

Explainability in Deep Reinforcement Learning, a Review into Current Methods and Applications

Thomas Hickling¹, Abdelhafid Zenati¹, Nabil Aouf¹ and Phillippa Spencer²

Department of Electrical Engineering, City University of London¹
Defence Science and Technology Laboratory (Dstl)²

Abstract

The use of Deep Reinforcement Learning (DRL) schemes has increased dramatically since their first introduction in 2015. Though uses in many different applications are being found they still have a problem with the lack of interpretability. This has bred a lack of understanding and trust in the use of DRL solutions from researchers and the general public. To solve this problem the field of explainable artificial intelligence (XAI) has emerged. This is a variety of different methods that look to open the DRL black boxes, they range from the use of interpretable symbolic decision trees to numerical methods like Shapley Values. This review looks at which methods are being used and what applications they are being used. This is done to identify which models are the best suited to each application or if a method is being underutilised.

1 Introduction

Machine Learning (ML) in commercial and public projects to optimise solutions has increased rapidly over the past decade. Tasks such as weather simulation, medical diagnosis, business optimisation and automation like driver-less cars have benefited from these new Artificial Intelligence methods. Some of these ML models are used in ways that their predictions can affect people's safety or commercial success. These models must be considered trustworthy with errors detected before it is too late to react.

One of these ML models is Neural Networks (NN) and specifically, Deep Neural Networks (DNN). Due to the nature of DNNs, the decisions they produce can seem arbitrary. The network acts as a black box where the operator cannot peer inside. It is especially true for the people who end up operating these networks but do not necessarily know how they work. The engineers understand the underlying mathematics, but that does not give insight enough into how the model will produce an answer without further analysis. The effort to explain these networks has become known as Explainable Artificial Intelligence (XAI). XAI techniques have varied in methods and applications. This paper will review how the methods are applied across industries to try and gain insight into the path that XAI is going to take.

Deep Reinforcement Learning (DRL) is a type of ML that instead of using a data set to train the model it is trained by using a Markov Decision Process that chooses a decision at every time step. This decision is picked by trying to maximise a reward function. Over many episodes the algorithm will learn which actions produce the highest rewards this will eventually lead to an optimal solution. These optimal solutions may be the same as what human operators would decide as the correct solution, this is seen when a DRL agent is used to play the Atari game Space Invaders where both machine and human players keep the aliens in a block formation. However this method has a benefit of being able to find innovative solutions to the problems it is tasked with that will differ from human led methods, this can be either by using novel strategies or by exploiting the game mechanics. As there is no guidance apart from a simple reward function the model is able to find the most efficient solution.

As the DRL method for creating models can lead to more innovative solutions that are not obvious to a human observer it is important that XAI can be used to check the solutions are correct and do not contain bias. These explanations can also be used to examine any novel strategies or solutions that these DRL models formulate, such as learning why a strategy in a video game has been chosen. As DRL models are used in time sensitive applications where a decision is made at every time step explanations must be delivered promptly to the user. Any safety critical operation of a model will require the trust of the operator with little knowledge of the underlying

mathematics, therefore the explanations have to be understandable by novices to be able to build that trust. With these applications that are safety dependent any accidents that take place under the control of a neural network XAI can be used to analyse the data post accident to determine the fault. Being able to access the fault that has caused could help with preventing future accidents or help with any insurance claims.

This review will first give an overview of the two subjects, DRL and XAI. After this a discussion of the state of the field as it stands at the moment will be done with a look at the work of two previous reviews. Following this will be the selection criteria that is used to choose the papers to be reviewed. After this there will be a summary of each paper before some concluding remarks.

2 Deep Reinforcement Learning

The original method for applying deep learning methods to reinforcement learning methods was the application of Q-learning formulated by Mnih et al.(2015) [38]. The Q-learning process begins with a Markov Decision Process (MDP) that takes the variables (S, A, R, P, γ) to interact with the environment.

- S: Set of the States.
- A: Set of the Actions.
- R: Set of all the possible Rewards from the actions taken. The rewards are denoted as being R_{t+1} after the action a_t and state s_t . The reward can also be described as $R(s)$ where this is the reward for reaching state s .
- P: Transition dynamics where $P(s'|s, a)$ is defined as the distribution of the next state s' where an action a has been taken from state s , where $s, s' \in S, a \in A$. The transition dynamics at the beginning is denoted as $P(s_0)$ or ρ_0 .
- γ : The discount factor with range $[0, 1]$.

To solve the MDP the DRL uses trial and error to map the policy between the states and the actions. The DRL uses the reward as a guide to what are good decisions. This is applied through Q-learning, a neural network performs as a function approximator. The addition of the neural network allows for schemes to conduct complex tasks.

To solve the MDP all these values need to be known but in many cases the reward function and the transition dynamics are not known. To solve this problem trial and error is used to explore the environment to build understanding of the inner information, this is what is known as reinforcement learning. The aim of the reinforcement learning is to map the policy π between the states and actions that maximises the expected discounted total reward over the agents lifetime. The action value function also known as the Q function describes this relationship as:

$$Q^\pi(s, a) = E^\pi \left[\sum_{t=0}^T \gamma^t R(s_t, a_t) \right] \quad (1)$$

In this equation E^π is the expected value by following policy π . This Q function can be solved using the Bellman equation:

$$Q^\pi(s, a) = R(s_t) + \gamma E[Q^\pi(s_{t+1}, a_{t+1})] \quad (2)$$

This equation allows for the Q-learning and SARSA recursive estimation procedures. It is possible to then improve the policy from π to $\hat{\pi}$ by greedily choosing a_t in each state:

$$\hat{\pi}(s_t) = \operatorname{argmax}_{a_t} Q^\pi(s_t, a_t) \quad (3)$$

This technique when paired with the use of deep neural networks that can perform as function approximators has led to the increase in the use of DRL for complex tasks. These neural networks are also the reason for the explainability problem that needs to be solved using XAI.

As DRL has matured there have been several innovations in the field that have created more models that can be used in a wider variety of situations. The field is split into model-based DRL and model-free DRL. Model-based systems use a defined model that can be queried to provide the future rewards and states allowing for the DRL agent to plan ahead to maximise the reward in the future. This was used by Google's AlphaZero model (Silver et al.(2016) [49]) to be able to

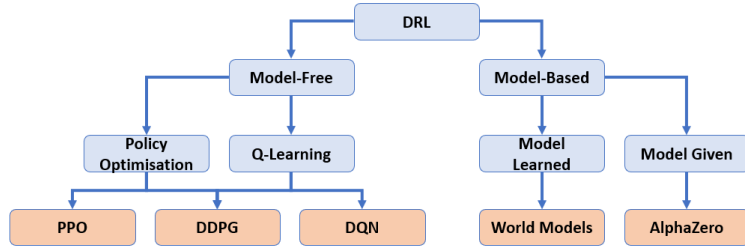


Figure 1: The types of DRL schemes with some examples.

strategise many moves in advance. In these model-based DRL modes, the model can either be learned or given to the agent.

In model-free agents, the process skips learning the model and just learns the policy and makes decisions based on previous actions rather than predicted rewards. These types of models have become more prevalent as they are easier to set up and can be used in a wider variety of areas. These agents come in three types, those that utilise Q-learning such as DQN, those that use policy optimisation such as PPO or Policy Gradient, or those that use a mixture such as DDPG and its derivatives. Most papers in this review will be using one of these types of DRL as they are more useful.

3 Explainable Artificial Intelligence

The problem of not understanding the ML models produced, there has been a race to develop tools that can explain model behaviour. Explainability has been a part of AI research since the 1970s with approaches such as decision trees though it has risen to importance in the last decade as ML has become more prominent. As the accuracy of AI has increased, its explainability has decreased. Therefore, the methods used to explain AI models have had to evolve.

