehmer, M., Emmert-Streib, F., Cucchiara, R., Augenstein, I., Del Ser, J., Samek, W., Jurisica, I., & Díaz-Rodríguez, N. (2022a). Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. *Information Fusion*, *79*(3), 263–278. https://doi.org/10.1016/j.inffus.2021.10.007

Holzinger, A., Malle, B., Saranti, A., & Pfeifer, B. (2021). Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai. *Information Fusion*, *71*(7), 28–37. https://doi.org/10.1016/j.inffus.2021.01.008

Holzinger, A., & Müller, H. (2021). Toward Human–AI Interfaces to Support Explainability and Causability in Medical AI [Conference Name: Computer]. *Computer*, *54*(10), 78–86. https://doi.org/10.1109/MC.2021.3092610

Holzinger, A., Saranti, A., Angerschmid, A., Retzlaff, C. O., Gronauer, A., Pejakovic, V., Medel, F., Krexner, T., Gollob, C., & Stampfer, K. (2022b). Digital transformation in smart farm and forest operations needs human-centered ai: Challenges and future directions. *Sensors*, *22*(8), 3043. https://doi.org/10.3390/s22083043

Holzinger, A., Saranti, A., Molnar, C., Biececk, P., & Samek, W. Explainable ai methods - a brief overview. In: In *Xxai - lecture notes in artificial intelligence lnai 13200*. Springer, 2022. https://doi.org/10.1007/978-3-031-04083-2_2.

Holzinger, A., Stampfer, K., Nothdurft, A., Gollob, C., & Kieseberg, P. (2022d). Challenges in artificial intelligence for smart forestry. *European Research Consortium for Informatics and Mathematics (ERCIM) News*, *130*(July), 40–41.

Hussein, A., Gaber, M. M., Elyan, E., & Jayne, C. (2017). Imitation Learning: A Survey of Learning Methods [Place: New York, NY, USA Publisher: Association for Computing Machinery]. *ACM Computing Surveys*, *50*(2). https://doi.org/10.1145/3054912

Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., & Levine, S. (2021). How to train your robot with deep reinforcement learning: Lessons we have learned. *The International Journal of Robotics Research*, *40*(4-5), 698–721. https://doi.org/10.1177/0278364920987859

Icarte, R. T., Klassen, T., Valenzano, R., & McIlraith, S. Using reward machines for high-level task specification and decomposition in reinforcement learning (J. Dy & A. Krause, Eds.). In: *Proceedings of the 35th international conference on machine learning* (J. Dy & A. Krause, Eds.). Ed. by Dy, J., & Krause, A. *80*. Proceedings of Machine Learning Research. PMLR. 2018, 2107–2116.

Icarte, R. T., Waldie, E., Klassen, T., Valenzano, R., Castro, M., & McIlraith, S. Learning reward machines for partially observable reinforcement learning (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett, Eds.). In: *Advances in neural information processing systems* (H. Wallach, H. Larochelle, A.

Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett, Eds.). Ed. by Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Garnett, R. *32*. 2019.

Jain, M., Lahlou, S., Nekoei, H., Butoi, V., Bertin, P., Rector-Brooks, J., Korablyov, M., & Bengio, Y. (2021). DEUP: direct epistemic uncertainty prediction. *arXiv preprint arXiv:2102.08501*.

Jiang, Y., Bharadwaj, S., Wu, B., Shah, R., Topcu, U., & Stone, P. Temporal-logic-based reward shaping for continuing reinforcement learning tasks. In: In *Proceedings of the 35th aaai conference on artificial intelligence*. 2021.

Jiang, Z., & Luo, S. Neural logic reinforcement learning (K. Chaudhuri & R. Salakhutdinov, Eds.). In: *Proceedings of the 36th international conference on machine learning* (K. Chaudhuri & R. Salakhutdinov, Eds.). Ed. by Chaudhuri, K., & Salakhutdinov, R. *97*. Proceedings of Machine Learning Research. PMLR. 2019, 3110–3119.

