



# Explainable reinforcement learning for broad-XAI: a conceptual framework and survey

Richard Dazeley<sup>1</sup> · Peter Vamplew<sup>2</sup> · Francisco Cruz<sup>3,4</sup>

Received: 16 March 2022 / Accepted: 21 February 2023 / Published online: 6 March 2023  
© The Author(s) 2023

## Abstract

*Broad-XAI* moves away from interpreting individual decisions based on a single datum and aims to provide integrated explanations from multiple machine learning algorithms into a coherent explanation of an agent's behaviour that is aligned to the communication needs of the explainee. Reinforcement Learning (RL) methods, we propose, provide a potential backbone for the cognitive model required for the development of Broad-XAI. RL represents a suite of approaches that have had increasing success in solving a range of sequential decision-making problems. However, these algorithms operate as black-box problem solvers, where they obfuscate their decision-making policy through a complex array of values and functions. EXplainable RL (XRL) aims to develop techniques to extract concepts from the agent's: perception of the environment; intrinsic/extrinsic motivations/beliefs; Q-values, goals and objectives. This paper aims to introduce the Causal XRL Framework (CXF), that unifies the current XRL research and uses RL as a backbone to the development of Broad-XAI. CXF is designed to incorporate many standard RL extensions and integrated with external ontologies and communication facilities so that the agent can answer questions that explain outcomes its decisions. This paper aims to: establish XRL as a distinct branch of XAI; introduce a conceptual framework for XRL; review existing approaches explaining agent behaviour; and identify opportunities for future research. Finally, this paper discusses how additional information can be extracted and ultimately integrated into models of communication, facilitating the development of Broad-XAI.

**Keywords** Reinforcement learning (RL) · Explainable Reinforcement learning (XRL) · Explainable artificial intelligence (XAI) · Broad XAI

## 1 Introduction

Successes, such as AlphaGo [1], autonomous vehicles [2] and playing Atari video games [3], saw the MIT Technology Review list Reinforcement Learning (RL) as one of the top ten technologies of 2017 [4]. However, while RL can often solve complex sequential decision-making problems, the algorithms currently operate as a black-box, where experts must analyse vast amounts of data and functions to determine why they make particular decisions. For example, during AlphaGo's second challenge against Lee Sedol (ranked 9-dan) AlphaGo's 37th turn surprised both commentators and Lee Sedol, which turned the course of the game in AlphaGo's favour [5]. David Silver, DeepMind researcher, reportedly had no insight into why

---

✉ Richard Dazeley  
richard.dazeley@deakin.edu.au

<sup>1</sup> School of Information Technology, Deakin University, Geelong, Australia

<sup>2</sup> Institute of Innovation, Science and Sustainability, Federation University, Ballarat, Australia

<sup>3</sup> School of Computer Science and Engineering, University of New South Wales (UNSW), Sydney, Australia

<sup>4</sup> Escuela de Ingeniería, Universidad Central de Chile, Santiago, Chile

AlphaGo made such a creative move until he had investigated the actual calculations made by the programme [6]. For these systems to go the next step and be used by everyday non-expert users that are not able to inspect an agent's internal representation of its policy, they must be able to provide explanations for their behaviour [7].

The term eXplainable Reinforcement Learning (XRL) has begun to emerge recently to cover research into explaining agent's decisions during temporally separated decision-making tasks. There have recently been a number of surveys [8–12] that have provided in-depth discussions of issues and abilities that reinforcement learning and embodied agents can provide. These papers pull together a number of papers that have explored the potential of explainable systems in interactive temporal agents.

In this paper, we aim to go beyond exploring the current work alone and instead put forward a *conceptual* framework that sets up a structure for providing Broad-XAI. The objective is to promote the research and development of systems that can explain behaviour from integrated systems built on a foundation of RL. Interactive temporal agents built on this framework would be able to explain decisions and outcomes that provide for the three key areas of human explanation [13]: contrastive explanation, attribution theory and explanation selection. This framework should be viewed in the same way Artificial General Intelligence (AGI) frameworks are sometimes suggested in the literature. It is not our intention to provide an implementation of the framework at this time. Extensive research is required to develop each of the components and this paper identifies possible *reach* targets to inspire researchers to pursue. This framework is tied to a psychological model of explanation that allows for user controlled and conversational levels of explanation [14]. In so doing this paper suggests that, if RL decisions can be explained using human models of explanation, then they can build more trust and social acceptance. In presenting this framework, this paper discusses plausible approaches to developing each component, as well as identify current work in each area.

This paper is structured with six further parts. The next section provides a background to XAI and argues how XRL presents a distinct domain to be pursued. Section 3 proposes the conceptual framework for XRL and discusses how this integrates with human models of explainability. Section 4 reviews the current approaches to the initial stage of the framework, while Sect. 5 identifies future research opportunities for the advanced stages of the framework. Section 6 discusses how the framework can be integrated into models of communication to better facilitate the development of *Broad-XAI*. Finally, Sect. 7 summarises the paper and its contributions.

## 2 Explainable artificial intelligence

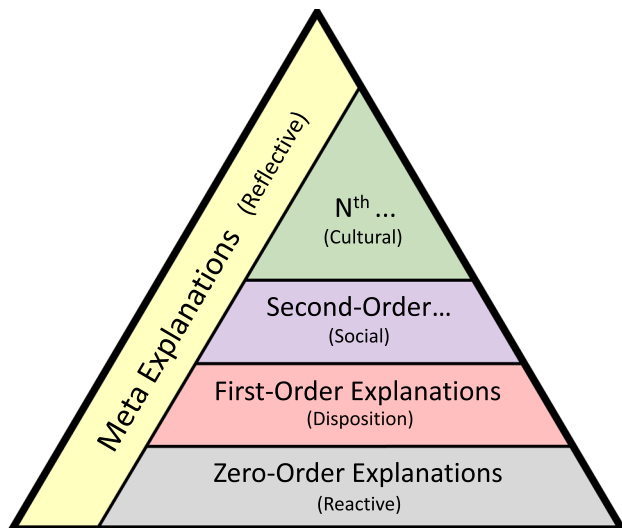
Harari (2016) [15] suggests that humans have always been a socially oriented species that has utilised their unique ability to articulate myths as integral to the social fabric. A myth is a story that aims to *explain* historical events or natural/social phenomena [16], which helps guide future behaviour. Explanation, therefore, is fundamental to human social interaction and trust, and therefore, key to the social acceptance of artificial intelligent agents. However, while human explanations have been studied by philosophers since Socrates, and over the last fifty years by psychologists and cognitive scientists, what it actually is, is still an open question [17]. As with the development of Artificial Intelligence, where research is hampered by people's poor understanding of intelligence, research into explainability is similarly restricted by a poor understanding of human explanation.

EXplainable Artificial Intelligence (XAI) is the general title given to the field of research aiming to generate explanations of AI systems that satisfies people's requirements in understanding and accepting the decisions made. There is a huge body of work providing a range of ways of interpreting black-box algorithms with mostly limited success. Various surveys have reviewed some of this work [14, 18–20]. Miller et al. (2017) [21], however, argue that the majority of researchers make XAI systems that are specific to their area of AI and that the primary aim behind these systems is to debug — rather than also considering the end-users' requirements. For instance, there are many explanation systems developed for image processing convolutional neural networks (CNN) that universally focus on identifying areas of an image or the parts of the network that contributed the most to a particular result [22]. Dazeley et al. (2021) [14] suggest that these 'narrow' XAI approaches that only focus on the individual task at hand do not provide the details required by users of the ever-increasing integrated intelligences currently appearing in the market. These emerging systems, such as autonomous cars, require *Broad-XAI* approaches that merge the decision-making of several integrated systems into a coherent explanation [14].

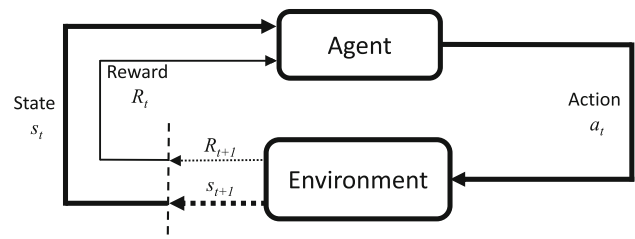
Dazeley et al. (2021) [14] suggest that most XAI research, often referred to as Interpretable Machine learning (IML), corresponds to zero-order (Reaction) explanations — where 'zero' refers to the absence of any explanation of the system's *intentionality*. Such approaches focus on explaining how the input just received was interpreted and how it affected the output. They argue that this foundational level is crucial to the development of *Broad-XAI*, but higher levels need to be developed for everyday users to accept decisions made by these systems.

Dazeley et al. (2021) [14] suggest a set of levels, reproduced in Fig. 1, that build up an explanation based on the level of intentionality utilised when making the decision. For instance, first-order (Disposition) details an agent’s intention such as its current goal or objective; second-order (Social) justifies its behaviour based on a prediction of other actors’ intentions; and N<sup>th</sup>-order (Cultural) provides an explanation of how it has modified its actions based on what it believes other actors’ expectations are of its behaviour. Interestingly, there have been several attempts to develop approaches for these higher levels. Dazeley et al’s (2021) [14] meta-survey identifies diverse subfields of XAI research, such as Explainable Agency [23], Goal-driven XAI [24, 25], Memory-aware XAI [26–28], Socially aware XAI [29–32], Cultural-aware XAI [33–36], Meta-explanation [37–39] and Utility-driven XAI [40–42].

These subfields, however, focus on developing approaches for explaining that individual component of an explanation, whereas, Broad-XAI requires an integrated approach across all levels that affect an agent’s decision. RL is a machine learning technique that potentially covers all these levels and offers a starting point for developing integrated explanations. However, currently, research in this space is relatively limited. Hence, the aim in this paper is to present a conceptual framework of how RL can be used to provide explanations across all levels of explainability and, thereby, provide a foundation for the development of Broad-XAI.



**Fig. 1** Levels of Explanation for XAI, as proposed by Dazeley et al. (2021) [14], indicates four levels of intentionality behind an agents behaviour should be explained. While Meta-explanations reflect on the processing process used in generating the explanation



**Fig. 2** Standard Reinforcement Learning model, as presented by [43], where an agent interacts with an environment through a series of discrete interactions

### 2.1 Explainable reinforcement learning: temporal explanations

Most introductory texts on machine learning (ML) identify three subfields: Supervised, Unsupervised and Reinforcement Learning (RL) methods. RL is often identified as separate and distinct to other ML because it utilises a fundamentally different approach to learning. In RL an agent learns by interacting with an environment using trial-and-error learning. While trialling a sequence of actions, it will occasionally receive feedback in the form of a positive or negative reward. This feedback will then be attributed to those actions taken providing a reinforcement to those behaviours to either increase or decrease their selection in the future.

This has similarities to supervised learning, in that an agent learns a mapping from input (state) to output (action), but unlike supervised approaches the reward can be distributed temporally, as it may not receive the reward until many actions have been taken. Formally, as defined by Sutton and Barto (2018) [43] and shown in Fig. 2, in the RL model, the agent and environment interact through a series of discrete time steps,  $t$ . Each time step the agent receives a representation of the environment’s current state,  $s_t \in \mathcal{S}$ , where  $\mathcal{S}$  is the set of all possible states. In a fully Markov Decision Process (MDP)<sup>1</sup> the agent uses only this state information to select an action,  $a_t \in \mathcal{A}(s_t)$ , where  $\mathcal{A}(s_t)$  represents the set of all possible actions in state,  $s_t$ . In the subsequent time step,  $t + 1$ , the agent receives a numerical reward,  $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$ , along with the new state,  $s_{t+1}$ .

Essentially, an RL agent learns a mapping from each state to an action, which expresses the agent’s behaviour. In model-based methods, the agent optimises the trajectory of its behaviour to minimise cost, while value-based methods maximise the reward explicitly through a value-

<sup>1</sup> RL is also often applied in environments that are not fully Markov. For instance, Semi-MDPs require information from previous states to determine outcomes taken in the future states. There is also significant work into Hidden, Partially Observable, Continuous-time, Multiobjective Markov processes, etc.

function. This mapping is commonly referred to as a *policy* and is denoted  $\pi$ , where  $\pi(s, a)$  represents an individual mapping from state,  $s$ , to action,  $a$ <sup>2</sup>.

There are numerous extensions to the basic RL approach that are frequently used in the literature. These are not the focus of this paper, but they do frequently present interesting information to the RL approach that allows for significantly improved explanations, and hence need some discussion. For instance, one difficulty with RL is as the state space grows so does the complexity of the agent's search for a solution. Hence, function approximation techniques such as neural networks are frequently utilised, giving rise to the field of Deep RL (DRL) [3, 44–48]. Secondly, while most RL assigns a single goal to an agent, such as pick up the rubbish in the room and put it in the bin, there is substantial work in multi-goal RL. In such systems, the agent not only must achieve its goal, but must also select the appropriate sub-goal to pursue [49–54]. Finally, a goal represents the agent's ultimate objective, and however, multiobjective RL (MORL) assumes that there can often be other conflicting objectives that also need to be balanced with the primary objective [55–57]. For instance, an agent may have the goal to tidy the room, and therefore, its primary objective is to do this efficiently; however, it may have a secondary objective to not damage any delicate things while accomplishing the primary objective [58, 59].

RL, from an explanation point of view, is of particular interest as it is often regarded as differing to that of supervised learning approaches [60]. Supervised learning techniques map each input to an output individually and so on their own the only explanation required is to identify the input components or the processes or processing stages that created the resulting classification. Each instance classified is regarded as a standalone instance and any *local* explanation is inherently based on this fact. Additionally, classifiers may provide *global* explanations that show how particular hyper-parameters or sets of training examples caused different outcomes for the classifier as a whole [61–63]. These causal explanations are important for system developers or designers to understanding issues like training bias [63], etc. However, supervised methods do not typically provide a mechanism for providing *local* causal explanations that explain individual decisions or behaviours of a system for non-technical end-users.

