Explainable Reinforcement Learning: A Survey

Erika Puiutta, Eric Veith

Explainable Reinforcement Learning: A Survey

Erika Puiutta[1 [0000−0003−3796−8931]] and Eric MSP Veith[1 [0000−0003−2487−7475]]

OFFIS – Institute for Information Technology, Escherweg 2, 26121 Oldenburg, Germany
erika.puiutta@offis.de, eric.veith@offis.de

Abstract. Explainable Artificial Intelligence (XAI), i.e., the development of more transparent and interpretable AI models, has gained increased traction over the last few years. This is due to the fact that, in conjunction with their growth into powerful and ubiquitous tools, AI models exhibit one detrimential characteristic: a performance-transparency trade-off. This describes the fact that the more complex a model's inner workings, the less clear it is how its predictions or decisions were achieved. But, especially considering Machine Learning (ML) methods like Reinforcement Learning (RL) where the system learns autonomously, the necessity to understand the underlying reasoning for their decisions becomes apparent. Since, to the best of our knowledge, there exists no single work offering an overview of Explainable Reinforcement Learning (XRL) methods, this survey attempts to address this gap. We give a short summary of the problem, a definition of important terms, and offer a classification and assessment of current XRL methods. We found that a) the majority of XRL methods function by mimicking and simplifying a complex model instead of designing an inherently simple one, and b) XRL (and XAI) methods often neglect to consider the human side of the equation, not taking into account research from related fields like psychology or philosophy. Thus, an interdisciplinary effort is needed to adapt the generated explanations to a (non-expert) human user in order to effectively progress in the field of XRL and XAI in general.

Keywords: Machine Learning · Explainable · Reinforcement Learning · Human-Computer Interaction · Interpretable.

1 Introduction

Over the past decades, AI has become ubiquitous in many areas of our everyday lives. Especially Machine Learning (ML) as one branch of AI has numerous fields of application, be it transportation [57], advertisement [46], or medicine [38]. Unfortunately, the more powerful and flexible those models are, the more opaque they become, essentially making them black boxes (see figure 1). This trade-off is referred to by different terms in the literature, e.g. readability-performance trade-off [12], accuracy-comprehensibility trade-off [16], or accuracy-interpretability trade-off [49]. This work aims to, first, establish the need for eXplainable Artificial Intelligence (XAI) in general and eXplainable Reinforcement Learning

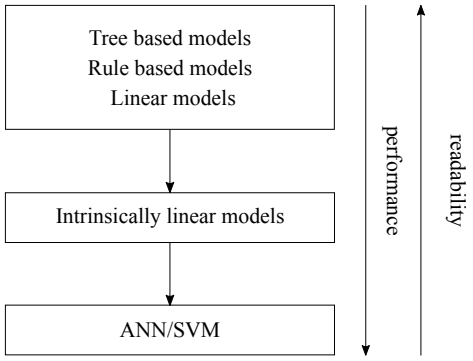

Fig. 1. Schematic representation of the performance-readability trade-off. Simpler, linear models are easy to understand and interpret, but suffer from a lack of performance, while non-linear, more flexible models are too complex to be understood easily. Adopted from Martens et al. [41].

(XRL) specifically. After that, the general concept of RL is briefly explained and the most important terms related to XAI are defined. Then, a classification of XAI models is presented and selected XRL models are sorted into these categories. Since there already is an abundance of sources on XAI but less so about XRL specifically, the focus of this work lies on providing information about and presenting sample methods of XRL models[1]. Thus, we present one method for each category in more detail and give a critical evaluation over the existing XRL methods. We will especially emphasize the need for XAI/XRL models that are adapted to a human (non-expert) end-user since that is what many XRL models are lacking but which we deem crucial.

1.1 The importance of explainability

Why is explainability so crucial? First, there is one obvious psychology-related reason: 'if the users do not trust a model or a prediction, they will not use it' [48, p. 1]. Trust can be defined as 'the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability' [32] and is an essential prerequisite of using a model or system [28, 12]. Bearing in mind the fact that transparency has been identified as one key component both in increasing users' trust [18, 67], as well as users' acceptance of a system [24], the interest in 'trustworthy AI' is not surprising, especially in the context of uncertainty [66]. Thus, in order to confidently use a system, it needs to be trusted, and in order to be trusted, it needs to be transparent and its decisions need to be justifiable (for a formal definition of transparency and related terms, see section 1.3).

[1] Please note that, while there is a distinction between Reinforcement Learning and Deep Reinforcement Learning (DRL), for the sake of simplicity, we will refer to both as just Reinforcement Learning going forward.

Second, AI technologies have become an essential part in almost all domains of Cyber-Physical Systems (CPSs). Reasons include the thrive for increased efficiency, business model innovations, or the necessity to accommodate volatile parts of today's critical infrastructures (e.g. a high share of renewable energy sources). In time, AI technologies evolved from being an additional input to an otherwise soundly defined control system, over fully decentralized, but still rule-governed systems (e.g. the *Universal Smart Grid Agent* [60]), to a system where all behavior originates from ML (e.g. *AlphaGo Zero* and *MuZero* [51]). For CPS analysis and operation, however, Adversarial Resilience Learning (ARL) has emerged as a novel methodology based on DRL [15, 59]. It is specifically designed to analyse and control critical infrastructures; obviously, explainability is tantamount here.

There is also a legal component to be considered; the EU General Data Protection Regulation (GDPR) [14], which came into effect in May 2018, aims to ensure a 'right to explanation' [19, p. 1] concerning automated decision-making and profiling. It states that '[...] such processing should subject to suitable safeguards, which should include [...] the right to obtain human intervention [...] [and] an explanation of the decision reached after such assessment' [14, recital 71]. Additionally, the European Commission set out an AI strategy with transparency and accountability as important principles to be respected [55], and in their Guidelines on trustworthy AI [56] they state seven key requirements, with transparency and accountability as two of them.