Karalus, J., & Lindner, F. (2021). Accelerating the convergence of human-in-the-loop reinforcement learning with counterfactual explanations. *arXiv preprint arXiv:2108.01358*.

Kartoun, U., Stern, H., & Edan, Y. (2010). A human-robot collaborative reinforcement learning algorithm. *Journal of Intelligent & Robotic Systems*, *60*(2), 217–239. https://doi.org/10.1007/s10846-010-9422-y

Kashyap, S., Karargyris, A., Wu, J., Gur, Y., Sharma, A., Wong, K. C. L., Moradi, M., & Syeda-Mahmood, T. Looking in the Right Place for Anomalies: Explainable Ai Through Automatic Location Learning [ISSN: 1945-8452]. In: In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. ISSN: 1945-8452. 2020, April, 1125–1129. https://doi.org/10.1109/ISBI45749.2020.9098370.

Keane, M. T., Kenny, E. M., Delaney, E., & Smyth, B. (2021, February). If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques [arXiv:2103.01035 [cs]]. https://doi.org/10.48550/arXiv.2103.01035

Khatib, O., Yokoi, K., Brock, O., Chang, K., & Casal, A. Robots in human environments. In: In *Proceedings of the first workshop on robot motion and control. romoco'99 (cat. no.99ex353)*. IEEE, 1999, 213–221. https://doi.org/10.1109/ROMOCO.1999.791078.

Knox, W. B., & Stone, P. Tamer: Training an agent manually via evaluative reinforcement. In: In *2008 7th ieee international conference on development and learning*. IEEE, 2008, 292–297. https://doi.org/10.1109/DEVLRN.2008.4640845.

Knox, W. B., & Stone, P. Combining manual feedback with subsequent mdp reward signals for reinforcement learning. In: In *9th international conference on autonomous agents and multiagent systems (aamas) 2010*. 2010, 5–12.

Knox, W. B., & Stone, P. Reinforcement learning from simultaneous human and mdp reward. In: In *11th international conference on autonomous agents and multiagent systems (aamas) 2012. 1004*. Valencia. 2012, 475–482.

Knox, W. B., Stone, P., & Breazeal, C. Training a robot via human feedback: A case study (G. Herrmann, M. J. Pearson, A. Lenz, P. Bremner, A. Spiers, & U. Leonards, Eds.). In: *International conference on social robotics* (G. Herrmann, M. J. Pearson, A. Lenz, P. Bremner, A. Spiers, & U. Leonards, Eds.). Ed. by Herrmann, G., Pearson,

M. J., Lenz, A., Bremner, P., Spiers, A., & Leonards, U. Springer. 2013, 460–470. https://doi.org/10.1007/978-3-319-02675-6_46.

Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, *32*(11), 1238–1274. https://doi.org/10.1177/0278364913495721

Koditschek, D. E. (2021). What is robotics? why do we need it and how can we get it? *Annual Review of Control, Robotics, and Autonomous Systems*, *4*(1), 1–33. https://doi.org/10.1146/annurev-control-080320-011601

Kulick, J., Toussaint, M., Lang, T., & Lopes, M. Active learning for teaching a robot grounded relational symbols. In: In *Proceedings of the twenty-third international joint conference on artificial intelligence.* IJCAI '13. Beijing, China: AAAI Press, 2013, 1451–1457.

Kulkarni, A., & Gkountouna, O. (2021). Demonstrating react: A real-time educational ai-powered classroom tool. *arXiv preprint arXiv:2108.07693.*

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people [Publisher: Cambridge University Press]. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

Lambert, N., Castricato, L., von Werra, L., & Havrilla, A. (2022). Illustrating Reinforcement Learning from Human Feedback (RLHF). *Hugging Face Blog.*

Le, H., Jiang, N., Agarwal, A., Dudik, M., Yue, Y., & Daumé III, H. Hierarchical imitation and reinforcement learning (J. Dy & A. Krause, Eds.). In: *Proceedings of the 35th international conference on machine learning* (J. Dy & A. Krause, Eds.). Ed. by Dy, J., & Krause, A. *80.* Proceedings of Machine Learning Research. PMLR, 2018, 2917–2926.