RL-based systems, however, have an implicit relationship between each instance. This is because the next state has only been visited because of the action taken in the previous state<sup>3</sup>. This creates a temporal dependency

between states, actions and subsequent states. These temporal dependencies, typically referred to as *transitions* and denoted as  $T(s_t, a, s_{t+1})$ , provide an implied causation for that individual transition. A sequence of transitions, either when reflecting on past transitions or a prediction of future transitions, can potentially provide causal networks that can be used to explain a number of details such as why actions were chosen according to some long-term goal [64]. So, while an individual transition is similar to an individual classification in supervised methods, the temporal sequence of transitions allows us to provide causal-based temporally extended explanations.

Additionally, supervised learning uses the learnt mapping to provide a classification or regression value with the aim of getting the 'right' answer, whereas RL aims to maximise a reward signal, which symbolises the goal or objective of the agent. Many approaches to RL have been developed to identify sub-goals [65–71], or may have alternative objectives that it can switch between [55, 59, 72]. These approaches mean the aim of the agent that guides its behaviour is not automatically going to be known to people affected by an agent operating in a human-agent shared environment. However, these approaches provide developers the ability to explain an agent's intentionality behind its behaviour, and thus, facilitate the provision of first-order explanations [14].

These fundamental differences between RL and supervised approaches to machine learning require us to think differently about explanation than simple interpretation — the common approach in machine learning. Of interest is the ability to provide introspective, causal and contrastive explanations within a single platform. RL is an approach that allows us to potentially develop broad-XAI systems. The aim of the remainder of this paper is to develop and present a conceptual framework for the development of Broad-XAI utilising RL as the basic backbone. Within the context of this framework this paper surveys current attempts to provide explanations (Sect. 4) and discuss potential approaches, not yet attempted, that will promote further research and development in to Broad-XAI (Sect. 5).

### 3 Conceptual framework for explainable reinforcement learning (XRL)

People interpret the world through explanations — either by attributing explanations to others' behaviour or by explaining their own behaviour to themselves or others.

<sup>2</sup> Methods not used for prediction rather than control may often learn a value per state rather than state/action pairs.

<sup>3</sup> Assuming a purely deterministic MDP. Clearly, in many situations, the new state only partially results from the previous state/action and

Footnote 3 continued

sometimes other factors could also contribute to the resulting state, such as wheel slippage and other actors in the environment.

When moving away from simply interpreting the decision-making process, as done by IML, developers need to consider how people tend to assign causes to behaviour [73–76]. Attribution theory, based on Heider's (1958) [73] seminal work, attempts to understand the process to which people attribute causal explanations for events [77]. Most events are usually categorised as either being dispositional or situational. Dispositional attribution assigns the cause to the internal disposition of the person such as their personality, motives or beliefs. In contrast, situational attribution assigns cause outside the person's control such as accidents or external events. More recently, researchers have shown that people instead tend to attribute behaviour towards the person's intention, goal, motive or disposition [78–81].

Drawing on knowledge structures such as scripts, plans, goals, and themes suggested by Schank and Abelson (2013) [82], Böhm and Pfister (2015) [83] extended ideas in attribution theory to develop a Casual Explanation Network (CEN), Fig. 3, based on the actual explanations provided by people. This model emphasises preconceptions about a causal relationship when providing explanations of behaviour. It builds on the idea that people will often want to explain others' behaviour not only in terms of why a particular behaviour occurred, but also what happened before to cause that behaviour and what is likely to happen in the future. Böhm and Pfister (2015) [83] propose a taxonomy that classifies both behaviour and explanations and is built around the intentionality that lead to the behaviour.

The CEN, Fig. 3, identifies seven categories that are relevant when considering the causal thinking about an actor's behaviour. This network is represented with a directed graph consisting of two sources and one sink. The end point, or sink, is the *outcome*, which is the final result of any behaviours. These outcomes are a result of either a person's intentional goal-directed *actions* or as a result of unintentional and uncontrolled *events*, such as tripping over. A person's *goal* represents the future states that the person is striving for, which can be caused by higher-order goals. The goal can also be caused by the *temporary state* or what can be thought of as their momentary disposition based on emotions, evaluations, mental states, motivational states, or bodily states (e.g. hunger, pain). This temporary state (momentary disposition) is in-turn affected by the person's personality traits or attitudes, which refers to as *disposition*, that are the result of long-term ingrained culturally based behaviours. The temporary state can also be caused by *stimulus attributes*, representing the features of the person or object that their behaviour was directed. For example, a person explaining the outcome of only passing an exam may state that it was too difficult (stimulus

attribute) causing them to be upset (temporary state) so they altered their goal to make sure they at least passed.

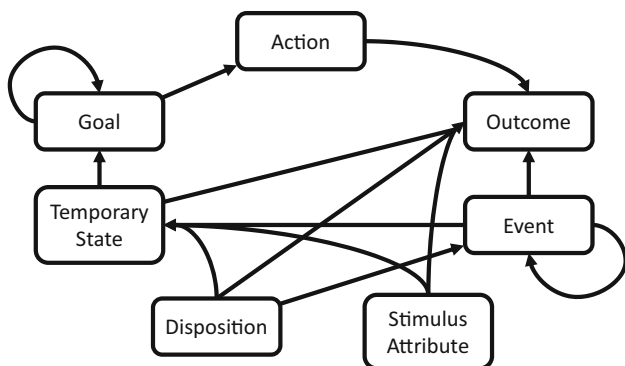
Figure 3 shows causal lines between these nodes indicating the causal directions provided in a person's explanations. These do not necessarily reflect the full and direct sequence of causes for outcomes, but they do represent the causal explanations that people *typically* use [83]. For instance, if a person trips (event) they may explain that they are clumsy (disposition) and that fearing injury (temporary state), attempt to arrest their fall (goal), by reaching out their hand (action) resulting in scratches on their hand. When asked what happened to their hand, they may provide the full causal path or simply explain the shortened causal path indicating they had tripped. This allows the explainee to fill in the gaps with their own general understanding of probable causes. Similar choices are provided for causal paths between other nodes. In this way, an explanation does not always require the full causal path from event, stimulus attribute or disposition through goal and action. This approach elegantly agrees with Lombrozo's (2007) [84] suggestion that an explanation should rely on as few causes (simple) as possible that covers the outcomes.

The CEN's focus on causal behaviour being the basis of explanations of intentionality aligns with Dazeley et al's (2021) [14] suggested levels of explanation for XAI. These levels were built upon Animal Ethology's idea of explaining behaviour through levels of intentionality [85]. Furthermore, the taxonomy of causal behaviour suggested in the CEN aligns well with the operating paradigm of an RL agent, and therefore, its application to XRL would be useful in providing structure to the generation of causal explanations from an RL agent. This paper proposes to merge these ideas from Dazeley et al (2021) [14] with the CEN, suggested by Böhm and Pfister (2015) [83], to form a framework, referred to as the Causal XRL Framework (CXF), and taxonomy for how XRL can generate causal explanations.

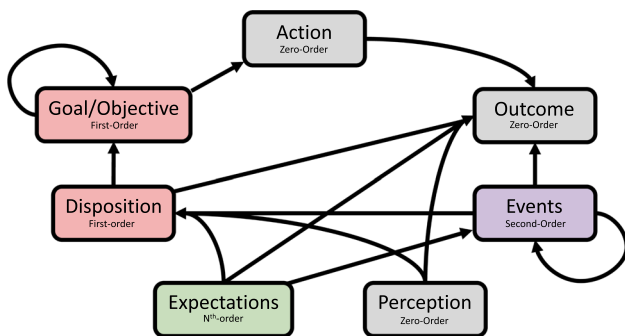
Figure 4 is an adaptation of Fig. 3 to facilitate the same causal pathways for explanation, but with categories aligned to RL and those indicated by Dazeley et al's (2021) [14] suggested levels of explanation. Included in this diagram is a mapping of XAI levels indicating the degree of intentionality that can be provided at each category of behaviour. This causal structure is intended to operate in a similar way to that suggested by Böhm and Pfister (2015) [83]. An *outcome*, represented by changes in the environment or the agent itself, is caused by either an intentional *action* by the agent or by an unintended or uncontrolled sequence of events. These *events* could be due to stochastic actions, such as wheel slippage or external actors.

In RL, an action is caused by an agent pursuing a particular *goal* or *objective*. This may be a single goal or a





**Fig. 3** A reproduction of the Causal Explanation Network (CEN) model for human lay causal explanations as suggested by Böhm and Pfister (2015) [83]. Each node represents a component used by people when explaining a person’s behaviour, while the arcs between nodes indicate the causal links between these concepts when people provide an explanation



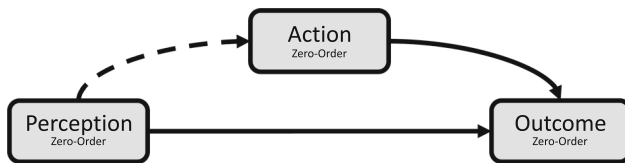
**Fig. 4** Conceptual Framework for Explainable Reinforcement Learning, referred to as the Causal XRL Framework (CXF) is based on the CEN given in Fig. 3. Each node, coloured and labelled to indicate the level of explanation (see Fig. 1), represents a process used by an agent when deciding on its behaviour. Each arc, joining nodes, represent the causal relationships that should be utilised when generating an explanation of an agent’s behaviour

hierarchy of goals, each of which can be cycled through to generate the explanation of its behaviour. A goal may be aligned to a single objective or to multiple objective that must be balanced [86]. The agent switches between these goals/objectives due to internal changes in priorities or progression in solving a larger goal. These internal changes are what Böhm and Pfister (2015) [83] describe as *temporary states*, however, this name could be confused with the perceived state of the RL agent, and hence, is avoided in CXF framework. Dazeley et al. (2021) [14], on the other hand, refer to this same concept as disposition — referring to an agent’s internal disposition. Therefore, to align with Dazeley et al. (2021) [14] a *disposition* in this sense is the same as a temporary states in the CEN model and represents temporary internal motivations such as a change in parameter, simulated emotion or safety threshold being passed.

Similarly, Böhm and Pfister (2015) [83] CEN model referred to disposition as an overarching set of long-term personality traits about how a person responds to situations. While there is no direct reference to responses to perceived cultural expectation, it is clear that disposition is the node where this would be best captured. As the temporal state node was renamed to be disposition, the disposition node has also been renamed to align with Dazeley et al. (2021) [14] notion of cultural expectations. Therefore, in this model an *expectation* refers to the ultimate aim of the agent to achieve what is expected of it. Dazeley et al. (2021) [14] suggest that expectations refer to a range of cultural conditions placed on an agent’s operation. In essence, expectations in this framework are the same as dispositions in the CEN. Finally, an agent’s current disposition, and therefore its goal/objective and ultimately the outcome, are caused by what is perceived by the agent. *Perception* is both the literal state, but also the result of any feature extraction, inference placed over what is perceived, or belief state in a Partially Observable MDP (POMDP).

Additionally, this framework is readily applicable to Multiagent Reinforcement Learning (MARL) domains [87]. For example, a MARL agent operating globally can simply use this framework directly with the understanding that the action space is a vector of actions that are similarly derived from its goals and higher-order influences. This aligns and extends current state-of-the-art explainable MARL [88]. However, a decentralised model presents a larger problem for the provision of explanations. A decentralised model requires agents to act independently of each other, and therefore, provide explanations of their behaviour independently. However, these agents require a sophisticated communication model between the agents to allow them to adjust their behaviour based on the other agents [87]. The CXF framework directly facilitates this MARL model. For example, when an agent changes its behaviour because of another agent’s communication or action then the CXF model allows us to incorporate this behaviour as a causal event that potentially alters the agent’s intrinsic disposition and goals. This approach allows for sophisticated models of explanation that incorporate teamwork directly into the causal framework. This decentralised model can be further extended to AI-Human collaborative teams [89, 90] where we require an explanation of an agent’s action in response to events caused by the human collaborators.

Ultimately, this framework is aimed at promoting future directions of research into explaining RL behaviour, but it also provides a lens for examining the current state of the art. The framework described in Fig. 4 is beyond the majority of current XRL research. Hence, this paper also presents a Simplified Conceptual Framework, which captures the majority of current XRL work. The simplified



**Fig. 5** Simplified Conceptual Framework for Explainable Reinforcement Learning, referred to as the Simplified-CXF, representing causal explanations in traditional RL. This framework includes a causal link between perception and action that is not included in the full model. This link replaces several behavioural components representing deeper causal paths that are assumed to not be modelled or explained in this Simplified-CXF. Note, this simplified model only includes zero-order explanations [14], indicated by grey only boxes, and therefore does not include any explanation of intentionality

framework, Fig. 5, shows the types of behaviours that can be explained when using a traditional approach to RL, as described in Sect. 2.1. As can be seen, this model only includes behaviours caused by what is perceived and the actions taken by the agent. It can also be observed that these behaviours all align with zero-order explanations [14] and, therefore, do not include any explanation of intentionality.

In this simplified model, it is assumed that an agent has a single preset goal and the objective is to maximise the reward in achieving that goal. This assumption is based on the standard RL framework [43] and constitutes the majority of RL research<sup>4</sup>. In such a situation the goal is often known to the user or can be observed over-time through observation of behaviour [91]. When utilising a predefined goal as its only objective its actions are directed towards achieving that goal. Its explanation of those actions is the target of that behaviour. Dazeley et al. (2021) [14] argue that there is no need to explain that the action is aimed at accomplish the goal in such a system. In situations where the goal itself is possible unknown to an end-user then the developers can incorporate details of the preset goal directly into any explanation of its behaviour. Equally, if the agent cannot alter its goal then no change to disposition or expectation can affect the goal being pursued. The aim of the *Goal* node in Fig. 4 is aimed at identifying how the current goal affected the action selected and why that is the current goal based on the agents current dispositions or expectation. Hence, the simplified model has no need to include causal explanations of these higher-level intentions. Similarly, the general RL model makes no attempt to model events outside its control making explanations of these also irrelevant. With the removal of goal/objectives, dispositions, expectations and events an RL agent cannot utilise those causal paths, and therefore, the simplified

<sup>4</sup> Subfields of research, such as multi-goal, hierarchical, and multi-objective, may have multiple goals or the goals may be unknown and provide possible approaches for use in the full framework discussed in Sect. 5.

framework must include a causal path from perception to action skipping those behaviours included in the full framework. Because this causal path is not part of the full framework, this is included only as a dotted line.