There are a few definitions of explainability. The first is that explainability is using processes to help make an AI model more transparent to the user by producing explanations. These explanations describe the actions or decisions that an AI may have taken. Explainability may also be an AI system that can be questioned or visualised.

These methods can be categorised depending on whether they describe part of the model or the whole model and when they provide the explanation. Explanations can be intrinsic or post-hoc. Post-hoc is where the explanations are derived from an already trained model and can be model-agnostic or model-specific. Intrinsic explanations are where the model is already explainable such as a decision tree. These types of explanations are always specific to that model. The scope can either be local or global. The global explanations will explain all the model’s actions, while a local explanation will only explain one action.

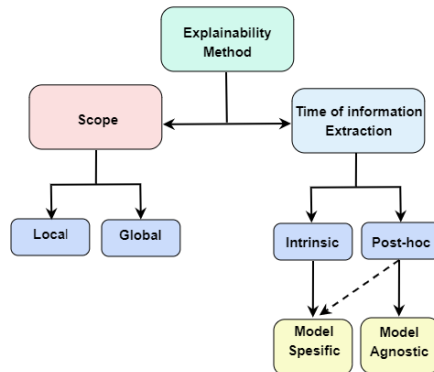


Figure 2: The two branches that define the type of explainability method.

The methods of XAI fall into a few different categories. The first is to create symbolic decision trees that are intrinsically explainable, which learn to mimic the actions of a more complex network. The second type is for models that have a visual input. Saliency maps show areas of interest to the model and thus affect the decisions the model takes. The third is to assign values to the

model inputs to find how important it is to the output using mathematical methods. The two most frequently used are Linear Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). The review will look at examples of all these types of XAI and their applications.

4 State of the Field

Recently the publication rate of XAI papers has increased exponentially; therefore, the methods and areas to that XAI is applied are in constant motion. It is paramount for research into XAI to apply to the correct DRL application. It is vital that XAI methods are understood and that the implementation by research teams is appropriate. It is imperative to keep an overview of a research field undergoing a rapid rate of advancement.

4.1 Previous Reviews on the use of XAI in DRL

There have been reviews done previously that cover this subject. The first of two to be looked at is by Wells and Bednarz (2021) [57]. They reviewed 25 papers published between 2019 and 2017. Most of these papers were from 2019 as the rate of studies covering both XAI and DRL rose quickly over the selected years. The authors looked at 5 different areas of XAI, human collaboration, visualisation, policy summarisation, query-based explanations, and verification. Rather than focusing on the methods used to generate the explanations, the authors looked at the types of explanations that were given. Due to the papers being selected from an earlier period they do miss out on some new methods that have become important in XAI. These include LIME and SHAP, these methods are used in 22.8% of the papers covered in this review. This highlights the fast pace at which this field of study is moving. The authors did not look at the trends that these XAI methods have been applied, this is where this review will be different. This paper concluded that there was still a way to go to make DRL agents interpretable. They found that an issue with XAI solutions shown in the papers they reviewed is the demonstrations were either on simplistic models or lacked scalability. Having explanations to models that have no bearing on real world problems are not useful. Though they did point out that research in this area is in its infancy so simple models are to be expected. The authors also were disappointed with the lack of testing. When there was testing it was of limited scope. Their final major conclusion was that the explanations ended up giving too much information. This over saturation of information can lead to confusion in what the explanations are suggesting.

The second review is by Vouros (2022) [55], they looked at 4 problems affecting interpretability in DRL and then used papers to propose solutions to those problems. The problems they suggested were model inspection, policy explanation, objectives explanation, and outcome explanation. Their review covered 31 studies in depth with most papers coming from 2019 or 2020. This review also did not focus on the applications for the explainability in DRL. This review found that it is difficult to quantify what constitutes a good explanation as there are no studies looking at this with relation to AI. They have a list of suggestions that should be implemented as to increase the level of interpretability. They include building an XDRL tool box to give explanations that are comprehensive as needed as well as transparent. This needs to be done in conjunction with defining interpretability, explainability and transparency.

5 Problem Statement

In light of the previous explanations of XAI and DRL in previous sections, the case has been made for maintaining a proper understanding. This review will look at the current fields in which researchers apply XAI to DRL models? What implementations of XAI are researchers using? How practical are these implementations of XAI? And what are the shortcomings of these studies? It will also look at any promising fields researchers have yet to explore.

5.1 Selection criteria for review papers

To select the papers to review Google Scholar is searched for the most recent academic papers that included Explainability and DRL as the main focus of their studies. The search parameters were using the terms "Explainability", "Interpretability", "XAI", "DRL", "Deep-Q", and "Deep Reinforcement Learning". Papers from the previous three years were preferred to limit overlap with other reviews. The selection criteria were as follows, the neural network should include an

XAI method to explain parts or the whole of a decision made by non-intrinsic AI networks, and the neural network must be a Deep Reinforcement Learning network. A necessary distinction since some studies use a DRL network to generate their explanations. These XAI methods are not in this review's purview.

A total of sixty papers were selected for review. These papers were classified by the field in which they apply. Four research papers were discarded for either being withdrawn by the authors or found to have a small contribution to the topics needed. The reviews done before into XAI and DRL have usually split the papers into categories based on the type of XAI used. This paper is looking to categorise based on the field in which it applies. Categorising in this manner will allow for an analysis in which areas certain types of XAI methods excel. It will also highlight where there needs to be more research done in applying XAI to DRL methods or implementing a particular XAI technique in a research field.

6 Research Fields with XAI applications

The applications of XAI in DRL found when gathering the relevant papers are as follows. The most prevalent use of XAI with DRL is those that use a video game to test the use of explanations. These video games give a non-compute intensive method to train and test a DRL model. The games are also simple to adapt to a DRL type neural network as the reward function can use the in-game score to judge success. These video games also allow for the altering of the difficulty. As these video games are complex, multiple strategies can be tried to maximise the in-game score. The DRL agents can even be used to find novel strategies that out-perform human strategies.

The second field that has seen the adoption of the DRL scheme and use of XAI is the use for guidance and navigation tasks. These guidance and navigation tasks cover multiple vehicle types. They share commonalities so have been combined. The types of vehicles seen are Uncrewed Aerial Vehicles (UAVs), automobiles and ships. UAVs were the most common vehicle type in the papers reviewed. With affordable off-the-shelf models of UAVs available real-life testing is possible for most research teams. There are also good software packages that allow research teams to simulate UAVs which is most useful when training neural networks. Autonomous ground vehicles were the second most common application in this field, varying from small battery-powered vehicles to full-sized cars. This variation allows a low barrier of entry on the smaller end. The larger vehicles are most representative of a high growth market. Simulations also make these vehicles highly accessible.

The need for explanation is vital in this field as these applications, especially autonomous cars are the most likely to interact with the general public and thus need to garner trust to be accepted for use.

The third field is that of system control. System control covers topics such as power system management. These are critical systems in which knowing the reasons behind a decision taken by a DRL model that is running one is vital to maintaining safety. These operators are more likely to be experts than those using autonomous vehicles; therefore, the XAI techniques will have different requirements. The explanations may require to be generated in real-time. This means computing time may become an important factor.

The fourth field is for Robotic Manipulators. Studies in this field look at how XAI can increase the understanding and performance of DRL controlled manipulators. Robotic manipulators have seen usage in manufacturing since the 1960s; however, these robots had a set action from which they could not deviate. The new DRL controlled manipulators can adapt to a variety of tasks. As the technology grows into more assembly lines the agent's behaviour must be known for safe operation.

The final field is using DRL networks to optimise mobile network solutions. Only a couple of papers covered this topic in the period for review. Suggesting this is a new field to exploit. It is imperative to build understanding as these networks are so vital to how modern life runs. Knowledge of bandwidth usage in certain situations can allow the engineers to construct redundancies into the correct parts of the networks.