Lee, K., Smith, L., & Abbeel, P. (2021). Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091.*

Li, B., Fang, G., Yang, Y., Wang, Q., Ye, W., Zhao, W., & Zhang, S. (2023, April). Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness [arXiv:2304.11633 [cs]]. https://doi.org/10.48550/arXiv.2304.11633

Li, G., Gomez, R., Nakamura, K., & He, B. (2019). Human-centered reinforcement learning: A survey. *IEEE Transactions on Human-Machine Systems*, *49*(4), 337–349. https://doi.org/10.1109/THMS.2019.2912447

Li, Y. (2017). Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274.*

Liang, H., Yang, L., Cheng, H., Tu, W., & Xu, M. Human-in-the-loop reinforcement learning. In: In *2017 chinese automation congress (cac).* 2017, 4511–4518. https://doi.org/10.1109/CAC.2017.8243575.

Likmeta, A., Metelli, A. M., Tirinzoni, A., Giol, R., Restelli, M., & Romano, D. (2020). Combining reinforcement learning with rule-based controllers for transparent and general decision-making in autonomous driving. *Robotics and Autonomous Systems*, *131*, 103568. https://doi.org/10.1016/j.robot.2020.103568

Lin, J., Ma, Z., Gomez, R., Nakamura, K., He, B., & Li, G. (2020). A review on interactive reinforcement learning from human social feedback. *IEEE Access*, *8*, 120757–120765. https://doi.org/10.1109/ACCESS.2020.3006254

Liu, G., Schulte, O., Zhu, W., & Li, Q. Toward interpretable deep reinforcement learning with linear model u-trees. In: In *Joint european conference on machine learning and knowledge discovery in databases*. Springer. 2018, 414–429. https://doi.org/10.1007/978-3-030-10928-8_25.

Liu, H., & Abbeel, P. (2020). Unsupervised active pre-training for reinforcement learning [Paper was rejected due to lack of novelty]. *ICLR 2021*.

Liu, Y., Mishra, N., Sieb, M., Shentu, Y., Abbeel, P., & Chen, X. (2022, October). Autoregressive Uncertainty Modeling for 3D Bounding Box Prediction [arXiv:2210.07424 [cs]]. https://doi.org/10.48550/arXiv.2210.07424

Liu, Z., Guo, Y., & Mahmud, J. (2021). When and why does a model fail? a human-in-the-loop error detection framework for sentiment analysis. *arXiv preprint arXiv:2106.00954*.

Lutjens, B., Everett, M., & How, J. P. (2018). Safe reinforcement learning with model uncertainty estimates. *arXiv preprint arXiv:1810.08700*.

Lyu, D., Yang, F., Liu, B., & Gustafson, S. Sdrl: Interpretable and data-efficient deep reinforcement learning leveraging symbolic planning. In: In *Proceedings of the aaai conference on artificial intelligence*. *33*. 2019, 2970–2977. https://doi.org/10.1609/aaai.v33i01.33012970.

MacGlashan, J., Ho, M. K., Loftin, R., Peng, B., Wang, G., Roberts, D. L., Taylor, M. E., & Littman, M. L. Interactive learning from policy-dependent human feedback (D. Precup & Y. W. Teh, Eds.). In: *Proceedings of the 34th international conference on machine learning* (D. Precup & Y. W. Teh, Eds.). Ed. by Precup, D., & Teh, Y. W. *70*. Proceedings of Machine Learning Research. PMLR, 2017, 2285–2294.

Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2020a). Distal explanations for model-free explainable reinforcement learning. *arXiv preprint arXiv:2001.10284*.

Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. Explainable reinforcement learning through a causal lens. In: In *Proceedings of the aaai conference on artificial intelligence*. *34*. 2020, 2493–2500. https://doi.org/10.1609/aaai.v34i03.5631.

Mandel, T., Liu, Y.-E., Brunskill, E., & Popović, Z. Where to add actions in human-in-the-loop reinforcement learning. In: In *Proceedings of the thirty-first aaai conference on artificial intelligence*. AAAI'17. San Francisco, California, USA, 2017, 2322—2328.

Mao, H., Chen, S., Dimmery, D., Singh, S., Blaisdell, D., Tian, Y., Alizadeh, M., & Bakshy, E. (2020, August). Real-world Video Adaptation with Reinforcement Learning [arXiv:2008.12858 [cs]]. https://doi.org/10.48550/arXiv.2008.12858

Martínez, D., Alenya, G., & Torras, C. (2017). Relational reinforcement learning with guided demonstrations [Special Issue on AI and Robotics]. *Artificial Intelligence*, *247*, 295–312. https://doi.org/10.1016/j.artint.2015.02.006

Martínez, D., Alenya, G., Torras, C., Ribeiro, T., & Inoue, K. Learning relational dynamics of stochastic domains for planning. In: In *Proceedings of the international conference on automated planning and scheduling*. *26*. 2016, 235–243.

Mathewson, K. W., & Pilarski, P. M. (2022). A brief guide to designing and evaluating human-centered interactive machine learning. *arXiv preprint arXiv:2204.09622*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning [Publisher: ACM New York, NY, USA]. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Mikhaylov, M. N., & Lositskii, I. A. Control and navigation of forest robot. In: In *2018 25th Saint Petersburg International Conference on Integrated Navigation Systems (ICINS)*. 2018, May, 1–2. https://doi.org/10.23919/ICINS.2018.8405881.

Milani, S., Topin, N., Veloso, M., & Fang, F. (2022). A survey of explainable reinforcement learning. *arXiv preprint arXiv:2202.08434*.

Miller, T. (2019a). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Miller, T. (2019b). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Mowshowitz, A., Tominaga, A., & Hayashi, E. (2018). Robot Navigation in Forest Management. *Journal of Robotics and Mechatronics*, *30*(2), 223–230. https://doi.org/10.20965/jrm.2018.p0223

Nair, A., McGrew, B., Andrychowicz, M., Zaremba, W., & Abbeel, P. Overcoming exploration in reinforcement learning with demonstrations. In: In *2018 ieee international conference on robotics and automation (icra)*. IEEE. 2018, 6292–6299. https://doi.org/10.1109/ICRA.2018.8463162.

Nascimento, E. d. S., Ahmed, I., Oliveira, E., Palheta, M. P., Steinmacher, I., & Conte, T. Understanding Development Process of Machine Learning Systems: Challenges and Solutions [ISSN: 1949-3789]. In: In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. ISSN: 1949-3789. 2019, September, 1–6. https://doi.org/10.1109/ESEM.2019.8870157.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2022, May). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI [arXiv:2201.08164 [cs]]. Retrieved December 6, 2022, from http://arxiv.org/abs/2201.08164

Ng, A. Y., Harada, D., & Russell, S. J. Policy invariance under reward transformations: Theory and application to reward shaping. In: In *Proceedings of the sixteenth international conference on machine learning. 99*. ICML '99. Morgan Kaufmann Publishers Inc., 1999, 278–287.

Nguyen, H., & La, H. Review of deep reinforcement learning for robot manipulation. In: In *2019 third ieee international conference on robotic computing (irc)*. IEEE, 2019, 590–595. https://doi.org/10.1109/IRC.2019.00120.

