Journal of Ambient Intelligence and Humanized Computing manuscript No.
(will be inserted by the editor)

A Conceptual Framework for Externally-influenced Agents: An Assisted Reinforcement Learning Review

Adam Bignold[1,*] · **Francisco Cruz**[2,3,*] · **Matthew E. Taylor**[4] · **Tim Brys**[5] · **Richard Dazeley**[2] · **Peter Vamplew**[1] · **Cameron Foale**[1]

Received: date / Accepted: date

Abstract A long-term goal of reinforcement learning agents is to be able to perform tasks in complex real-world scenarios. The use of external information is one way of scaling agents to more complex problems. However, there is a general lack of collaboration or interoperability between different approaches using external information. In this work, while reviewing externally-influenced methods, we propose a conceptual framework and taxonomy for assisted reinforcement learning, aimed at fostering collaboration by classifying and comparing various methods that use external information in the learning process. The proposed taxonomy details the relationship between the external information source and the learner agent, highlighting the process of information decomposition, structure, retention, and how it can be used to influence agent learning. As well as reviewing state-of-the-art methods, we identify current streams of reinforcement learning that use external information in order to improve the agent's performance and its decision-making process. These include heuristic reinforcement learning, interactive reinforcement learning, learning from demonstration, transfer learning, and learning from multiple sources, among others. These streams of reinforcement learning operate with the shared objective of scaffolding the learner agent. Lastly, we discuss further possibilities for future work in the field of assisted reinforcement learning systems.

Keywords Assisted reinforcement learning · Externally-influenced agents · Assistance taxonomy.

1 Introduction

Reinforcement learning (RL) (Sutton and Barto, 2018) is a learning approach in which an agent uses sequential decisions to interact with its environment trying to find a (near-) optimal policy to perform an intended task. RL agents have the ability to improve while operating, to learn without supervision, and to adapt to changing circumstances (Kaelbling et al, 1996). By exploring, a standard agent learns solely from the signals it receives from the environment. The RL approach has shown success in domains such as robotics (Kitano et al, 1997; Kober et al, 2013; Cruz et al, 2018c; Contreras et al, 2020), game-playing (Tesauro, 1994; Barros et al, 2020), inventory management (Giannoccaro and Pontrandolfo, 2002), and cloud computing (Shakarami et al, 2020; Shahidinejad and Ghobaei-Arani, 2020; Ghobaei-Arani et al, 2018), among others.

Like many machine learning techniques, RL faces the problem of high-dimensionality spaces. As environments

This work has been partially supported by the Australian Government Research Training Program (RTP) and the RTP Fee-Offset Scholarship through Federation University Australia. Moreover, this work has taken place in part in the Intelligent Robot Learning Lab at the University of Alberta, which is supported in part by research grants from the Alberta Machine Intelligence Institute (Amii); CIFAR; a Canada CIFAR AI Chair, Amii; Compute Canada; and NSERC.

[1] School of Engineering, IT and Physical Sciences, Federation University, Ballarat, Australia.
[2] School of Information Technology, Deakin University, Geelong, Australia.
[3] Escuela de Ingeniería, Universidad Central de Chile, Santiago, Chile.
[4] Department of Computing Science and The Alberta Machine Intelligence Institute (Amii), University of Alberta, Edmonton, AB, Canada.
[5] Action Research Associates, Beirut, Lebanon.
[*] Both authors contributed equally to this manuscript.
Corresponding e-mails:
{a.bignold, p.vamplew, c.foale}@federation.edu.au,
{francisco.cruz, richard.dazeley}@deakin.edu.au,
matthew.e.taylor@ualberta.ca,
tbrys@actionresearchassociates.org.

Cite as: Bignold, A., Cruz, F., Taylor, M., Brys, T., Dazeley, R., Vamplew, P., Foale, C. A Conceptual Framework for Externally-influenced Agents: An Assisted Reinforcement Learning Review. Journal of Ambient Intelligence and Humanized Computing (JAIHC). In press (2021). This version of the article has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1007/s12652-021-03489-y. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/acceptedmanuscript-terms

become larger, the agent's learning time increases and finding the optimal solution becomes impractical (Cassandra and Kaelbling, 2016). Early research on this topic (Kaelbling et al, 1996; Lin, 1991) argued that for RL to successfully scale into real-world scenarios, then the use of information external to the environment would be needed. Different RL strategies using this approach have emerged in order to speed up the learning process. They use external information to assist either the process of generalising the environment representation (Price and Boutilier, 2003), the agent's decision-making process (Griffith et al, 2013), or in providing more focused exploration (Fernández and Veloso, 2006).

In this article, we refer to *external information* as any kind of information provided to the agent originating from outside of the agent's representation of the environment. This may include demonstrations (Konidaris et al, 2012; Rozo et al, 2013; Chen et al, 2019), advice and critiques (Knox and Stone, 2010; Griffith et al, 2013), initial bias based on previously gathered data (Taylor and Stone, 2009), or highly-detailed domain-specific shaping functions (Randløv and Alstrøm, 1998). Additionally, in this work, we use independently the concepts of RL approach, method, and technique to refer to the underlying learning algorithm. These concepts have been previously used mostly equally by the RL research community.

In this regard, we define *Assisted reinforcement learning* (ARL) as a range of techniques that use external information, either before, during, or after training, to improve the performance of the learner agent, as well as to scale RL to larger and more complex scenarios. While a relevant characteristic of RL is its ability to endow agents with new skills from the ground up, ARL also makes use of existing information and/or previously learned behaviour. Some methods for improving the agent's performance using external information include: directly altering weights for actions and states (biasing) (Vlassis et al, 2012); altering the state or action space (Erez and Smart, 2008); critiquing past or advising on future decision-making (Thomaz and Breazeal, 2007); dynamically altering reward functions (Knox and Stone, 2010); directly modifying the policy (Griffith et al, 2013); guiding exploration and action selection (Fernández and Veloso, 2006); and, creating information repositories/models to supplement the environmental information (Price and Boutilier, 2003). Figure 1 captures all of these methods in a basic view of the ARL conceptual framework used in this work. The classic RL approach is shown within the figure where an agent performs an action on the environment reaching a new state and obtaining a reward. In ARL, the response of the environment is also shared with the external information source from where advice is given to the agent or changes sometimes made directly to the environment (Xu et al, 2020).

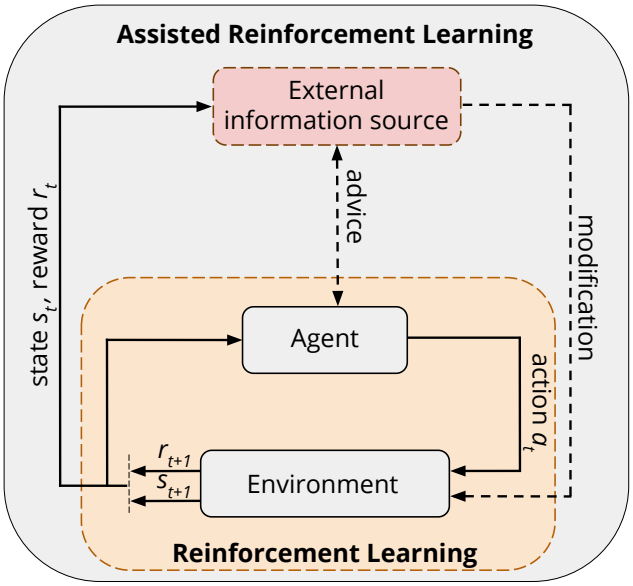

Fig. 1 Assisted reinforcement learning simplified framework. In autonomous reinforcement learning, an agent performs an action a_t from a state s_t and the environment produces an answer leading the agent to a new state s_{t+1} and receiving a reward r_{t+1}. Assisted reinforcement learning adds an external information source, referred to as a trainer, teacher, advisor or assistant, that observes the environment and the agent in order to generate advice. The trainer may advise the learner agent or sometimes directly modify the environment. Moreover, the agent may also actively ask advice to the external information source.

To date, many methods using external information have been proposed aiming to speed up the learning process for an autonomous agent (Arzate Cruz and Igarashi, 2020; Lin et al, 2020; Da Silva et al, 2020b; Zhuang et al, 2020). Usually, they have been organized according to the technique employed, e.g., heuristic, interactive, or transfer learning, among others. Nevertheless, there is an important lack of understanding of how these techniques are related and what characteristics they share. Therefore, in this review, we present a conceptual framework and a taxonomy to be used to describe the practice of using external information. A standardised ARL taxonomy will foster collaboration between different RL communities, improve comparability, allow a precise description of new approaches, and assist in identifying and addressing key questions for further research.

2 A Conceptual Framework for Assisted Reinforcement Learning

In this section, we give more details about the ARL approach including some introductory examples of works in which external information sources have been used. Moreover, we define a conceptual framework identifying the different parts that comprise the underlying process used in ARL techniques. Based on this conceptual framework, in the following section, we define a more detailed taxonomy for ARL approaches.

2.1 Assisted Reinforcement Learning

The main strength of RL is its ability for endowing an agent with new skills given no initial knowledge about the environment. With an appropriate reward function and enough interaction with its environment, an RL agent can learn (near-) optimal behaviour (Sutton and Barto, 2018). The agent's behaviour at every step is defined by its policy. The reward function promotes desirable behaviour and sometimes penalises undesirable behaviour. In the traditional view of RL, the reward function, and the rewards it produces, are internal to the environment (Kaelbling et al, 1996). Traditional RL, in which the environment is the sole provider of information to the agent, has been demonstrated to perform well in many different domains, especially when facing small and bounded problems (Sutton and Barto, 2018). However, RL has some difficulties when scaling up to large, unbounded environments, particularly regarding the time needed for the agent to learn the optimal policy (Cruz et al, 2016a, 2018b). In RL, one approach to tackling this issue is to use external information to supplement the information that the environment provides (Suay and Chernova, 2011; Millán et al, 2019).

Information is considered external if it originates from outside of the agent's interactions with the environment. In this regard, internal information is determined solely through interactions and observations with the environment. For example, in the case of a human the internal information would be anything the person can observe from the environment using their senses (Niv, 2009). The external information would be any information provided by peers, advisors, the internet, books, maps, and tutelage. In RL, anything external to the agent is usually considered part of the environment. In this regard, if an agent is learning in an environment, a person can be considered as part of it, therefore, the agent could model that person or communicate with them (Sert et al, 2020). Although it is possible that external sources of information could be just treated as part of the environment, this is handicapping the agent in an unnecessary way. There are external sources of information that might not necessarily be treated as part of the environment because they are socially advantaged. For instance, if an external source is providing action advice using directions as 'left' and 'right', the agent does not have to learn the meaning of these words from the ground up, or learn how to react to these instructions. Instead, we assume the agent knows that advice is coming, what it means, and how to use it. For example, if a person eats some berries and later becomes sick, the person may determine that those berries are poisonous. In this case, this would be internal information obtained by interaction with the environment. If instead, a peer had previously advised the person that eating those berries will make them sick, that would be external information provided by an extrinsic source.

In this work, we refer to methods using externally-influenced agent learning as as assisted reinforcement learning. The ARL framework is defined to include any type of RL that uses external information to supplement agent learning and the decision-making process. Some common practices include the direct alteration of the agent's understanding of the environment (Price and Boutilier, 2003), focusing exploration efforts through critique and advice (Thomaz and Breazeal, 2007), or assisting the agent in the decision-making process (Fernández and Veloso, 2006). For instance, existing ARL techniques include interactive reinforcement learning (Amershi et al, 2014; Cruz et al, 2017), learning from demonstration (Argall et al, 2007; Nair et al, 2018), and transfer learning (Taylor and Stone, 2009; Shao et al, 2018), among others.

The previously mentioned RL approaches are just examples of ARL methods that use external information to supplement the agent's decision-making process and learning. Additional details of these and other approaches and how they use an external information source to assist the agent (in terms of our ARL framework) are addressed in Section 4. The external information source is most commonly a human or another artificial agent. Regardless of the source, the use of external information has often been shown to improve an agent's ability and learning speed. In the next section, we present a more detailed conceptual framework for ARL which is the base for the taxonomy we propose subsequently.