Of the 56 studies read for this paper, 26 of them were studies that used a video game to train their networks and run their tests on the explainability. The second most common was the study of XAI in Vehicle Guidance with 14 papers reviewed. The third most common with six research papers were studies using XAI in control systems such as traffic light controllers. Next, there were six papers written on the use of Robotic Manipulators. Two papers were reviewed on the application in the field of Networks. The final two research papers will be handled separately as they did not fit the previous categories.

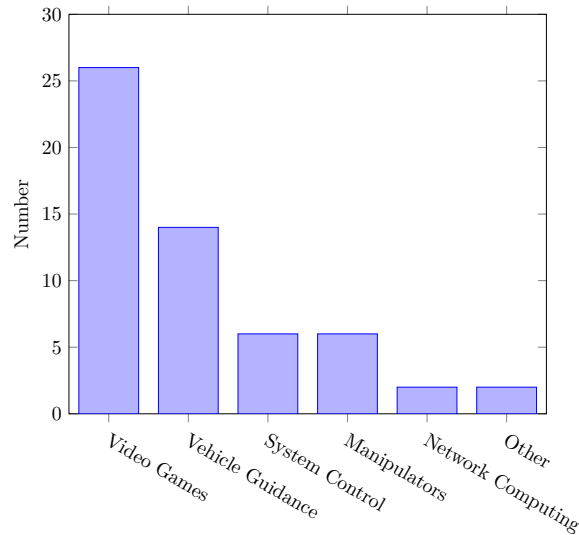


Figure 3: Areas of Studies for the Reviewed Papers

6.1 Video Game Simulations

Video games are a good testing ground for XAI techniques. There is no surprise that there are many methods to be reviewed. These include mathematical methods where the model inputs are varied to find the effect on the output like Shapley Values. The use of symbolic networks to emulate a trained DRL network where the decision tree itself is explainable. The use of visualisation techniques such as saliency mapping to see where the agent is focusing its attention. Finally, methods that use separate neural networks or added layers to the agent network to generate explanations. This section will go through these different XAI techniques to show the implementation.

To start with papers that used visual explanation to describe the actions of the DRL network. There are 7 of these papers. The first to be reviewed is Joo et al.(2019) [26]. The authors described the Gradient-weighted Class Activation Mapping (Grad-CAM) system for classification tasks that use Convolutional Neural Networks. They apply this Grad-CAM to the A3C DRL developed by DeepMind and run through a series of Atari video games.

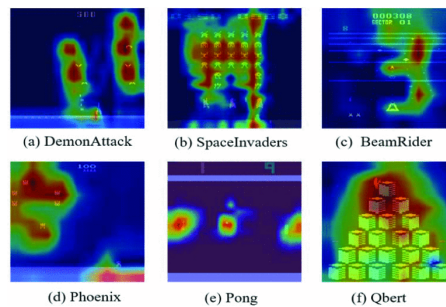


Figure 4: Grad-CAM images generated

The team managed to create Grad-CAM results for various Atari games shown in figure 3. The results generally show that the focus of the DRL agent is on the more crucial parts of the screen. In the paper, they conclude that the use of the Grad-CAM could lead to insights for expert users rather than the general public and will allow a deeper understanding of how DRL agents learn and operate.

Next, the paper to review is by Douglas et al.(2019) [13]. Their paper uses a saliency map generation technique and then uses that to create understandable visualisations for the end-users. The video game they used to train their agent was the Pommerman benchmark domain. They had to adjust the usual saliency algorithms by moving from a pixel by pixel method to focusing on each game square instead. The second modification was to record the change in the saliency values produced by each action. Normalising these changes to between [0,1]. The result was a compromise between emphasising changes in magnitude and changes in positive and negative values.

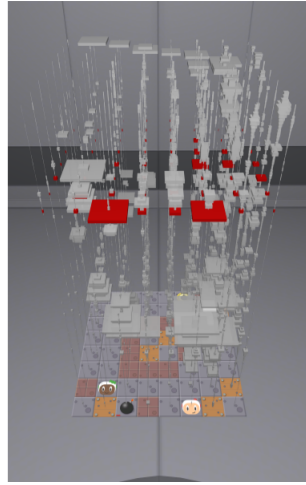


Figure 5: One of the generated "Towers of Saliency", the game board shows the current positions and the highlighted red areas are the current Saliency map.

These saliency maps were then taken for a whole game and arranged into a 3D saliency map that the paper has taken to call "Towers of Saliency". With these towers, the visualisation of the progress of the importance of the game squares is achieved. A Virtual Reality (VR) headset can view these visualisations. This allows an innovative view that can help with understanding. Overall this paper generated a visualisation method for use in video games where each game is over a short time frame. The visualisation is then able to show the entire progress of the game. The authors believe that the system can drive greater understanding into the field.

Next, the paper by Guo et al.(2021) [18] looks into the differences in attention between human players and DRL agents. They do this by producing saliency maps for the AI players while tracking the eyes of the human players. They found that the attention maps became more similar as the agents trained. The similarity of the attention maps was useful for predicting the model's success. Their work also looked at discount factors and found longer-term outlooks that are too long can lead to the model becoming too distracted by long term goals. They also found that these maps could be used to detect the reason for a failure state, either by the model having its attention on the wrong object or making the incorrect decision. Finally, they found that this could extend to many different types of DRL agents.

In Dammann (2021) [9], they looked into using saliency maps to give insight into the training of a deep-Q-Learning Atari Agent, insight into transfer learning between agents, and how MATS as a modular approach to functions operates. The first part found that Saliency maps gave an insight into how the attention the model focused on certain parts of the screen changed as the model progressed through training. Using saliency maps they could determine if the learning of a particular task was complete. They found that transfer learning was only a partial success in one situation. However, they did find saliency maps could predict the success of transfer learning. The final part looked at the Modular Autoencoder-based Task Splitting (MATS). MATS is where the autoencoder compresses the image to a state vector combined with four other states and is fed to the CNN and the Q-estimation to generate the action. This approach learned the MsPacman game but at a lower level than the regular DDQN. They theorised that optimising the network could give equal performance and use it to study the impact of the separate sections of the network.

In the next paper, Anderson et al.(2019) [1], looked at explaining the DRL with the use of Saliency maps and reward-decomposition bars. They tested these explanations with 124 naive human participants and found that they needed both the saliency maps and reward-decomposition bars to gain an understanding of how the agent was working. They also found that different situations and different people required different explanations to understand the models. Therefore they found it is best to give multiple explanations to cover more bases.

In Dao et al.(2021) [10], the authors used Grad-CAM to visualise snapshots in Atari video games. They then analysed how many snapshots are required to understand the DRL agent. They did this by utilising state-value approximations to construct sparse-bias space. They reduced the number needed for analysis by grouping states holding similar attention in the image. They found that these reduced the number of snapshots required for interpretation. The selection of snapshots chosen to be interpreted or discarded gave insight into how the model behaved.