Finally, there are important practical reasons to consider; despite the increasing efficiency and versatility of AI, its incomprehensibility reduces its usefulness, since 'incomprehensible decision-making can still be effective, but its effectiveness does not mean that it cannot be faulty' [33, p. 1]. For example, in [54], neural nets successfully learnt to classify pictures but could be led to misclassification by (to humans) nearly imperceptible perturbations, and in [45], deep neural nets classified unrecognizable images with >99% certainty. This shows that a high level of effectiveness (under standard conditions) or even confidence does not imply that the decisions are correct or based on appropriately-learnt data. There exists a number of 'AI explainability toolkits' [2] that have common explainability methods implemented and enable their (easy) use with a pre-built framework, but, as far as we know, there is none specifically for RL.

Bearing this in mind, and considering the fact that, nowadays, AI can act increasingly autonomous, explaining and justifying the decisions is now more crucial than ever, especially in the domain of RL where an agent learns by itself, without human interaction.

1.2 Reinforcement Learning

Reinforcement Learning is a trial-and-error learning algorithm in which an autonomous agent tries to find the optimal solution to a problem through automated learning [52]. It is usually introduced as a Markov Decision Process

[2] E.g. the AI Explainability 360 (AIX360) as the currently most comprehensive one [4](see also for a list of other toolkits).

(MDP) if it satisfies the Markov property: the next state depends only on the current state and the agent's action(s), not on past states [30].

The learning process is initiated by an agent randomly performing an action which leads to a certain environmental state. This state has a reward assigned to it depending on how desirable this outcome is, set by the designer of the task. The algorithm will then learn a policy, i.e., an action-state-relation, in order to maximize the cumulative reward and be able to select the most optimal action in each situation. For more information on RL, see also [52, 34].

1.3 Definition of important terms

As already mentioned in section 1, the more complex a systems becomes, the less obvious its inner workings become. Additionally, there is no uniform term for this trade-off in the literature; XAI methods use an abundance of related, but distinct terms like transparency, reachability, etc... This inconsistency can be due to one or both of the following reasons: a) different terms are used in the same sense due to a lack of official definition of these terms, or b) different terms are used because the authors (subjectively) draw a distinction between them, without an official accounting of these differences. In any case, a uniform understanding and definition of what it means if a method is described as 'interpretable' or 'transparent' is important in order to clarify the potential, capacity and intention of a model. This is not an easy task, since there is no unique definition for the different terms to be found in the literature; even for 'interpretability', the concept which is most commonly used, 'the term [...] holds no agreed upon meaning, and yet machine learning conferences frequently publish papers which wield the term in a quasi-mathematical way' [35]. In Doshi-Velez and Kim [11, p. 2], interpretability is 'the ability to explain or to present in understandable terms to a human', however, according to Kim et al. [31, p. 7] 'a method is interpretable if a user can correctly and efficiently predict the method's result'. Some authors use transparency as a synonym for interpretability [35], some use comprehensibility as a synonym [16], then again others draw a distinction between the two [10] (for more information on how the different terms are used in the literature, we refer the reader to [35, 36, 11, 16, 31, 10, 44, 8]). If we tackle this issue in a more fundamental way, we can look at the definition of 'to interpret' or 'interpretation'. The Oxford Learners Dictionary[3] defines it as follows:

- to explain the meaning of something
- to decide that something has a particular meaning and to understand it in this way
- to translate one language into another as it is spoken
- the particular way in which something is understood or explained

Seeing that, according to the definition, interpretation contains an explanation, we can look at the definition for 'to explain'/'explanation':

[3] https://www.oxfordlearnersdictionaries.com/

- to tell somebody about something in a way that makes it easy to understand
- to give a reason, or be a reason, for something
- a statement, fact, or situation that tells you why something happened
- a statement or piece of writing that tells you how something works or makes something easier to understand

Both definitions share the notion of conveying the reason and meaning of something in order to make someone understand, but while an explanation is focused on *what* to explain, an interpretation has the additional value of considering *how* to explain something; it translates and conveys the information in a way that is more easily understood. And that is, in our opinion, essential in the frame of XAI/XRL: not only extracting the necessary information, but also presenting it in an appropriate manner, translating it from the 'raw data' into something humans and especially laypersons can understand.

So, because we deem a shared consensus on the nomenclature important, we suggest the use of this one uniform term, *interpretability*, to refer to the ability to not only extract or generate explanations for the decisions of the model, but also to present this information in a way that is understandable by human (non-expert) users to, ultimately, enable them to predict a model's behaviour.

2 XAI Taxonomy

While there are slight differences between the different taxonomy approaches [2, 4, 7], XAI methods can be broadly categorized based on two factors; first, based on when the information is extracted, the method can be intrinsic or post-hoc, and second, the scope can be either global or local (see figure 2, and figure 3 for examples).

Global and local interpretability refer to the scope of the explanation; global models explain the entire, general model behaviour, while local models offer explanations for a specific decision [43]. Global models try to explain the whole logic of a model by inspecting the structures of the model [2, 13]. Local explanations try to answer the question: 'Why did the model make a certain prediction/decision for an instance/for a group of instances?' [43, 2]. They also try to identify the contributions of each feature in the input towards a specific output [13]. Additionally, global interpretability techniques lead to users trusting a model, while local techniques lead to trusting a prediction [13].

Intrinsic vs. post-hoc interpretability depend on the time when the explanation is extracted/generated; An intrinsic model is a ML model that is constructed to be inherently interpretable or self-explanatory at the time of training by restricting the complexity of the model [13]. Decision trees, for example, have a simple structure and can be easily understood [43]. Post-hoc interpretability, in contrast, is achieved by analyzing the model after training by creating a second, simpler model, to provide explanations for the original model [13, 43]. Surrogate models or saliency maps are examples for this type [2]. Post-hoc interpretation models can be applied to intrinsic interpretation models, but not necessarily vice versa. Just like the models themselves, these interpretability models also suffer