2.2 Conceptual Framework

The proposed ARL framework is built to improve the classification, the comparability, and the discussion on different externally-influenced RL methods. To achieve

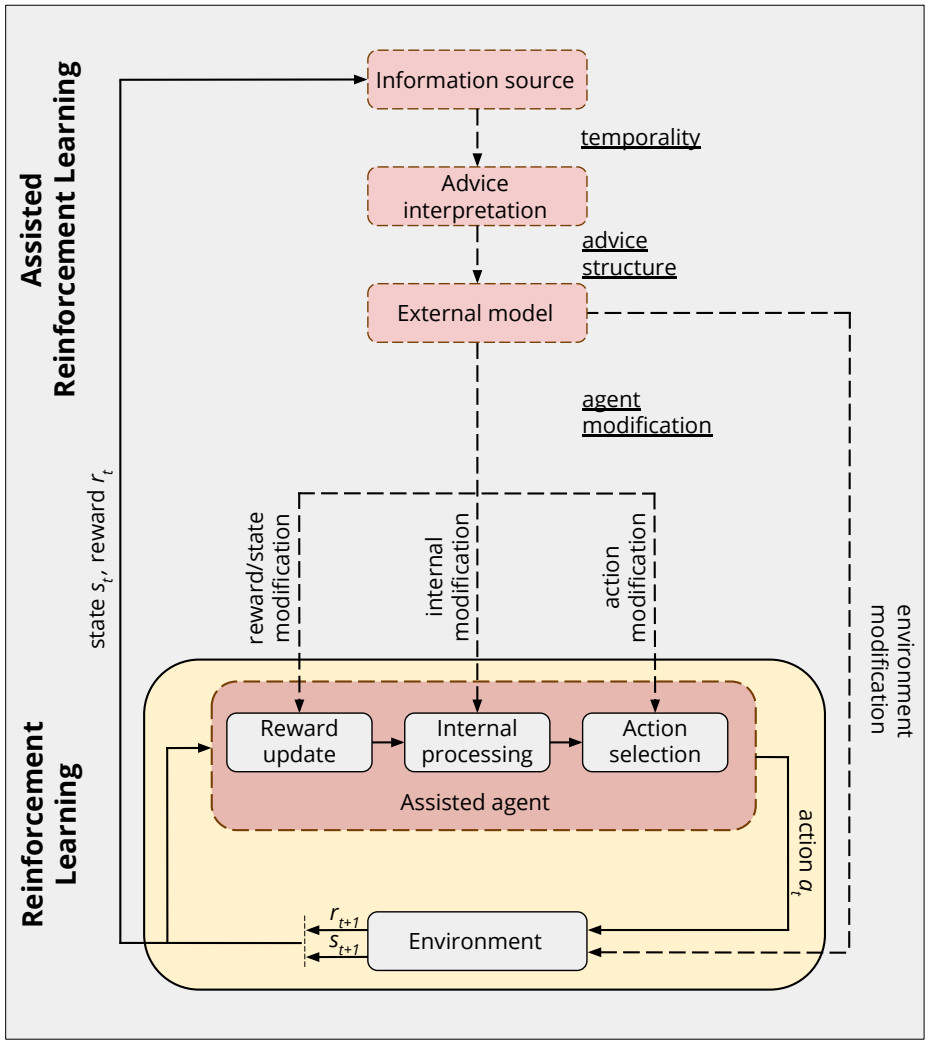

Fig. 2 Detailed view of the assisted reinforcement learning framework. The diagram includes four processing components shown as dashed red boxes. Inside the assisted agent, one can observe three different points where it can receive possible modifications from the external model. Additionally, three communication links are shown with underlined text. This framework is subsequently used to further discuss the proposed ARL taxonomy.

this aim, the framework has been designed using insights and observations drawn from many different ARL approaches. The result is a framework that can describe existing methods while also being flexible enough to include future research. The framework details are shown in Figure 2.

The proposed ARL framework comprises four processing components shown using red boxes in the diagram, i.e., information source, advice interpretation, external model, and the assisted agent itself. The external information source may not have perfect observability and also may not know details about the RL agent (algorithms, weights, hyperparameters, etc.), or make assumptions, e.g., value-based learners (Taylor et al, 2005). The processing components are responsible for providing, transforming, and storing information. We do include the agent as part of the processing components since it is part of the RL process as well. However, an agent using ARL generally behaves as a traditional RL agent, i.e., it interacts with the environment by exploring/exploiting actions. Inside the agent, there are three different stages: reward update, internal processing, and action selection. Each of those stages may be altered by the external model using reward/state modifications, internal modifications, or action modifications respectively. Moreover, the ARL framework also comprises three communication links that connect the four processing components and are labelled: temporality, advice structure, and agent modification. These links are shown between the processing components and represent the communication lines in Figure 2 that connect the processing components together. The communication links

convey information or denote constraints on the data such as where or when to provide information.

The ARL framework describes the transmission, modification, and modality of sourced information. In this regard, we consider the ARL framework as a whole unit, comprising traditional autonomous RL plus the components and links for assistance. Thus, the taxonomy is a part of the framework and oriented to describe the assisted learning section. Although the framework has been developed on how ARL is usually built, not all ARL approaches use all the proposed components and links. Below, we briefly describe each of the components and links of the framework. They are subsequently used in the next section to describe in detail the proposed taxonomy.

- **Information source**: is the origin of the assistance being provided to the agent. The source may be a human, a repository, or another agent. There may be multiple information sources providing assistance to an agent.
- **Temporality**: determines both the time at which information is provided to the agent, and the frequency with which it is provided. Information may be provided, before, during, or after agent training, and occur multiple times through the learning process. Therefore, it is also responsible for how the information source communicates temporal issues to the advice interpreter.
- **Advice interpretation**: denotes the process of transforming incoming information into a format better suited for the agent. This may involve extracting key frames from video, converting audio samples to rewards, or mapping information to states.
- **Advice structure**: represents the structure of the advice after translation in a form suitable for the external model. Some approaches may not have an explicit external model, therefore, this structure might instead be directly used to modify the agent.
- **External model**: is responsible for retaining and relaying the information between the source and the agent. The model may retain the received information in the learning model, using it for later decisions, or it may discard the received information as soon as it has been used.
- **Agent modification**: denotes the approach that the agent uses to benefit from the incoming information. The most common modification approaches may use information to alter the environmental reward signal or modify the agent's behaviour or the decision-making process directly.
- **Assisted Agent**: is the RL agent receiving the external information or advice while learning a new task. The agent needs to work out how to incorporate the provided information with its own learning. If a different action is suggested by the trainer then the agent may decide if it should follow to that advice or not.

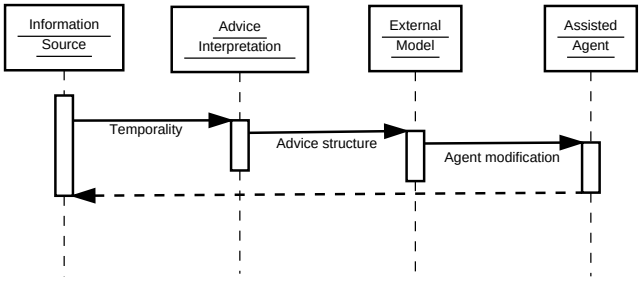

Fig. 3 Relation between the processing components and the communication links as a UML sequence diagram.

Figure 3 shows in a UML sequence diagram the interaction between the processing components and communication links according to Figure 2.

3 Assisted Reinforcement Learning Taxonomy

In this section, we describe the processing components and communication links included in the proposed framework within an ARL taxonomy[1] and give more details of each of them. Figure 4 shows all the elements of the proposed ARL taxonomy including examples for each processing component and communication link. In the taxonomy, we include the agent as a component being the one that receives the advice. Each of the seven elements, i.e., processing components and communication links, is described in detail in the following subsections. In our work, the concept of taxonomy is used to classify the different elements within a class of problems, i.e., ARL problems. In this regard, our proposal is represented by a general ontology where the class is ARL, the properties are the processing components and the communication links, and the relations between the properties are as shown in Figure 4.

3.1 Information Source

The external information source is the main factor that sets ARL apart from traditional RL approaches. It is responsible for introducing new information about the task to the agent, supplementing or replacing the information the agent receives from the environment. The

[1] In this context, we refer the taxonomy as a classification of the different elements of the ARL framework, i.e., processing components and communication links, and not as a way to classify each ARL method.

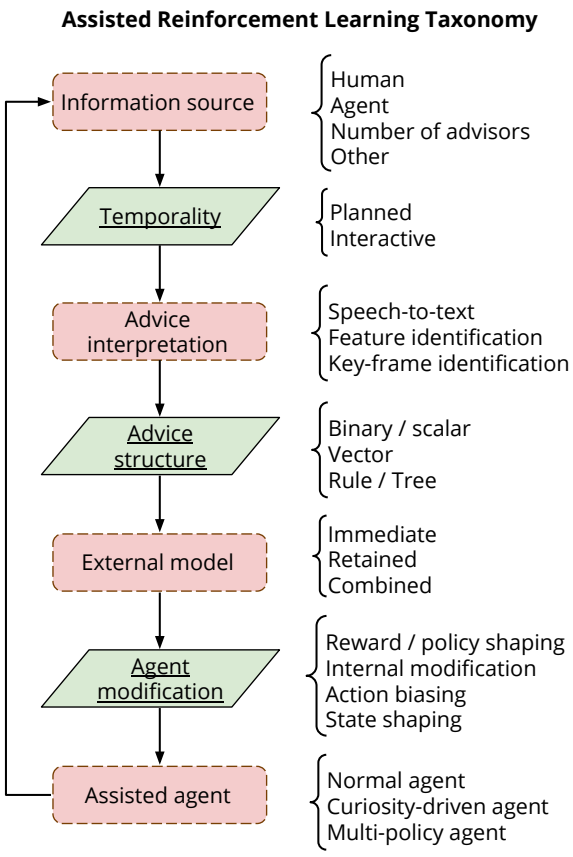

Fig. 4 The assisted reinforcement learning taxonomy. This figure shows the four processing components as dashed red boxes and the communication links as green parallelograms using underlined text. Examples for each component and method are included at the right.

source is external to the agent and the environment, providing information that either the agent may not have had access to, or would have eventually learned itself. The information source may be able to observe the environment, the agent, or the agent's decision-making process. The objective of the information source is to assist the agent in achieving its goal faster.

There may be multiple information sources communicating with an agent. This may be humans, agents, other digital sources, or any combination of the three (Isbell et al, 2000). The use of multiple sources offers a wider range of available information to the agent. However, more complex modification methods may be required to manage the information and handle conflicting advice (Kamar et al, 2012).

There are many examples of external information sources in current ARL literature, the most common of which are humans and additional reward functions (Ng et al, 1999; Thomaz et al, 2006a; Millán et al, 2019). For instance, RLfD and IntRL use human guidance to provide the agent with a generalised view of the solution (Cobo et al, 2014; Subramanian et al, 2016). Moreover, the use of additional reward functions is one of the earliest examples of ARL. In such cases, the designer of the agent encodes some further information about the environment or goal as an additional reward, supplementing the original reward given by the environment.

An example of the use of additional reward functions can be found in Randløv and Alstrøm's bicycle experiment (Randløv and Alstrøm, 1998), in which, they teach an agent to ride a bicycle towards a goal point. Without additional assistance, the RL agent would only receive a reward upon reaching the termination state. Randløv and Alstrøm encoded some of their knowledge as a shaping reward signal external to the environment, providing the agent with additional rewards if it is cycling towards the goal point. In this scenario, the system designers acted as an external information source, providing extra information to the RL agent. The use of this external information results in the agent learning the solution faster than using the traditional RL approach.

Some other information sources include behaviours from past experiences or other agents, repositories of labelled data or examples, or distribution tables for initialising/biasing agent behaviour (Cruz et al, 2017). Video, audio, and text sources may be used as well (Cruz et al, 2016b). However, these sources may require substantial amounts of interpretation and preprocessing to be of use.

The accuracy, availability, or consistency of the information source can affect the maximum utility of the information (Torrey and Taylor, 2013; Taylor et al, 2014). Identifying in advance inaccurate information given to the agent can significantly improve performance (Cruz et al, 2016a, 2018a). While the information source may perform the validation and the verification of the given advice, the primary duty remains simply to act as a supplementary source of information. In this regard, both validation and verification of information are functions better suited for the external model or the assisted agent.

3.2 Temporality

The temporal component, or temporality, refers to the time at which information is communicated by the information source. The information may be provided in full to the agent at a set time (either before, during, or after training). This is referred to as *planned assistance* (Partalas et al, 2008; Cheng et al, 2013). Alternatively, the information may be provided at any

time during the agent's operation, referred to as *interactive assistance* (Pilarski and Sutton, 2012; Stahlhut et al, 2015).

Planned assistance, on the one hand, is common in ARL methods. Some examples are predefined additional shaping functions, agent policy initialisation based on either prior experience or a known distribution, and the creation of subgoals that lead the way to a final solution (Partalas et al, 2008). These methods let the experiment designer endow the agent with initial information about the environment or the goal to be achieved. By providing this initial knowledge, the designer can reduce the agent's need for exploration.

The bicycle experiment discussed in the previous section is an example of planned assistance. As mentioned, the agent is learning to control a bicycle and must learn to steer it towards a goal (Randløv and Alstrøm, 1998). Before the experiment, the designers give the agent additional information in the form of a reward signal that correlates to the direction of the goal state. This planned assistance approach helps the agent to narrow the search space by giving it extra information about the environment. This small yet beneficial initial information results in a significant improvement in the agent's learning speed.

Another example of planned assistance is found in heuristic RL. Heuristic RL is a method of applying advice to agent decision-making. One example is an experiment which implements heuristic RL in the RoboCup soccer domain (Celiberto Jr et al, 2007), a domain known for its large state space and continuous state range. In this environment, one team attempts to score a goal, while the other team tries to block the first team from scoring, such as in half-field offence (Kalyanakrishnan et al, 2006; Hausknecht et al, 2016). In this experiment using heuristic RL, the defending team is given initial advice before training. This advice consists of two rules: if the agent is not near the ball then move closer, and if the agent is near the ball then do something with it. The experiment results show that a team that uses planned assistance performs better than a team that is given no initial knowledge (Celiberto Jr et al, 2007).