OpenAI, Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., Józefowicz, R., Gray, S., Olsson, C., Pachocki, J., Petrov, M., Pinto, H. P. d. O., Raiman, J., Salimans, T., … Zhang, S. (2019, December). Dota 2 with Large Scale Deep Reinforcement Learning [arXiv:1912.06680 [cs, stat]]. Retrieved July 13, 2023, from http://arxiv.org/abs/1912.06680

Parisi, S., Rajeswaran, A., Purushwalkam, S., & Gupta, A. The unsurprising effectiveness of pre-trained vision models for control. In: In *International conference on machine learning, ICML 2022, 17-23 july 2022, baltimore, maryland, USA*. PMLR, 2022, 17359–17371. https://proceedings.mlr.press/v162/parisi22a.html

Pearl, J. (2009). *Causality: Models, reasoning, and inference (2nd edition)*. Cambridge University Press.

Pearl, J., et al. (2000). Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, *19*, 2.

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., & Gao, J. (2023, March). Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback [arXiv:2302.12813 [cs]]. https://doi.org/10.48550/arXiv.2302.12813

Peng, B., MacGlashan, J., Loftin, R., Littman, M. L., Roberts, D. L., & Taylor, M. E. A Need for Speed: Adapting Agent Action Speed to Improve Task Learning from Non-Expert Humans. In: In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. International Foundation for Autonomous Agents; Multiagent Systems, 2016, 957–965.

Penkov, S., & Ramamoorthy, S. Learning programmatically structured representations with perceptor gradients. In: In *Proceedings of the 7th international conference on learning representations*. OpenReview.net, 2019. https://openreview.net/forum?id=SJggZnRcFQ

Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. *Science advances*, *4*(7), eaap7885. https://doi.org/10.1126/sciadv.aap7885

Puiutta, E., & Veith, E. M. Explainable reinforcement learning: A survey. In: In *International cross-domain conference for machine learning and knowledge extraction*. Springer. 2020, 77–95. https://doi.org/doi.org/10.1007/978-3-030-57321-8_5.

Ragunath, P., Velmourougan, S, Davachelvan, P, Kayalvizhi, S, & Ravimohan, R. (2010). Evolving A New Model (SDLC Model-2010) For Software Development Life Cycle (SDLC). *International Journal of Computer Science and Network Security*, 8.

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Rodriguez, A., Tabassum, A., Cui, J., Xie, J., Ho, J., Agarwal, P., Adhikari, B., & Prakash, B. A. (2021). Deepcovid: An operational deep learning-driven framework for explainable real-time covid-19 forecasting. *medRxiv*, 2020–09. https://doi.org/10.1101/2020.09.28.20203109

Roy, N., Posner, I., Barfoot, T., Beaudoin, P., Bengio, Y., Bohg, J., Brock, O., Depatie, I., Fox, D., Koditschek, D., et al. (2021). From machine learning to robotics: Challenges and opportunities for embodied intelligence. *arXiv preprint arXiv:2110.15245*.

Saba, W. S. (2023, May). Towards Explainable and Language-Agnostic LLMs: Symbolic Reverse Engineering of Language at Scale [arXiv:2306.00017 [cs]]. https://doi.org/10.48550/arXiv.2306.00017

Saranti, A., Eckel, G., & Pirró, D. Quantum harmonic oscillator sonification (S. Ystad, M. Aramaki, R. Kronland-Martinet, & K. Jensen, Eds.). In: *Auditory display* (S. Ystad, M. Aramaki, R. Kronland-Martinet, & K. Jensen, Eds.). Ed. by Ystad, S., Aramaki, M., Kronland-Martinet, R., & Jensen, K. Springer, 2009, pp. 184–201. https://doi.org/10.1007/978-3-642-12439-6_10.

Saranti, A., Taraghi, B., Ebner, M., & Holzinger, A. Insights into learning competence through probabilistic graphical models (A. Holzinger, P. Kieseberg, A. M. Tjoa, &

E. Weippl, Eds.). In: *International cross-domain conference for machine learning and knowledge extraction* (A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl, Eds.). Ed. by Holzinger, A., Kieseberg, P., Tjoa, A. M., & Weippl, E. Springer. Springer, 2019, 250–271. https://doi.org/10.1007/978-3-030-29726-8_16.