The following paper for review is by Huber et al.(2019) [23]. The authors used a variation on the Layer-wise Relevance Propagation (LRP) scheme. This scheme uses only the most relevant

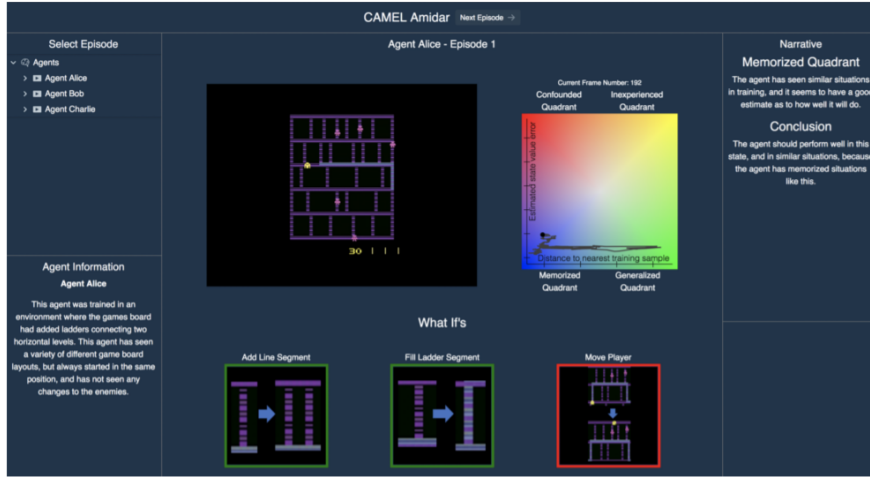


Figure 6: The user interface used to show the different types of explanation.

neurons in the CNN to generate saliency maps that highlight the paramount areas for the agent’s decision making. Using a duelling Deep Q-Network (DQN) they tested the scheme on three Atari games and found that the new scheme managed to highlight more pertinent information on the screen. The authors were able to show that this scheme worked on the most current versions of the DQN algorithm.

Next, the paper by Druce et al. (2021) [14] produces three explanations for the actions taken by the agent through a graphical depiction of the performance in the game state, a measure of how the agent would perform in similar environments, and a text-based explanation of what these two explanations imply. The Value Estimated Error (VEE) measures how well the agent estimates its state. A lower value means that the agent has seen a similar state in training. The second value is the Distance to Nearest Training Sample (DNTS). They created a user interface for this that shows the various explanations, shown in figure 5.

The visualisation was tested on the Atari video game Amidar, with participants in the experiments shown several scenarios and then asked about how they trusted the agent. The study found that their visualisation software increased trust in the agent. The authors were disappointed that the rise was only one standard deviation.

The following study by Druce et al.(2021) [15] looked at using human-machine teaming for the AI to work in conjunction with a human teammate and then use explanations to justify the AI’s actions. The authors used the StarCraft 2 video game to run their DRL agent. They used the CAMEL to generate the explanations and then shown to the participants. They found that the participants gave flawed explanations for the AI behaviour by adding human-like motivations, such as the AI getting scared of certain enemies. They also found that when a reasonable explanation was not producible because the agent was acting unreasonably.

The next few papers move away from the use of visualisation to explain the agents and instead use modified neural network architecture to increase the explainability of the networks. The first of these is by Lyu et al.(2019) [36]. They leveraged symbolic planning to enable the DRL model to maintain its high-dimensional sensory inputs while enabling task-level interpretability. It does this by linking the symbolic actions to options and using the architecture to create subtasks that the DRL learns and a meta-controller that suggests new rewards. They found SDRL framework successfully utilised symbolic planning to improve task-level interpretability.

In Guo et al. (2021) [20], the authors proposed a novel self-explainable model EDGE that modifies a Gaussian process with a customised kernel process and an interpretable predictor. They also developed a method to improve learning efficiency using a parameter learning procedure that leverages inducing points and variational inference. Testing these solutions on Atari and MuJoCo games, they found that they could form strategy level explanations by predicting final rewards generated by the agent and extracting time step importance in the levels. They found that using the explanations generated for several tasks like discovering vulnerabilities or errors in the policy was possible. They also found that this scheme could defend against adversarial attacks.

Landajuela et al.(2021) [30], to help with the interpretability of DRL agents, came up with a deep symbolic policy. This approach directly searches the space for symbolic policies. Using an autoregressive Recurrent Neural Network (RNN) to generate the control policies that can be represented by tractable mathematical expressions. To maximise the performance of the generated

expressions they used a risk-seeking policy gradient. Across eight video game tasks, the symbolic policies outperformed the state of the art DRL schemes they tested against while being readily interpretable. The authors felt that this is a viable alternative to neural networks when working with systems that need to be interpretable but have deployment constraints such as memory or latency limits.

The following paper is by Sieusahai et al.(2021) [48]. They used an interpretable surrogate model that predicts how the primary model acts. This surrogate model works by transforming the pixels that are input into the DRL model into an interpretable, percept-like input representation. They found that their trained model could accurately predict the actions of the primary model on a wide range of simple 2D games at about 90%. The authors decided to show both models adversarial examples to see if the surrogate model would predict the right action. This experiment showed that the surrogate model changed the same as the primary model showing that they used the same features to find the following actions.

In the following paper, Barros et al.(2020) [3] used the Moody framework to explain the behaviour of a DRL agent in a competitive card game. The Moody framework tries to create a representation of the agents by using the Pleasure/Arousal model. The Moody framework works in competitive environments by measuring the confidence of the opposing player after the agent’s actions. These confidence measures transform into a pleasure/arousal scale representing two scales pleasing to unpleasing and excited to calm. The authors found that this framework gave the model an enriching explanation of its actions. They also managed to assess the accuracy in understanding the confidence of the opposing player.

In Climent et al.(2021) [6], the authors used Policy Graphs (PG) to create a framework for explanations. These PGs represent the agent’s behaviour as this team tries to measure the similarity between the explanations and the actions. The policy graph is a generated list of statements that describe the system’s state. In this case, its usage is to balance a virtual cart pole. They found that controlling the actions using the policy graph they created almost reproduced the success of the DRL agent at controlling the cart pole. The success rate at holding the cart pole vertical did drop by a small amount but can still be considered a success. More research is required as this study was only limited to one example.

In this following paper by van Rossum et al.(2021) [52], they used a Curious Sub-Goal focused agent (CSG) to attempt to explain a DRL agent. The agent is split into three parts. First, they used a Generative Adversarial Network (GAN) to generate the next state. This GAN can build an understanding of the object interaction and actions the agent takes within the environment. The second part was the traditional DRL agent that dealt with the player navigation. For the final section, the team broke down the task into easily understood sub-goals that allowed for easier understanding than trying to understand the whole strategy. The authors found that this method successfully broke down the tasks into understandable sub-goals and allowed for a greater understanding of how the agent behaved. They did note that this was a simple task limited to discrete observation spaces with basic dynamics. They did speculate that a more advanced DRL could be integrated and suggested this for future work.

These papers take the DRL agents and try to use Decision Trees (DT) to emulate them and turn the black-box agent into an interpretable model. In Ding et al.(2021) [12], the authors used cascading decision trees to realise a DRL agent in an explainable way. The use of decision trees to learn the actions of a DRL agent and mimic them. Decision trees are inherently interpretable as the rules are easy to follow along the nodes. This paper suggests using Cascading Decision Trees (CDT), which cascades a feature learning DT and a decision making DT into one model. The CDT has a better function approximation in both training and operation. The reduced number of parameters while keeping the tree prediction accuracy was also achieved. The authors also found that while the CDTs had good model approximation in the imitation learning setting, they did struggle with instability in different imitators in the tree structures that affected interpretability.

The authors of the following paper Vasic et al.(2021) [53], used a Mixture of Expert Trees (MoET) to produce verifiable reinforcement learning. MoET is a model that consists of decision tree experts and a generalised linear model gating function. The new scheme allows decision boundaries with hyperplanes of arbitrary orientation instead of the usual axis-perpendicular hyperplanes seen with other schemes. The authors also designed a scheme that they called MoET_h that bases decisions on a single expert chosen by the gating function, this change allows for easy decomposition into a set of logical rules to generate explanations. The authors found that this scheme outperformed other decision tree techniques in model reproduction and still produced the means of model verification.