Interactive assistance, on the other hand, refers to information provided by the source repeatedly throughout the agent's learning. Information sources that assist interactively often can observe the agent's current state, or the environment the agent is operating in. In current literature, humans are more commonly used as information sources for interactive assistance (Thomaz et al, 2006b; Subramanian et al, 2011). The human can observe how the agent is performing and its current state in the environment, and provides guidance or critiques of the agent's behaviour (Bignold et al, 2020).

For example, Sophie's Kitchen (Thomaz and Breazeal, 2007) presents an IntRL based agent, called Sophie, which attempts to bake a cake by interacting with the items and ingredients found in a kitchen. In this experiment, the agent will receive a reward if it successfully bakes the cake. At any point during the agent's training, an observing human can provide the agent with an additional reward to supplement the reward signal given by the environment. If the agent performs an undesirable action, such as forgetting to add eggs to the cake, the human can punish the agent by providing an immediate negative reward. The human can also reward the agent for performing desirable actions, such as adding ingredients in the correct order. In this experiment, the human advisor is acting as an interactive information source.

Although the agent could learn the task without any assistance, the addition of the human advisor and interactive feedback allows the agent to learn the desired behaviour faster in comparison to autonomous RL (Thomaz and Breazeal, 2007). The benefit of using interactive advice rather than planned advice is that the information source can react to the current state of the agent. Additionally, an interactive information source does not need to encode all possibly useful advice up front. Instead, it can choose to provide relevant information only when required. This approach does have a significant cost; the information source needs to be constantly observing the agent and determining what information is relevant. For instance, an approach using inverse RL through demonstrations may also consider providing failed examples to show the agent what not to do (Shiarlis et al, 2016).

3.3 Advice Interpretation

The advice interpretation stage of the taxonomy denotes what transformations need to occur on the incoming information. The source provides information for the agent to use that may need to be translated into a format that the agent can understand. The information source may provide their assistance in many different forms. Some examples include audio (Cruz et al, 2015), video (Cruz et al, 2016b), text (Liu et al, 2019), distributions and probabilities (Millán et al, 2019), or prior learned behaviour from a different task or agent (Da Silva et al, 2020b). This information needs to be adapted for use by the agent for the current task. The product of the advice interpretation stage depends on the structure that the agent or external model requires.

A field where the interpretation of incoming advice is crucial is Transfer Learning (TL). The goal of TL is to use behaviour learned in a prior task to improve performance in a new, previously unseen task (Da Silva

and Costa, 2019). A critical step in TL is the mapping of states and observations between the old and new domains. The information source provides information to the agent that does not fully align with its current task. Therefore, it is crucial that the information provided can be correctly interpreted, so as to be useful to the current domain. More commonly, this interpretation stage in TL is performed by hand. However, there has also been effort attempting to automate this stage (Taylor et al, 2008; Narvekar et al, 2016).

Another example of the use of the advice interpretation stage is with the sourcing of feedback for RL agents. In the Sophie's Kitchen experiment (Thomaz and Breazeal, 2007), discussed in the previous section, the agent can be given positive or negative feedback by a human regarding its choice of actions. In this experiment, the human creates either a green (positive) or a red (negative) bar to represent the desired feedback to be given to the agent. This bar is used to interpret the reward signal to give to the agent, with the colour of the bar designating whether the reward is positive or negative, and the size of the bar designating the magnitude of the reward. This type of feedback can also be extended to audio, where recording phrases such as 'Good' or 'Well Done' are interpreted as positive rewards and 'Bad' or 'Try Again' are interpreted as negative rewards (Tenorio-Gonzalez et al, 2010).

These methods can also be combined into a multi-model architecture to provide advice to an RL robotic agent using audiovisual sensory inputs, such as work by Cruz et al. (Cruz et al, 2016b). In this experiment, a simulated robot learns how to clean a table using a multi-modal associative function to integrate auditory and visual cues into a single piece of advice which is used by the RL algorithm. In this scenario, the external information source is a human trainer and the RL algorithm represents the integrated advice as a state-action pair.

3.4 Advice Structure

The advice structure component refers to the form that the agent or external model requires incoming information to take. The information that the agent uses can be represented in a number of ways. Some examples of advice structures include: Boolean values denoting positive or negative feedback; rules determining action selection; matrices for mapping prior experiences to new states; case-based reasoning structures for the agent to consult with; or, hierarchical decision trees to represent options for the agent to take (Subramanian et al, 2011; Kaplan et al, 2002).

The simplest form of structure is binary, in which the information takes only one from two options, such as 'Good' or 'Bad'. An example of the use of a binary structure is the TAMER-RL agent (Knox and Stone, 2009). TAMER-RL is an IntRL agent that uses binary feedback from an observing human. At any time step, the human can agree or disagree with the agent about its last action. In this case, the feedback is a binary structure indicating agree or disagree.

A more complex advice structure is used in case-based RL agents (Sharma et al, 2007). A case in this context represents a generalised area of the state space and provides information about which actions to take in that state. The use of a case-based structure allows the agent to gain more information from the information source compared to a binary structure, at a cost of more complex sourcing and interpretation approaches.

One of the more common advice structures is a simple state-action pair. A state-action pair consists of a single state and an associated piece of advice. The associated advice may be an additional scalar reward or a recommended action. Using a state-action pair, sourced information is interpreted to provide advice for a given state. In the cleaning-table robot task (Cruz et al, 2016b), discussed in the previous section, the external trainer using multi-modal advice provides an action to be performed in specific states. Once the advice is processed using the multi-modal integration function, the proposed action is given to the RL agent to be executed as a state-action pair considering the agent's current state. This state-action structure has also been used for other methods including TAMER-RL (Knox and Stone, 2009), Sophie's Kitchen (Thomaz and Breazeal, 2007), and policy-shaping approaches (Griffith et al, 2013).

A novel rule-based interactive advice structure is introduced in (Bignold et al, 2021b). Interactive RL methods rely on constant human supervision and evaluation, requiring a substantial commitment from the advice-giver. This constraint restricts the user to providing advice relevant to the current state and no other, even when such advice may be applicable to multiple states. Allowing users to provide information in the form of rules, rather than per-state action recommendations, increases the information per interaction, and does not limit the information to the current state. Rules can be interactively created during the agent's operation and be generalised over the state space while remaining flexible enough to handle potentially inaccurate or irrelevant information. The learner agent uses the rules as persistent advice allowing the retention and reuse of the information in the future. Rule-based advice significantly reduces human guidance requirements while improving agent performance.

3.5 External Model

The external model is responsible for retaining and relaying information between the information source and the agent. The external model receives interpreted information from the information source and may either retain the information for use by the agent when required or pass it to the agent immediately.

A *retained model* is an external model that stores all information provided by the information source (Fernández and Veloso, 2006). A retained model may be used if the cost of acquiring information is greater than the cost of storing it, if the information provided is general or applies to multiple states, or if the information is gathered incrementally. In instances where information is gathered incrementally, using a retained model allows the agent to build up a knowledge base over time. The agent may consult with the model at any time to determine if a reward signal is to be altered, or if there is any extra information that may assist with decision-making.

An *immediate model* passes the information directly to the agent (Moreira et al, 2020). In this case, the information received is only relevant to the current time step, or the cost of reacquiring the information from the source is less than that of retaining the information.

Approaches can also combine this by incorporating both a retained model as well as passing some information through directly, such as (Cruz et al, 2016a). In this work, an RL agent uses a combination of interactive feedback and contextual affordances (Cruz et al, 2016c) to speed up the learning process of a robot performing a domestic task. On the one hand, contextual affordances are learned at the beginning of autonomous RL and are readily available from there on to avoid the so-called failed-states, which are states from where the robot is not able to finish the task successfully anymore. On the other hand, interactive feedback is provided by an external advisor and used to suggest actions to perform when the robot is learning the task. This advice is given to the robot to be used in the current state and it is discarded immediately after.

The external model may have different functions depending on its implementation. For instance, heuristic RL hosts a model that stores rules and advice that generalise over sections of the state space (Dorigo and Gambardella, 2014). In TL, the external model may hold information regarding past experiences and policies from problems similar to the current domain (Taylor and Stone, 2009; Banerjee, 2007), or in inverse RL, the external model is a substitute for the reward function (Abbeel and Ng, 2004).

3.6 Agent Modification

The modification stage of the framework denotes how the information that the external model contains is used to assist the agent in achieving its goal. It is responsible for supplementing the agent's reward, altering the agent's policy, or helping with the decision-making process. A popular method for injecting external information into agent learning is shaping (Skinner, 1975). Shaping is a common method for altering agent performance by modifying parameters in the learning process. Erez and Smart (Erez and Smart, 2008) propose a list of techniques in which shaping can be applied to RL agents. These include altering the reward, the agent's policy, agent learning parameters, and environmental dynamics (Xu et al, 2020).

Altering the reward the agent receives is a straightforward method for influencing an agent's learning (Churamani et al, 2016). It is known as reward-shaping, in which the external information is used to bias the agent's learning (Ng et al, 1999). Special care must be taken to ensure that any modification of the reward signal remains zero-sum to avoid the agent exploiting the shaped reward in ways that do not align with the desired goal. This can be achieved by ensuring that additional rewards are potential-based, meaning that they are derived from the difference in the values of a potential function at the current and successor states (Harutyunyan et al, 2015). However, recent work by (Behboudian et al, 2020) shows a flaw in the previous method when transforming non-potential-based reward-shaping into potential-based. Alternatively, the authors introduce a policy invariant explicit shaping algorithm allowing for arbitrary advice, confirming that it ensures convergence to the optimal policy when the advice is misleading and also accelerates learning when the advice is useful (Behboudian et al, 2020). Shaping techniques have also been used to alter state-action pairs (Wiewiora et al, 2003), for dynamic situations (Harutyunyan et al, 2015; Devlin and Kudenko, 2012), and for multi-agent systems (Devlin and Kudenko, 2011).

Policy-shaping is the modification of the agent's behaviour (Griffith et al, 2013). This modification can be done either by influencing how the agent makes decisions or by directly altering the agent's learned behaviour. A simple method of policy-shaping involves forcing it to take certain actions if advice from the information source has recommended them (Grizou et al, 2013; Navidi, 2020). Human-in-the-loop techniques may be beneficial to address complex RL problems with the help of domain experts, e.g., in health informatics (Holzinger, 2016). This allows the external information source to guide the agent and take direct control

over exploration/exploitation. Alternatively, the information source can choose to alter the agent's behaviour directly by changing Q-values or installing rules that override the actions for chosen states (Knowles and Wermter, 2008). This method of modification can improve agent performance rapidly, as it can give the agent partial solutions.

Internal modification is a method of altering the parameters of the agent that are essential to its learning. Parameters such as the learning rate (α), discount factor (γ), and exploration percentage (ϵ), are all internal to the RL agent and may be altered to affect its performance (Tesauro, 2004). For example, if an advisor observes that an agent is repeating actions and not exploring enough then the exploration percentage or learning rate may be temporarily increased. Internal modification is a simple method to implement. However, it can be difficult at times to know which parameters to adjust, and to what degree they are to be adjusted.

Environmental modification is an indirect method for influencing an RL agent. Altering the environment is not always achievable and may be a technique better suited for digital or simulated environments. Some examples of modifying the environment include altering or reducing the state space and observable information (Kerzel et al, 2018; Breyer et al, 2019), reducing the action space (Sridharan et al, 2017), modifying the agent's starting state (Dixon et al, 2000), or altering the dynamics of the environment to make the task easier to solve (Millán-Arias et al, 2021) Below, we further describe these environmental modifications.

Reducing the state space can speed up the agent's learning as there is less of the environment to search. While the agent cannot fully solve the task with an incomplete environment representation, it allows the agent to learn the basic behaviour. The level of detail in the state representation can then be increased, allowing the agent to refine its policy towards the correct behaviour (Kerzel et al, 2018; Breyer et al, 2019). Reducing the action space is similar to the previous. The agent's available actions are limited, and the agent attempts to learn the best behaviour it can with the actions it has available. Once a suitable behaviour has been achieved, new actions can be provided, and the agent can begin to learn more complex solutions (Sridharan et al, 2017). Modifying the agent's starting space alters where in the environment the agent begins learning. Using this approach, the agent can begin training close to the goal. As the agent learns how to navigate to the goal, the starting state is incrementally moved further away. This allows the agent to build upon its past knowledge of the environment (Dixon et al, 2000). Altering the dynamics of the environment involves changing how the environment operates to make the task easier for the agent to learn (Xu et al, 2020). By altering attributes of the environment such as reducing gravity, lowering maximum driving speed, or reducing noise, the agent may learn the desired behaviour faster or more safely. After the agent learns a satisfactory behaviour, the environment dynamics can be changed to more typical levels (Millán-Arias et al, 2020).

3.7 Assisted Agent

The final component of the proposed ARL taxonomy is the RL agent. A key aspect of the taxonomy is that the agent, in the absence of any external information, should operate the same as any RL agent would. Given no external information, the agent should continue to explore and interact autonomously with its environment and attempt to achieve its goal.

In the next section, we present an in-depth look at some ARL techniques and describe them in terms of the taxonomy that has been presented in this section.