## 4 Simplified framework: reviewing explainable reinforcement learning

The term eXplainable Reinforcement learning (XRL) only appeared in research publications recently and is often published as Interpretable Machine Learning (IML). However, the aim of this paper is to show that the idea of explaining the behaviour of an RL agent, while sometimes related, is often quite distinct and separate to traditional IML; provides opportunities for deeper explanations to provide user trust and acceptance; already has a substantial body of research; and still has significant avenues for future work. This section represents the second substantive component of this paper, which reviews current work and discusses opportunities for future research. Rather than using a traditional taxonomy of approaches, it will review the literature in the light of the Simplified-CXF discussed in Sect. 3.

This review is not a systematic review and does not attempt to provide any form of meta-analysis of the topic [92]. The aim is to review and discuss the literature using a narrative approach [93] in the context of how it aligns with the CXF. Articles were identified through a combination of approaches, including: known references from papers from prior XAI surveys; searches using terms including “Explainable”, “Interpretable”, “Broad-XAI” combined with “Reinforcement Learning” or “Machine Learning”; and, the use of forward and backward snowballing from each previously identified paper.

The following subsections discuss each of the processes used by an agent to influence its choice of behaviour. This includes a discussion of the possible types of causal explanations that each process can contribute. Finally, for each type of causal explanation pathway, this paper discusses current approaches to explaining that causal link, as well as suggesting additional approaches that could be utilised. In discussing these points it will start with the nodes represented in the Simplified-CXF and discuss the opportunities available in the more advanced components of the full CXF in Sect. 5. The first Sect. 4.1 discusses explanations of what the agent has perceived and how that perception has affected the actions and outcomes. Section 4.2 discusses explanations based on why actions are selected and how they caused the resulting outcomes.

## 4.1 Explanation of perceptions

At its fundamental level, an RL algorithm is learning to do two things: receive information about the environment and use this to decide on an action to make in response. These two fundamental operations of an RL system represent the first two types of XRL discussed in this paper. The fundamental nature of these operations is also indicated in Fig. 5, with them being recognised as providing zero-order explanations [14]. That is, these operations represent a purely reactionary level of processing with zero intentionality. The first of these operations is to perceive the environment, which represents a significant amount of research in XRL. This section briefly overviews this class of XRL and discusses some example approaches. As identified by the simplified conceptual framework, Fig. 5, the perceptual stage, not only explains what it has perceived, but also how that perception resulted in the action taken and the outcome observed. Therefore, explanations of an agent's perception aim to detail one or more of the following:

1. *Perception*: what did the agent perceive as the current environment?
2. *Introspective*: how the perceived state contributed to the action being selected?
3. *Contrastive*: why didn't the perceived state cause some other action to be selected?
4. *Counterfactual*: what changes in perception would be required to cause an alternative action to be selected?
5. *Influenced*: how did the perceived state affect the outcome?

In the simple discrete RL situation each state or state/action pair can be represented with a mapping directly to the preferred action. However, in most realistic problems the state dimensionality for a direct mapping is too complex or continuous preventing a direct mapping. Instead, one of several approaches can be used such as function approximation [94], hierarchical representations [49, 52, 95], state aggregation [96, 97], relational methods [98] or options [99, 100].

To perform these approximations RL researchers generally utilise a range of traditional supervised learning approaches. For instance, the utilisation of Deep Neural Networks (DNN) is so common that a separate branch of research, known as Deep RL (DRL), has emerged, which now represent approximately 35% of RL papers published in 2022<sup>5</sup>. DRL methods utilise a DNN to map large state spaces to Q-values (regression) or directly to actions

(classification) [101]. In many cases, the supervised learning model used requires some level of adaptation to handle the temporal aspects of RL. For instance, DRL methods frequently utilise various forms of experience replay to improve convergence [102]. However, regardless of the learning process, the perception of the environment at any single moment is essentially the same process used in the supervised version.

### 4.1.1 XRL-perception with interpretable machine learning (IML)

Reliance on traditional supervised learning for function approximation means that XRL-Perception is essentially the process of interpreting the function used to model the state. Therefore, XRL-Perception is closely aligned with *Interpretable Machine Learning* (IML) methods [18, 19, 103–105]. IML is a well-established field with substantial work already having been done. The aim of this paper is not to resurvey IML work in detail — except to discuss how this work can be related to XRL specifically. According to Molnar (2019) [104], there are several approaches to interpreting machine learning models, as shown in Fig. 6. This suggests that IML typically produces one or more of the following types of interpretation:

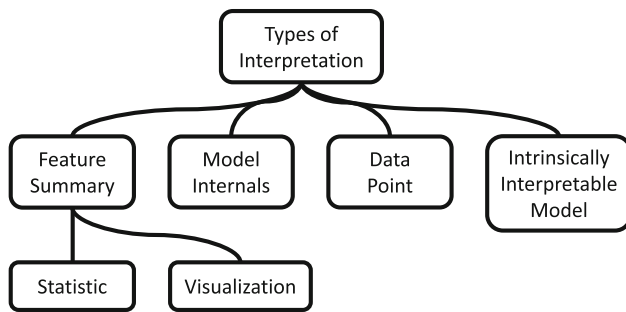
- a *feature summary*, using *statistics* or *visualisations*, showing the features and their relationships that were of most importance when reaching the outcome.
- a representation of the *internal model's* operation, such as the rules or neurons that fired, or pathways through the evaluation process that were followed.
- through the identification of similar or related *data points*, such as an image from the same class.
- through the construction of a secondary *intrinsically interpretable model*, which may then use one of the above methods to provide an interpretation.

Deep learning methods for IML tend to focus on visualisations of features found in the input (feature summaries) and neuron/layer activity (internal models); with some examples of using specifically designed neural networks for the provision of interpretations — see Gilpin et al (2018) [105] for a detailed discussion. Regardless of the approach used to interpret these models they can all be utilised to provide an interpretation of the perception of current state of an RL model.

<sup>5</sup> At time of writing, using Google Scholar, the number of papers with a title including the phrase “Deep Reinforcement Learning” was 4,690 and the number of “Reinforcement Learning” titled papers was 13,500. This is a crude estimation and almost certainly lower than the

Footnote 5 continued  
true percentage as many researchers assume DRL when discussing RL.





**Fig. 6** Types of Interpretation that can be generated from an Interpretable Machine Learning model. This is an original diagram derived from a taxonomy textually described by Molnar (2019) [104]

#### 4.1.2 Introspective XRL-perception

Due to the alignment of RL perception and traditional IML, there has been limited research specifically on perception in the context of XRL [8, 106]. However, there are two primary issues that make perception in XRL distinct from traditional IML. The first is that spatially similar states may often still require different control rules making generalisation difficult. This contrasts with most traditional supervised approaches that can afford local generalisation. Secondly, the perceptually similar states may in fact be significantly temporally separated [107]. This problem is attributed to why pooling layers are often absent in many DRL approaches as these are used to identify local generalisable patterns [3, 108, 109]. Therefore, research into XRL-perception has largely focused on providing explanations that can help developers to better understand the learning process, improve interpretation of policy, and for debugging/parameter tuning [107].

One approach used by both Mnih et al. (2015) [3] and Zahavy et al. (2016) [107] employed t-Distributed Stochastic Neighbour Embedding (t-SNE) on recorded neural activations [3, 107] to identify and visualise the similarity of states. Zahavy et al. (2016) [107] also displayed hand-crafted policy features over the low-dimensional t-SNE to better describe what each sub-manifold represents. A second approach used by Wang et al. (2015) [110] and Zahavy et al. (2016) [107] was to use Jacobian Saliency Maps [111] to better analyse how different features affect the network. Shi et al. (2020) [112] use a self-supervised interpretable network (SSINet) to locate causal features most used by an agent in its action selection.

These approaches are complex to understand and do not easily provide a reasonable explanation to a non-expert user. Saliency maps provide a reasonable level of understandability when using image-based state spaces but the Jacobian approach, borrowed from IML, can provide poor results as they have no relationship with the physical meaning of entities in the image. This problem can be

exacerbated in an RL agent due to the spatial similarity of states. Greydanus et al. (2017) [113] improved this approach by utilising the unique dual use of networks in the Asynchronous Advantage Actor-Critic (A3C) algorithm to separately represent both the critic's value assignment and the actor's actions. Greydanus et al. (2017) [113] then used these more accurate maps to visualise an agent's perception over time during the training process. This approach provides an important example for detecting features and identifying which features caused the agent to take a particular action, and separately, which ones were associated with particular outcomes, such as the highest rewards.

Verma et al. (2018) [114] presented a unique approach to performing introspection of an RL agent's perception by altering the RL framework itself. This work introduced the Programmatically Interpretable RL (PIRL) approach, where policies are initially learnt using DRL. This network is then used to direct a search over programmatic policies using Neurally Directed Program Synthesis (NDPS). During this repeated search process, a set of interesting perception patterns are maintained that minimise the distance between the DRL and NDPS (oracle) models. The completed oracle can then be inspected to identify causal links between feature vectors and actions taken and/or outputs.

#### 4.1.3 Results of XRL-perception

Perceiving the state is of particular interest to developers when validating a system's operation. Reassuring a non-expert user that the important features being used are also of importance, provided this is combined with the resulting effect of what was perceived. Simply informing the user of the action and the resulting change in the environment is implied in the previously discussed approaches as these are generally easily observed and do not require an explanation. However, the ability of a system to provide either contrastive or counterfactual explanations can be very valuable to a non-expert user and not easily observable from the agent's behaviour. Such explanation facilities aim to not only identify the features that led to the selected action, but also suggest why another action was not selected (contrastive), or what features needed to be observed to result in a different action/outcome being selected (counterfactual).

Conceptually counterfactual thinking and contrastive explanations are viewed as very different concepts. However, they are really just different views of the same predictive mechanism [115]. A counterfactual focuses on a prediction of what would happen under different initial circumstances, whereas a contrastive explanation details what change was needed to get a particular outcome. A counterfactual can be derived by providing a case study, or example fictitious state (sometimes referred to as a

‘distractor’ state or image), and observing the result. The real outcome, along with the fictitious outcome, can then be compared to provide the counterfactual explanation [116–118].

Contrastive explanations, however, are not as simple because there is no specific start state, but instead have a specific result that is of interest. The approaches in the last section cannot readily provide such explanations. For example, generating a contrastive explanation requires us to identify the features that are missing from the input space. One approach is to present multiple distractors and find the closest to the required conclusion [117]. This, however, is highly computational, and impossible when there are infinite possible distractors, such as continuous state problems. Recent methods for generating missing features, such as the Contrastive Explanation Method (CEM) [119, 120], have been proposed. These systems effectively identify absent pixels using a perturbation variable [119] or through Contrastive Layer-wise Relevance Propagation (CLRP) [120].

In RL, however, there is a temporal relationship between states and the outcomes that can be used to map a sequence of changes over time. This creates additional possibilities for providing contrastive explanations, and thereby, through extension counterfactual explanations as well. One approach is to identify those states that are critical to a human understanding the result of an agent, such as the Huang et al. (2017) [121] utilisation of DBSCAN [122] to identify such states. Alternative approaches [123, 124], especially in non-image-based inputs, use hand-crafted state features specifically identified for being semantically meaningful to humans. For instance, Hayes and Shah (2017) [123] use a vector of features to generate a list of predicates that could be searched to identify subsets of actions commonly associated. These approaches can explain why an action was selected in terms of features perceived by the agent. These approaches are not, however, readily usable in large state spaces or where hand-crafted features could not be provided. There is potential in using an agent’s perception to generate contrastive and counterfactual explanation, for instance, some works [125–128] have utilised grey box methods like decision trees and SHAP to identify the perception boundaries used by the actual decision-making neural network. However, most XRL focus has been around explaining the choice of actions and performing causal analysis of those choices, which are further discussed in Sect. 4.2.

## 4.2 Explanation of actions

While the provision of explanations of an agent’s perception is interesting, and in many cases required by the explainee, such explanations are not particularly unique to

RL. In fact, in reviewing the literature above, very little referred to RL specifically. As discussed in Sect. 2.1, the reason XRL is different to IML is due to the temporal nature of RL. This temporality is evident when considering how an action taken by an agent affects the outcome. These explanations are inherently *temporal explanations* as they detail a prediction of the expected future efficacy of an action. Temporal explanations detail relations between temporal constraints such as delays between causes and effects and were first investigated in temporal abductive reasoning [129] and recommendation systems [130, 131]. The CXF, Fig. 4, and simplified CXF, Fig. 5, indicate that an explanation can include why an agent took particular actions and how those actions caused particular results. Therefore, explanations of an agent’s actions aim to detail either:

1. *Introspective*: why was an action chosen?
2. *Contrastive*: why wasn’t another action chosen?
3. *Influenced*: how the action taken affected the outcome?
4. *Counterfactual*: what prior behaviour would have resulted in a particular alternative action being selected?

The first point addresses an explainee’s requirement to understand the choice of action and why the agent predicts it is a better choice than the alternatives. This can be presented in one of two forms, either: providing a visual representation of the path; or, by stating how the action leads to the eventual aim. For example, imagine an agent takes an action a user wants justified. It could present a map showing where the agent is currently located and the path it plans to follow, where the user can see the selected action follows this path. They could also be shown the best path should an alternative action be taken. This approach is of course regularly used in navigation recommendation systems such as Google Maps. Non-navigation in discrete tasks can also use this approach by representing the MDP as a graph using nodes and arcs to represent concepts the explainee will understand. An alternative approach is to state that the agent has selected a particular action because it has a measurably better result of a desirable quality as defined by the reward function, such as a higher chance of success, reduced cost, safer, and smoother. In either case the agent is being asked to make a prediction about both its future behaviour and how it expects the environment to respond.