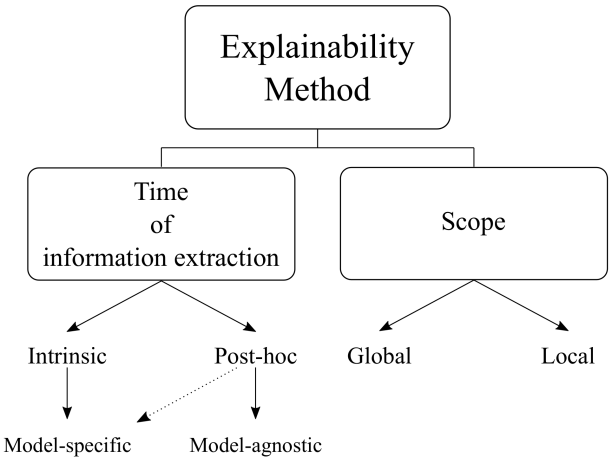

Fig. 2. A pseudo ontology of XAI methods taxonomy. Adapted from Adadi and Berrada [2].

from a transparency-accuracy-trade-off; intrinsic models usually offer accurate explanations, but, due to their simplicity, their prediction performance suffers. Post-hoc interpretability models, in contrast, usually keep the accuracy of the original model intact, but are harder to derive satisfying and simple explanations from [13].

Another distinction, which usually coincides with the classification into intrinsic and post-hoc interpretability, is the classification into model-specific or model-agnostic. Techniques are model-specific if they are limited to a specific model or model class [43], and they are model-agnostic if they can be used on any model [43]. As you can also see in figure 2, intrinsic models are model-specific, while post-hoc interpretability models are usually model-agnostic.

Adadi and Berrada [2] offer an overview of common explainability techniques and their rough (i.e., neither mutually exclusive nor exhaustive) classifications into these categories. In section 3, we follow their example and provide classifications for a list of selected XRL method papers.

3 Non-exhaustive list of XRL methods

A literature review was conducted using the database Google Scholar. Certain combinations of keywords were used to select papers; first, 'explainable reinforcement learning', and 'XRL' together with 'reinforcement learning' and 'machine learning' were used. Then, we substituted 'explainable' for common variations used in literature like 'explainable', 'transparent', and 'understandable'. We then scanned the papers for relevance and consulted their citations and reference lists for additional papers. Because we only wanted to focus on current methods, we

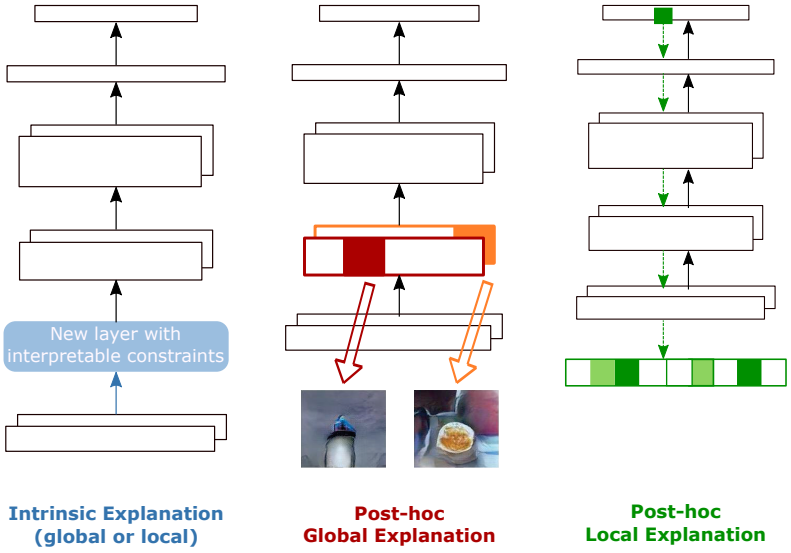

Fig. 3. An illustration of global vs. local, and intrinsic vs. post-hoc interpretable machine learning techniques, with a deep neural network as an example. On the left, the model and the layers' constraints are built in a way that is inherently interpretable (intrinsic interpretability). The middle and right column show post-hoc interpretability, achieved by a global and local explanation, respectively. The global explanation explains the different representations corresponding to the different layers in general, while the local explanation illustrates the contribution of the different input features to a certain output. Adopted from Du et al. [13].

restricted the search to papers from 2010-2020. Table 1 shows the list of selected papers and their classification according to section 2 based on our understanding.

For a more extensive demonstration of the different approaches, we chose the latest paper of each quadrant [4] and explain them in more detail in the following sections as an example for the different XRL methods.

3.1 Method A: Programmatically Interpretable Reinforcement Learning

Verma et al. [61] have developed 'PIRL', a Programmatically Interpretable Reinforcement Learning framework, as an alternative to DRL. In DRL, the policies are represented by neural networks, making them very hard (if not impossible) to interpret. The policies in PIRL, on the other hand, while still mimicking the ones from the DRL model, are represented using a high-level, human-readable

[4] With the exception of method C in section 3.3, where we present a Linear Model U-Tree method although another paper with a different, but related method was published slightly later. See the last paragraph of that section for our reasoning for this decision.

Table 1. Selected XRL methods and their categorization according to the taxonomy described in section 2.

Time \ Scope	Global	Local
Intrinsic	• **PIRL** (Verma et al. [61]) • Fuzzy RL policies (Hein et al. [22])	• **Hierarchical Policies** (Shu et al. [53])
Post-hoc	• Genetic Programming (Hein et al. [23]) • Reward Decomposition (Juozapaitis et al. [29]) • Expected Consequences (van der Waa et al. [62]) • Soft Decision Trees (Coppens et al. [9]) • Deep Q-Networks (Zahavy et al. [64]) • Autonomous Policy Explanation (Hayes and Shah [21]) • Policy Distillation (Rusu et al. [50]) • **Linear Model U-Trees** (Liu et al. [37])	• Interestingness Elements (Sequeira and Gervasio [52]) • Autonomous Self-Explanation (Fukuchi et al. [17]) • **Structural Causal Model** (Madumal et al. [40]) • Complementary RL (Lee [33]) • Expected Consequences (van der Waa et al. [62]) • Soft Decision Trees (Coppens et al. [9]) • **Linear Model U-Trees** (Liu et al. [37])

Notes. Methods in bold are presented in detail in this work.