4 Illustrative Approaches with Components and Links from the Taxonomy

This section presents an in-depth analysis of some popular and well-known ARL approaches. Each illustrative approach is described as an instance of the proposed taxonomy shown in Section 3, in some cases using a specific approach and in other cases a set of them. Therefore, for each presented ARL approach, we show how each processing component and each communication link particularly adapts to the ARL taxonomy using current literature in the respective field for concrete examples.

4.1 Heuristic Reinforcement Learning

Heuristic RL uses pieces of information that generalise over an area of the state space. The information is used to assist the agent in decision-making and reduce the searchable state space (Bianchi et al, 2015; Yang et al, 2019). An example of a heuristic is a rule. A rule can cover multiple states, making its use efficient at delivering advice to an agent. In Section 3.2, we have introduced a heuristic RL experiment applied to the RoboCup soccer domain (Celiberto Jr et al, 2007). In the RoboCup soccer domain, one team actively tries to score a goal, while the other team tries to block it. As mentioned, the defending team is given initial advice before training, consisting of two predefined rules. The following is an analysis of this heuristic RL example applied as the ARL taxonomy.

- **Information source**: The information source for the RoboCup experiment is a person. In this case, the person has previously experimented with the robot soccer domain and can advise the agent with some rules that will speed up learning.
- **Temporality**: The advice for the agent is given before training begins. Once training has begun the person does not interact with the agent again. This is an example of planned assistance, where information is given to the agent at a fixed time, and the information is known by the information source in advance.
- **Advice interpretation**: The information needs to be understandable by the agent. In the robot soccer domain, the person gives two rules; (i) if not near the ball then move towards the ball, and (ii) if near the ball do something with the ball. These rules are understandable by the human but need to be translated into machine code so that agent can use them. This is usually a task easily performed by a knowledgeable human operator. The result is conditional-like rules as: (i) **IF NOT** close_to_ball() **THEN** target_and_move(), and (ii) **IF** close_to_ball() **THEN** kick_ball().
- **Advice structure**: The structure of the advice after being interpreted is a new rule. The rule needs to be compatible with the agent, including the ability to substitute variables and evaluate expressions.
- **External model**: The external model used by the heuristic RL agent is a rule set. The external model retains all rules given to it. The model may also retain statistics about the rule relating to confidence, number of uses, and state space covered.
- **Agent modification**: Heuristic RL uses the rule set to assist the agent in its decision-making. If a rule applies to the current state, then the action that the rule recommends is taken by the agent. This is a form of policy-shaping as the agent's decision-making is directly manipulated by the external information.
- **Assisted Agent**: The RL agent operates as usual. When it is time to decide on an action to take it consults the external model. The external model tests all the rules it has and checks to see if any applies to the current state, otherwise, the agent's default decision-making mechanism is used.

Figure 5 shows how the heuristic RL approach fits into the proposed ARL taxonomy taking into consideration the previous definitions of processing components and communication links from the RoboCup soccer domain.

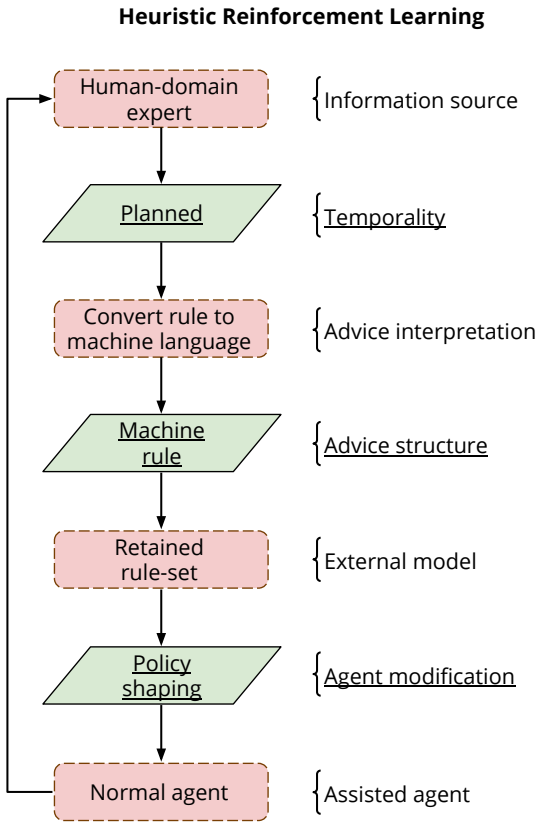

Fig. 5 Heuristic RL components according the proposed ARL taxonomy. The particular processing components and communication links illustrate a technique used in the RoboCup soccer domain (Celiberto Jr et al, 2007).

4.2 Interactive Reinforcement Learning

IntRL is another application of ARL. Most commonly, the information source is an observing human or a substitute for a human, such as an oracle, a simulated user, or another agent (Thomaz et al, 2005). The human provides assessment and advice to the agent, reinforcing the agent's past actions and guiding future decisions. The human can assess past actions in two ways, by stating that the agent's chosen action is somehow correct or incorrect, or by telling the agent what the correct action to take is in that instance. Alternatively, the human can advise the agent on what actions to take in the future (Li et al, 2019). The human can recommend actions to take or to avoid, or provide more information about the current state to assist the agent in its decision-making (Cruz et al, 2018b).

IntRL applications include having a human to provide additional reward information (Knox and Stone, 2012b,a), and having a human or agent provide action advice (Zhan et al, 2016; Amir et al, 2016). All of

these methods work in real-time and similarly, differing mainly in the agent modification stage. The following is an analysis of these IntRL approaches applied as the ARL taxonomy.

- **Information source**: The information source is a human or simulated user. A simulated user is a program, analogous to a human, that acts how a human would in a given situation. The human can observe the agent's current and past states, past actions taken, and what action the agent recommends it takes (Bignold et al, 2021a).
- **Temporality**: IntRL agents operate interactively. The advisor can provide information to the agent before, during, or after learning, and repeatedly throughout the learning process. This allows the advisor to react to current information and supply the agent with relevant advice.
- **Advice interpretation**: The advisor provides either an assessment of past actions taken, recommendations about actions to take, or a reward signal. Computer simulated agents can receive this information as key presses. However, physical agents may receive this information through audio or video inputs (Cruz et al, 2016b). In the case of audio inputs, these may be simple commands such as 'Correct' or 'Go Right', which can be translated to a form the agent can understand (Cruz et al, 2015). Supporting input modalities such as natural language makes systems based on IntRL more accessible to users who are not themselves familiar with RL.
- **Advice structure**: A common structure of advice the agent requires is simply a state-action pair. Using this structure the human can assign advice to a state for the agent to use, such as: In this state, do this (Ayala et al, 2019).
- **External model**: Either retained or immediate models are commonly used (Fernández and Veloso, 2006; Knox et al, 2012). A retained model tracks what advice/feedback has been received for each state (Fernández and Veloso, 2006). The agent can use this model to determine the human's accuracy, consistency, and discount for each piece of advice received. The model acts as a lookup table for the agent, if advice exists for the current state, then the agent can use it. Alternative methods may not retain information given by the human and only use it for the current state (Knox et al, 2012).
- **Agent modification**: The most common methods of using the advice to modify the agents learning process are reward- and policy-shaping (Li et al, 2019). Reward-shaping uses assessment/critique gathered from the advisor to alter the reward given to the agent. If the advisor disagrees with a past action, then the reward received for that state-action pair is decreased. If the advisor recommends an action to take in the future, then policy-shaping can be used to override the agent's usual action selection mechanism. One method of implementing policy-shaping for interactive advice is probabilistic policy reuse (Fernández and Veloso, 2006).
- **Assisted Agent**: Most of the time, the RL agent operates as any other RL agent would, i.e., it performs actions in the environment by exploiting/ exploring. The agent should continue to do so even if no advice from the trainer is given. Although a trainer could proactively provide advice to the learner, sometimes the student could decide to request such advice, and the trainer may or may not respond to that request. For instance, heuristics have been used to decide if the trainer should provide advice and/or if the learner should ask for it (Amir et al, 2016). In contrast, recent work estimates the learner's uncertainty in its current state, asking for advice in case the level of uncertainty is above a predefined threshold (Da Silva et al, 2020a).

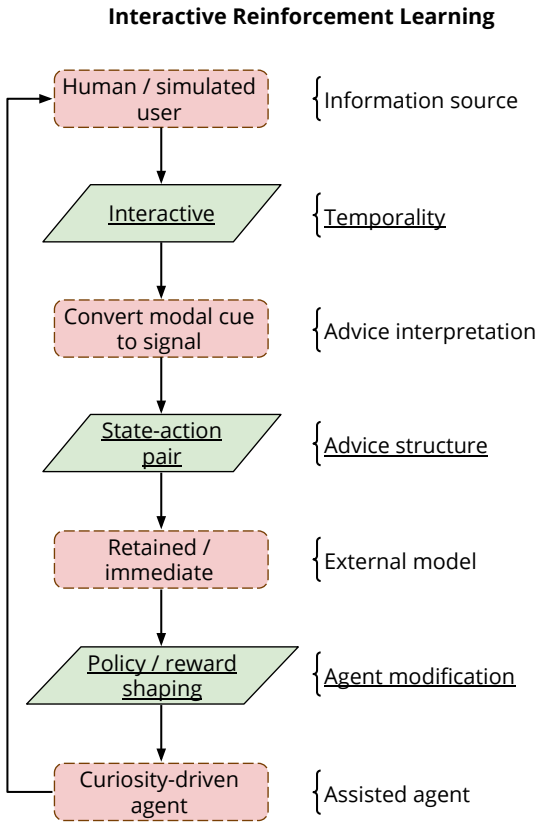

Fig. 6 Interactive RL as the proposed ARL taxonomy. In this approach, interactive advice is given by the user and more commonly used as policy and reward shaping.

Figure 6 shows how the IntRL approach is adapted to the proposed ARL taxonomy taking into account the previous definitions of processing components and communication links.

4.3 Reinforcement Learning from Demonstration

RLfD is a term coined by Schaal (Schaal, 1997). It refers to the setting where both a reward signal and demonstrations are available to learn from, combining the best of the fields of RL and Learning from Demonstration (LfD). Since RL presents an objective evaluation of behaviour, optimal behaviour can be achieved. Such an objective evaluation of behaviour is not present in LfD (Argall et al, 2009b), where only expert demonstrations are available to be mimicked and generalised. The student can thus not surpass its master. Nevertheless, LfD is typically much more sample efficient than RL. Therefore, the aim is to combine the fast LfD method with objective behaviour evaluation and theoretical guarantees from RL.

Two different approaches have been proposed to use demonstrations in an RL setting. The first is the generation of an initial value-function for temporal-difference learning by using the demonstrations as passive learning experiences for the RL agent (Smart and Kaelbling, 2002). The second approach derives an initial policy from the demonstrations and uses that to kickstart the RL agent (Brys et al, 2015; Suay et al, 2016). In this regard, Taylor et al. propose the Human-Agent Transfer (HAT) algorithm (Taylor et al, 2011), which consists of three steps: (i) demonstration: the agent performs the task teleoperated and records all state-action transitions, (ii) policy summarising: in order to bootstrap autonomous learning, policy rules are derived from the recorded state-action transitions, and (iii) independent learning: autonomous reinforcement learning using the policy summary to bias the learning. Below we use the HAT algorithm to describe how RLfD fits into the ARL taxonomy.

- **Information source**: An expert of the task (human or otherwise) can provide sample behaviour by demonstrating its execution of the task. Preferably these demonstrations are efficient and successful executions of the task.
- **Temporality**: It uses planned assistance. Demonstrations are recorded and given to the learning agent before it starts training.
- **Advice interpretation**: The received demonstrations must be first transformed into the agent's perspective by encoding them as sequences of state-action pairs. These are then processed using a classifier, which serves as the LfD component, creating an approximation of the demonstrator's policy using rules.
- **Advice structure**: The information is encoded as a classifier that maps states to the actions which the demonstrator is hypothesised to execute in those states.
- **External model**: The generated rules are stored in the external model and not modified anymore. The external model can be queried with a state and responds with the hypothesised demonstrator action in that state.
- **Agent modification**: The action proposed by the demonstrator can be integrated into the agent through three action biasing methods: (i) attributing a value bonus to the Q-value for that state-action pair, (ii) extending the agent's action set with an action that executes the hypothesised demonstrator action, and (iii) probabilistically choosing to execute the action suggested by the model.

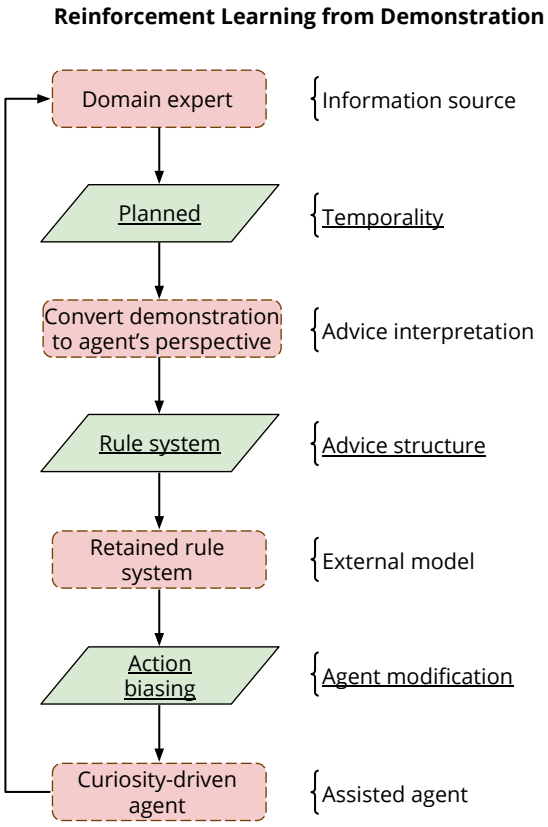

Fig. 7 RL from demonstration as the proposed ARL taxonomy. In this case, the processing components and communication links are defined from the HAT algorithm (Taylor et al, 2011), which combines RL and LfD.