In Liu et al.(2021) [34], they suggest creating a Represent And Mimic framework (RAMi). This model can capture the independent factors of variation for the objects using identifiable

latent representation, using a mimic tree that measures the impact on the action values by the latent features. A novel Minimum Description Length (MDL) objective based on the Information Bottleneck (IB) principle was used to optimise the tree’s fidelity and maintain its simplicity. They used a Monte Carlo Regression Tree Search (MCRTS) algorithm to find an optimal decision tree. These optimised trees showed strong approximations with fewer nodes than against the baseline models. The interpretability was tested by showing latent traversals, decision rules, causal impacts and human evaluation results. In the human evaluation, they found that 83% of respondents preferred the mimic tree solution for interpretability compared to saliency and superpixels.

In the following paper, Dhebar et al.(2021) [11] use Non-Linear Decision Trees (NLDT) to provide interpretability for DRL agents. They build on previous work using NLDTs by introducing an evolutionary computation function to optimise the decision tree. Data from a pre-trained DRL model for the moon lander game is used to train an open-loop version of the NLDT. Training is done using a recursive bilevel evolutionary algorithm. The top part of this open-loop NLDT is then trimmed off and then trained using an evolutionary optimisation in a closed loop. This process gives the final NLDT. This method created an interpretable model that operated on discrete action problems. This study did not go into the interpretability of the model and just relied on the inherent interpretability of decision trees.

Liu et al.(2018) [33] proposed using Linear Model U-Trees (LMUTs) to approximate the DRL predictions. The team put forward a novel training method for the decision tree where the mimic learner observes the ongoing interactions between the environment and the DRL agent using an online algorithm. U-Trees are a natural choice for modelling DRL agents as they are essentially regression trees for value functions. These are slow in training leading the team to use an LMUT that allows the nodes to be linear models. This linearity allows for a better approximation and a smaller tree that is better for interpretation. They tested on three simple games and found that their LMUT could approximate the Q value function learned by the DRL, meaning that the black box DRL became more interpretable as a result.

Finkelstein et al.(2021) [16] use a taxi navigation game to look into the use of decision trees to help explain the actions of a DRL agent. They propose to use Markov Decision Processes (MDP) transforms to explain the behaviours. These transforms to the MDP are searched for examples where the actor’s behaviour is as expected. If the change from one agent to the new agent is meaningful, the resulting change is identified as an explanation. The team found that these model transforms produced interpretable explanations.

In Jaunet et al.(2020) [25] the authors looked at the memory that the DRL agent was using to determine the decisions that the agent was taking, they called this system DRLViz. The memory stores a vector of a timestep that can be viewed using their developed software, this can be compared to other situations and similar vectors can be seen for similar situations. These explanations are limited to expert users as a naive user would gain no insight from the memory data. To validate they used three experts from the field to give their thoughts on the DRLViz package, they stated that it provided good explanations and two out of the three thought the software was easy to navigate while the third took some time to become comfortable with it. Overall the authors developed a tool to look at explanations for decisions by DRL agents in a novel way.

The next paper by Karalus et al.(2021) [28] uses counterfactual explanations along with human-in-loop training to try to expedite the training of the agent. In their research they use TAMER which replaces the automatically generated reward in the training of the DRL with a human-generated reward, they are specifically using the DeepTAMER architecture that applies a deep neural network to approximate the H function that is used to approximate the human-based reward. The counterfactuals are generated by suggesting an action to be taken by the agent if some other input is changed in some way, these counterfactuals were only shown during negative rewards to teach the agent how to move to positive reward scenarios. In their experiments they found using this scheme did increase the training rate seen especially during the early stages of the training when there were large differences between the actions and the counterfactuals.

In Cruz et al.(2019) [7] the authors proposed a memory-based reinforcement learning (MXRL) approach. The DRL agent is using episodic memory to make explanations using the probability of success and the transactions needed to reach the goal. The questions for the agent to answer were the why? and why not? format. Each state-action pair is recorded on a list where the number of actions needed to reach the goal and the probability of success is calculated for the action by using the number of steps to success and total time steps. Using a simple grid game where the agent had to move a goal position. The team found that this scheme generated useful explanations that non-expert users could understand though they recognised that this was only a limited situation and more work needs to be done to be able to apply this to a more complicated problem.

One technique to be talked about is the use of Shapley Values to generate explanations. Shapley

values assign a weighting to an input depending on how it affects the output so for example in a game of pong the pixels that denote where the paddle is would have high Shapley Values. Only one paper covered the topic of Shapley values and that was Heuillet et al. (2021) [22]. In their paper, they showed that the use of Shapley values, or an approximation of such values using a Monte Carlo algorithm, could be a pertinent way to evaluate the contribution of different players in a multi-agent RL context. They proposed three research questions:

- Can Shapley value be used to determine how much each agent contributes to the global reward?
- Does the proposed Monte Carlo based algorithm empirically offer a good approximation of Shapley Values?
- What is the best method to replace an agent missing from the coalition (e.g., a random action, an action chosen randomly from another player or the "no operation" action)?

To do this they used two games where different agents had to cooperate to complete the task. The first game used was a predator/prey environment where the agents acted as the predators. The second game was a social game where the agents had to work together to produce as many apples as possible. They compared the Shapley values across up to six agents that played the games where these agents had to cooperate to generate the largest global reward possible, to decipher which agents were producing the highest local rewards they could look to the Shapley values.

The researchers found that these values were a useful form for explaining the multi-agent models that they were studying. They also found that the Monte Carlo derived approximations could be used as a suitable substitute for properly derived Shapley Values. They felt that these explanations were more useful to the researchers and developers but they could be of use to the general public if they were described as the intrinsic value of an agent. They also felt that these values could also be a method to detect bias in the RL model when training as they allow the analysis of the individual agents and how they are completing the tasks. As for their final research question, they found that using a no-operation action was the most neutral, and interaction-free method for this possibility as using a random action it was probable that a high negative reward could be the result.

6.2 Vehicle Guidance

The next area of study involves DRL networks and explainability in the guidance of vehicles. The vehicles are drones, cars or ships. The use of DRL networks in these tasks has a huge benefit as they operate in a continuous action space in which these types of neural networks are specialised. It is also crucial to build trust in these networks if they are to be accepted by the general public and regulators alike. That means that explainability will have to be able to explain to the general public and experts. These non-experts need to understand the decisions while they are under the control of these networks to feel comfortable, while experts will want to be able to query decisions if something should go wrong. This section will look at the three types of vehicles used in guidance, starting with the UAVs.

The first paper is by He et al.(2021) [21]. In this team's study, they set up a UAV to autonomously navigate around a simulated environment and then used Shapley Values to set up an explanation method. They used the Shapley Values to generate textual responses that justify the DRL agent's response to the goal and objects. Using these Shapley values, they could diagnose which section of the network influenced the action taken by the agent. For example, an obstacle seen by the drone would activate the CNN layers producing a text response that of the action. The experiment reveals the importance of each section of the network to the agent's actions. The study also did a real-world test where it was confirmed to work. The SHAP CNN explainer was successful at describing the actions of the drone. The authors found that this scheme for generating produced good explanations that would be understandable to novice users.

Next up is a paper by Guo et al.(2020) [19] where they were investigating how to use UAVs to provide coverage for 5G networks while maintaining efficiency in energy usage and signal optimisation. The insight into the explainability of this system is small as it extracts features from the hidden layers, thus giving some form of interpretability. They propose a Double Dueling Deep Q-learning Neural Network (DDDQN) with a Prioritised Experience Replay (PER) and a fixed Q-target that allows the model to maintain stability and prevent over-fitting while also offering performance gains. The team was able to show that this scheme was able to have more efficient UAV usage than typical UAV autonomous flight. They were also able to use the interpretability from examining the hidden layers to find the optimal drone deployment. The authors felt that the

next step was to dive further into the explainability to understand the propagation of the features in the DDDQN.