programming language. Here, the problem stays the same as in traditional RL (i.e., finding a policy that maximises the long-term reward), but in addition, they restrict the vast amount of target policies with the help of a *(policy) sketch*. To find these policies, they employ a framework which was inspired by imitation learning, called *Neurally Directed Program Search (NDPS)*. This framework first uses DRL to compute a policy which is used as a neural 'oracle' to direct the policy search for a policy that is as close as possible to the neural oracle. Doing this, the performances of the resulting policies are not as high than the ones from the DRL, but they are still satisfactory and, additionally, more easily interpretable. They evaluate this framework by comparing its performance with, among others, a traditional DRL framework in The Open Racing Car Simulator (TORCS) [63]. Here, the controller has to set five parameters (acceleration, brake, clutch, gear and steering of the car) to steer a car around a race track as fast as possible. Their results show that, while the DRL leads to quicker lap time, the NDPS still outperforms this for several reasons: it shows much smoother driving (i.e., less steering actions) and is less perturbed by noise and blocked sensors. It also is easier to interpret and is better at generalization, i.e., it performs better in situations (in this case, tracks) not encountered during training than a DRL model.

Concerning restrictions of this method, it is worth noting that the authors only considered environments with symbolic inputs, not perceptual, in their experiments. They also only considered deterministic policies, not stochastic policies.

3.2 Method B: Hierarchical and Interpretable Skill Acquisition in Multi-task Reinforcement Learning

Shu et al.[53] proposed a new framework for multi-task RL using hierarchical policies that addressed the issue of solving complex tasks that require different skills and are composed of several (simpler) subtasks. It is based on and extends multi-task RL with modular policy design through a two-layer hierarchical policy [3] by incorporating less assumptions, and, thus, less restrictions. They trained and evaluated their model with object manipulation tasks in a Minecraft game setting (e.g. finding, getting, or stacking blocks of a certain color), employing advantage actor-critic as policy optimization using off-policy learning. The model is hierarchical because each top-level policy (e.g., 'stack x') consists of several lower levels of actions ('find x'→ 'get x'→ 'put x', see also figure 4). The novelty of this method is the fact that each task is described by a human instruction (e.g. 'stack blue'), and agents can only access learnt skills through these descriptions, making its policies and decisions inherently human-interpretable.

Additionally, a key idea of their framework is that a complex task could be decomposed into several simpler subtasks. If these sub-tasks could be fulfilled by employing an already learnt 'base policy', no new skill had to be learnt; otherwise, it would learn a new skill and perform a different, novel action. To boost efficiency and accuracy, the framework also incorporated a stochastic temporal grammar model that was used to model temporal relationships and priorities of tasks (e.g., before stacking a block on top of another block, you must first obtain said block).

The resulting framework could efficiently learn hierarchical policies and representations in multi-task RL, only needing weak human supervision during training to decide which skills to learn. Compared to a flat policy that directly maps the state and instruction to an action, the hierarchical model showed a higher learning efficiency, could generalize well in new environments, and was inherently interpretable.

3.3 Method C: Toward Interpretable Deep Reinforcement Learning with Linear Model U-Trees

In Liu et al. [37], a mimic learning framework based on stochastic gradient descent is introduced. This framework approximates the predictions of an accurate, but complex model by mimicking the model's Q-function using Linear Model U-Trees (LMUTs). LMUTs are an extension of Continuous U-Trees (CUTs) which were developed to approximate continuous functions [58]. The difference between CUTs and LMUTs is that, instead of constants, LMUTs have a linear model at each leaf node which also improves its generalization ability. They also generally have fewer leaves and are therefore simpler and more easily understandable. The

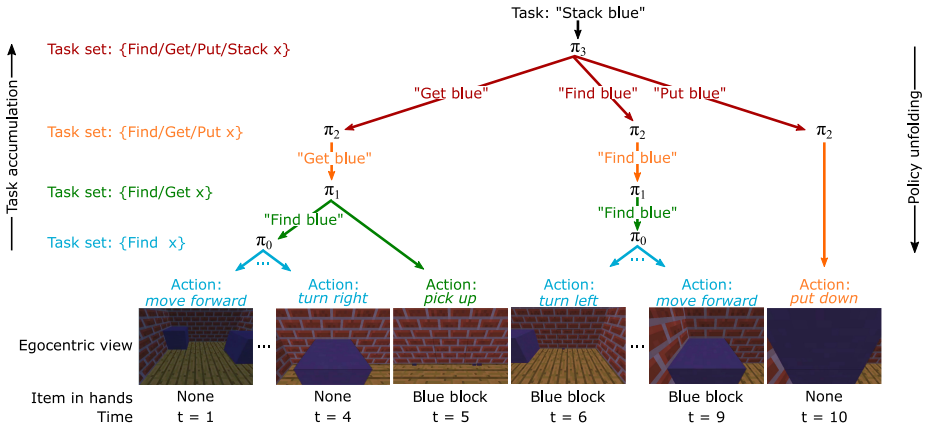

Fig. 4. Example for the mutli-level hierarchical policy for the task to stack two blue boxes on top of each other. The top-level policy (π_3, in red) encompasses the high-level plan 'get blue'→'find blue'→'put blue'. Each step (i.e., arrow) either initiates another policy (marked by a different color) or directly executes an action. Adopted from Shu et al. [53].

novelty of this method lies in the fact that other tree representations used for interpretations were only developed for supervised learning, not for DRL.

The framework can be used to analyze the importance of input features, extract rules, and calculate 'super-pixels' ('contiguous patch[es] of similar pixels' [48, p. 1]) in image inputs (see table 2 and figure 5 for an example). It has two approaches to generate data and mimic the Q-function; the first one is an *experience training setting* which records and generates data during the training process for batch training. It records the state-action pairs and the resulting Q-values as 'soft supervision labels' [37, p. 1] during training. In cases where the mimic learning model cannot be applied to the training process, the second approach can be used: *active play setting*, which generates mimic data by applying the mature DRL to interact with the environment. Here, an online algorithm is required which uses stochastic gradient descent to dynamically update the linear models as more data is generated.