– **Assisted agent**: During its decision-making (when and how depends on the implemented modification method) the agent has the option to consult the external model to obtain the action that the demonstrator is assumed to take. This kind of agent is sometimes referred to as curiosity-driven agent (Pathak et al, 2017). Otherwise, the agent acts as a usual RL agent.

Figure 7 shows how the RLfD approach is adapted to the proposed ARL taxonomy taking into account the previous definitions of processing components and communication links for the HAT algorithm.

4.4 Transfer Learning

The idea of transferring information between tasks (or between agents), rather than learning every task from the ground up seems to be obvious in retrospect. While transfer between different tasks has long been studied in humans, it has only gained popularity in RL settings in the last decade (Taylor and Stone, 2009). We consider three distinct settings where TL can be useful.

First, an agent may have learned how to perform a task and a new agent must learn to perform that same task or a variation on the task under different circumstances. Let us consider two agents with different state features, i.e., different sensors, or different action spaces (or different actuators). In this case, an inter-task mapping (Bou Ammar et al, 2011; Taylor et al, 2007a) can be hand specified or learned from data (Taylor et al, 2007b; Ammar et al, 2015) to relate the new target agent to the existing source agent. One of the simplest ways to reuse such knowledge is to embed it into the target task agent, e.g., directly reuse the Q-values that the source agent had learned (Taylor et al, 2007a).

Second, let us now consider that the world may be non-stationary. In TL settings, it is common to assume that the agent is notified when the world (or task in that world) changes. However, a TL agent sometimes does not need to detect changes (Hernandez-Leal et al, 2016) or worry about the slow world changes over time (Akila and Zayaraz, 2015). As in the previous setting, the agent may want to modify the information, e.g., by using an inter-task mapping, to relate the two tasks. In addition, the agent may decide not to use its prior knowledge at all, e.g., to avoid negative transfer because the tasks are too dissimilar (Taylor et al, 2007a).

Third, TL could be a critical step within a curriculum learning approach (Taylor, 2009; Bengio et al, 2009). For example, previous work has shown that learning a sequence of tasks that gradually increase in difficulty can be faster than directly training on the final (difficult) task (Taylor et al, 2007b; Eppe et al, 2019). In addition to curricula that are created by machine learning experts, curricula constructed by naive human participants have also been considered (Peng et al, 2017). Others have considered as a complementary problem a learning agent autonomously creating a curriculum (Narvekar et al, 2017; Da Silva and Costa, 2018). In all cases, the difficulty is scaffolding correctly so that the agent can learn quickly on a sequence of tasks. These approaches are distinct from multi-task learning (Fernández and Veloso, 2006), where the agent wants to learn over a distribution of tasks, and lifelong learning (Chen and Liu, 2016; Parisi et al, 2019), where learning a new task should also improve performance on previous tasks. The following is an analysis of TL methods in terms of the ARL taxonomy.

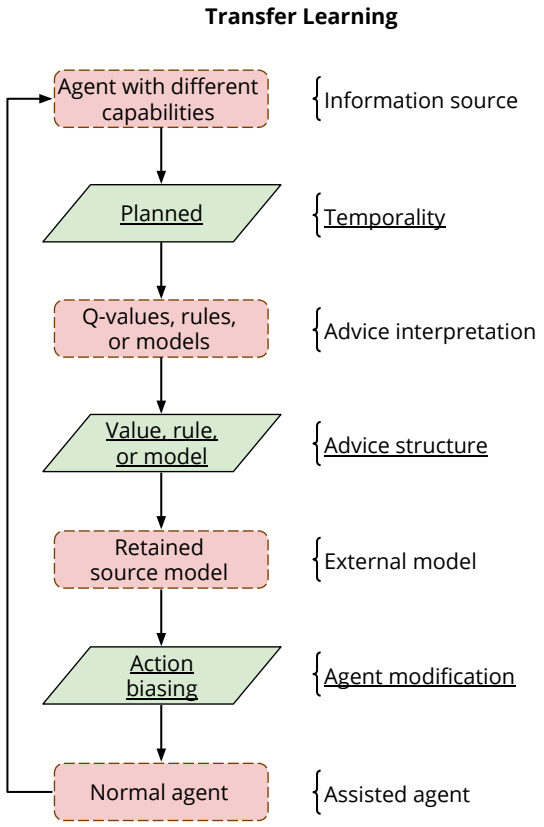

Fig. 8 Transfer learning as the proposed ARL taxonomy. In this case, an agent with different capabilities (or the same agent) provides the model of a source task which is transferred to a target task.

– **Information source**: The information comes from an agent with different capabilities or the same agent that has trained on a different task.
– **Temporality**: Transfer typically occurs when a task changes or when an agent first faces a novel task. In

both cases, it is planned assistance, i.e., the source agent transfers knowledge to the target agent before the target agent begins learning. If the inter-task mapping is initially unknown, some time may be spent trying to learn an inter-task mapping or estimate task similarity to previous tasks. However, the more time spent before the transfer, the less impact transfer can have.

- **Advice interpretation**: There are many types of information that can be transferred, including Q-values, rules, a model, etc. (Taylor et al, 2007a). TL methods assume the target agent has access to the source agent's 'brain', an assumption that may not always be true, e.g., if the designer of the source agent has not provided an API or if the source agent is a human.
- **Advice structure**: The structure of the transferred knowledge is as varied as the types of information that can be provided. This variety of information includes Q-values, rules, or a model, among others.
- **External model**: The source model is normally retained. Because the source task knowledge is not necessarily sufficient for optimal performance in the target task, it is important for the target agent to be able to learn to outperform the transferred information.
- **Agent modification**: The target task agent uses the transferred information to bias its learning. The transferred knowledge is not typically modified. Instead, the target task agent builds on top of the knowledge, learning when to ignore it and instead follow the knowledge it has learned from the environment.
- **Assisted Agent**: The agent is a typical RL agent that can take advantage of one or more types of prior knowledge.

Figure 8 shows how the TL approach can be represented within the proposed ARL taxonomy taking into account the previous definitions of processing components and communication links.

4.5 Multiple Information Sources

While the majority of work in ARL is based on a single source of advice, several researchers have considered scenarios where multiple sources of advice may exist (Brys et al, 2017; Da Silva, 2019; Gimelfarb et al, 2018; Yamagata et al, 2019). Although the use of multiple information sources is not an ARL approach by itself and could comprise sources utilising any of the previously mentioned approaches, we include it here to highlight how this multiple sources can be framed within

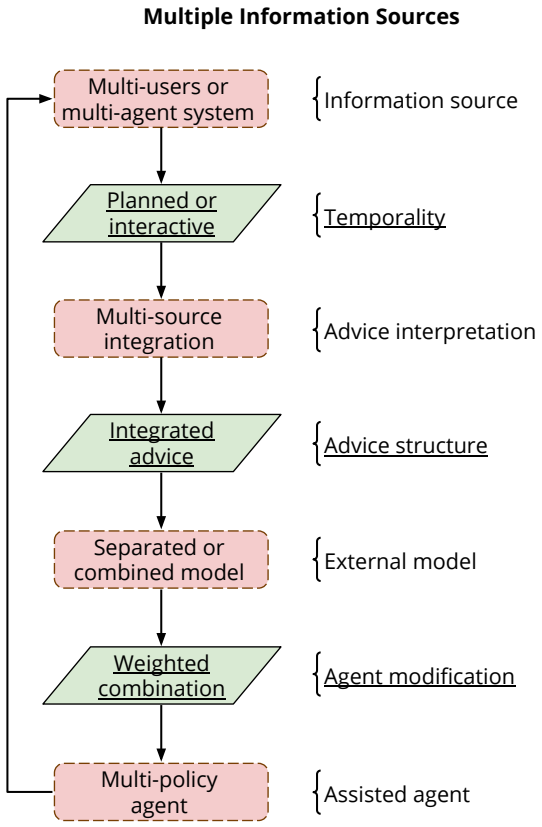

Fig. 9 Multiple information sources as the proposed ARL taxonomy. In this case, there could be multiple humans or multiple agents. One important aspect is to integrate the different pieces of advice. The agent may also learn multiple policies as in multi-objective RL.

the proposed taxonomy. The introduction of multiple advisors may have benefits for ARL agents, particularly in scenarios where each individual advisor has knowledge which is limited in some way (Shelton, 2001), e.g., individual advisors may have expertise covering different sub-areas of the problem domain. However, it also introduces additional problems for the agent, such as handling inconsistencies or direct conflicts between the guidance provided by different advisors, or learning to judge the reliability of each advisor, possibly in a state-sensitive manner (Zhan et al, 2016). In the extreme case, an agent may even need to be able to identify and ignore the advice provided by deliberately malicious advisors (Nunes and Oliveira, 2003). The following is an analysis of approaches using multiple information sources with respect to the proposed ARL taxonomy.

- **Information source**: Prior research has identified several scenarios in which an agent may have access to multiple sources of external information. Argall et al. (Argall et al, 2009a) argue that when robots

are applied to tasks within society in general, it is very likely that multiple users will interact with and guide the behaviour of a robot. In the context of TL, multiple sources of information may be derived either from experience on varying MDPs (Parisotto et al, 2015), or on alternative mappings from a single prior MDP to the current environment (Talvitie and Singh, 2007). In multi-agent systems, each agent may serve as a potential source of information for every other agent (Da Silva et al, 2017; Fachantidis et al, 2019).

- **Temporality**: Assistance may be planned or interactive. For instance, Argall et al. (Argall et al, 2009a) have considered two different sources of information, in the form of teacher demonstrations and teacher feedback on trajectories generated by the learner. The former may be provided in advance of learning consisting of complete state-action trajectories, i.e., planned assistance, while the latter occurs on an interactive basis during learning, and structurally consists of a subset of the learner's actions being flagged as correct by the teacher, i.e., interactive assistance.
- **Advice interpretation**: The majority of work so far on ARL from multiple information sources has assumed that these sources are homogeneous in terms of the timing and nature of the information provided. However, this need not be the case, and for heterogeneous information sources, some aspects of the advice may differ in terms of interpretation and structure. In this regard, the advice needs to be integrated considering either all possible sources (equally or non-equally contributing), some sources (with the information provided partially or fully considered), or only from one source at a time (Shelton, 2001).
- **Advice structure**: Each information source may use a different structure of advice. Therefore, individually all the aforementioned structures in previous sections are possible to be used, e.g., machine rule, state-action pair, rule system, value, or model. The final structure into a single piece of advice may be done by integrating the multiple information sources, for instance using a multi-modal integration function (Cruz et al, 2016b) or using graph structures (e.g., graph neural networks) using causal links between features for multi-modal causability (Holzinger et al, 2021).
- **External model**: An ARL agent must choose whether (i) to maintain a separate model for each information source, (ii) to combine the information from all sources into a single model, or (iii) a combination of both. An example of the latter approach is the inverse RL system presented in (Karlsson, 2014), which learns a model of each information source in the form of a feature-weighting function and then forms a combined feature-weighting via averaging. As noted by Karlsson (Karlsson, 2014), single-model approaches may encounter difficulties if dealing with information sources which are fundamentally incompatible with each other. An additional benefit of maintaining independent models is that these can also be augmented by additional data on characteristics of each information source, such as the reliability or consistency of its advice (Argall et al, 2009a; Talvitie and Singh, 2007).
- **Agent modification**: Any of the modification approaches discussed in the earlier sections of this paper may also be applied in the context of multiple information sources. For example, agent modification methods from LfD (Argall et al, 2009a), TL (Talvitie and Singh, 2007; Parisotto et al, 2015), reward-shaping (Brys et al, 2014; Knox et al, 2013) as well as inverse RL (Karlsson, 2014; Tanwani and Billard, 2013). The main additional consideration is how these methods may be affected by the presence of multiple external models. The main methods examined so far use a combination of the models, either weighted or unweighted (Argall et al, 2009a; Karlsson, 2014) or select a single best model to use (Talvitie and Singh, 2007).
- **Assisted Agent**: In most circumstances, the operation of the agent itself is largely unaffected by the presence of more than one information source. However, Tanwani and Billard (Tanwani and Billard, 2013) consider the task of performing inverse RL from multiple demonstrations provided by multiple experts, operating according to different strategies or preferences. To address the potential incompatibilities between these strategies, the agent attempts to learn a set of multiple policies, so as to be able to satisfy any policy expert strategy, including those not provided to the agent. This approach is closely related to multi-policy algorithms developed for multiobjective RL (Roijers et al, 2013).