Chang et al.(2020) [5] look into how autonomous UAVs interact with human operators and large flocks of UAVs. They proposed an agent-based task planner that will take a task and decompose it into a series of interpretable sub-goals. The team designed a simulation where drones had to avoid radar hot spots and with differing levels of human cooperation. They found that the swarm completed the complex tasks using the scheme. The explainability was improved by showing the decomposed tasks on a timeline.

Stefik et al.(2021) [51] describe the COmmon Ground Learning and Explanation (COGLE) system in this paper. COGLE is an XAI system that delivers supplies to field units in mountainous areas. The explanations come in the form of What, Why and Where questions. The why questions depend on counterfactuals. The what question depends upon the mission profile that the UAV has chosen. Finally, the question of where is answered by analysing the map to highlight risks. The answers to these questions come from a visualisation that gives narrative answers. Overall this method produced good explanations of why the UAV had chosen a particular action. It was especially beneficial to have pre-decision explanations so the operator could step in if needed.

Robotic navigation is the next area to be looked at. Nie et al.(2019) [39] look at visualising the behaviour of a swarm robot system. A swarm robotic system is a multi-robot system where many homogeneous autonomous robots act in tandem. They suggest using a Deconvolutional Network (Deconvnet) and Grad-CAM to visualise the decision-making process. The experiment was for the robots to visit two opposing places as much as possible in an allotted time. The team found that the two methods could interpret the policies learned by the DRL agent. They did point out that with actions coming from a DRL agent it is difficult to say whether it is wrong or correct as there is no absolute standard.

Roth et al.(2021) [45] use decision trees to give a layer of interpretability to the small robot they are guiding with a DRL agent. They go one step further with the transformation of their agent into a decision tree as they also apply performance improvements over what the DRL agent has. Some of these improvements are smoothing out oscillations and frequency of immobilisation. They manage to do this without retaining the model, which is a bonus. They demonstrated in their paper that they could provide interpretability and improve their agent with this technique.

Next, the paper by Josef et al.(2020) [27] looks at the guidance of a robot through an unknown obstacle course using a DRL agent while incorporating a self-attention module that provides explainability to the agent's actions. This paper mostly looked at the new methods they were implementing for designing a DRL agent for a robot. The use of self-attention for explainability was not a focus. It did find that the agent's attention was on the closest edge of a hazard to the robot, which is probably what most people would expect to learn from this feature.

In Xu et al.(2021) [59], they use a symbolic DRL framework to help with data efficiency and interpretability when guiding a robot around an area. The framework consists of a high-level agent, a sub-task solver and a symbolic transition model. The high-level policy generates goals that the agent can do to complete the overarching goal. By separating the high-level and low-level goals, the high-level policy can focus on longer-term goals and thus reducing sample complexity. The symbolic transition model allows for interpretability because the logic rules for the model are readable by the operator. The sub-task solver is just the DRL that interacts with the environment to solve the short term goal that the high-level policy has been tasked with. The authors found that this method did allow for high efficiency and an interpretable model.

The following papers will be looking into automotive guidance, an important area of research with the first attempts at commercialised self-driving cars starting to appear. The first of these papers to be reviewed is by Soares et al.(2020) [50]. In this paper, the authors attempt to provide explanations for their DRL model through rule-based approximation and visualisation. The rule-based approximations will be provided by a set of IF...THEN rules that will form an alternative interpretable model. A visualisation is provided to enhance the explanations. The experiments showed that this method was accurate and computationally efficient while allowing for good interpretability and thus validating the DRL agent.

As well as piloting a car, DNNs will also have to predict when accidents might occur to know when to avoid them. Bao et al.(2021) [2] try to use a DRL agent to achieve this and provide visual explanations. The team's DRIVE model uses both a top-down and bottom-up visual attention mechanism to make observations from the dash-cam footage. It uses this information to produce an explanation from the attentive areas for the decision taken by the agent. The experiments showed that the DRIVE model generates state-of-the-performance in real-world situations while providing good explainability.

Gajcin et al.(2021) [17] proposed using contrastive explanations to help decide which of the

two DRL agents in a situation to use. The best strategy can be hard to choose for a complex task when presented with options. These authors lay out a strategy for picking the better solution using contrastive explanations. They found they could find the preferences of the two models by comparing them. In their experiment, the researchers could see that the safer of the two models preferred to leave a more pronounced gap and travel at a slower speed. In future research, they wanted to move on to comparing more than two models at a time.

The paper by Liessner et al.(2021) [32] uses Shapley Additive ExPlanations (SHAP) to describe the control of a car’s speed. Shapley values have their origins in 1950s game theory which described how participants in a game could maximise their value. This concept was used to produce SHAP to generate explanations for neural networks. The car in this experiment is driving down a one-lane highway and has to follow the speed limit changes that it comes across. Unsurprisingly the SHAP values showed that the current speed limit and the velocity had the highest impact on the model’s choices. This method does give great understanding through visuals for this simple problem.

The following paper looks at three methods for explainability in the DRL agent that controls the kinematics of a car. N.B Carbone (2020) [4] uses three explainability techniques to analyse their agent, Shapley Additive ExPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Linear Model Trees (LMTs). LIME creates locally interpretable explanations for a single data point. By making these perturbations in the dataset, the importance of a feature can be deduced. Shapley values measure how much an individual feature in the neural network affects the output. Several SHAP methods approximate the values instead of direct calculation. The calculation of the exact values is computationally expensive. This paper uses the Kernel SHAP, which builds on the LIME framework. LMTs produce interpretable results by creating a decision tree model approximating the targeted agent. This decision tree is more interpretable as the nodes follow interpretable rules. Much like the other methods that have used LMTs to approximate the models they also found good performance, though they found it was a trade-off between the accuracy of the LMT and the interpretability as a deeper tree while more accurate loses its value as an explainer. They found that the use of SHAP helped analyse the behaviour of the model it didn’t produce the global policy insights into the agent that the LMT provided. The author then concludes that LMTs may become the gold standard in creating explanations for DRL agents.

There is only one paper in the time frame that looks at the use of explainable DRL agents by Løver et al.(2021) [35]. Again in this paper, the same three methods as the last are used, LIME, SHAP and LMTs. These three methods appear as the benchmark methods for applying explainability to DRL models in the field. The experiment was to dock a ship to a quayside using the DRL agent and then measure the performance of the three explainers. The experiments showed that a task that needs real-time explanations of the LMTs was the better solution. The two other types were computationally expensive. This meant the explanations were slow to be generated. They summarised that SHAP was better than LIME for post-hoc explanations though it may be susceptible to biased predictions.

6.3 System Control

The usefulness of DRL agents controlling time-continuous systems has been realised in the last few years. The ability of DRL agents to react quickly and effectively is impressive. These systems are most often safety-critical; therefore, trust in the decisions is vital. For this reason, explainability is a salient area of current research.

The first papers will be looking at the application of DRL agents in the control of traffic light systems. The usual method of controlling traffic lights on a timer can lead to inefficiencies if the level or pattern of traffic changes dramatically throughout the day; therefore, a system that can react to changing circumstances is crucial. The first paper on this subject is by Schreiber et al.(2021) [46]. In their study, they use the Shapley Additive Explanations (SHAP) framework to explain the policy of a traffic light control system. They measured against a fixed time cycle of the traffic lights. They based this on average speed score, average wait time and average travel time. They found that the DRL model could improve the traffic flow by about 25%. They also found that SHAP values described why the model was taking those actions, building trust in the model.

Next, traffic signal control by Rizzo et al.(2019) [44] also used the SHAP framework to help explain their DRL model. In this paper, the junction chosen was more complex when compared to the previous study. This junction was a four-way intersection with a roundabout and an underpass for one of the roads. These junctions are complex due to the many variables affecting traffic speed, such as any traffic that backs onto the main road from the slip road will slow traffic that doesn’t have to stop at the junction. The cumulative wait times using the DRL controlled lights were reduced from 3262s for the fixed short phase and 2594s for the fixed long phase down to 728s.