Table 2. Examples of feature influences in the Mountain Car and Cart Pole scenario, extracted by the LMUTs in Liu et al. [37]

	Feature	Influence
Mountain Car	Velocity	376.86
	Position	171.28
Cart Pole	Pole Angle	30541.54
	Cart Velocity	8087.68
	Cart Position	7171.71
	Pole Velocity At Tip	2953.73

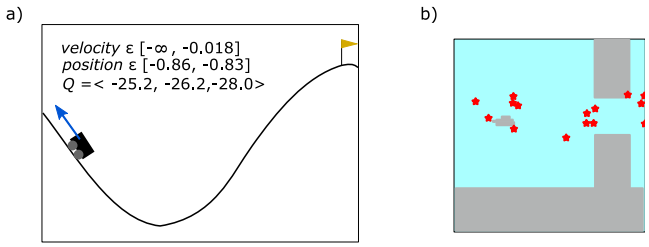

Fig. 5. Examples of a) rule extraction, and b) super-pixels extracted by the LMUTs in Liu et al. [37]. a) Extracted rules for the mountain Cart scenario. Values at the top are the range of velocity and position and a Q vector (Q_{move_left},Q_{no_push}, Q_{move_right}) representing the average Q-value). In this example, the cart is moving to the left to the top of the hill. The car should be pushed left (Q_{move_left} is highest) to prepare for the final rush to the target on the right side. b) Super-pixels for the Flappy Bird scenario, marked by red stars. This is the first of four sequential pictures where the focus lies on the location of the bird and obstacles (i.e., pipes). In later pictures the focus would shift towards the bird's location and velocity.

They evaluate the framework in three benchmark environments: Mountain Car, Cart Pole, and Flappy Bird, all simulated by the OpenAI Gym toolkit [6]. Mountain Car and Cart Pole have a discrete action space and a continuous feature space, while Flappy Bird has two discrete actions and four consecutive images as inputs which result in 80x80 pixels each, so 6400 features. The LMUT method is compared to five other tree methods: a CART regression tree [39], M5 trees [47] with regression tree options (M5-RT) and with model tree options (M5-MT), and Fast Incremental Model Trees (FIMT, [27]) in the basic version, and in the advanced version with adaptive filters (FIMT-AF). The two parameters *fidelity* (how well the predictions of the mimic model match those from the mimicked model) and *play performance* (how well the average return in the mimic model matches that of the mimicked model) are used as evaluation metrics. Compared to CART and FIMT (-AF), the LMUT model showed higher fidelity with fewer leaves. For the Cart Pole environment, LMUT showed the higest fidelity, while the M5 trees showed higher performance for the other two environments, although LMUT was comparable. Concerning the play performance, the LMUT model performs best out of all the models. This was likely due to the fact that, contrary to the LMUTs, the M5 and CART trees fit equally over the whole training experience which includes sub-optimal actions in the beginning of training, while the FIMT only adapts to the most recent input and thus cannot build linear models appropriately. In their work, this is represented by sorting the methods on an axis between 'data coverage' (when the mimic model matches the mimicked model on a large section of the state space) and 'data optimality' (when it matches the states most important for performance) with the LMUT at the, as they call it, 'sweet spot between optimality and coverage' (p. 12, see also figure 6).

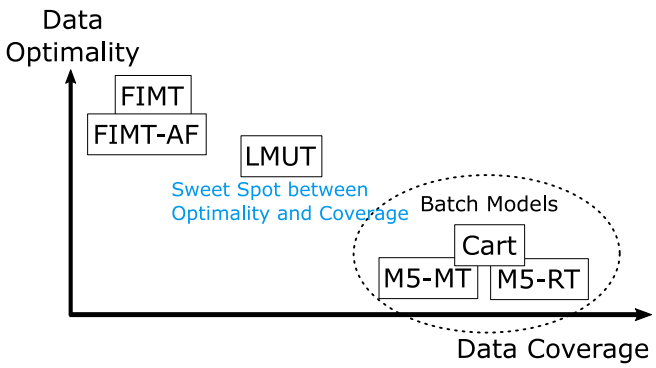

Fig. 6. Placement of the different tree models on the axes data coverage vs. data optimality. Adapted from Liu et al. [37].

There is a similar, newer tree method that uses Soft Decision Trees (SDTs) to extract DRL polices [9]. This method was not presented in this paper because, for one thing, it is less versatile (not offering rule extraction, for example), and for another, it was not clear whether the SDTs actually adequately explained the underlying, mimicked policy for their used benchmark.

3.4 Method D: Explainable RL Through a Causal Lens

According to Madumal et al. [40], not only is it important for a RL agent to explain itself and its actions, but also to bear in mind the human user at the receiving end of this explanation. Thus, they took advantage of the prominent theory that humans develop and deploy causal models to explain the world around them, and have adapted a structural causal model (SCM) based on Halpern [20] to mimic this for model-free RL. SCMs represent the world with random exogenous (external) and endogenous (internal) variables, some of which might exert a causal influence over others. These influences can be described with a set of structural equations.

Since Madumal et al. [40] focused on providing explanations for an agent's behaviour based on the knowledge of how its actions influence the environment, they extend the SCM to include the agent's actions, making it an *action influence model*. More specifically, they offer 'actuals' and 'counterfactuals', that is, their explanations answer 'Why?' as well as 'Why not?' questions (e.g. 'Why (not) action A?'). This is noticeable because, contrary to most XAI models, it not only considers actual events occured, but also hypothetical events that did not happen, but could have.

In more detail, the process of generating explanations consists of three phases; first, an action influence model in the form of a directed acyclic graph (DAG) is required (see figure 7 for an example). Next, since it is difficult to uncover the true structural equations describing the relationships between the variables, this problem is circumvented by only approximating the equations so that they are

exact enough to simulate the counterfactuals. In Madumal et al. [40], this is done by multivariate regression models during the training of the RL agent, but any regression learner can be used. The last phase is generating the explanations, more specifically, *minimally complete contrastive explanations*. This means that, first, instead of including the vectors of variables of ALL nodes in the explanation, it only includes the absolute minimum variables necessary. Moreover, it explains the actual (e.g. 'Why action A?') by simulating the counterfactual (e.g. 'Why not action B?') through the structural equations and finding the differences between the two. The explanation can then be obtained through a simple NLP template (for an example of an explanation, again, see figure 7).