Figure 9 shows how an approach using multiple information sources is adapted to the proposed ARL taxonomy taking into account the previous definitions of processing components and communication links. Moreover, Table 1 summarises how each of the ARL approaches and examples reviewed in this section is adapted to the proposed taxonomy.

Table 1 Summary of the reviewed assisted reinforcement learning approaches adapted to the proposed taxonomy.

Approach	Information source	Temporality	Advice interpretation	Advice structure	External model	Agent modification	Assisted agent
Heuristic reinforcement learning	Human-domain expert	Planned	Convert rule to machine language	Machine rule	Retained rule-set	Policy shaping	Normal agent
Interactive reinforcement learning	Human / simulated user	Interactive	Convert modal cue to signal	State-action pair	Immediate	Policy / reward shaping	Curiosity-driven agent
Reinforcement learning from demonstration	Domain expert	Planned	Convert demonstration to agent's perspective	Rule system	Retained rule system	Action biasing	Curiosity-driven agent
Transfer learning	Agent with different capabilities	Planned	Q-values, rules, or models	Value, rule, or model	Retained source model	Action biasing	Normal agent
Multiple information sources	Multi-users or multi-agent system	Planned or interactive	Multi-source integration	Integrated advice	Separated or combined model	Weighted or unweighted combination	Multi-policy agent

5 Future Directions and Open Challenges

In this section, we discuss open issues and propose further possibilities for future work in the field of ARL. These open questions have been identified from the current literature in the field. Many of these issues are shared with autonomous RL but it still remains open how they could be addressed within the ARL framework.

5.1 Incorrect Assistance

A common assumption that ARL methods make is that all external information that the agent receives is accurate (Efthymiadis et al, 2013). Accurate information is correct advice that assists the agent in completing its goal. However, the assumption that information will always be of use to the agent is wrong, especially when the information source is an observing human, as in RL from imperfect demonstrations (Gao et al, 2018; Jing et al, 2020). Humans may deliver advice late, and therefore the agent may relate it to a wrong state. The advice may be of short-term use to the agent but prevent it from achieving optimal performance. Moreover, the human trainer may even be malicious and actively attempting to sabotage the agent's performance.

Incorrect information can be introduced by other sources as well. Some examples for non-human incorrect advice include behaviour transferred from another domain that does not align correctly, rules that generalise over multiple states which may cover exception states, and noisy or missing information from audio-visual sources (Cruz et al, 2016b).

Information given to agents may be correct initially, but over time no longer be the optimal solution (Akila and Zayaraz, 2015). Other advice may be mostly accurate or correct for most states, however, there can exist states of exception to the advice. These exception states can be the critical difference between an ordinary solution and the optimal solution. There is a need for research on how to identify and mitigate incorrect information in these scenarios, especially considering that even a very small amount of incorrect advice may be really detrimental for the learning process (Cruz et al, 2018a).

5.2 Multiple Information Sources

As reviewed in the previous section, the use of multiple information sources may naturally arise on some application scenarios, and can increase the agent's knowledge of the environment, and increase confidence in decision-making if the different sources agree on an action. However, the use of multiple sources raises additional questions:

– What if the different sources disagree on the best action to take?

- How can the agent identify the best information source to listen to?
- How can the agent manage conflicting information?
- How can the agent measure trust in the different information sources?

Additionally, the use of multiple sources may be extended to crowdsourcing (Kamar et al, 2012). In this context, crowdsourcing refers to the enlistment and use of a large number of people, either paid or unpaid and can range in size from tens to tens of thousands. Typically, crowdsourcing is performed via the internet. This can raise challenges of malicious users, anonymity, and large uncertainty in the value and reliability of the information.

5.3 Explainability

Explainability refers to translating the agent's information into a form the human can understand (Cruz et al, 2019; Dazeley et al, 2021b). The reasons why an agent develops certain behaviours can sometimes be difficult to understand for non-expert end-users. Systems to measure the quality of explanations generated by AI-based systems have been previously introduced in order to build effective and efficient human-AI interaction (Holzinger et al, 2020). When combining the RL method with policy modification methods such as rules, expert assistance, external models, and policy-shaping, understanding why an agent chooses to take an action becomes even more difficult. Developing methods for understanding agent learning and its decision-making is important as it allows the human to remain informed of the agent's motivations and decisions, and keep track of the accountability of the actions taken (Dazeley et al, 2021a). This can be beneficial for artificial intelligence ethics, and human-computer teaching, among other fields.

5.4 Two-Way Communication

Two-way communication refers to the ability for the information source and the agent to converse with each other, perhaps multiple times before making a decision (Kessler Faulkner et al, 2019). Two-way communication can allow the information source, presumably human, and the agent to ask questions to each other, request more information, and to clarify decision-making and its reasoning. Although the proposed framework includes two-way communication, as shown in Figure 1, most current ARL methods do not have two-way communication to the extent that non-expert human advisors can interact with the agent freely. For two-way communication to apply to non-expert human advisors issues of explainability (as shown in the previous section), timing, and agent initiation need to be addressed.

Timing refers to the time it takes to communicate back and forth. Agents sometimes have a fixed time limit, during which they need to learn, communicate, and decide on the next action. Methods for reducing the time it takes to interact with the human and reducing the number of interactions needed with the human are two areas open for research. Agent initiation refers to the ability for the agent to initiate communication with the human source itself. The agent may choose to do this so to request clarification on information, or request assistance for decision-making. A challenge for agent initiation is to determine when and how often the agent should request assistance. The requests for assistance should be frequent enough to make use of the information source while not becoming a nuisance to the human, or detracting from learning time, and should consider the cost of the request, e.g., in paid crowdsourcing.

5.5 Other Challenges

There are also other challenges to be considered for future possibilities of ARL systems. Although many of the issues described in this section are also shared with autonomous RL (Mankowitz et al, 2019), we focus the discussion on how particularly externally-influenced agents may be affected in the context of the ARL framework. While we describe the essential implications on ARL systems for each of the following areas, we note that further and deeper discussion may be addressed for each of them.

- **Real-time policy inference**: Many RL systems need to be deployed in real-world scenarios and, therefore, policy inference must happen in real-time (Koenig and Simmons, 1993). Using ARL frameworks may lead to additional issues since the external information source should observe and react to the RL agent's state as fast as possible, otherwise the assistance may become unnecessary or incorrect for the new reached state.
- **Assistance delay**: There are RL systems where determining the state or receiving the reward signal may take even weeks, such as a recommender system where the reward is based on user interaction (Mann et al, 2018). In these contexts, the external information source may also lead to unknown delays in the system actuators, sensors, or rewards, making the assistance atemporal, either delayed or ahead, or even in some cases being conflicting or redundant considering the RL agent's autonomous operation.

- **Continuous states and actions**: When an RL agent works in high-dimensional continuous state and action spaces (Millán et al, 2019; Ayala et al, 2019) there could be issues for learning even in traditional RL (Dulac-Arnold et al, 2015). In an ARL framework, additional problems may be present as the agent uses external information which may be not accurate enough given the high dimensionality. In the presence of high-dimensional states and actions, even small differences in the received assistance may substantially slow the learning process since these differences may represent in essence a very different state or action.
- **Safety constraints**: In RL environments, there are safety constraints that should never or at least rarely be violated (Karimpanal et al, 2019). Special care is needed when receiving information from an external source since there could be situations that the advisor may repeatedly direct the agent to unsafe states and, in turn, lead to an increase in the time needed for learning.
- **Partially observable environments**: In practice, many RL problems are partially observable (Chen et al, 2018). For instance, partial observabilities may occur in non-stationary environments (Millán et al, 2019) or in presence of stochastic transitions (Cruz et al, 2021). If the external information source does not have observations to clearly infer the current state in the environment may lead to giving incorrect assistance to the learner agent.
- **Multi-objective reward**: In many cases, RL agents need to balance multiple and conflicting subgoals, therefore, they may use multi-dimensional reward functions (Vamplew et al, 2020). In this regard, an external information source may give priority to a particular subgoal over the others, unbalancing the global reward function. There could be also issues when multiple information sources are used covering or favouring different subgoals. Moreover, when using a multi-objective reward in TL, there could only be some subgoals from the source task which are relevant in the target task, therefore, the RL agent should also coordinate and filter relevant information.
- **Multi-agent systems**: There could be multiple agents learning a task and multiple external information sources. In this case, if an information source provides advice it could be generalised to all of them or it could be pointed specifically to an agent. Moreover, advice useful for one agent may be detrimental to another, depending on the state, the agent's current knowledge, or its particular reward function. Using multiple information sources, if an agent consults an external source, it may be necessary to discriminate which one is the best for the particular state. Additionally, the teacher-student approach usually integrated into ARL requires the teacher to be an expert in the learning domain. In this regard, multiple learning agents may also advise each other while learning in a common environment (Da Silva et al, 2017).

6 Conclusions

In this article, we have reviewed ARL methods and presented an ARL framework, comprising all RL techniques that use external information. ARL methods use external information to supplement the information the agent receives from the environment to improve performance and decision-making.

To describe the different ARL methods, we propose a taxonomy to classify the different functions of an externally-influenced RL agent. Through the analysis of the current literature, we have found seven key features that make up an ARL technique. They are divided into four processing components and three communication links. A definition and examples of each of these seven features have been presented.

The contribution of this paper is twofold: the review of state-of-the-art ARL methods and the ARL taxonomy as an additional level of abstraction. However, future work framed into our proposed ARL taxonomy can also make use of the different concepts here defined, either processing components or communication links. In this regard, it is essential to understand that not each ARL method must necessarily use all the proposed concepts. In some cases, simplified models may also be a representation of the ARL framework.

Additionally, we demonstrated the applicability of the framework on different ARL fields. These areas include heuristic RL, IntRL, RLfD, TL, and multiple information sources. Each of these fields has been analysed and described as applied to the presented taxonomy. Finally, we also present some ideas about areas for future research in order to extend the ARL field.

Compliance with ethical standards

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

Data availability statements

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

References

Abbeel P, Ng AY (2004) Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the International Conference on Machine learning ICML, ACM, pp 1–8

Akila V, Zayaraz G (2015) A brief survey on concept drift. In: Intelligent Computing, Communication and Devices, Springer, pp 293–302

Amershi S, Cakmak M, Knox WB, Kulesza T (2014) Power to the people: The role of humans in interactive machine learning. AI Magazine 35(4):105–120

Amir O, Kamar E, Kolobov A, Grosz B (2016) Interactive teaching strategies for agent training. In: Proceedings of the International Joint Conference on Artificial Intelligence IJCAI, pp 804–811

Ammar HB, Eaton E, Ruvolo P, Taylor ME (2015) Unsupervised Cross-Domain Transfer in Policy Gradient Reinforcement Learning via Manifold Alignment. In: Proceedings of the Association for the Advancement of Artificial Intelligence conference AAAI

Argall B, Browning B, Veloso M (2007) Learning by demonstration with critique from a human teacher. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction HRI, ACM, pp 57–64

Argall BD, Browning B, Veloso M (2009a) Automatic weight learning for multiple data sources when learning from demonstration. In: Proceedings of the IEEE International Conference on Robotics and Automation ICRA, IEEE, pp 226–231

Argall BD, Chernova S, Veloso M, Browning B (2009b) A survey of robot learning from demonstration. Robotics and autonomous systems 57(5):469–483

Arzate Cruz C, Igarashi T (2020) A survey on interactive reinforcement learning: Design principles and open challenges. In: Proceedings of the 2020 ACM Designing Interactive Systems Conference, pp 1195–1209

Ayala A, Henríquez C, Cruz F (2019) Reinforcement learning using continuous states and interactive feedback. In: Proceedings of the International Conference on Applications of Intelligent Systems, pp 1–5

Banerjee B (2007) General game learning using knowledge transfer. In: Proceedings of the International Joint Conference on Artificial Intelligence IJCAI, pp 672–677

Barros P, Tanevska A, Cruz F, Sciutti A (2020) Moody learners-explaining competitive behaviour of reinforcement learning agents. In: 2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), IEEE, pp 1–8

Behboudian P, Satsangi Y, Taylor ME, Harutyunyan A, Bowling M (2020) Useful policy invariant shaping from arbitrary advice. In: AAMAS Adaptive and Learning Agents Workshop ALA 2020, p 9

Bengio Y, Louradour J, Collobert R, Weston J (2009) Curriculum learning. In: Proceedings of the International Conference on Machine learning ICML, ACM, New York, NY, USA, pp 41–48

Bianchi RA, Celiberto Jr LA, Santos PE, Matsuura JP, de Mantaras RL (2015) Transferring knowledge as heuristics in reinforcement learning: A case-based approach. Artificial Intelligence 226:102–121

Bignold A, Cruz F, Dazeley R, Vamplew P, Foale C (2020) Human engagement providing evaluative and informative advice for interactive reinforcement learning. arXiv preprint arXiv:200909575

Bignold A, Cruz F, Dazeley R, Vamplew P, Foale C (2021a) An evaluation methodology for interactive reinforcement learning with simulated users. Biomimetics 6(1):13

Bignold A, Cruz F, Dazeley R, Vamplew P, Foale C (2021b) Persistent rule-based interactive reinforcement learning. Neural Computing and Applications pp 1–18