Schneeberger, D., Stoeger, K., & Holzinger, A. The european legal framework for medical ai. In: In *International cross-domain conference for machine learning and knowledge extraction, springer lncs 12279*. Springer, 2020, pp. 209–226. https://doi.org/10.1007/978-3-030-57321-8_12.

Scurto, H., Kerrebroeck, B. V., Caramiaux, B., & Bevilacqua, F. (2021). Designing deep reinforcement learning for human parameter exploration. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *28*(1), 1–35. https://doi.org/10.1145/3414472

Serradilla, O., Zugasti, E., Cernuda, C., Aranburu, A., de Okariz, J. R., & Zurutuza, U. Interpreting Remaining Useful Life estimations combining Explainable Artificial Intelligence and domain knowledge in industrial machinery [ISSN: 1558-4739]. In: In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. ISSN: 1558-4739. 2020, July, 1–8. https://doi.org/10.1109/FUZZ48607.2020.9177537.

Shneiderman, B., Plaisant, C., Cohen, M. S., Jacobs, S., Elmqvist, N., & Diakopoulos, N. (2016). *Designing the user interface: Strategies for effective human-computer interaction*. Pearson.

Silva, A., Gombolay, M., Killian, T., Jimenez, I., & Son, S.-H. Optimization methods for interpretable differentiable decision trees applied to reinforcement learning (S. Chiappa & R. Calandra, Eds.). In: *Proceedings of the twenty third international conference on artificial intelligence and statistics* (S. Chiappa & R. Calandra, Eds.). Ed. by Chiappa, S., & Calandra, R. PMLR. 2020, 1855–1865.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge [Number: 7676 Publisher: Nature Publishing Group]. *Nature*, *550*(7676), 354–359. https://doi.org/10.1038/nature24270

Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D. S., & Findlater, L. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In: In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, April, 1–13. https://doi.org/10.1145/3313831.3376624.

Song, W., Duan, Z., Yang, Z., Zhu, H., Zhang, M., & Tang, J. (2019). Explainable knowledge graph-based recommendation via deep reinforcement learning. *arXiv preprint arXiv:1906.09506*.

Sreedharan, S., Kulkarni, A., & Kambhampati, S. (2022). Explainable Human–AI Interaction: A Planning Perspective [Publisher: Morgan & Claypool Publishers]. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *16*(1), 1–184. https://doi.org/10.2200/S01152ED1V01Y202111AIM050

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. F. Learning to summarize with human feedback (H. Larochelle, M.

Ranzato, R. Hadsell, M. F. Balcan, & H. Lin, Eds.). In: *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin, Eds.). Ed. by Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., & Lin, H. *33*. Curran Associates, Inc., 2020, 3008–3021. https://proceedings.neurips.cc/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf

Stoeger, K., Schneeberger, D., & Holzinger, A. (2021). Medical artificial intelligence: The european legal perspective. *Communications of the ACM*, *64*(11), 34–36. https://doi.org/10.1145/3458652

Sun, S.-H., Wu, T.-L., & Lim, J. J. Program guided agent. In: In *International conference on learning representations*. 2019.

Surmann, H., Jestel, C., Marchel, R., Musberg, F., Elhadj, H., & Ardani, M. (2020, May). Deep Reinforcement learning for real autonomous mobile robot navigation in indoor environments [arXiv:2005.13857 [cs]]. https://doi.org/10.48550/arXiv.2005.13857

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tabrez, A., & Hayes, B. Improving human-robot interaction through explainable reinforcement learning. In: In *2019 14th acm/ieee international conference on human-robot interaction (hri)*. IEEE. 2019, 751–753. https://doi.org/10.1109/HRI.2019.8673198.

Taylor, M. E., & Stone, P. (2009). Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, *10*(7), 1633–1685.