### 4.2.1 Model-based XRL-behaviour

Early research into explaining why an action is preferred when accomplishing a particular task can be traced back to some of the earliest work in explaining the reasoning of expert systems [132–135]. An expert system generates a

conclusion through a series of inferences. These inferences represent a sequence of reasoning steps that can be considered actions during a problem-solving process. Explaining these involved providing either a rule trace of the inferences/actions taken or a trace of key, previously identified, decision points. These early ideas were later extended in domains such as Bayesian Networks (BN) [136], where explanations were generated from the relations between variables [137, 138] or through visual representations of relations between nodes [139]. Decision Networks or influence diagrams further extended BNs through the incorporation of utility nodes. These models help the decision process by selecting the path with the maximum utility, where explanations have been generated by reducing the optimal decision table [140].

An MDP, as used in RL, can be considered to be a *dynamic* decision network [141, 142]. Similar approaches have been applied in deterministic or decision theoretic planning [143] because these have a model of the environment that they can use to trace the entire decision path followed. XAI-Planning (XAIP) approaches are well placed to provide explanations for planning tasks with MDPs. Fox (2017) [144] provides a roadmap for the development of XAIP. These methods have a model of the environment in which they operate and can use this directly in their explanations to provide greater transparency. These approaches allow a more direct utilisation of the historical BN and DN methods. For instance, Krarup et al. (2019) [145] use waypoints for explanation, where this use of an execution-trace is a similar approach to that of rule traces and tracing nodes through a BN or DN. Similar approaches of generating explanations from actions using a model can be seen in other recent research [146–158]. Fox et al. (2017) [144] identify several questions that XAIP can answer. Ignoring questions regarding if and when to replan, which are specific to XAIP, these questions align with the previously mentioned aims for explaining actions. While planning approaches are not the focus of this paper, Chakraborti et al. (2020) [159] provide an extensive and recent survey of XAIP identifying the recent growth.

#### 4.2.2 Introspective XRL-behaviour

A direct adaptation of the BN and DN approaches is not as evident in value-based RL. Cruz et al. (2019) [160], Hayes and Shah (2017) [123], and Lee (2019) [161] could be considered as attempts to do this by essentially developing a model of the environment during exploration. The models built can then be used to generate an explanation, such as a prediction of the likelihood of reaching a goal, and how long until it was reached, from each state/action pair. Hayes and Shah (2017) [123] learn its model entirely separately from the agent, while Cruz et al. (2019) [160]

build the model internally. The approaches are inherently still RL as the model is not used for planning purposes and the agent still learns entirely from experience. However, these approaches of building a model of the environment presents allow an RL agent to present a similar level and range of transparency exhibited by the model-based approaches.

These learnt-model based approaches can also be used to provide users with an overview of the model through Policy Summarization or similar approaches [91, 162–167]. These global explanation approaches learn key state/action pairs that globally characterise the agent's behaviour. Using Inverse RL techniques, a policy can be inferred, and a summary formed from multiple examples of agent behaviour. The intuition is that policy summaries, like waypoints, can help people generalise and anticipate agent behaviour [162]. Another approach is to abstract away from low-level decision and provide explanations from this higher level. Beyret et al. (2019) [168] used Hierarchical RL to perform these layered abstractions and recognised their applicability to providing explanations and Acharya et al. (2020) [169] used a decision tree classifier to learn which state features were most likely to predict particular behaviours.

Ultimately, without a model, value-based approaches are hampered in their ability to explain an action in terms of the eventual aim. While people may think their aim is to achieve a goal, it is in fact only to maximise the long-term average reward. Schroeter et al (2022) [170] and Cruz et al. (2021) [64] extended [160] to provide the same explanations without requiring the memory overhead of learning a model, thereby providing the ability to provide these explanations in larger environments, including those requiring deep learning-based function approximation. To do this Cruz et al. (2021) [64] proposed two approaches: learning-based and introspection-based. The first approach was to directly learn a probability value  $P$  during training, while the second approach, referred to as introspection-based, was to infer the value directly from the agent's  $Q$ -value using a numerical transformation. These approaches allow an agent to explain why one action is preferred over another in terms of outcomes in a similar way as XAIP approaches.

What is interesting about these approaches is that rather than learning a model they use introspection of available information to provide explanations. *Introspection* is the utilisation of internal data for explanation as opposed to external frameworks that explain through observation. This introspective approach has also been utilised by Sequeira and Gervasio (2020) [166], which actively builds a database of historical interactions, allowing for simple information like, observations, actions and transitions; along with inferred probabilities such as the prediction error.

While this work as presented is not built to specifically answer questions, it does provide details that can provide additional analytics to the user and could easily utilise these statistics to provide such answers. This work has since been extended to provide short video highlights of key interactions [166].

### 4.2.3 Results of XRL-behaviour

When providing an explanation using the above techniques the system can simply state the reason the action selected is a good choice for achieving its goal. This, however, will often result in a relatively meaningless explanation that it chose the best, fastest, cheapest, etc., depending on the choice of reward. Instead, as discussed in Sect. 4.1.3, an explanation aiming to improve trust and acceptance would ideally be presented in contrast to an alternative action. These *contrastive explanations* are presented as *fact* and *foil* [171, 172], where the same fact, the action selected, can have multiple foils, any one of the actions not selected. Explaining contrastive and counterfactual of XRL-Behaviour involves comparing outcomes from alternative transitions paths through the MDP.

The most common approach to providing these explanations is to develop a model of the agent's behaviour using a separate observer that learns the agent's behaviour. There have been several generic explanation facilities that can perform this task, such as Pocius et al. (2019) [173], which extends Local Interpretable Model-Agnostic Explanations (LIME) [174] and can provide contrastive explanations of any type of agents' behaviour — not solely an RL agent. These generic explanation facilities can predict behaviour, but do not explain the agent's internal reasoning for its behaviour.

Extending Hayes and Shah (2017) [123], van der Waa et al. (2018) [175] provide contrastive explanations based on the result of transitions. The approach uses a provided model of the transition network but acknowledges this can be learnt through the observation of behaviour, by translating state features and actions to a predefined domain-specific ontology. The system then compares a user-selected foil to the taken actions to provide explanations on outcome differences. Cahmore et al. (2019) [176] provide a generic planning wrapper that builds on Fox et al's (2017) [144] roadmap for XAIP, to provide these contrastive explanations for known MDPs as a service. Rather than an *a priori* model Madumal et al. (2019) [177] used a learnt model to extensively study the generation of both contrastive and counterfactual explanations for explaining recommendations in the game of Starcraft II [178]. Madumal et al. (2019) [177] learn a Structural Causal Model (SCM) during training and analyse this model to understand how states led to different outcomes.

To investigate the ability to provide a value-based approach, Cruz et al. (2021) [64] illustrate that contrastive explanations on the likely success or failure of actions and the time to a result can be provided by an agent using the introspection-based approach to transforming the Q-values directly. Khan et al. (2009) [179] developed an approach to generate explanations for why a recommendation has been provided to a user, called a Minimal Sufficient Explanation (MSE). In this approach, a recommendation equates to an action and the approach tries to explain why that action is regarded as optimal. It takes one step beyond simply saying the action selected has the highest Q-value and thus is the optimal action, and instead provides reasons according to templated justifications about frequency of expected future rewards.

Two possible approaches to providing contrastive explanations are through the utilisation of either reward decomposition [180] or multi-objective Reinforcement Learning (MORL) [55, 58, 181]. Reward Decomposition separates each of the different rewards into semantically meaningful reward types allowing actions to be explained in terms of trade-offs between the separate rewards [180].

One avenue to providing contrastive explanation that has only recently been attempted is through the utilisation of multiobjective RL (MORL) [55, 58, 181]. MORL approaches maintain a vector of Q-values for each reward and at any given time there may be several Pareto-optimal policies offering different trade-offs between the objectives. Such approaches, such as reward decomposition, allow an agent to compare the known results of these policies that aligned with different actions. Sukkerd et al. (2018) [182] and its extension Sukkerd et al. (2020) [183] along with work by Juozapaitis et al. (2019) [180] are the first papers to directly pursue this approach to contrastive explanation. This model-based approach generates quality-attribute-based contrastive explanations to compare actions against alternative objectives. In value-based RL there is also one known attempt to use multiple objectives using reward decomposition<sup>6</sup> [184]. This approach is performing RL in an Adaptive-Based Programming formalism that allows annotations of decision points with ontological information for explanation. Currently, there is significant opportunity to pursue explainable MORL approaches for contrastive and counterfactual explanations.

The above approaches assume there is only one foil, alternative action, or that the user knows which foil they want the agent to compare with the selected action. However, this can be tedious, difficult or sometimes impossible

<sup>6</sup> The authors do not identify their work as MORL, and strictly speaking decomposed rewards are not necessarily conflicting (a generally accepted component of an MORL problem), but the approach uses the same principles that would underpin an explainable MORL approach.



for the user to provide. For instance, in an autonomous car, it is not practical to go through all alternative angles a steering wheel could have been turned to observe alternative results. Therefore, deriving the foil from the context is part of the explanation facility's task. However, apart from some attempts in IML [115, 175, 185, 186], this has not been widely discussed in the context of RL. Erwig et al. (2020) [187], while not the focus of their work in dynamic programming, did find that the context for contrastive explanations could be anticipated by identifying principal and minor categories and using these to anticipate user questions through value decomposition. As yet, foil prediction does not appear to have been transferred to value-based RL.

## 5 Full framework: opportunities for explainable reinforcement learning

Explaining perception, action and the causal outcomes of each, discussed above, represent the majority of current XRL research. These explanation facilities are important but focus primarily on providing debugging style explanations for developers [13, 21]. Dazeley et al. (2021) [14] argued that this represents only a zero-order, or reactionary level, of explanation and does not provide the broad-XAI required to develop user trust and acceptance. While there is still plenty of scope for interesting advances in the above simplified-CXF, this paper suggests there are significant possibilities for higher-level explanations built on an RL foundation. This section discusses each of the remaining components of the full framework and how existing extensions to RL can be utilised to provide Broad-XAI facilities in XRL.

### 5.1 Explanation of goals

Explaining an agent's goal and how it caused the selected action has been recognised as a potential future direction of research for XRL [14]. *Goal-driven explanation*, also referred to as *eXplainable Goal-Driven AI (XGDAI)*, is an emerging area of importance in the XAI literature with recent papers surveying the concept [14, 24, 25, 188]. This recent work shows a growing recognition that the only way people will accept an agent's behaviour is if the system provides details around the context in which its decision was based [189]. Langley et al. (2017) [23] describe this as *explainable agency*, that Dazeley et al. (2021) [14] considers a first-order explanation, where the aim is to communicate the agent's Theory of Mind [190]. Goal-driven explainability is primarily focused on Belief, Desire, Intention (BDI) agents [191], or potentially in multiobjective optimisation [192]. The potential for explainable

agency in RL has only been recognised recently [14, 188]. In particular, Sado et al. (2020) [188] accept approaches to explaining actions, discussed in Sect. 4.2, as a post hoc and domain independent approach to explaining behaviour.

The difficulty is that RL does not explicitly project the effect of their actions and associate them with a goal. Therefore, when there is no model RL is essentially learning a habit, rather than a goal [193]. For most applications, this distinction is trivial as there is only a single goal and the agent learns a habit for how to solve it. Beyond informing the user of what the goal is, explaining the choice of goal (when there is only one to choose from) is relatively meaningless. Therefore, for XRL to provide meaningful goal explanation it should have multiple goals that it could be pursuing at any given time. This utilisation of multiple goals, while not part of the standard RL framework, is a well-established approach with extensions to RL such as hierarchical [49, 52, 69, 95, 194], multi-goal [65, 66, 70, 71], and multi-objective [55, 58, 181]. This paper argues that more meaningful goal-based explanations can be provided if RL utilise these methods more readily.

As shown in Sect. 4.2, the first attempts to utilise MORL to provide contrastive explanations [182–184, 187] have been published. The aim for a goal-based explanation though would be to extend this initial work and answer questions about the XRL-goal being selected and how that goal affected the action selection. For instance, Karimpanal and Wilhelm (2017) [195] identify 'interesting states' and learn how to find them using off-policy learning while focusing on its primary objective. Attaching a goal-based explanation to this would allow explanation about how actions could also lead to/or avoid alternative objectives. A second example would be when an agent is performing a primary task, but has an alternative objective to avoid dangerous situations [196], then an explanation can identify contrastive explanations for an action on the basis of the primary or secondary objective, e.g. "While X was the fastest action, I chose Y because it was safer".

Multi-goal [65, 66, 70, 71] and Hierarchical [49, 52, 69, 95, 194] RL provide mechanisms for identifying alternative or sub-goals to problems and switching or progressing through these during a problem-solving process. At this stage, there does not appear to be any attempts to provide explanations based on the currently selected goal as a means of providing better contextual information to a user. However, this paper has suggested the provision of such explanations would be a valuable area of pursuit. For instance, Beyret et al's (2019) [168] approach could be extended to provide an explanation for the currently active goal through a tree traversal of potential goals using waypoints during the inference process.

## 5.2 Explanation of disposition

Agents that are changing their goals and/or objectives do not generally do so randomly. Some agents may do so because they have learnt that a sequence of sub-goals is required to achieve its primary goal [49–52]. Others may have multiple conflicting objectives [55], such as achieving a task while maintaining a safe working environment [58, 59]. This process of changing goals or objectives is a result of variations made in an agent's internal disposition [14]. It is important that an agent is, therefore, able to include in its explanation how its current internal disposition has influenced the current choice of goal or objective.

This change can be caused by: an observation that the prior goal was no longer appropriate for achieving its primary goal; an observed change in the environment, possibly by an external actor; or, a change in an internal simulation of an emotion, belief or desire. In Cognitive Science the Theory of Motivated control investigates how behaviour is coordinated to achieve meaningful outcomes [197]. In particular, Pezzulo et al. (2018) [194] discuss the multidimensional and hierarchical nature of goals when decision making. Essentially, people weigh-up conflicting objectives through a hierarchy of goals [194]. Through careful introspection it is possible for an RL agent to identify these changes in its internal disposition and provide an explanation for these changes. Such an explanation would represent a first-order explanation [14] and provide a valuable insight into an agent's reasoning for a human observer.