Bou Ammar H, Taylor ME, Tuyls K, Weiss G (2011) Reinforcement learning transfer using a sparse coded inter-task mapping. In: European Workshop on Multi-Agent Systems, Springer, pp 1–16

Breyer M, Furrer F, Novkovic T, Siegwart R, Nieto J (2019) Comparing task simplifications to learn closed-loop object picking using deep reinforcement learning. IEEE Robotics and Automation Letters 4(2):1549–1556

Brys T, Nowé A, Kudenko D, Taylor ME (2014) Combining multiple correlated reward and shaping signals by measuring confidence. In: Proceedings of the Association for the Advancement of Artificial Intelligence conference AAAI, pp 1687–1693

Brys T, Harutyunyan A, Suay HB, Chernova S, Taylor ME, Nowé A (2015) Reinforcement learning from demonstration through shaping. In: Proceedings of the International Joint Conference on Artificial Intelligence IJCAI, p 26

Brys T, Harutyunyan A, Vrancx P, Nowé A, Taylor ME (2017) Multi-objectivization and ensembles of shapings in reinforcement learning. Neurocomputing 263:48–59

Cassandra AR, Kaelbling LP (2016) Learning policies for partially observable environments: Scaling up. In: Proceedings of the International Conference on Machine Learning ICML, Morgan Kaufmann, p 362

Celiberto Jr LA, Ribeiro CH, Costa AH, Bianchi RA (2007) Heuristic reinforcement learning applied to robocup simulation agents, Springer, pp 220–227

Chen H, Yang B, Liu J (2018) Partially observable reinforcement learning for sustainable active surveillance. In: Proceedings of the International Conference on Knowledge Science, Engineering and Management, Springer, pp 425–437

Chen SA, Tangkaratt V, Lin HT, Sugiyama M (2019) Active deep Q-learning with demonstration. Machine Learning pp 1–27

Chen Z, Liu B (2016) Lifelong Machine Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers

Cheng ST, Chang TY, Hsu CW (2013) A framework of an agent planning with reinforcement learning for e-pet. In: Proceedings of the International Conference on Orange Technologies ICOT, IEEE, pp 310–313

Churamani N, Cruz F, Griffiths S, Barros P (2016) iCub: learning emotion expressions using human reward. In: Proceedings of the Workshop on Bio-inspired Social Robot Learning in Home Scenarios, IEEE/RSJ IROS, p 2

Cobo LC, Subramanian K, Isbell Jr CL, Lanterman AD, Thomaz AL (2014) Abstraction from demonstration for efficient reinforcement learning in high-dimensional domains. Artificial Intelligence 216:103–128

Contreras R, Ayala A, Cruz F (2020) Unmanned aerial vehicle control through domain-based automatic speech recognition. Computers 9(3):75

Cruz F, Twiefel J, Magg S, Weber C, Wermter S (2015) Interactive reinforcement learning through speech guidance in a domestic scenario. In: Proceedings of the International Joint Conference on Neural Networks IJCNN, IEEE, pp 1341–1348

Cruz F, Magg S, Weber C, Wermter S (2016a) Training agents with interactive reinforcement learning and contextual affordances. IEEE Transactions on Cognitive and Developmental Systems 8(4):271–284

Cruz F, Parisi GI, Twiefel J, Wermter S (2016b) Multi-modal integration of dynamic audiovisual patterns for an interactive reinforcement learning scenario. In: Proceedings fo the IEEE/RSJ International Conference on Intelligent Robots and Systems IROS, IEEE, pp 759–766

Cruz F, Parisi GI, Wermter S (2016c) Learning contextual affordances with an associative neural architecture. In: Proceedings of the European Symposium on Artificial Neural Network. Computational Intelligence and Machine Learning ESANN, UCLouvain, pp 665–670

Cruz F, Wüppen P, Magg S, Fazrie A, Wermter S (2017) Agent-advising approaches in an interactive reinforcement learning scenario. In: Proceedings of the Joint IEEE International Conference on Development and Learning and Epigenetic Robotics ICDL-EpiRob, IEEE, pp 209–214

Cruz F, Magg S, Nagai Y, Wermter S (2018a) Improving interactive reinforcement learning: What makes a good teacher? Connection Science 30(3):306–325

Cruz F, Parisi GI, Wermter S (2018b) Multi-modal feedback for affordance-driven interactive reinforcement learning. In: Proceedings of the International Joint Conference on Neural Networks IJCNN, IEEE, pp 5515–5122

Cruz F, Wüppen P, Fazrie A, Weber C, Wermter S (2018c) Action selection methods in a robotic reinforcement learning scenario. In: 2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI), IEEE, pp 1–6

Cruz F, Dazeley R, Vamplew P (2019) Memory-based explainable reinforcement learning. In: Proceedings of the Australasian Joint Conference on Artificial Intelligence, Springer, pp 66–77

Cruz F, Dazeley R, Vamplew P, Ithan M (2021) Explainable robotic systems: Understanding goal-driven actions in a reinforcement learning scenario. Neural Computing and Applications pp 1–17

Da Silva FL (2019) Integrating agent advice and previous task solutions in multiagent reinforcement learning. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, International Foundation for Autonomous Agents and Multiagent Systems, pp 2447–2448

Da Silva FL, Costa AHR (2018) Object-oriented curriculum generation for reinforcement learning. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, International Foundation for Autonomous Agents and Multiagent Systems, pp 1026–1034

Da Silva FL, Costa AHR (2019) A survey on transfer learning for multiagent reinforcement learning systems. Journal of Artificial Intelligence Research 64:645–703

Da Silva FL, Glatt R, Costa AHR (2017) Simultaneously learning and advising in multiagent reinforcement learning. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, pp 1100–1108

Da Silva FL, Hernandez-Leal P, Kartal B, Taylor ME (2020a) Uncertainty-aware action advising for deep

reinforcement learning agents. In: Proceedings of the Association for the Advancement of Artificial Intelligence conference AAAI, pp 5792–5799

Da Silva FL, Warnell G, Costa AHR, Stone P (2020b) Agents teaching agents: a survey on inter-agent transfer learning. Autonomous Agents and Multi-Agent Systems 34(1):9

Dazeley R, Vamplew P, Cruz F (2021a) Explainable reinforcement learning for broad-xai: A conceptual framework and survey. arXiv preprint arXiv:210809003

Dazeley R, Vamplew P, Foale C, Young C, Aryal S, Cruz F (2021b) Levels of explainable artificial intelligence for human-aligned conversational explanations. Artificial Intelligence p 103525

Devlin S, Kudenko D (2011) Theoretical considerations of potential-based reward shaping for multi-agent systems. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, International Foundation for Autonomous Agents and Multiagent Systems, pp 225–232

Devlin S, Kudenko D (2012) Dynamic potential-based reward shaping. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, International Foundation for Autonomous Agents and Multiagent Systems, pp 433–440

Dixon K, Malak RJ, Khosla PK (2000) Incorporating prior knowledge and previously learned information into reinforcement learning agents. Carnegie Mellon University, Institute for Complex Engineered Systems

Dorigo M, Gambardella L (2014) Ant-Q: A reinforcement learning approach to the traveling salesman problem. In: Proceedings of International Conference on Machine Learning ICML, pp 252–260

Dulac-Arnold G, Evans R, van Hasselt H, Sunehag P, Lillicrap T, Hunt J, Mann T, Weber T, Degris T, Coppin B (2015) Deep reinforcement learning in large discrete action spaces. arXiv preprint arXiv:151207679

Efthymiadis K, Devlin S, Kudenko D (2013) Overcoming erroneous domain knowledge in plan-based reward shaping. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, International Foundation for Autonomous Agents and Multiagent Systems, pp 1245–1246

Eppe M, Magg S, Wermter S (2019) Curriculum goal masking for continuous deep reinforcement learning. In: Proceedings of the Joint IEEE International Conference on Development and Learning and Epigenetic Robotics ICDL-EpiRob, IEEE, pp 183–188

Erez T, Smart WD (2008) What does shaping mean for computational reinforcement learning? In: Proceedings of the IEEE International Conference on Development and Learning ICDL, IEEE, pp 215–219

Fachantidis A, Taylor ME, Vlahavas I (2019) Learning to teach reinforcement learning agents. Machine Learning and Knowledge Extraction 1(1):21–42

Fernández F, Veloso M (2006) Probabilistic policy reuse in a reinforcement learning agent. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, ACM, pp 720–727

Gao Y, Xu H, Lin J, Yu F, Levine S, Darrell T (2018) Reinforcement learning from imperfect demonstrations. arXiv preprint arXiv:180205313

Ghobaei-Arani M, Rahmanian AA, Shamsi M, Rasouli-Kenari A (2018) A learning-based approach for virtual machine placement in cloud data centers. International Journal of Communication Systems 31(8):e3537

Giannoccaro I, Pontrandolfo P (2002) Inventory management in supply chains: a reinforcement learning approach. International Journal of Production Economics 78(2):153–161

Gimelfarb M, Sanner S, Lee CG (2018) Reinforcement learning with multiple experts: A Bayesian model combination approach. In: Advances in Neural Information Processing Systems, pp 9528–9538

Griffith S, Subramanian K, Scholz J, Isbell C, Thomaz AL (2013) Policy shaping: Integrating human feedback with reinforcement learning. In: Advances in Neural Information Processing Systems, pp 2625–2633

Grizou J, Lopes M, Oudeyer PY (2013) Robot learning simultaneously a task and how to interpret human instructions. In: Proceedings of the Joint IEEE International Conference on Development and Learning and Epigenetic Robotics ICDL-EpiRob, IEEE, pp 1–8

Harutyunyan A, Devlin S, Vrancx P, Nowé A (2015) Expressing arbitrary reward functions as potential-based advice. In: Proceedings of the Association for the Advancement of Artificial Intelligence conference AAAI, pp 2652–2658

Hausknecht M, Mupparaju P, Subramanian S, Kalyanakrishnan S, Stone P (2016) Half field offense: An environment for multiagent learning and ad hoc teamwork. In: AAMAS Adaptive and Learning Agents Workshop ALA 2016

Hernandez-Leal P, Zhan Y, Taylor ME, Sucar LE, Munoz de Cote E (2016) Efficiently detecting switches against non-stationary opponents. Autonomous Agents and Multi-Agent Systems pp 1–23

Holzinger A (2016) Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Informatics 3(2):119–131

Holzinger A, Carrington A, Müller H (2020) Measuring the quality of explanations: the system causability scale (scs). KI-Künstliche Intelligenz pp 1–6

Holzinger A, Malle B, Saranti A, Pfeifer B (2021) Towards multi-modal causability with graph neural net-

works enabling information fusion for explainable ai. Information Fusion 71:28–37

Isbell CL, Kearns M, Kormann D, Singh S, Stone P (2000) Cobot in LambdaMOO: A social statistics agent. In: Proceedings of the Association for the Advancement of Artificial Intelligence conference AAAI, pp 36–41

Jing M, Ma X, Huang W, Sun F, Yang C, Fang B, Liu H (2020) Reinforcement learning from imperfect demonstrations under soft expert guidance. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 5109–5116

Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: A survey. Journal of artificial intelligence research pp 237–285

Kalyanakrishnan S, Liu Y, Stone P (2006) Half field offense in RoboCup soccer: A multiagent reinforcement learning case study. In: Robot Soccer World Cup, Springer, pp 72–85

Kamar E, Hacker S, Horvitz E (2012) Combining human and machine intelligence in large-scale crowdsourcing. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, International Foundation for Autonomous Agents and Multiagent Systems, pp 467–474

Kaplan F, Oudeyer PY, Kubinyi E, Miklósi A (2002) Robotic clicker training. Robotics and Autonomous Systems 38(3):197–206

Karimpanal TG, Rana S, Gupta S, Tran T, Venkatesh S (2019) Learning transferable domain priors for safe exploration in reinforcement learning. In: Proceedings of the International Joint Conference on Neural Networks IJCNN, pp 1–8

Karlsson J (2014) Learning to play games from multiple imperfect teachers. Master's thesis, Chalmers University of Technology, Gothenburg, Sweden

Kerzel M, Mohammadi HB, Zamani MA, Wermter S (2018) Accelerating deep continuous reinforcement learning through task simplification. In: Proceedings of the International Joint Conference on Neural Networks IJCNN, IEEE, pp 1–6

Kessler Faulkner T, Gutierrez RA, Short ES, Hoffman G, Thomaz AL (2019) Active attention-modified policy shaping: socially interactive agents track. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, International Foundation for Autonomous Agents and Multiagent Systems, pp 728–736

Kitano H, Asada M, Kuniyoshi Y, Noda I, Osawa E, Matsubara H (1997) RoboCup: A challenge problem for AI. AI magazine 18(1):73

Knowles MJ, Wermter S (2008) The hybrid integration of perceptual symbol systems and interactive reinforcement learning. In: Proceedings of the International Conference on Hybrid Intelligent Systems, IEEE, pp 404–409

Knox WB, Stone P (2009) Interactively shaping agents via human reinforcement: The TAMER framework. In: Proceedings of the International Conference on Knowledge Capture, ACM, pp 9–16

Knox WB, Stone P (2010) Combining manual feedback with subsequent MDP reward signals for reinforcement learning. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, International Foundation for Autonomous Agents and Multiagent Systems, pp 5–12