A significant improvement over the timed lights. For the explainability, the SHAP values gave a detailed explanation of the decision to change the colour of the lights. As this information is only for the engineer of the traffic lights it can be more expert focused than if it was for the general public, so this level of explanations from the SHAP values is at the right level.

The last paper on traffic light control uses a different system, which is knowledge compilation. Wollenstein-Betech et al.(2020) [58] propose using knowledge compilation, which is a technique to build a Directed Acyclic Graph (DAG). DAG is a representation of the logical theory. The knowledge compilation assembles the unorganized logical theory of the DRL into a structured one which is more explainable. This DAG will use the deterministic Decomposable Negation Normal Form (d-DNNF). This method is computationally expensive though it only has to be run once. For the experiment, they used a four-lane highway intersection with the simulation time being 1.5 hours. They found that the d-DNNF DAG allowed them to get the likelihood of a state change in the system. They posited that using this tool in debugging a black-box system would let them check that the controller is logically sound.

In Zhang et al.(2021) [60], they look to use a DRL agent to control a power system’s emergency procedures and explain the actions of this system using SHAP. They use the Deep-SHAP algorithm built upon the DeepLIFT algorithm, which takes a sample of states to approximate the Shapley values. They trained their DRL on a power system simulation. Using the SHAP values to make visualisations they found an increased understanding of which variables drove the state changes. They did find that raw SHAP values are hard to understand as they are just values but changing them into probabilities made them more understandable.

Next, the paper by Nunes et al.(2021) [40] described a system to help manage air traffic using Solution Space Diagrams (SSD). This system alerts the air traffic controller to any aircraft getting into proximity to each other and suggests an exit vector to the aircraft to avoid any collisions. Using this SSD as an input, the DRL agent outputs the required exit vector. This paper didn’t talk about the explainability aspects of this technology. The visualisations used for the SSD did explain the DRL’s actions without it explaining the DRL directly.

The final paper for review in the Systems control section is Kotevska et al.(2020) [29]. They suggest a method for introducing interpretability to a Heating, Ventilation and Air Conditioning (HVAC) control system. To provide the explanations they used the LIME algorithm. The DRL model improved the efficiency of the HVAC system. However, the team found that this increase in efficiency was not enough to build trust in the system, though their use of the LIME algorithm allowed them to see inside the decision making of the control system and see what environmental factors were important in changing the system.

6.4 Robotic Manipulators

The next application to be looked at is robotic manipulators controlled by DRL agents. The use of these networks has grown in popularity recently as the ability of these control systems allows the agent to adapt to situations that it has necessarily seen before. These features would allow for more generalised robots in assembly lines that could do different jobs that would have been impossible before.

The first three papers in this section use the SHAP algorithm to describe their DRL agents. The first is by Remman et al.(2021) [43]. Their study looks at how the robotic level manipulator system can be described using SHAP. Their application of SHAP is an alteration of the Kernel-SHAP as that only describes the direct effect that features have on the output. The team suggests creating a Causal SHAP that looks at the indirect effects that a feature has on the output. To accomplish this Causal SHAP alters the sampling method used by Kernel SHAP by using a partial causal ordering that captures these indirect effects. The robotic manipulator to be used is the OpenMANIPULATOR-X which they tasked with moving a lever from one position to another. They showed that Causal SHAP describes both the direct and indirect effects that a feature can have on the output. They also showed that using Casual SHAP can generate better explanations. They did recognise that these explanations are not well suited to non-expert users but are helpful for data scientists and researchers.

The second robotic manipulator SHAP paper is by Remman et al.(2021) [42]. This paper has a similar experimental setup to the previous study but focuses on using a Deep Deterministic Policy Gradient (DDPG) algorithm with a Hindsight Experience Replay. They use the regular SHAP implementation and try to test their DRL agent. They make some conclusions on using SHAP values with this DRL agent. They found that some variables featured prominently in the SHAP values as expected, whereas others, namely the joint variables did not. They theorise that the lack of assumed independence between states may lead to this as there is a correlation between states.

The final SHAP paper for this section is from Wang et al.(2020) [56]. Their study applies the SHAP method to describe an automated crane system. The mass that the crane is lifting can swing, which complicates this situation. The DRL agent has to account for the momentum that this object will have and will have to learn to be smooth as possible. This paper was too short for any grand conclusions though they found that the SHAP values did generate the explanations as predicted.

Next in the paper, Iucci et al.(2021) [24] addresses XAI in Human-Robot Collaboration (HRC) scenarios. They do this by combining two methods. First, they use Reward Decomposition to break down the reward function to give insight into the factors that influenced the agent’s decisions. The second is an Autonomous Policy Explanation (APE) that produces natural language responses to queries about the robot’s behaviour. To differentiate between the rewards an action generated was classified into five sub-types. The APE uses a series of logic statements placed onto a grid to produce a natural language explanation. Using these approaches, the authors found that they could produce satisfactory explanations. These types of explanations were useful when debugging the DRL agent.

The following paper is by Schwaiger et al.(2021) [47]. This study investigated how to extract explainability from the DRL agent of a robotic manipulator performing a pick and place task. The aim was to find the robot arms’ dimensions from the DRL agent to open up the black box. The authors theorise that the DRL agent must find these dimensions to learn a task. If extracting the dimensions is possible, then generating explanations could be done. They found that they could accurately pick out the lengths with the highest average error for a section being a quarter of a per cent. They felt that this justified their hypothesis.

The final paper in the robotic manipulator section is by Cruz et al.(2021) [8]. They propose using goal-driven explanations to add interpretability to their DRL agent. They suggest these goal-driven explanations to try to create explanations that would be understandable by a non-expert user. They look to be able to answer the questions of why? And why not? Using the probability of success (P_s) for an action, they generate an explanation of the favouring of one action over another. The three ways of producing this probability of success are the Memory, Learning, and Introspection-based approaches.

The memory-based approach uses a list of state-action pairs that produce the P_s . The method has a downside of the size of the memory growing as the number of episodes grows.

The learning-based approach uses the agent’s learning process to generate the P_s . The algorithm learns as the agent does by using P-values instead of Q-values which are usually the output of a training process. This process does add some overhead in the memory and does increase computing time during the learning phase.

The introspection-based approach tries to remove all memory overhead by calculating the P_s by using the Q-value directly by using a numerical transform. They use the Q-value to approximate the reward at that point to find how far the value is from the total reward. This can be changed into P_s by applying a transform to get the correct shape of the curve.

Testing these methods was done using three robotic tasks, two simulations and one real-world test. They found that all three explainability methods produced similar results and thus validated their introspection method as a good alternative with a light computing load. They found that the explanations were not perfect. The authors suggest this could be an issue later where the user becomes too trusting in the explanations or wrong explanations destroy all trust between the agent and the user, especially in non-expert scenarios. In future, they want to look at reward decomposition to generate more reward signals to produce explanations.

6.5 Network Solutions

The ubiquity of mobile networks has become part of modern life. With the introduction of 5g as the latest standard, this is becoming more prevalent. The need for more and more data bandwidth has focused researchers’ on efficiency gains. So the use of DRL agents to find these optimisations is proving to be a rich vein for researchers. Using DRL agents leads to the question of explainability, especially in equipment that has become so important in modern life. With the needs of the regulators and end-users, this is an area that requires research. As this is quite a niche application, there are currently only two papers covered by the time frame covered in this research.