Currently, there have not been any examples of explaining such dispositional RL systems, but there are numerous examples of agent-based systems, including RL, that adapt their goal autonomously during their operation. Intrinsically motivated RL has been researched for two decades, where agents construct a hierarchy of reusable skills dynamically [53, 198, 198]. These agents change their operating goal due to internal changes such as motivations [199]. While methods such as Beyret et al. (2019) [168] explain an action relative to a goal, they could be extended to explain the motivation behind its choice of goal and skill.

Disposition and motivation are not just hierarchical, but also multidimensional [194]. For instance, Vamplew et al. (2017) [72] and Vamplew et al. (2015) [200] used an algorithm referred to as Q-steering to provide the agent the ability to switch between objectives autonomously. When objectives are in conflict the agent can have an internal desire to focus on one over another and while it pursues that objective the desire to switch to an alternative objective often increases until that change is made. This approach has potential in several domains where

autonomous balancing of objectives is required. An explanation for identifying the reason behind switching between policies would provide a user valuable information.

The recently emerging research in Emotion-aware Explainable AI (EXAI) methods illustrates an interest in providing explanations for agent's internal dispositions [26]. This work focuses on self-explaining emotions and can identify important beliefs and desires. While this work is based in on a BDI framework, Dazeley et al. (2021) [14] argue that this can be extended to XRL. One example of this approach in RL is Barros et al. (2020) [201] which uses Cruz et al's (2021) [64] introspection-based approach to identify an explanation, which is used to provide a self-explanation so that it can self-determine its intrinsic 'mood' concerning its performance in competitive games. This approach uses an explanation that informs the agent's behaviour directly. However, Barros et al's (2020) [201] approach does not currently provide an explanation for how this dispositional change has affected its current goal. Currently providing such an explanation is not evident in the XRL literature and represents an opportunity for future research.

## 5.3 Explanation of events

In many real-world applications, an RL agent will be required to deal with stochastic and dynamic environments [202]. In such environments unplanned events will occur potentially creating unexpected outcomes. An explainable agent, in such an environment, will be expected to explain how that event caused an outcome, or provide a full causal path detailing how the event caused any changes in the agent's disposition, goal or action selection. For an agent to provide such an explanation it must be able to predict the future states that would arise independent of the presence and actions of other actors within the environment. The agent's response in terms of disposition, goals and actions of the expected state and the actual state can then be compared to provide such an explanation. An extension of this model would also be able to explain what the event was that changed the environment from that which was expected. Therefore, it must be able to model the nature of stochastic events or model external actor's behaviour to understand how they may affect the environment. Therefore, this type of explanation requires the agent to perform a second-order, or social, level of explanation [14].

There is a range of value-based approaches to optimising an agent's behaviour in such environments. For instance, Robust RL [203], and specially designed training mechanisms [204] can provide value-based solutions for learning and adapting in stochastic and dynamic environments. However, these approaches rarely predict the future

state or model changes in the environment explicitly. Therefore, providing an explanation facility with such approaches is unlikely to provide suitable results. The nature of requiring an explicit prediction for such an explanation excludes the direct application of value-based RL methods without some form of separate predictive model being developed. For instance, one approach used for this is the utilisation of generative adversarial networks (GANs) [205–207] and even recurrent generative adversarial networks (RGANs) [208, 209]. In RL, these methods are more frequently being referred to as Predictive State Encoders [210, 211] and are used to generate future states, also called belief states, and to predict dynamic actor's behaviour [212, 213]. Similarly, in model-based methods, there has been significant work in developing multiple models of a domain, where prediction errors are used to select the controller or policy [214–216].

There is evidence of this being a valuable form of explanation by work in BDI agents [191, 217]. As the name suggests, BDI agents use knowledge engineering principles to explicitly model beliefs, desires and intentions for an agent. Using knowledge-based graph traversals the beliefs about events and external actors can be a component of an integrated broad explanation of the system's behaviour. Similarly, outside of BDI research, there have been planning methods developed for providing such an explanation. Based on knowledge engineering principles, these approaches utilise abductive reasoning to generate explanations [218, 219]. Molineaux et al. (2011) [220] present a particularly interesting method that learns an event-model to explain anomalies through generative abductive reasoning over historical observation in partially observable dynamic environments. Currently explaining events in stochastic and dynamic environments has not been done in the RL space. As XRL research moves away from debugging style explanation towards non-expert focused explanations, providing event-based explanations is an important future research direction.

#### 5.4 Explanation of expectations

Traditional RL has a single goal that is generally defined by the reward engineer implementing the solution. This goal is an articulation of the expectation being placed on the agent. Due to the 'hard coded' nature of this expectation there is little need to explain how this expectation has caused its behaviour. However, this approach only allows for the development of very narrowly defined agents and is not applicable as agents become increasingly societal integrated with society. Such a system must adapt to their dynamic surroundings; changing their disposition and goal based on the cultural expectations of the external society with which they are integrated. Any such system must also

be able to explain what expectations it is using to modify its behaviour. Dazeley et al. (2021) [14] extensively discussed these  $N^{\text{th}}$ -order explanations and the need for an autonomous agent operating in a human-AI integrated environment to model the cultural expectations that other actors may have on how the agent should behave.

Expectations may be easily codifiable rules such as government-enforced laws, military rules of engagement, ethical guidelines or business rules or they can be more abstract, learnt, or niche rules such as staying out of the doctor's way when they are rushing through an emergency ward. To meet these expectations an agent is required to change their behaviour away from their primary objective, whatever that might be. These changes in behaviour represent an area that must be explained as it may not be obvious to observers why an agent behaved in the way that it did. In particular, it should be able to articulate what expectation the agent is pursuing at any given time, why it selected that expectation, and how that changed its behaviour.

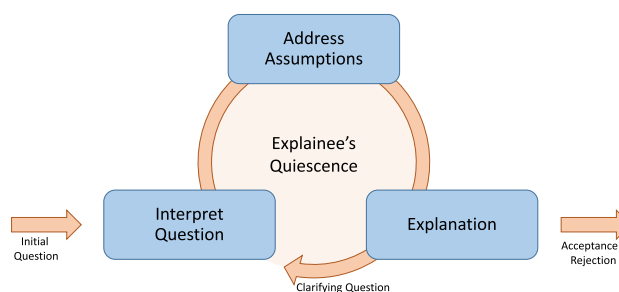
Only agents that actively maintain a model of the expectations being placed upon it would require such explanations and currently this can only be done in RL through the incorporation of secondary systems. For instance, behaviour modelling has been studied in several fields such as BDI-based Normative Agents [221–224]; Game Theory [1, 225–228]; Emotion-Driven or Emotion Augmentation learning [29, 229–234]; and, most directly by Social Action research, which models the external demands placed on an agent that affect its goals or actions [235–237]. Direct use of expectation in RL is evident where some systems are designed to incorporate social and cultural awareness in to their action selection mechanism, such as pedestrian and crowd avoidance systems [238–243].

Like explanations of events, there are currently no known examples of XRL research into providing explanations of such systems. One particularly interesting recent study by Kampik et al. (2019) [33] uses the idea of Explicability [244], where an agent can perform actions and make decisions based on human expectations. Kampik et al. (2019) [33] developed an approach and taxonomy for sympathetic actions that incorporate a utility for socially beneficial behaviour at the detriment of the agent's own personal gain. This system then provided explanations for the agent's behaviour resulting from these expectations. Furthermore, Kampik et al. (2019) [33] recognise the relevance and applicability of this approach to RL-based systems. Identifying papers in this space is, however, difficult as there is no defined research domain for this research and papers are often published under more generic fields such as understandability [34], transparency [35], and predictability [36].

## 6 Using the causal explainable reinforcement learning framework

The work discussed previously focused on how each of the individual components of the Causal XRL Framework (CXF) has, or could be, implemented. This section briefly looks at how the CXF can be implemented and used. To some extent this can simply involve implementing all of the approaches in a single explanation facility for an agent. For instance, a system could initially use a technique such as Greydanus et al. (2017) [113] to identify the active features of a state. These key features could then also be used to learn causal links between actions and outcomes creating a model similar to those developed by Madumal et al. (2019) [177], Khan et al. (2009) [179] or Cruz et al. (2019) [160]. The combination of these approaches could provide answers to many of the reactive explanations required of the Simple-CXF. Extending these approaches to incorporate multiple objectives [182, 183], or reward decomposition [184, 187] would allow expressive contrastive and counterfactual explanations that would also facilitate the explanation of the goals and dispositions behind those choices. Second and  $N^{\text{th}}$ -order explanations of events and expectations would require an agent to construct models of other actors in dynamic environment using approaches such as Predictive State Encoders [210, 211], Emotion Augmentation [29, 229–234], or Social Action [235–237]. Methods would then need to be developed to explain how these models affected the agent's expectations, disposition or its interpretation of an event. Such an approach combining all these elements would accomplish the idea of explaining the full details of the decision.

One important example illustrating the potential of this combined approach was a study of non-experts carried out by Anderson (2019) [245]. This study of 124 participants found that there was a significant improvement in the explainee's mental model of the agents behaviour when both XRL-Perception and XRL-Behavioural explanations were provided, when compared to only providing one or no explanation. This suggests that the combined approach was of value in improving peoples understanding. However, Anderson (2019) [245] also found that the combined explanation created disproportionately high cognitive loads for the explainee. This suggests that providing explanations across all categories would be unwieldy and difficult for most people to understand — simply because there is too much, potentially conflicting, information. This result aligns with Lombrozo's (2007) [84] suggestion that an explanation needs as few causes (simple) that cover as many events (general) and maintain consistency with peoples' prior knowledge (coherent) [246]. Therefore, simply merging explanation facilities brings us no closer to



**Fig. 7** A conversational model for explanation, proposed by Dazeley et al. (2021) [14], where an agent iterates through three stages until the explainee is satisfied. The agent starts at the highest level of explanation and progresses down the pyramid, Fig. 1 to more specific explanations

presenting explanations that improve understanding, and hence trust and social acceptance.

Dazeley et al. (2021) present a model for conversational interaction for explaining an AI agent's behaviour, reproduced in Fig. 7. The proposed model suggests that the agent presents the explanations incrementally over a sequence of interaction cycles. It is suggested that such a model would start at the highest level of intentionality in its explanation ( $N^{\text{th}}$ -order) and progress down the pyramid, Fig. 1, until the explainee reaches a point of *Quiescence* (state of being quiet) representing a measure of stability in the user's understanding and acceptance and no longer requires deeper explanations. This model of conversational explanation aligns with the CXF proposed in this paper. Due to each of the CXF categories being aligned to the levels of explanation [14], this paper proposes that an implementation of the CXF can use this model to break the range of explanation types down and only present those that are required for the user at that time.

In this model, the user would initially pose a query concerning an agent's decision, either explicitly or implicitly, which is first interpreted. The second stage attempts to identify and clarify any assumptions. This stage allows the agent to skip higher levels of explanation and go straight to the lower-level explanations to address any assumptions if required. For example, if the user asks: 'why didn't you catch the ball?' There is an assumption that the agent was aware that there was a ball, or that it did not succeed in catching the ball. Therefore, in resolving such assumption the agent should first determine if it was aware of a ball, and secondly, whether the outcome was in fact that no ball was caught. In the event that the assumptions are incorrect, e.g. there was no ball in its perception, then the explanation provided in the last stage skips the higher levels and provides the relevant lower-level explanation. If there are no assumptions or if they are correct then the agent provides the highest level explanation as this is the most general. In the final stage the agent,



using ontological models or visualisations, etc., provides a causal explanation at the determined level.

This approach ensures that the explanation is coherent, focused on the explainee's context, while otherwise being as general as possible. In the event that the explanation does not satisfy the user they will either ask a follow-up question, or through body language, indicate they are not satisfied. In such situations, the agent simply progresses to the next lower-level explanation. The process ends once: the user expresses satisfaction; they change their questions to a new topic; or, all available explanations have been provided. This interactive approach to communicating explanations to a user represents a process where the agent aims to facilitate the development of a shared mental model with a human. This shared mental model is key in many situations, particularly in team-based and socially integrated domains. Development of these shared mental models has been previously explored by Tabrez and Hayes (2019) [247] where an agent uses a process referred to as Reward Augmentation and Repair through Explanation (RARE), based on inverse RL, to infer the most likely 'reward' function used by a human collaborator and explain how that differs from the optimal function. In other words, this project is similar to this paper's approach in that it is providing an explanation in the context of the explainee's current understanding.

Currently, there is no attempt to build a facility like the CXF in the XRL literature, apart from some attempts to combine perception with behaviour [177, 245] and suggested extensions to combine actions with goals [64, 177, 182, 183]. The approach has been extended more thoroughly outside of XRL, using generic explanation facilities. These systems observed interactions by the agent and use the learnt model to provide explanations, such as Local Interpretable Model-Agnostic Explanations (LIME) [174] and Black Box Explanations through Transparent Approximations (BETA) [248]. Both of these approaches provide explanations across a subset of the components in the CXF.

One particularly notable example is Neerincx et al. (2018) [217], which extended LIME and separated perceptual explanations from the cognitive processing to provide holistic explanations. The cognitive processing component incorporated goal and dispositional explanation based on emotion-based explanations. Finally, the approach incorporated ontological and interaction design patterns to communicate explanations. This approach represents the most advanced implementation utilising [14] levels of explanation based on intentionality and could be interpreted using multiple sections of the CXF.

## 7 Conclusion

Reinforcement Learning (RL) is widely acknowledged as one of three subfields of Machine Learning, where an agent learns through interaction with the environment using trial-and-error. However, research in eXplainable RL (XRL) are often published under the area of Interpretable Machine Learning (IML), along with supervised learning approaches to explanation. This categorisation, however, misrepresents the possibilities that XRL presents. This paper's aim was to articulate how XRL is distinct when compared to IML, and that it offers the potential to go well beyond simply interpreting decisions. More importantly, that XRL could be the foundation to the development of truly *Broad-XAI* [14] systems that are capable of providing trusted and socially acceptable AI systems to the wider public. In order to illustrate this point, this paper provides a conceptual framework, referred to as the Causal XRL Framework (CXF), that highlights the range of explanations that can be provided. This framework was used to review the current extent of research that has been carried out and to identify opportunities for future research.