Knox WB, Stone P (2012a) Reinforcement learning from human reward: Discounting in episodic tasks. In: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication RO-MAN, IEEE, pp 878–885

Knox WB, Stone P (2012b) Reinforcement learning from simultaneous human and MDP reward. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, pp 475–482

Knox WB, Glass BD, Love BC, Maddox WT, Stone P (2012) How humans teach agents. International Journal of Social Robotics 4(4):409–421

Knox WB, Stone P, Breazeal C (2013) Training a robot via human feedback: A case study. In: Proceedings of the International Conference on Social Robotics, Springer, pp 460–470

Kober J, Bagnell JA, Peters J (2013) Reinforcement learning in robotics: A survey. The International Journal of Robotics Research 32(11):1238–1274

Koenig S, Simmons RG (1993) Complexity analysis of real-time reinforcement learning. In: Proceedings of the Association for the Advancement of Artificial Intelligence conference AAAI, pp 99–107

Konidaris G, Kuindersma S, Grupen R, Barto A (2012) Robot learning from demonstration by constructing skill trees. The International Journal of Robotics Research 31(3):360–375

Li G, Gomez R, Nakamura K, He B (2019) Human-centered reinforcement learning: A survey. IEEE Transactions on Human-Machine Systems 49(4):337–349

Lin J, Ma Z, Gomez R, Nakamura K, He B, Li G (2020) A review on interactive reinforcement learning from human social feedback. IEEE Access 8:120757–120765

Lin LJ (1991) Programming robots using reinforcement learning and teaching. In: Proceedings of the Association for the Advancement of Artificial Intelligence conference AAAI, pp 781–786

Liu X, Deng R, Choo KKR, Yang Y (2019) Privacy-preserving reinforcement learning design for patient-

centric dynamic treatment regimes. IEEE Transactions on Emerging Topics in Computing

Mankowitz DJ, Dulac-Arnold G, Hester T (2019) Challenges of real-world reinforcement learning. In: ICML Workshop on Real-Life Reinforcement Learning, p 14

Mann TA, Gowal S, Jiang R, Hu H, Lakshminarayanan B, Gyorgy A (2018) Learning from delayed outcomes with intermediate observations. arXiv preprint arXiv:180709387

Millán C, Fernandes B, Cruz F (2019) Human feedback in continuous actor-critic reinforcement learning. In: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning ESANN, ESANN, pp 661–666

Millán-Arias C, Fernandes B, Cruz F, Dazeley R, Fernandes S (2020) A robust approach for continuous interactive reinforcement learning. In: Proceedings of the 8th International Conference on Human-Agent Interaction, pp 278–280

Millán-Arias C, Fernandes B, Cruz F, Dazeley R, Fernandes S (2021) A robust approach for continuous interactive actor-critic algorithms. IEEE Access pp 104242–104260

Moreira I, Rivas J, Cruz F, Dazeley R, Ayala A, Fernandes B (2020) Deep reinforcement learning with interactive feedback in a human–robot environment. Applied Sciences 10(16):5574

Nair A, McGrew B, Andrychowicz M, Zaremba W, Abbeel P (2018) Overcoming exploration in reinforcement learning with demonstrations. In: Proceedings of the IEEE International Conference on Robotics and Automation ICRA, IEEE, pp 6292–6299

Narvekar S, Sinapov J, Leonetti M, Stone P (2016) Source task creation for curriculum learning. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, pp 566–574

Narvekar S, Sinapov J, Stone P (2017) Autonomous task sequencing for customized curriculum design in reinforcement learning. In: Proceedings of the International Joint Conference on Artificial Intelligence IJCAI

Navidi N (2020) Human AI interaction loop training: New approach for interactive reinforcement learning. arXiv preprint arXiv:200304203

Ng AY, Harada D, Russell S (1999) Policy invariance under reward transformations: Theory and application to reward shaping. In: Proceedings of the International Conference on Machine Learning ICML, vol 99, pp 278–287

Niv Y (2009) Reinforcement learning in the brain. Journal of Mathematical Psychology 53(3):139–154

Nunes L, Oliveira E (2003) Exchanging advice and learning to trust. Cooperative Information Agents VII pp 250–265

Parisi GI, Kemker R, Part JL, Kanan C, Wermter S (2019) Continual lifelong learning with neural networks: A review. Neural Networks

Parisotto E, Ba JL, Salakhutdinov R (2015) Actor-mimic: Deep multitask and transfer reinforcement learning. arXiv preprint arXiv:151106342

Partalas I, Vrakas D, Vlahavas I (2008) Reinforcement learning and automated planning: A survey. In: Artificial Intelligence for Advanced Problem Solving Techniques, IGI Global, pp 148–165

Pathak D, Agrawal P, Efros AA, Darrell T (2017) Curiosity-driven exploration by self-supervised prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 16–17

Peng B, MacGlashan J, Loftin R, Littman ML, Roberts DL, Taylor ME (2017) Curriculum Design for Machine Learners in Sequential Decision Tasks (Extended Abstract). In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS

Pilarski PM, Sutton RS (2012) Between instruction and reward: human-prompted switching. In: AAAI Fall Symposium Series: Robots Learning Interactively from Human Teachers, pp 45–52

Price B, Boutilier C (2003) Accelerating reinforcement learning through implicit imitation. Journal of Artificial Intelligence Research 19:569–629

Randløv J, Alstrøm P (1998) Learning to drive a bicycle using reinforcement learning and shaping. In: Proceedings of the International Conference on Machine Learning ICML, pp 463–471

Roijers DM, Vamplew P, Whiteson S, Dazeley R (2013) A survey of multi-objective sequential decision-making. Journal of Artificial Intelligence Research 48:67–113

Rozo L, Jiménez P, Torras C (2013) A robot learning from demonstration framework to perform force-based manipulation tasks. Intelligent service robotics 6(1):33–51

Schaal S (1997) Learning from demonstration. Advances in neural information processing systems 9:1040–1046

Sert E, Bar-Yam Y, Morales AJ (2020) Segregation dynamics with reinforcement learning and agent based modeling. Scientific reports 10(1):1–12

Shahidinejad A, Ghobaei-Arani M (2020) Joint computation offloading and resource provisioning for edge-cloud computing environment: A machine learning-based approach. Software: Practice and Experience 50(12):2212–2230

Shakarami A, Ghobaei-Arani M, Masdari M, Hosseinzadeh M (2020) A survey on the computation offloading approaches in mobile edge/cloud computing environment: a stochastic-based perspective. Journal of Grid Computing pp 1–33

Shao K, Zhu Y, Zhao D (2018) Starcraft micromanagement with reinforcement learning and curriculum transfer learning. IEEE Transactions on Emerging Topics in Computational Intelligence 3(1):73–84

Sharma M, Holmes MP, Santamaría JC, Irani A, Isbell Jr CL, Ram A (2007) Transfer learning in real-time strategy games using hybrid cbr/rl. In: Proceedings of the International Joint Conference on Artificial Intelligence IJCAI, vol 7, pp 1041–1046

Shelton CR (2001) Balancing multiple sources of reward in reinforcement learning. In: Advances in Neural Information Processing Systems, pp 1082–1088

Shiarlis K, ao Messias J, Whiteson S (2016) Inverse reinforcement learning from failure. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, pp 1060–1068

Skinner BF (1975) The shaping of phylogenic behavior. Journal of the Experimental Analysis of Behavior 24(1):117–120

Smart WD, Kaelbling LP (2002) Effective reinforcement learning for mobile robots. In: Proceedings of the IEEE International Conference on Robotics and Automation ICRA, IEEE, vol 4, pp 3404–3410

Sridharan M, Meadows B, Gomez R (2017) What can I not do? towards an architecture for reasoning about and learning affordances. In: Proceedings of the International Conference on Automated Planning and Scheduling, pp 461–469

Stahlhut C, Navarro-Guerrero N, Weber C, Wermter S, WTM VKS (2015) Interaction is more beneficial in complex reinforcement learning problems than in simple ones. In: Proceedings of the Interdisziplinärer Workshop Kognitive Systeme (KogSys), pp 142–150

Suay HB, Chernova S (2011) Effect of human guidance and state space size on interactive reinforcement learning. In: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication RO-MAN, IEEE, pp 1–6

Suay HB, Brys T, Taylor ME, Chernova S (2016) Learning from demonstration for shaping through inverse reinforcement learning. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, International Foundation for Autonomous Agents and Multiagent Systems, pp 429–437

Subramanian K, Isbell C, Thomaz A (2011) Learning options through human interaction. In: IJCAI Workshop on Agents Learning Interactively from Human Teachers (ALIHT), Citeseer

Subramanian K, Isbell Jr CL, Thomaz AL (2016) Exploration from demonstration for interactive reinforcement learning. In: Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems AAMAS, pp 447–456

Sutton RS, Barto AG (2018) Reinforcement learning: An introduction. MIT press

Talvitie E, Singh SP (2007) An experts algorithm for transfer learning. In: Proceedings of the International Joint Conference on Artificial Intelligence IJCAI, pp 1065–1070

Tanwani AK, Billard A (2013) Transfer in inverse reinforcement learning for multiple strategies. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems IROS, IEEE, pp 3244–3250

Taylor ME (2009) Assisting Transfer-Enabled Machine Learning Algorithms: Leveraging Human Knowledge for Curriculum Design. In: The AAAI 2009 Spring Symposium on Agents that Learn from Human Teachers

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

Taylor ME, Stone P, Liu Y (2005) Value functions for rl-based behavior transfer: A comparative study. In: Proceedings of the Association for the Advancement of Artificial Intelligence conference AAAI, vol 5, pp 880–885

Taylor ME, Stone P, Liu Y (2007a) Transfer learning via inter-task mappings for temporal difference learning. Journal of Machine Learning Research 8(1):2125–2167

Taylor ME, Whiteson S, Stone P (2007b) Transfer via Inter-Task Mappings in Policy Search Reinforcement Learning. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, pp 156–163

Taylor ME, Kuhlmann G, Stone P (2008) Autonomous transfer for reinforcement learning. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, International Foundation for Autonomous Agents and Multiagent Systems, pp 283–290

Taylor ME, Suay HB, Chernova S (2011) Integrating reinforcement learning with human demonstrations of varying ability. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS, International Foundation for Autonomous Agents and Multiagent Systems, pp 617–624

Taylor ME, Carboni N, Fachantidis A, Vlahavas I, Torrey L (2014) Reinforcement learning agents providing

advice in complex video games. Connection Science 26(1):45–63

Tenorio-Gonzalez AC, Morales EF, Villaseñor-Pineda L (2010) Dynamic reward shaping: training a robot by voice. In: Advances in Artificial Intelligence–IBERAMIA 2010, Springer, pp 483–492

Tesauro G (1994) TD-Gammon, a self-teaching backgammon program, achieves master-level play. Neural computation 6(2):215–219

Tesauro G (2004) Extending Q-learning to general adaptive multi-agent systems. In: Advances in neural information processing systems, pp 871–878

Thomaz AL, Breazeal C (2007) Asymmetric interpretations of positive and negative human feedback for a social learning agent. In: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication RO-MAN, IEEE, pp 720–725

Thomaz AL, Hoffman G, Breazeal C (2005) Real-time interactive reinforcement learning for robots. In: AAAI 2005 Workshop on Human Comprehensible Machine Learning

Thomaz AL, Breazeal C, et al (2006a) Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In: Proceedings of the Association for the Advancement of Artificial Intelligence conference AAAI, Boston, MA, vol 6, pp 1000–1005

Thomaz AL, Hoffman G, Breazeal C (2006b) Reinforcement learning with human teachers: Understanding how people want to teach robots. In: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication RO-MAN, IEEE, pp 352–357

Torrey L, Taylor ME (2013) Teaching on a Budget: Agents Advising Agents in Reinforcement Learning. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems AAMAS

Vamplew P, Foale C, Dazeley R (2020) A demonstration of issues with value-based multiobjective reinforcement learning under stochastic state transitions. arXiv preprint arXiv:200406277

Vlassis N, Ghavamzadeh M, Mannor S, Poupart P (2012) Bayesian reinforcement learning. Reinforcement Learning pp 359–386

Wiewiora E, Cottrell G, Elkan C (2003) Principled methods for advising reinforcement learning agents. In: Proceedings of the International Conference on Machine learning ICML, pp 792–799

Xu H, Bector R, Rabinovich Z (2020) Teaching multiple learning agents by environment-dynamics tweaks. In: AAMAS Adaptive and Learning Agents Workshop ALA 2020, p 8

Yamagata T, Santos-Rodríguez R, McConville R, Elsts A (2019) Online feature selection for activity recognition using reinforcement learning with multiple feedback. arXiv preprint arXiv:190806134

Yang MC, Samani H, Zhu K (2019) Emergency-response locomotion of hexapod robot with heuristic reinforcement learning using q-learning. In: Proceedings of the International Conference on Interactive Collaborative Robotics, Springer, pp 320–329

Zhan Y, Ammar HB, Taylor ME (2016) Theoretically-Grounded Policy Advice from Multiple Teachers in Reinforcement Learning Settings with Applications to Negative Transfer. In: Proceedings of the International Joint Conference on Artificial Intelligence IJCAI

Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q (2020) A comprehensive survey on transfer learning. Proceedings of the IEEE 109(1):43–76