Madumal et al. [40]'s evaluations of the action influence model show promising results; in a comparison between six RL benchmark domains measuring accuracy ('Can the model accurately predict what the agent will do next?') and performance (training time), the model shows reasonable task prediction accuracy and negligible training time. In a human study, comparing the action influence model with two different models that have learnt how to play Starcraft II (a real-time strategy game), they assessed task prediction by humans, explanation satisfaction, and trust in the model. Results showed that the action influence model performs significantly better for task prediction and explanation satisfaction, but not for trust. The authors propose that, in order to increase trust, further interaction might be needed. In the future, advancements to the model can be made including extending the model to continuous domains or targeting the explanations to users with different levels of knowledge.

4 Discussion

In this paper, inspired by the current interest in and demand for XAI, we focused on a particular field of AI: Reinforcement Learning. Since most XAI methods are tailored for supervised learning, we wanted to give an overview of methods employed only on RL algorithms, since, to the best of our knowledge, there is no work present at the current point in time addressing this.

First, we gave an overview over XAI, its importance and issues, and explained related terms. We stressed the importance of a uniform terminology and have thus suggested and defined a term to use from here on out. The focus, however, lay on collecting and providing an overview over the aforementioned XRL methods. Based on Adadi and Berrada [2]'s work, we have sorted selected methods according to the scope of the method and the time of information extraction. We then chose four methods, one for each possible combination of those categorizations, to be presented in detail.

Looking at the collected XRL methods, it becomes clear that post-hoc interpretability models are much more prevalent than intrinsic models. This makes sense, considering the fact that RL models were developed to solve tasks without human supervision that were too difficult for un-/supervised learning and are thus highly complex; it is, apparently, easier to simplify an already existing, complex model than it is to construct it to be simple in the first place. It

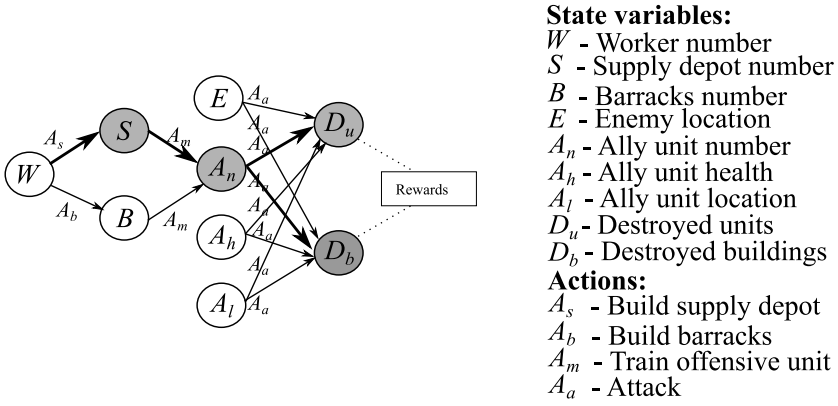

Fig. 7. Action influence graph of an agent playing Starcraft II, a real-time strategy game with a large state and action space, reduced to four actions and nine state variables for the purpose of generating the explanations. In this case, the causal chain for the actual action 'Why A_s?' is shown in bold, and the chain for the counterfactual action 'Why not A_b?' would be $B \rightarrow A_n \rightarrow [D_u, D_b]$. The explanation to the question 'Why not build_barracks (A_b)?' would be 'Because it is more desirable to do action build_supply_depot (A_s) to have more Supply Depots (S) as the goal is to have more Destroyed Units (D_u) and Destroyed buildings (D_b)'. Adapted from Madumal et al. [40].

seems that the performance-interpretability trade-off is present not only for the AI methods themselves, but also for the explainability models applied to them.

The allocation to global vs. local scope, however, seems to be more or less balanced. Of course, the decision to develop a global or a local method is greatly dependent on the complexity of the model and the task being solved, but one should also address the question if one of the two is more useful or preferable to human users. In van der Waa et al.'s study [62], for example, 'human users tend to favor explanations about policy rather than about single actions' (p. 1).

In general, the form of the explanation and the consideration of the intended target audience is a very important aspect in the development of XAI/XRL methods. Holzinger et al. [26] emphasize the need for *causability*: a combination of causality - referring to the human reliance on causal models for explanations - and usability - a term from the area of human-computer interaction defined as 'a function of the ease of use [...] and the acceptability of the product' [5, p. 2]. Just as usability measures the 'quality of use', causability measures the 'quality of explanation'. In Holzinger et al. [25], they have developed a *System Causability Scale*, an explanation interface evaluating explanations along specific dimensions such as efficiency and completeness. XAI methods also need to exhibit *context-awareness*: adapting to environmental and user changes like the level of experience, cultural or educational differences, domain knowledge, etc., in order to be more human-centric [2]. The form and presentation of the explanation is essential as XAI 'can benefit from existing models of how people define,

generate, select, present, and evaluate explanations' [42, p. 59]. For example, research shows that (causal) explanations are contrastive, i.e., humans answer a 'Why X?' question through the answer to the – often only implied – counterfactual 'Why not Y instead?'. This is due to the fact that a complete explanation *for* a certain event (instead of an explanation *against* the counterevent) involves a higher cognititve load [42]. Not only that, but a layperson also seems to be more receptive to a contrastive explanation, finding it 'more intuitive and more valuable' [42, p. 20]). While there are some XAI studies focusing on the human side of the equation (e.g. [66, 41, 65, 67]), especially in XRL, this is often neglected [1].

Out of the papers covered in this work, we highlight Madumal et al.'s work [40], but also Sequeira and Gervasio [52] and van der Waa et al. [62]; of all thirteen selected XRL methods, only five evaluate (non-expert) user satisfaction and/or utility of a method [52, 29, 62, 17, 40], and only three of these offer contrastive explanations [40, 52, 62]. So, of *all* selected papers, only these free provide a combination of both, not only offering useful contrastive explanations, but also explicitly bearing in mind the human user at the end of an explanation.