Taylor, M. E., Suay, H. B., & Chernova, S. Integrating reinforcement learning with human demonstrations of varying ability. In: In *The 10th international conference on autonomous agents and multiagent systems-volume 2*. AAMAS '11. Citeseer. Richland, SC: International Foundation for Autonomous Agents; Multiagent Systems, 2011, 617–624.

Thomaz, A. L., & Breazeal, C. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In: In *Aaai*. *6*. Boston, MA. 2006, 1000–1005.

Tickle, A., Andrews, R., Golea, M., & Diederich, J. (1998). The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Transactions on Neural Networks*, *9*(6), 1057–1068. https://doi.org/10.1109/72.728352

Tomar, M., Mishra, U. A., Zhang, A., & Taylor, M. E. (2021). Learning representations for pixel-based control: What matters and why? *arXiv preprint arXiv:2111.07775*.

Topin, N., Milani, S., Fang, F., & Veloso, M. Iterative bounding mdps: Learning interpretable policies via non-interpretable methods. In: In *Proceedings of the aaai conference on artificial intelligence*. *35*. 2021, 9923–9931.

Torrey, L., & Taylor, M. Teaching on a budget: Agents advising agents in reinforcement learning. In: In *Proceedings of the 2013 international conference on autonomous agents and multi-agent systems*. Richland, SC: International Foundation for Autonomous Agents; Multiagent Systems, 2013, 1053–1060. https://doi.org/10.5555/2484920.2485086.

Vecerik, M., Hester, T., Scholz, J., Wang, F., Pietquin, O., Piot, B., Heess, N., Rothörl, T., Lampe, T., & Riedmiller, M. (2017). Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817*.

Verma, A., Le, H., Yue, Y., & Chaudhuri, S. Imitation-projected programmatic reinforcement learning (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett, Eds.). In: *Advances in neural information processing systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett, Eds.). Ed. by Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Garnett, R. *32*. Curran Associates, Inc., 2019.

Verma, A., Murali, V., Singh, R., Kohli, P., & Chaudhuri, S. Programmatically interpretable reinforcement learning (J. Dy & A. Krause, Eds.). In: *Proceedings of the 35th international conference on machine learning* (J. Dy & A. Krause, Eds.). Ed. by Dy, J., & Krause, A. *80*. Proceedings of Machine Learning Research. PMLR, 2018, 5045–5054. https://proceedings.mlr.press/v80/verma18a.html

Vu, M. N., & Thai, M. T. (2020). Pgm-explainer: Probabilistic graphical model explanations for graph neural networks. *arXiv preprint arXiv:2010.05788*.

Wang, X., Lee, K., Hakhamaneshi, K., Abbeel, P., & Laskin, M. Skill preferences: Learning to extract and execute robotic skills from human feedback. In: In *Conference on robot learning*. PMLR. 2022, 1259–1268.

Wells, L., & Bednarz, T. (2021a). Explainable ai and reinforcement learning — a systematic review of current approaches and trends. *Frontiers in artificial intelligence*, *4*, 48. https://doi.org/10.3389/frai.2021.550030

Wells, L., & Bednarz, T. (2021b). Explainable AI and Reinforcement Learning—A Systematic Review of Current Approaches and Trends. *Frontiers in Artificial Intelligence*, *4*. Retrieved July 15, 2023, from https://www.frontiersin.org/articles/10.3389/frai.2021.550030

Wu, J., Huang, Z., Huang, C., Hu, Z., Hang, P., Xing, Y., & Lv, C. (2021a). Human-in-the-loop deep reinforcement learning with application to autonomous driving. *arXiv preprint arXiv:2104.07246*.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2021b). A survey of human-in-the-loop for machine learning. *arXiv preprint arXiv:2108.00941*.

Xin, D., Ma, L., Liu, J., Macke, S., Song, S., & Parameswaran, A. Accelerating human-in-the-loop machine learning: Challenges and opportunities. In: In *Proceedings of the second workshop on data management for end-to-end machine learning*. New York, NY, USA: Association for Computing Machinery, 2018, 1–4. https://doi.org/10.1145/3209889.3209897.