The Causal XRL Framework (CXF), presented in Fig. 4, is based on the Casual Explanation Network (CEN) suggested by Böhm and Pfister (2015) [83]. The CEN presents a cognitive science view of how people explain behaviour and extends prior work in attribution theory [82]. Like the CEN, the CXF identifies seven components to causal thinking about an actor's behaviour. This directed graph of causal relationships includes a single sink node representing the outcome. This outcome is caused by either an intentional action by an agent, or from an unintended or uncontrolled sequence of events. These events could be a result of stochastic actions or external actors. The agent's actions are caused by a goal which in turn may be altered by its internal and temporary disposition, such as a change in parameter, simulated emotion or safety threshold being passed. Finally, the disposition can be affected by external cultural expectations placed upon the agent or by its perception of the world. A simplified framework, referred to as the Simplified-CXF, containing only perception and action causes for an outcome, was also provided, which represents the majority of current research in XRL.

In surveying the current state-of-the-art research into XRL this paper discussed how most of the XRL-Perception was derived directly from IML research. This connection with IML is due to RL's utilisation of standard supervised learning approaches for function approximation and state feature extraction. However, there were also several examples moving beyond straight IML; providing both model and value-based extensions that were specific to XRL. These methods use introspection of the RL

framework to identify causal relationships between the perceived features and either the action selected or the outcome that resulted. Of particular interest was the emergence of methods used for generating counterfactual and contrastive explanations based on these causal links between state-features and outcomes. XRL-behaviour, as a sub-branch of XRL where the explanation aims to clarify the agent's choice of action and the effect it has on the outcome, was explored in detail.

Finally, in discussing the full framework, this paper identified several opportunities for further research in XRL, such as using hierarchical, multi-goal, multi-objective and intrinsically motivated RL techniques in Goal-driven explanation and emotion-aware explainable AI (EXAI). This paper also discussed hypothetical approaches to the development of event-based and expectation-based explanations, such as utilising predictive state encoders and explicability in RL. Currently, these areas have been studied by other fields of explanation, such as BDI agents and Social Action, but represent the fringe of XRL research. This paper suggests these are exciting areas of future study that this field should pursue so that RL can be more widely used in real-world human-agent mixed application domains.

**Funding** Open Access funding enabled and organized by CAUL and its Member Institutions.

**Data availability** Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M et al (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587):484–489
- Huval B, Wang T, Tandon S, Kiske J, Song W, Pazhayampallil J, Andriluka M, Rajpurkar P, Migimatsu T, Cheng-Yue R, et al (2015) “An empirical evaluation of deep learning on highway driving.” <http://arxiv.org/abs/1504.01716>
- Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Belle-mare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G et al (2015) Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533
- Knight W (2017) “Reinforcement learning: By experimenting, computers are figuring out how to do things that no programmer could teach them.”. accessed: 2019-10-06
- Metz C (2017) “In two moves, AlphaGo and Lee Sedol redefined the future.”. accessed: 2019-10-06
- Metz C (2017) “How google's AI viewed the move no human could understand.”. 2019-10-06
- Cruz F, Young C, Dazeley R, Vamplew P (2022) “Evaluating human-like explanations for robot actions in reinforcement learning scenarios,” In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 894–901, IEEE
- Heuillet A, Couthouis F, Rodríguez ND (2020) “Explainability in deep reinforcement learning.” <http://arxiv.org/abs/2008.06693>
- Wallkötter S, Tulli S, Castellano G, Paiva A, Chetouani M (2021) Explainable embodied agents through social cues: a review. *ACM Trans Human-Robot Interact* 10:1–24
- Vouros GA (2022) Explainable deep reinforcement learning: state of the art and challenges. *ACM Comput Survays (CSUR)* 55:1–39
- Milani S, Topin N, Veloso M, Fang F (2022) “A survey of explainable reinforcement learning,” <http://arxiv.org/abs/2202.08434>
- Qing Y, Liu S, Song J, Song M (2022) “A survey on explainable reinforcement learning: Concepts, algorithms, challenges,” <http://arxiv.org/abs/2211.06665>
- Miller T (2017) “Explanation in artificial intelligence: insights from the social sciences,” <http://arxiv.org/abs/1706.07269>
- Dazeley R, Vamplew P, Foale C, Young C, Aryal S, Cruz F (2021) Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artif Intell* 299:103525
- Harari YN (2016) *Homo Deus: a brief history of tomorrow*. Harvill Secker, London
- Merriam-Webster, “Dictionary,” 2020. accessed: 2020-03-12
- Woodward J (2017) “Scientific explanation,”. accessed: 2019-10-06
- Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M (2018) “Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda,” In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI'18
- Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6:52138–52160
- Kumari M, Chaudhary A, Narayan Y (2023) “Explainable ai (xai): A survey of current and future opportunities,” In: explainable edge AI: a futuristic computing perspective, pp. 53–71, Springer
- Miller T, Howe P, Sonenberg L (2017) “Explainable AI: Beware of inmates running the asylum,” In: IJCAI-17 Workshop on Explainable AI (XAI), p. 36
- Zhang Q-S, Zhu S-C (2018) Visual interpretability for deep learning: a survey. *Front Inf Technol Electron Eng* 19(1):27–39

23. Langley P, Meadows B, Sridharan M, Choi D (2017) “Explainable agency for intelligent autonomous systems,” In: Twenty-Ninth IAAI Conference
24. Anjomshoae S, Najjar A, Calvaresi D, Främling K (2019) “Explainable agents and robots: Results from a systematic literature review,” In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, pp. 1078–1088, International Foundation for Autonomous Agents and Multiagent Systems
25. Anjomshoae S, Främling K (2019) Intelligible explanations in intelligent systems
26. Kaptein F, Broekens J, Hindriks K, Neerincx M (2017) “The role of emotion in self-explanations by cognitive agents,” In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp. 88–93, IEEE
27. Rorty AO (1978) Explaining emotions. *J Philos* 75(3):139–161
28. O’Rorke P, Ortony A (1994) Explaining emotions. *Cognit Sci* 18(2):283–323
29. Hao M, Cao W, Liu Z, Wu M, Yuan M (2019) “Emotion regulation based on multi-objective weighted reinforcement learning for human-robot interaction,” In: 2019 12th Asian Control Conference (ASCC), pp. 1402–1406, IEEE
30. Mathews SM (2019) “Explainable artificial intelligence applications in NLP, biomedical, and malware classification: A literature review,” In: Intelligent Computing-Proceedings of the Computing Conference, pp. 1269–1292, Springer
31. Weitz K, Schiller D, Schlagowski R, Huber T, André E (2019) “Do you trust me?: Increasing user-trust by integrating virtual agents in explainable AI interaction design,” In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, pp. 7–9, ACM
32. Sindlar M, Dastani M, Meyer J-J (2011) “Programming mental state abduction,” In: The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 1, pp. 301–308, International Foundation for Autonomous Agents and Multiagent Systems
33. Kampik T, Nieves JC, Lindgren H (2019) “Explaining sympathetic actions of rational agents,” In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, pp. 59–76, Springer
34. Hellström T, Bensch S (2018) Understandable robots-what, why, and how. *J Behav Robotics* 9(1):110–123
35. Wortham RH, Theodorou A (2017) Robot transparency, trust and utility. *Connect Sci* 29(3):242–248
36. Dragan AD, Lee KC, Srinivasa SS (2013) “Legibility and predictability of robot motion,” In: Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction, pp. 301–308, IEEE Press
37. Pitrat J et al (2006) “Meta-explanation in a constraint satisfaction solver,” In: Information Processing and Management of Uncertainty in Knowledge-based Systems IPMU, pp. 1118–1125, Citeseer
38. Galitsky B (2016) “Formalizing theory of mind,” in Computational Autism, pp. 95–176, Springer
39. Galitsky BA, de la Rosa i Esteva JL, Kovalerchuk B (2010) “Explanation versus meta-explanation: What makes a case more convincing,” In: FLAIRS Conference
40. Ehsan U, Tambwekar P, Chan L, Harrison B, Riedl MO (2019) “Automated rationale generation: a technique for explainable AI and its effects on human perceptions,” In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 263–274, ACM
41. Ehsan U (2019) On design and evaluation of human-centered explainable AI systems
42. McLaughlin J (1988) “Utility-directed presentation of simulation results,” In: Proceedings of the Annual Symposium on Computer Application in Medical Care, p. 292, American Medical Informatics Association,
43. Sutton RS, Barto AG (2018) Reinforcement learning: an introduction (Second Edition). A Bradford Book, adaptive computation and machine learning series, London
44. Li Y (2017) “Deep reinforcement learning: an overview,” <http://arxiv.org/abs/1701.07274>
45. Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA (2017) “A brief survey of deep reinforcement learning,” <http://arxiv.org/abs/1708.05866>
46. Duan Y, Chen X, Houthoofd R, Schulman J, Abbeel P (2016) “Benchmarking deep reinforcement learning for continuous control,” In: International Conference on Machine Learning, pp. 1329–1338
47. Hossain M, Sohel F, Shiratuddin MF, Laga H (2019) A comprehensive survey of deep learning for image captioning. *ACM Comput Surveys (CSUR)* 51(6):118
48. Van Hasselt H, Doron Y, Strub F, Hessel M, Sonnerat N, Modayil J (2018) “Deep reinforcement learning and the deadly triad,” <http://arxiv.org/abs/1812.02648>
49. Barto AG, Mahadevan S (2003) Recent advances in hierarchical reinforcement learning. *Discr Event Dyn Syst* 13(1–2):41–77
50. Gebhardt C, Oulasvirta A, Hilliges O (2020) “Hierarchical reinforcement learning as a model of human task interleaving,” <http://arxiv.org/abs/2001.02122>
51. Zhou W-J, Yu Y (2020) “Temporal-adaptive hierarchical reinforcement learning,” <http://arxiv.org/abs/2002.02080>
52. Botvinick MM (2012) Hierarchical reinforcement learning and decision making. *Curr Opin Neurobiol* 22(6):956–962
53. Barto AG, Singh S, Chentanez N (2004) “Intrinsically motivated learning of hierarchical collections of skills,” In: Proceedings of the 3rd International Conference on Development and Learning, pp. 112–19,
54. Florensa C, Held D, Geng X, Abbeel P (2017) “Automatic goal generation for reinforcement learning agents,” <http://arxiv.org/abs/1705.06366>
55. Roijers DM, Vamplew P, Whiteson S, Dazeley R (2013) A survey of multi-objective sequential decision-making. *J Artif Intell Res* 48:67–113
56. Hayes CF, Rădulescu R, Bargiacchi E, Källström J, Macfarlane M, Reymond M, Verstraeten T, Zintgraf LM, Dazeley R, Heintz F et al (2022) A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents Multi-Agent Syst* 36(1):1–59
57. Vamplew P, Smith BJ, Källström J, Ramos G, Rădulescu R, Roijers DM, Hayes CF, Heintz F, Mannion P, Libin PJ et al (2022) Scalar reward is not enough: a response to silver, singh, precup and sutton (2021). *Autonomous Agents Multi-Agent Syst* 36(2):1–19
58. Vamplew P, Dazeley R, Foale C, Firmin S, Mummery J (2018) Human-aligned artificial intelligence is a multiobjective problem. *Ethics Inf Technol* 20(1):27–40
59. Vamplew P, Foale C, Dazeley R (2020) Potential-based multi-objective reinforcement learning approaches to low-impact agents for AI safety (submitted). *Eng Appl Artif Intell* 100:104186
60. Alharin A, Doan T-N, Sartipi M (2020) Reinforcement learning interpretation methods: a survey. *IEEE Access* 8:171058–171077
61. Schwab P, Karlen W (2019) Explain: causal explanations for model interpretation under uncertainty. *Adv Neural Inf Process Syst* 87:10220–10230