4.1 Conclusion

For practical, legal, and psychological reasons, XRL (and XAI) is a quickly advancing field in research that has to address some key challenges to prove even more beneficial and useful. In order to have a common understanding about the goals and capabilities of an XAI/XRL model, a ubiquitous terminology is important; due to this, we suggest the term *interpretability* to be used from here on out and have defined it as 'the ability to not only extract or generate explanations for the decisions of the model, but also to present this information in a way that is understandable by human (non-expert) users to, ultimately, enable them to predict a model's behaviour'. This is closely related to *Causability*, refering to the quality of an explanation [26]. Different approaches are possible to achieve interpretability, depending on the scope (global vs. local) and the time of information extraction (intrinsic vs. post-hoc). Due to the complexity of a RL model, post-hoc interpretability seems to be easier to achieve than intrinsic interpretability: simplifying the original model (for example with the use of a surrogate model) instead of developing a simple model in the first place seems to be easier to achieve, but comes at the cost of accuracy/performance.

However, despite some existing research, especially XRL models often lack to consider the human user at the receiving end of an explanation and to adapt the model to them for maximum benefit. Research shows that contrastive explanations are more intuitive and valuable [42], and there is evidence that human users favor a global approach over a local one [62]. A context-aware system design is also important in order to cater to users with different characteristics, goals, and needs [2]. Especially considering the growing role of AI in critical infrastructures (for example analyzing and controlling power grids with models such as ARL [15, 59]), where the AI model might have to act autonomously or in cooperation

with a human user, being able to explain and justify the model's decisions is crucial.

To achieve this and be able to develop human-centered models for optimal and efficient human-computer interaction and cooperation, a bigger focus on interdisciplinary work is necessary, combining efforts from the fields of AI/ML, psychology, philosophy, and human-computer interaction.

5 Acknowledgements

This work was supported by the German Research Foundation under the grant GZ: JI 140/7-1. We thank our colleagues Stephan Balduin, Johannes Gerster, Lasse Hammer, Daniel Lange and Nils Wenninghoff for their helpful comments and contributions.

Bibliography

[1] Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and trajectories for explainable, accountable and intelligible systems. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18. ACM Press (2018)

[2] Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160 (2018), https://doi.org/10.1109/access.2018.2870052, 10.1109/access.2018.2870052

[3] Andreas, J., Klein, D., Levine, S.: Modular multitask reinforcement learning with policy sketches. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. p. 166–175. ICML'17, JMLR.org (2017)

[4] Arya, V., Bellamy, R.K.E., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K.R., Wei, D., Zhang, Y.: One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques (2019)

[5] Bevana, N., Kirakowskib, J., Maissela, J.: What is usability. In: Proceedings of the 4th International Conference on HCI. Citeseer (1991)

[6] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W.: Openai gym (2016)

[7] Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics 8(8), 832 (2019), https://doi.org/10.3390/electronics8080832

[8] Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R.M., Kelley, T.D., Braines, D., Sensoy, M., Willis, C.J., Gurram, P.: Interpretability of deep learning models: A survey of results. In: 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE (2017)

[9] Coppens, Y., Efthymiadis, K., Lenaerts, T., Nowé, A., Miller, T., Weber, R., Magazzeni, D.: Distilling deep reinforcement learning policies in soft decision trees. In: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence. pp. 1–6 (2019)
[10] Doran, D., Schulz, S., Besold, T.R.: What does explainable ai really mean? a new conceptualization of perspectives (2017)
[11] Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017)
[12] Dosilovic, F.K., Brcic, M., Hlupic, N.: Explainable artificial intelligence: A survey. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE (2018), https://doi.org/10.23919/mipro.2018.840004
[13] Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. Communications of the ACM 63(1), 68–77 (2019), https://doi.org/10.1145/3359786
[14] European Commission, Parliament: Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 1–88 (2016)
[15] Fischer, L., Memmen, J.M., Veith, E.M., Tröschel, M.: Adversarial resilience learning—towards systemic vulnerability analysis for large and complex systems. In: The Ninth International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies (ENERGY 2019). vol. 9, pp. 24–32 (2019)
[16] Freitas, A.A.: Comprehensible classification models. ACM SIGKDD Explorations Newsletter 15(1), 1–10 (2014)
[17] Fukuchi, Y., Osawa, M., Yamakawa, H., Imai, M.: Autonomous self-explanation of behavior for interactive reinforcement learning agents. In: Proceedings of the 5th International Conference on Human Agent Interaction - HAI '17. ACM Press (2017)
[18] Glass, A., McGuinness, D.L., Wolverton, M.: Toward establishing trust in adaptive agents. In: Proceedings of the 13th international conference on Intelligent user interfaces - IUI '08. ACM Press (2008)
[19] Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". AI Magazine 38(3), 50–57 (2017)
[20] Halpern, J.Y.: Causes and explanations: A structural-model approach. part II: Explanations. The British Journal for the Philosophy of Science 56(4), 889–911 (2005)
[21] Hayes, B., Shah, J.A.: Improving robot controller transparency through autonomous policy explanation. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17. ACM Press (2017)
[22] Hein, D., Hentschel, A., Runkler, T., Udluft, S.: Particle swarm optimization for generating interpretable fuzzy reinforcement learning poli-