Xiong, Z., Eappen, J., Zhu, H., & Jagannathan, S. (2020). Robustness to adversarial attacks in learning-enabled controllers. *arXiv preprint arXiv:2006.06861*.

Xu, W., Wang, D., Pan, L., Song, Z., Freitag, M., Wang, W. Y., & Li, L. (2023, May). INSTRUCTSCORE: Towards Explainable Text Generation Evaluation with Automatic Feedback [arXiv:2305.14282 [cs]]. https://doi.org/10.48550/arXiv.2305.14282

Yang, S., Nachum, O., Du, Y., Wei, J., Abbeel, P., & Schuurmans, D. (2023, March). Foundation Models for Decision Making: Problems, Methods, and Opportunities [arXiv:2303.04129 [cs]]. https://doi.org/10.48550/arXiv.2303.04129

Yang, T., Hao, J., Meng, Z., Zhang, Z., Hu, Y., Chen, Y., Fan, C., Wang, W., Liu, W., Wang, Z., & Peng, J. Efficient deep reinforcement learning via adaptive policy transfer. In: In *Proceedings of the twenty-ninth international conference on international joint* *conferences on artificial intelligence*. 2021, 3094–3100. https://doi.org/10.24963/ijcai.2020/428.

Zagal, J. C., del Solar, J. R., & Vallejos, P. (2004). Back to reality: Crossing the reality gap in evolutionary robotics [IFAC/EURON Symposium on Intelligent Autonomous Vehicles, Lisbon, Portugal, 5-7 July 2004]. *IFAC Proceedings Volumes*, *37*(8), 834–839. https://doi.org/10.1016/S1474-6670(17)32084-0

Zanzotto, F. M. (2019). Viewpoint: Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, *64*, 243–252. https://doi.org/10.1613/jair.1.11345

Zhang, C., Yong, L., Chen, Y., Zhang, S., Ge, L., Wang, S., & Li, W. (2019a). A Rubber-Tapping Robot Forest Navigation and Information Collection System Based on 2D LiDAR and a Gyroscope [Number: 9 Publisher: Multidisciplinary Digital Publishing Institute]. *Sensors*, *19*(9), 2136. https://doi.org/10.3390/s19092136

Zhang, J., & Bareinboim, E. Can humans be out of the loop? In: In *First conference on causal learning and reasoning (clear 2022*. 2020. https://openreview.net/forum?id=P0f91v5fTK

Zhang, P., Hao, J., Wang, W., Tang, H., Ma, Y., Duan, Y., & Zheng, Y. Kogun: Accelerating deep reinforcement learning via integrating human suboptimal knowledge. In: In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*. 2021, 2291–2297.

Zhang, R., Torabi, F., Guan, L., Ballard, D. H., & Stone, P. Leveraging human guidance for deep reinforcement learning tasks. In: In *Twenty-eighth international joint conference on artificial intelligence (ijcai-19)*. International Joint Conferences on Artificial Intelligence Organization, 2019, 6339–6346. https://doi.org/10.24963/ijcai.2019/884.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P. S., & Sun, L. (2023, May). A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT [arXiv:2302.09419 [cs]]. https://doi.org/10.48550/arXiv.2302.09419

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, *10*(5), 593. https://doi.org/10.3390/electronics10050593

Zhu, G., Wang, J., Ren, Z., Lin, Z., & Zhang, C. Object-oriented dynamics learning through multi-level abstraction. In: In *Proceedings of the aaai conference on artificial intelligence*. *34*. 2020, 6989–6998. https://doi.org/10.1609/aaai.v34i04.6183.

Zini, J. E., & Awad, M. (2022). On the Explainability of Natural Language Processing Deep Models. *ACM Computing Surveys*, *55*(5), 103:1–103:31. https://doi.org/10.1145/3529755