62. O'Shaughnessy M, Canal G, Connor M, Rozell M, Davenport M (2020) Generative causal explanations of black-box classifiers. *Adv Neural Inf Process Syst* 33:87
63. Zhang J, Bareinboim E (2018) Equality of opportunity in classification: a causal approach. *Adv Neural Inf Process Syst* 31:3671–3681
64. Cruz F, Dazeley R, Vamplew P, Moreira I (2021) Explainable robotic systems: understanding goal-driven actions in a reinforcement learning scenario. *Neural Comput Appl* 45:1–18
65. Vattam S, Klenk M, Molineaux M, Aha DW (2013) "Breadth of approaches to goal reasoning: a research survey," tech. rep., Naval Research Lab Washington DC
66. Plappert M, Andrychowicz M, Ray A, McGrew B, Baker B, Powell G, Schneider J, Tobin J, Chociej M, Welinder P, Kumar V, Zaremba W (2018) "Multi-goal reinforcement learning: Challenging robotics environments and request for research," *CoRR*, vol. <http://arxiv.org/abs/1802.09464>
67. Caruana R (1997) Multitask learning. *Mach Learn* 28(1):41–75
68. Barto AG, Mahadevan S (2003) Recent advances in hierarchical reinforcement learning. *Discr Event Dyn Syst* 13(1):41–77
69. Al-Emran M (2015) Hierarchical reinforcement learning: a survey. *Int J Comput Digital Syst* 4:2
70. Teh Y, Bapst V, Czarnecki WM, Quan J, Kirkpatrick J, Hadsell R, Heess N, Pascanu R (2017) "Distral: Robust multitask reinforcement learning," In: *Advances in Neural Information Processing Systems* 30 (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 4496–4506, Curran Associates, Inc
71. Kaelbling LP (1993) "Learning to achieve goals," in In: *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pp. 1094–1098, Morgan Kaufmann
72. Vamplew P, Issabekov R, Dazeley R, Foale C, Berry A, Moore T, Creighton D (2017) Steering approaches to Pareto-optimal multiobjective reinforcement learning. *Neurocomputing* 263:26–38
73. Heider F (1958) *The psychology of interpersonal relations*. Psychology Press, England
74. Jones EE, Davis KE (1965) From acts to dispositions the attribution process in person perception. *Adv Exp Social Psychol* 2:219–266
75. Kelley HH (1967) "Attribution theory in social psychology.," In: *Nebraska symposium on motivation*. University of Nebraska Press
76. Kelley HH (1973) The processes of causal attribution. *Am Psychol* 28(2):107
77. Fiske ST, Taylor SE (1991) *Social cognition*. McGraw-Hill Book Company, New York
78. Malle BF (1999) How people explain behavior: a new theoretical framework. *Personality Soc Psychol Rev* 3(1):23–48
79. Malle BF (2006) *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT Press, Cambridge
80. Malle BF, Knobe J, O'Laughlin MJ, Pearce GE, Nelson SE (2000) Conceptual structure and social functions of behavior explanations: beyond person-situation attributions. *J Personality Soc Psychol* 79(3):309
81. Kamrath LK, Mendoza-Denton R, Mischel W (2005) Incorporating if... then... personality signatures in person perception: beyond the person-situation dichotomy. *J Personality Soc Psychol* 88(4):605
82. Schank RC, Abelson RP (2013) *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press, London
83. Böhm G, Pfister H-R (2015) How people explain their own and others' behavior: a theory of lay causal explanations. *Front Psychol* 6:139
84. Lombrozo T (2007) Simplicity and probability in causal explanation. *Cognit Psychol* 55(3):232–257
85. Cheney DL, Seyfarth RM (1990) *How monkeys see the world: inside the mind of another species*. University of Chicago Press, Chicago and London
86. Hayes CF, Rădulescu R, Bargiacchi E, Källström J, Macfarlane M, Reymond M, Verstraeten T, Zintgraf LM, Dazeley R, Heintz F et al (2021) "A practical guide to multi-objective reinforcement learning and planning," <http://arxiv.org/abs/2103.09568>
87. Chu T, Wang J, Codecà L, Li Z (2019) Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Trans Intell Transp Syst* 21(3):1086–1095
88. Kazhdan D, Shams Z, Liò P (2020) "Marleme: A multi-agent reinforcement learning model extraction library," In: *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE
89. Lewis M, Sycara K, Walker P (2018) "The role of trust in human-robot interaction," In: *Foundations of trusted autonomy*, pp. 135–159, Springer, Cham
90. Bethel CL, Carruth D, Garrison T (2012) "Discoveries from integrating robots into swat team training exercises," In: *2012 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pp. 1–8, IEEE
91. Huang SH, Held D, Abbeel P, Dragan AD (2019) Enabling robots to communicate their objectives. *Autonomous Robots* 43(2):309–326
92. Russo MW (2007) How to review a meta-analysis. *Gastroenterol Hepatol* 3(8):637
93. Baumeister RF, Leary MR (1997) Writing narrative literature reviews. *Rev Gener Psychol* 1(3):311–320
94. Busoniu L, Babuska R, De Schutter B, Ernst D (2010) *Reinforcement learning and dynamic programming using function approximators*, vol 39. CRC Press, USA
95. Botvinick MM, Niv Y, Barto AG (2009) Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition* 113(3):262–280
96. Singh SP, Jaakkola T, Jordan MI (1995) Reinforcement learning with soft state aggregation. *Adv Neural Inf Process Syst* 45:361–368
97. Hutter M (2014) "Extreme state aggregation beyond MDPs," In: *International Conference on Algorithmic Learning Theory*, pp. 185–199, Springer
98. Van Otterlo M (2005) *A survey of reinforcement learning in relational domains*, Centre for Telematics and Information Technology (CTIT) University of Twente. Tech, Rep
99. Sutton RS, Precup D, Singh S (1999) Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. *Artif Intell* 112(1–2):181–211
100. Bacon P-L, Harb J, Precup D (2017) "The option-critic architecture," In: *Thirty-First AAAI Conference on Artificial Intelligence*
101. Mirowski P, Pascanu R, Viola F, Soyer H, Ballard AJ, Banio A, Denil M, Goroshin R, Sifre L, Kavukcuoglu K et al (2016) "Learning to navigate in complex environments," <http://arxiv.org/abs/1611.03673>
102. Schaul T, Quan J, Antonoglou I, Silver D (2015) "Prioritized experience replay," <http://arxiv.org/abs/1511.05952>
103. Doshi-Velez F, Kim B (2017) "Towards a rigorous science of interpretable machine learning," <http://arxiv.org/abs/1702.08608>
104. Molnar C (2019) *Interpretable machine learning*. Lulu. com
105. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) "Explaining explanations: An overview of interpretability of machine learning," In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89, IEEE



106. Puiutta E, Veith E (2020) “Explainable reinforcement learning: A survey,” arXiv preprint <http://arxiv.org/abs/2005.06247>
107. Zahavy T, Ben-Zrihem N, Mannor S (2016) “Graying the black box: Understanding dqns,” In: International Conference on Machine Learning, pp. 1899–1908
108. Levine S, Finn C, Darrell T, Abbeel P (2016) End-to-end training of deep visuomotor policies. *J Mach Learn Res* 17(1):1334–1373
109. Andrulis J, Meyer O, Schott G, Weinbach S, Gruhn V (2020) “Domain-level explainability—a challenge for creating trust in superhuman ai strategies,” <http://arxiv.org/abs/2011.06665>
110. Wang Z, Schaul T, Hessel M, Van Hasselt H, Lanctot M, De Freitas N (2015) “Dueling network architectures for deep reinforcement learning,” <http://arxiv.org/abs/1511.06581>
111. Simonyan K, Zisserman A (2014) “Very deep convolutional networks for large-scale image recognition,” <http://arxiv.org/abs/1409.1556>
112. Shi W, Wang Z, Song S, Huang G (2020) “Self-supervised discovering of causal features: Towards interpretable reinforcement learning,” <http://arxiv.org/abs/2003.07069>
113. Greydanus S, Koul A, Dodge J, Fern A (2017) “Visualizing and understanding atari agents,” <http://arxiv.org/abs/1711.00138>
114. Verma A, Murali V, Singh R, Kohli P, Chaudhuri S (2018) “Programmatically interpretable reinforcement learning,” <http://arxiv.org/abs/1804.02477>
115. Robeer MJ (2018) “Contrastive explanation for machine learning,” Master’s thesis,
116. Chang C-H, Creager E, Goldenberg A, Duvenaud D (2018) “Explaining image classifiers by counterfactual generation,” <http://arxiv.org/abs/1807.08024>
117. Goyal Y, Wu Z, Ernst J, Batra D, Parikh D, Lee S (2019) “Counterfactual visual explanations,” <http://arxiv.org/abs/1904.07451>
118. Atrey A, Clary K, Jensen D (2019) “Exploratory not explanatory: Counterfactual analysis of saliency maps for deep rl,” <http://arxiv.org/abs/1912.05743>
119. Dhurandhar A, Chen P-Y, Luss R, Tu C-C, Ting P, Shanmugam K, Das P (2018) Explanations based on the missing: towards contrastive explanations with pertinent negatives. *Adv Neural Inf Process Syst* 12:592–603
120. Gu J, Yang Y, Tresp V (2018) “Understanding individual decisions of cnns via contrastive backpropagation,” In: Asian Conference on Computer Vision, pp. 119–134, Springer
121. Huang SH, Bhatia K, Abbeel P, Dragan AD (2017) “Leveraging critical states to develop trust,” In: RSS 2017 Workshop: Morality and Social Trust in Autonomous Robots
122. Ester M, Kriegel H-P, Sander J, Xu X et al (1996) “A density-based algorithm for discovering clusters in large spatial databases with noise
123. Hayes B, Shah JA (2017) “Improving robot controller transparency through autonomous policy explanation,” In: 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 303–312, IEEE
124. Lin Z, Lam K-H, Fern A (2020) “Contrastive explanations for reinforcement learning via embedded self predictions,” <http://arxiv.org/abs/2010.05180>
125. Kuhnle A, May MC, Schaefer L, Lanza G (2022) Explainable reinforcement learning in production control of job shop manufacturing system. *Int J Prod Res* 60(19):5812–5834
126. Jiang X, Zhang J, Wang B (2022) Energy-efficient driving for adaptive traffic signal control environment via explainable reinforcement learning. *Appl Sci* 12(11):5380
127. Kumar S, Vishal M, Ravi V (2022) “Explainable reinforcement learning on financial stock trading using shap,” <http://arxiv.org/abs/2208.08790>
128. Guo W, Wei P (2022) “Explainable deep reinforcement learning for aircraft separation assurance,” In: 2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC), pp. 1–10, IEEE
129. Brusoni V, Console L, Terenziani P, Dupré DT (1997) “An efficient algorithm for temporal abduction,” In: Congress of the Italian Association for Artificial Intelligence, pp. 195–206, Springer
130. Bresina JL, Morris PH (2006) “Explanations and recommendations for temporal inconsistencies,” *Proc. Int. Work. on Planning and Scheduling for Space*
131. Bharadhwaj H, Joshi S (2018) Explanations for temporal recommendations. *KI-Künstliche Intelligenz* 32(4):267–272
132. Shortliffe EH, Buchanan BG (1975) A model of inexact reasoning in medicine. *Math Biosci* 23(3–4):351–379
133. Davis R, Buchanan B, Shortliffe E (1977) Production rules as a representation for a knowledge-based consultation program. *Artif Intell* 8(1):15–45
134. Swartout WR (1983) XPLAIN: a system for creating and explaining expert consulting programs. *Artif Intell* 21(3):285–325
135. Chandrasekaran B, Tanner MC, Josephson JR (1988) “Explanation: the role of control strategies and deep models,” *Expert Systems: The User Interface*, pp. 219–247
136. Lacave C, Díez FJ (2002) A review of explanation methods for bayesian networks. *Knowl Eng Rev* 17(2):107–127
137. Druzdzel MJ (1996) “Explanation in probabilistic systems: is it feasible? will it work,” Citeseer
138. Renooij S, Van Der Gaag LC (1998) “Decision making in qualitative influence diagrams,” In: FLAIRS Conference, pp. 410–414
139. Lacave C, Atienza R, Díez FJ (2000) “Graphical explanation in bayesian networks,” In: International Symposium on Medical Data Analysis, pp. 122–129, Springer
140. Bielza C, Fernández del Pozo JA, Lucas P (2003) “Optimal decision explanation by extracting regularity patterns,” In: Coenen F., Preece A., Macintosh A. (eds) *Research and Development in Intelligent Systems XX*. SGAI 2003, pp. 283–294, Springer
141. Elizalde F, Sucar LE, Reyes A, Debuen P (2007) “An MDP approach for explanation generation
142. Elizalde F, Sucar E, Noguez J, Reyes A (2009) “Generating explanations based on markov decision processes,” In: Mexican International Conference on Artificial Intelligence, pp. 51–62, Springer
143. Hanheide M, Göbelbecker M, Horn GS, Pronobis A, Sjöö K, Aydemir A, Jensfelt P, Gretton C, Dearden R, Janicek M et al (2017) Robot task planning and explanation in open and uncertain worlds. *Artif Intell* 247:119–150
144. Fox M, Long D, Magazzeni D (2017) “Explainable planning,” <http://arxiv.org/abs/1709.10256>
145. Krarup B, Cashmore M, Magazzeni D, Miller T (2019) Model-based contrastive explanations for explainable planning
146. Dodson T, Mattei N, Goldsmith J (2011) “A natural language argumentation interface for explanation generation in markov decision processes,” In: International Conference on Algorithmic Decision Theory, pp. 42–55, Springer
147. Elizalde F (2008) Policy explanation in factored markov decision processes
148. Kasenberg D, Roque A, Thielstrom R, Chita-Tegmark M, Scheutz M (2019) “Generating justifications for norm-related agent decisions,” <http://arxiv.org/abs/1911.00226>
149. Kasenberg D, Roque A, Thielstrom R, Scheutz M (2019) “Engaging in dialogue about an agent’s norms and behaviors,” <http://arxiv.org/abs/1911.00229>