cies. Engineering Applications of Artificial Intelligence 65, 87–98 (2017), https://doi.org/10.1016/j.engappai.2017.07.005
[23] Hein, D., Udluft, S., Runkler, T.A.: Interpretable policies for reinforcement learning by genetic programming. Engineering Applications of Artificial Intelligence 76, 158–169 (2018)
[24] Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM conference on Computer supported cooperative work - CSCW '00. ACM Press (2000)
[25] Holzinger, A., Carrington, A., Müller, H.: Measuring the quality of explanations: The system causability scale (SCS). KI - Künstliche Intelligenz 34(2), 193–198 (2020), https://doi.org/10.1007/s13218-020-00636-z
[26] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery 9(4) (2019), https://doi.org/10.1002/widm.1312
[27] Ikonomovska, E., Gama, J., Džeroski, S.: Learning model trees from evolving data streams. Data Mining and Knowledge Discovery 23(1), 128–168 (2010)
[28] Israelsen, B.W., Ahmed, N.R.: "dave...i can assure you ...that it's going to be all right ..." a definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. ACM Computing Surveys 51(6), 1–37 (2019)
[29] Juozapaitis, Z., Koul, A., Fern, A., Erwig, M., Doshi-Velez, F.: Explainable reinforcement learning via reward decomposition. In: Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence. pp. 47–53 (2019)
[30] Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey (1996)
[31] Kim, B., Khanna, R., Koyejo, O.O.: Examples are not enough, learn to criticize! criticism for interpretability. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29. pp. 2280–2288. Curran Associates, Inc. (2016), http://papers.nips.cc/paper/6300-examples-are-not-enough-learn-to-criticize-criticism-for-interpretability.pdf
[32] Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society 46(1), 50–80 (2004), https://doi.org/10.1518/hfes.46.1.50_30392
[33] Lee, J.H.: Complementary reinforcement learning towards explainable agents (2019)
[34] Li, Y.: Deep reinforcement learning (2018)
[35] Lipton, Z.C.: The mythos of model interpretability (2016)
[36] Lipton, Z.C.: The mythos of model interpretability. Communications of the ACM 61(10), 36–43 (2018)
[37] Liu, G., Schulte, O., Zhu, W., Li, Q.: Toward interpretable deep reinforcement learning with linear model u-trees. In: Machine Learning and Knowledge Discovery in Databases, pp. 414–429. Springer International Publishing (2019)

[38] Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P.Q., Corrado, G.S., Hipp, J.D., Peng, L., Stumpe, M.C.: Detecting cancer metastases on gigapixel pathology images (2017)
[39] Loh, W.Y.: Classification and regression trees. WIREs Data Mining and Knowledge Discovery 1(1), 14–23 (2011)
[40] Madumal, P., Miller, T., Sonenberg, L., Vetere, F.: Explainable reinforcement learning through a causal lens (2019)
[41] Martens, D., Vanthienen, J., Verbeke, W., Baesens, B.: Performance of classification models from a user perspective. Decision Support Systems 51(4), 782–793 (2011)
[42] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267, 1–38 (2019)
[43] Molar, C.: Interpretable machine learning (2018), https://christophm.github.io/interpretable-ml-book/, [Retrieved: 2020-03-31]
[44] Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing 73, 1–15 (2018)
[45] Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
[46] Nguyen, T.T., Hui, P.M., Harper, F.M., Terveen, L., Konstan, J.A.: Exploring the filter bubble. In: Proceedings of the 23rd international conference on World wide web - WWW '14. ACM Press (2014)
[47] Quinlan, J.R., et al.: Learning with continuous classes. In: 5th Australian joint conference on artificial intelligence. vol. 92, pp. 343–348. World Scientific (1992)
[48] Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?". In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. ACM Press (2016)
[49] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1(5), 206–215 (2019)
[50] Rusu, A.A., Colmenarejo, S.G., Gulcehre, C., Desjardins, G., Kirkpatrick, J., Pascanu, R., Mnih, V., Kavukcuoglu, K., Hadsell, R.: Policy distillation (2015)
[51] Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al.: Mastering ATARI, go, chess and shogi by planning with a learned model (2019)
[52] Sequeira, P., Gervasio, M.: Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations (2019)
[53] Shu, T., Xiong, C., Socher, R.: Hierarchical and interpretable skill acquisition in multi-task reinforcement learning (2017)
[54] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks (2013)

[55] The European Commission: Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. The European Commission (2018), https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe, article; accessed 27.03.2020
[56] The European Commission: Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission. The European Commission (2018), https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe, article; accessed 27.04.2020
[57] Tomzcak, K., Pelter, A., Gutierrez, C., Stretch, T., Hilf, D., Donadio, B., Tenhundfeld, N.L., de Visser, E.J., Tossell, C.C.: Let tesla park your tesla: Driver trust in a semi-automated car. In: 2019 Systems and Information Engineering Design Symposium (SIEDS). IEEE (2019)
[58] Uther, W.T., Veloso, M.M.: Tree based discretization for continuous state space reinforcement learning. In: Aaai/iaai. pp. 769–774 (1998)
[59] Veith, E., Fischer, L., Tröschel, M., Nieße, A.: Analyzing cyber-physical systems from the perspective of artificial intelligence. In: Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control. ACM (2019)
[60] Veith, E.M.: Universal Smart Grid Agent for Distributed Power Generation Management. Logos Verlag Berlin GmbH (2017)
[61] Verma, A., Murali, V., Singh, R., Kohli, P., Chaudhuri, S.: Programmatically interpretable reinforcement learning. PMLR 80:5045-5054 (2018)
[62] van der Waa, J., van Diggelen, J., van den Bosch, K., Neerincx, M.: Contrastive explanations for reinforcement learning in terms of expected consequences. IJCAI-18 Workshop on Explainable AI (XAI). Vol. 37. 2018 (2018)
[63] Wymann, B., Espié, E., Guionneau, C., Dimitrakakis, C., Coulom, R., Sumner, A.: Torcs, the open racing car simulator. Software available at http://torcs. sourceforge. net 4(6), 2 (2000)
[64] Zahavy, T., Zrihem, N.B., Mannor, S.: Graying the black box: Understanding dqns (2016)
[65] Zhou, J., Chen, F. (eds.): Human and Machine Learning. Springer International Publishing (2018), https://doi.org/10.1007/978-3-319-90403-0
[66] Zhou, J., Chen, F.: Towards trustworthy human-ai teaming under uncertainty. In: IJCAI 2019 Workshop on Explainable AI (XAI) (2019)
[67] Zhou, J., Hu, H., Li, Z., Yu, K., Chen, F.: Physiological indicators for user trust in machine learning with influence enhanced fact-checking. In: Lecture Notes in Computer Science, pp. 94–113. Springer International Publishing (2019), https://doi.org/10.1007/978-3-030-29726-8_7