150. Chen S, Boggess K, Feng L (2020) “Towards transparent robotic planning via contrastive explanations,” <http://arxiv.org/abs/2003.07425>
151. Hoffmann J, Magazzeni D (2019) “Explainable AI planning (XAIP): Overview and the case of contrastive explanation,” In: Reasoning Web. Explainable Artificial Intelligence, pp. 277–282, Springer
152. Chakraborti T, Fadnis KP, Talamadupula K, Dholakia M, Srivastava B, Kephart JO, Bellamy RK (2017) “Visualizations for an explainable planning agent,” <http://arxiv.org/abs/1709.04517>
153. Gopalakrishnan S, Kambhampati S (2019) Tge-viz: Mixed initiative plan visualization
154. Chakraborti T, Fadnis KP, Talamadupula K, Dholakia M, Srivastava B, Kephart JO, Bellamy RK (2019) Planning and visualization for a smart meeting room assistant. *AI Commun* 32(1):91–99
155. Bongartz IN (2018) Explaining unsolvable planning tasks
156. Wang N, Pynadath DV, Hill SG (2016) “The impact of POMDP-generated explanations on trust and performance in human-robot teams,” In: Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, pp. 997–1005
157. Seegebarth B, Müller F, Schattenberg B, Biundo S (2012) “Making hybrid plans more clear to human users—a formal approach for generating sound explanations,” In: Twenty-second International Conference on Automated Planning and Scheduling
158. Hein D, Udluft S, Runkler TA (2018) Interpretable policies for reinforcement learning by genetic programming. *Eng Appl Artif Intell* 76:158–169
159. Chakraborti T, Sreedharan S, Kambhampati S (2020) “The emerging landscape of explainable AI planning and decision making,” <http://arxiv.org/abs/2002.11697>
160. Cruz F, Dazeley R, Vamplew P (2019) “Memory-based explainable reinforcement learning,” In: The 32nd Australasian Joint Conference on Artificial Intelligence (AusAI-19), pp. 66–77
161. Lee JH (2019) “Complementary reinforcement learning towards explainable agents,” <http://arxiv.org/abs/1901.00188>
162. Lage I, Lifschitz D, Doshi-Velez F, Amir O (2019) “Toward robust policy summarization,” in Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, pp. 2081–2083, International Foundation for Autonomous Agents and Multiagent Systems
163. Amir O, Doshi-Velez F, Sarne D (2018) “Agent strategy summarization,” In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, pp. 1203–1207, International Foundation for Autonomous Agents and Multiagent Systems
164. Amir D, Amir O (2018) “Highlights: Summarizing agent behavior to people,” In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, pp. 1168–1176, International Foundation for Autonomous Agents and Multiagent Systems
165. Lage I, Lifschitz D, Doshi-Velez F, Amir O (2019) “Exploring computational user models for agent policy summarization,” <http://arxiv.org/abs/1905.13271>
166. Sequeira P, Gervasio M (2020) Interestingness elements for explainable reinforcement learning: understanding agents’ capabilities and limitations. *Artif Intell* 288:103367
167. Huang SH, Bhatia K, Abbeel P, Dragan AD (2018) “Establishing appropriate trust via critical states,” In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3929–3936, IEEE
168. Beyret B, Shafti A, Faisal AA (2019) “Dot-to-dot: Explainable hierarchical reinforcement learning for robotic manipulation,” In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5014–5019, IEEE
169. Acharya A, Russell R, Ahmed NR (2020) “Explaining conditions for reinforcement learning behaviors from real and imagined data,” <http://arxiv.org/abs/2011.09004>
170. Schroeter N, Cruz F, Wermter S (2022) “Introspection-based explainable reinforcement learning in episodic and non-episodic scenarios,” In: Proceedings of the Australian Conference on Robotics and Automation (ACRA 2022)
171. Lipton P (1990) Contrastive explanation. *R Instit Philos Suppl* 27:247–266
172. Miller T (2018) “Contrastive explanation: A structural-model approach,” <http://arxiv.org/abs/1811.03163>
173. Pocius R, Neal L, Fern A (2019) Strategic tasks for explainable reinforcement learning. Proceedings of the AAAI Conference on Artificial Intelligence 33:10007–10008
174. Ribeiro MT, Singh S, Guestrin C (2016) “Why should I trust you?” explaining the predictions of any classifier,” In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144
175. van der Waa J, van Diggelen J, Bosch Kvd, Neerinx M (2018) “Contrastive explanations for reinforcement learning in terms of expected consequences,” <http://arxiv.org/abs/1807.08706>
176. Cashmore M, Collins A, Krarup B, Krivic S, Magazzeni D, Smith D (2019) “Towards explainable AI planning as a service,” <http://arxiv.org/abs/1908.05059>
177. Madumal P, Miller T, Sonenberg T, Vetere F (2019) “Explainable reinforcement learning through a causal lens,” <http://arxiv.org/abs/1905.10958>
178. Vinyals O, Ewalds T, Bartunov S, Georgiev P, Vezhnevets AS, Yeo M, Makhzani A, üttler HK, Agapiou J, Schrittwieser J et al (2017) “Starcraft II: a new challenge for reinforcement learning,” <http://arxiv.org/abs/1708.04782>
179. Khan OZ, Poupart P, Black JP (2009) “Minimal sufficient explanations for factored markov decision processes,” In Nineteenth International Conference on Automated Planning and Scheduling
180. Juozapaitis Z, Koul A, Fern A, Erwig M, Doshi-Velez F (2019) “Explainable reinforcement learning via reward decomposition,” In: IJCAI/ECAI Workshop on Explainable Artificial Intelligence
181. Vamplew P, Dazeley R, Berry A, Issabekov R, Dekker E (2011) Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Mach Learn* 84(1–2):51–80
182. Sukkerd R, Simmons R, Garlan D (2018) “Toward explainable multi-objective probabilistic planning,” In: 2018 IEEE/ACM 4th International Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS), pp. 19–25, IEEE
183. Sukkerd R, Simmons R, Garlan D (2020) Tradeoff-focused contrastive explanation for MDP planning
184. Erwig M, Fern A, Murali M, Koul A (2018) Explaining deep adaptive programs via reward decomposition,” In: IJCAI/ECAI Workshop on Explainable Artificial Intelligence
185. Sokol K, Flach P (2020) “One explanation does not fit all,” *KI-Künstliche Intelligenz*, pp. 1–16
186. Rathi S (2019) Generating counterfactual and contrastive explanations using SHAP,” <http://arxiv.org/abs/1906.09293>
187. Erwig M, Kumar P, Fern A (2020) “Explanations for dynamic programming,” In: International Symposium on Practical Aspects of Declarative Languages, pp. 179–195, Springer
188. Sado F, Loo CK, Kerzel M, Wermter S (2020) “Explainable goal-driven agents and robots—a comprehensive review and new framework,” <http://arxiv.org/abs/2004.09705>
189. Dazeley R, Kang BH (2008) Epistemological approach to the process of practice. *Minds Mach* 18(4):547–567

190. Leslie AM (1987) Pretense and representation: the origins of theory of mind. *Psychol Rev* 94(4):412
191. Rao AS, Georgeff MP et al (1995) BDI agents: from theory to practice. *ICMAS* 95:312–319
192. Khan SU, Min-Allah N (2012) A goal programming based energy efficient resource allocation in data centers. *J Supercomput* 61(3):502–519
193. Doll BB, Simon DA, Daw ND (2012) The ubiquity of model-based reinforcement learning. *Curr Opin Neurobiol* 22(6):1075–1081
194. Pezzulo G, Rigoli F, Friston KJ (2018) Hierarchical active inference: a theory of motivated control. *Trends Cognit Sci* 22(4):294–306
195. Karimpanal TG, Wilhelm E (2017) Identification and off-policy learning of multiple objectives using adaptive clustering. *Neurocomputing* 263:39–47
196. Vamplew P, Foale C, Dazeley R, Bignold A (2021) Potential-based multiobjective reinforcement learning approaches to low-impact agents for ai safety. *Eng Appl Artif Intell* 100:104186
197. Shapiro DH Jr, Schwartz CE, Astin JA (1996) Controlling ourselves, controlling our world: Psychology's role in understanding positive and negative consequences of seeking and gaining control. *Am Psychol* 51(12):1213
198. Chentanez N, Barto AG, Singh SP (2005) Intrinsically motivated reinforcement learning. *Adv Neural Inf Process Syst* 24:1281–1288
199. Kulkarni TD, Narasimhan K, Saeedi A, Tenenbaum J (2016) Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. *Adv Neural Inf Process Syst* 87:3675–3683
200. Vamplew P, Issabekov R, Dazeley R, Foale R (2015) “Reinforcement learning of pareto-optimal multiobjective policies using steering.” In: *Australasian Joint Conference on Artificial Intelligence*, pp. 596–608, Springer
201. Barros P, Tanevska A, Cruz F, Sciutti A (2020) “Moody learners - explaining competitive behaviour of reinforcement learning agents.” In: *Proceedings of the IEEE International Conference on Development and Learning (ICDL-EpiRob 2020)*
202. Wiering MA (2001) “Reinforcement learning in dynamic environments using instantiated information.” In: *Machine Learning: Proceedings of the Eighteenth International Conference (ICML2001)*, pp. 585–592
203. Morimoto J, Doya K (2005) Robust reinforcement learning. *Neural Comput* 17(2):335–359
204. Pieters W, Wiering MA (2016) “Q-learning with experience replay in a dynamic environment.” In: *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8, IEEE
205. Aissa FB, Mejdoub M, Zaied M (2020) “A survey on generative adversarial networks and their variants methods.” In: *Twelfth International Conference on Machine Vision (ICMV 2019)*, vol. 11433, p. 114333N, International Society for Optics and Photonics
206. Gui J, Sun Z, Wen Y, Tao D, Ye J (2020) “A review on generative adversarial networks: Algorithms, theory, and applications,” <http://arxiv.org/abs/2001.06937>
207. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. *Adv Neural Inf Process Syst* 45:2672–2680
208. Mardani M, Monajemi H, Papyan V, Vasanawala S, Donoho D, Pauly J (2017) “Recurrent generative adversarial networks for proximal learning and automated compressive image recovery,” <http://arxiv.org/abs/1711.10046>
209. Salehinejad H, Sankar S, Barfett J, Colak E, Valaee E (2017) “Recent advances in recurrent neural networks,” <http://arxiv.org/abs/1801.01078>
210. Venkatraman A, Rhinehart N, Sun W, Pinto L, Hebert M, Boots B, Kitani K, Bagnell J (2017) “Predictive-state decoders: Encoding the future into recurrent networks,” In: *Advances in Neural Information Processing Systems*, pp. 1172–1183
211. Gregor K, Papamakarios G, Besse F, Buesing F, Weber T (2018) “Temporal difference variational auto-encoder,” <http://arxiv.org/abs/1806.03107>
212. Gupta A, Johnson J, Fei-Fei L, Savarese S, Alahi A (2018) “Social gan: Socially acceptable trajectories with generative adversarial networks,” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2255–2264
213. Kuefler A, Morton J, Wheeler T, Kochenderfer M (2017) “Imitating driver behavior with generative adversarial networks,” In: *2017 IEEE Intelligent Vehicles Symposium (IV)*, pp. 204–211, IEEE
214. Doya K, Samejima K, Katagiri K-I, Kawato M (2002) Multiple model-based reinforcement learning. *Neural Comput* 14(6):1347–1369
215. Clavera I, Rothfuss J, Schulman J, Fujita Y, Asfour T, Abbeel P (2018) “Model-based reinforcement learning via meta-policy optimization,” <http://arxiv.org/abs/1809.05214>
216. Vuong T-L, Tran K (2019) “Uncertainty-aware model-based policy optimization,” <http://arxiv.org/abs/1906.10717>
217. Neerincx MA, van der Waa J, Kaptein F, van Diggelen J (2018) “Using perceptual and cognitive explanations for enhanced human-agent team performance,” In: *International Conference on Engineering Psychology and Cognitive Ergonomics*, pp. 204–214, Springer
218. Molineaux M, Kuter M, Klenk M (2011) “What just happened? explaining the past in planning and execution,” tech. rep., *NAVAL RESEARCH LAB MONTEREY CA*
219. Friedman S, Forbus KD, Sherin B (2011) “Constructing and revising commonsense science explanations: A metareasoning approach,” In: *2011 AAAI Fall Symposium Series*
220. Molineaux M, Aha DW, Kuter U (2011) “Learning event models that explain anomalies,” tech. rep., *NAVY CENTER FOR APPLIED RESEARCH IN ARTIFICIAL INTELLIGENCE WASHINGTON DC*
221. Adam C, Gaudou B (2016) BDI agents in social simulations: a survey. *Knowl Eng Rev* 31(3):207–238
222. Santos JS, Zahn JO, Silvestre EA, Silva VT, Vasconcelos WW (2017) Detection and resolution of normative conflicts in multi-agent systems: a literature survey. *Autonomous agents and multi-agent systems* 31(6):1236–1282
223. Hollander CD, Wu AS (2011) The current state of normative agent-based systems. *J Artif Soc Soc Simul* 14(2):6
224. Beheshti R (2014) “Normative agents for real-world scenarios,” In: *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 1749–1750, International Foundation for Autonomous Agents and Multi-agent Systems
225. Myerson RB (2013) *Game theory*. Harvard University Press, Harvard
226. Camerer CF (2011) *Behavioral game theory: experiments in strategic interaction*. Princeton University Press, Princeton
227. Suleiman R, Troitzsch KG, Gilbert N (2012) *Tools and techniques for social science simulation*. Springer Science & Business Media, Cham
228. Silver D, Hassabis D (2016) *AlphaGo: mastering the ancient game of go with machine learning*. Res Blog 9:7
229. Marinier RP, Laird JE (2008) “Emotion-driven reinforcement learning,” In: *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 30
230. Elliott R (1998) A model of emotion-driven choice. *J Market Manage* 14(1–3):95–108

231. Marinier RP III, Laird JE, Lewis RL (2009) A computational unification of cognitive behavior and emotion. *Cognit Syst Res* 10(1):48–69
232. Hoey J, Schröder T, Alhothali A (2016) Affect control processes: intelligent affective interaction using a partially observable markov decision process. *Artif Intell* 230:134–172
233. Gadanho SC, Hallam J (2001) Robot learning driven by emotions. *Adapt Behav* 9(1):42–64
234. Yu H, Yang P (2019) “An emotion-based approach to reinforcement learning reward design,” In: 2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC), pp. 346–351, IEEE
235. Castelfranchi C (1998) Modelling social action for AI agents. *Artif Intell* 103(1–2):157–182
236. Conte R, Castelfranchi C et al (2016) *Cognitive and social action*. Garland Science, New York
237. Poggi I, D’Errico F (2010) “Cognitive modelling of human social signals,,” In: *SSPW@ MM*, pp. 21–26
238. Charalampous K, Kostavelis I, Gasteratos A (2017) Recent trends in social aware robot navigation: a survey. *Robot Autonomous Syst* 93:85–104
239. Chen YF, Everett M, Liu M, How JP (2017) “Socially aware motion planning with deep reinforcement learning,” In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1343–1350, IEEE
240. Triebel R, Arras K, Alami R, Beyer L, Breuers S, Chatila R, Chetouani M, Cremers D, Evers V, Fiore M, et al (2016) “Spencer: A socially aware service robot for passenger guidance and help in busy airports,” In: *Field and service robotics*, pp. 607–622, Springer
241. Kim B, Pineau J (2016) Socially adaptive path planning in human environments using inverse reinforcement learning. *Int J Soc Robot* 8(1):51–66
242. Vasquez D, Okal B, Arras KO (2014) “Inverse reinforcement learning algorithms and features for robot navigation in crowds: an experimental comparison,” In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1341–1346, IEEE
243. Ritschel H (2018) “Socially-aware reinforcement learning for personalized human-robot interaction,” In: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1775–1777, International Foundation for Autonomous Agents and Multiagent Systems
244. Kulkarni A, Zha Y, Chakraborti T, Vadlamudi SG, Zhang Y, Kambhampati S (2019) “Explicable planning as minimizing distance from expected behavior,” In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2075–2077, International Foundation for Autonomous Agents and Multiagent Systems
245. Anderson AA (2019) Mental models of mere mortals with explanations of reinforcement learning
246. Thagard P (1989) Explanatory coherence. *Behav Brain Sci* 12(3):435–467
247. Tabrez A, Hayes B (2019) “Improving human-robot interaction through explainable reinforcement learning,” In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 751–753, IEEE
248. Lakkaraju H, Kamar E, Caruana R, Leskovec J (2017) “Interpretable & explorable approximations of black box models,” <http://arxiv.org/abs/1707.01154>

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.