The first paper by Li et al.(2021) [31] defends their network against threats by using explainable DRL agents. The study looks at the edge devices as they are most susceptible to attack and builds an agent to respond to the threats and allocate resources correctly. The agent used to perform this defence is a Dueling Deep-Q Network (DDQN). To generate the explanations for the DDQN they are using LIME. LIME uses perturbations to the features to see the impact on the output

and generates a probability of an outcome. In this paper, they managed to produce a model that did help to deal with attacks on a network and used the LIME package successfully explain the actions that this model chose.

The second paper is by Vijay et al.(2019) [54]. Here they look to overcome the inherent security issues in a 5g network. The authors look to develop a Main Aggregator Server (MAS) with a Deep Q-Learning (DQN) that aggregates the responses and makes the final decision in allocations in the network. Though the paper suggests that this approach will improve the use of XAI in this field, there is no suggestion on how to implement XAI into this particular type of agent and that is left up to further research.

6.6 Other Applications

Two reviewed studies didn't fit the other areas of application, therefore are reviewed here. The first is a suggestion for a method to create an explainable DLR agent but offers no experiments to assign it to an application. The second does have some use of Atari games for preliminary studies. The paper, however, focuses on using DRL agents to diagnose HIV patients and maintain a level of explainability to experts. There being only one paper on applying XAI to health sciences is surprising as this has been an area of interest for AI.

The paper that looks for a new explainable agent is by Parra-Ullauri et al.(2021) [41]. The authors propose Event-driven TEmporal MOdels for eXplanations (ETeMoX). ETeMoX is an architecture based upon Temporal Models (TMs) that look at the reasoning that the model makes through time and extract explanations that are history aware when asked. They use Complex Event Processing (CEP) and TMs to generate the explanations and present them to the end-user through plots and graphs. Using ETeMoX in three case studies that leveraged three different types of DRL algorithms, they found that the system provided explanations for all three agents. They were explanations that both tracked the evolution of a metric and the relationships between them. They were also able to look at time windows to see how interest changed through that period. For future work, the team proposed to look at how the system could be made more efficient in computation. They also suggested looking at using these explanations to improve human-in-the-loop situations.

Mishra et al.(2021) [37] tackled XAI by designing a PolicyExplainer visualisation, using a decision tree classifier to create the visualisation. This decision tree was able to track the features that led to classification results. They proposed a text-based solution so the software could answer the questions why? Why not? And when? The authors felt that this is the best way to explain the agent's behaviour to an operator unfamiliar with machine learning. Using the visualiser in three tasks, the first two were simple games. The third was a study into the prediction of HIV infection. They found that the PolicyExplainer gave good explanations in the first two tasks and then used experts in the HIV domain to use the software to judge its usefulness. The experts found that the software gave good explanations that made sense and were informative. The authors felt that the interface currently is limited by state and action scalability. The plan is for further research to solve this problem by looking at ways to select the most crucial state features.

7 Limitations of the Current Methods

From these papers covered in the review, there are some limitations. First, the applications for some of these XAI methods are working on simple models that might not have real-world uses. These simple models may not scale or are of limited value by being too specific to one task. For example, an explanation that works in an Atari game might not hold up in a more complex game with more actions available and more consequences to those actions. This problem appears more in the decision tree based XAI methods as a more complex system produces a large decision tree that becomes more uninterpretable as the tree grows. The numerical XAI methods SHAP and LIME can be used for more complex systems though they have to deal with high compute loads limiting them to situations where the explanations don't have to be instantaneous.

There is also the problem of judging the explanations provided by the schemes described here, with only a few of the fifty-six papers using some form of the population to test the explanations gained with even fewer large enough samples to make conclusions. Applying this to experts would be harder to gain a large enough sample size due to the smaller population due to their specialised knowledge. With these trust and understanding tests, it is much easier to gain trust in a model performing a simple task that has no bearing on the person observing. To build the trust required for AI to be accepted by the general public by producing explanations with real impact on the

individual. With the models still being simple, it is harder to get to that stage of building a deeper level of trust that is impactful.

8 Future areas of study

Though there are limitations in these studies, they leave room for researchers to carry on with their research. Future studies must apply to more complex systems to see if solutions are scalable and produce viable explanations. One area where it was surprising that there was a lack of research was health science. Only one paper looked at this field when it seems like an area that could benefit from XAI. It is crucial to trust models directly linked to people’s health. In future, it would be good to see more papers covering this.

Furthermore, testing the explanations on the people they are designed for should be carried out, or comparisons on which types of explanations certain groups prefer would be interesting. Studies that cover this would allow the people designing systems that interact with the outside world to understand what approaches work when choosing the appropriate XAI solution.

9 Conclusion

The applications of the different XAI methods in the different fields of study can now be summarised. In the video game examples, there was the widest range of different methods. A reason for this is that it is a research field with little real-world application. This means that researchers can try out many different methods in a variety of different scenarios without worrying about applying them to the real world. Among the XAI methods, the construction of symbolic networks in the form of decision trees was a popular method to mimic the DRL agents. The use of these symbolic networks is expected as video games can have simplistic state spaces and limited actions that can be approximated easily. Another method that was popular with these types of applications was using saliency methods. With the visual element in these video games, a purely visual method like saliency maps is expected to be a popular method. It allows for a person looking for an explanation to look directly where on the screen is important to the agent when making a decision. This is one of the better ways of providing a succinct explanation.

In the field of vehicle guidance and navigation the XAI methods used numerical methods such as SHAP values more than with video games and in these more complex spaces, there were fewer symbolic networks used. There were still visual methods used on the images recovered from the sensors. These methods were sometimes mixed with other methods like with He [21] as they used GRAD-CAM and SHAP values to explain what was important that the drone saw but also what decisions affected the drone.

For system control, the use of SHAP values seemed to be the most popular method. There are no visual inputs to these systems so methods such as saliency maps have no value. There were no symbolic networks either so this might be an area that could be exploited in the future. The researchers felt that SHAP and LIME that give importance to inputs allowed for good explanations in this particular application on DRL agents.

Robotic manipulators also heavily used SHAP values for explanations, this suggests that particularly in applications without a visual element SHAP values are a great tool to deliver explanations to DRL agents. Along with SHAP values, there were methods used that extracted the features of the robotic arm from the network to break the black box directly.

In the other fields, it is hard to make any conclusions about what XAI methods are best as there is such a limited sample. The recommendation is for further research into these areas by applying some of the XAI methods that have been described in other applications. In the field of health science, there are a great many methods that would seem to apply well.

As machine learning and deep reinforcement learning become prevalent, it is imperative to give explanations for the actions they take. The studies shown here show good progress in several fields to provide explanations that can be useful for experts and non-experts alike. Working with decision trees has shown to work well in many different situations though scalability needs addressing. The use of metrics such as SHAP and LIME has shown to be effective in many different scenarios. For example, in applications with no visual input like traffic light systems. Finally, visualisation methods for generating saliency maps such as Grad-CAM have shown to be very effective in the video games tested.

There is further work in proving the scalability of these ideas and proving that they can be understood well by their respective audiences. Supplying explanations for AI solutions must continue to improve. Improvements from better explanations, using better visualisation methods and

producing well-articulated explanations. With these explanations, the public can accept the role of machine learning and its adoption can begin.

References

- [1] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Alan Fern, and Margaret Burnett. Explaining reinforcement learning to mere mortals: An empirical study. *arXiv preprint arXiv:1903.09708*, 2019.
- [2] Wentao Bao, Qi Yu, and Yu Kong. Drive: Deep reinforced accident anticipation with visual explanation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7619–7628, 2021.
- [3] Pablo Barros, Ana Tanevska, Francisco Cruz, and Alessandra Sciutti. 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)*, pages 1–8. IEEE, 2020.
- [4] Nicolas Blystad Carbone. Explainable ai for path following with model trees. Master’s thesis, NTNU, 2020.
- [5] WANG Chang, WU Lizhen, YAN Chao, WANG Zhichao, LONG Han, and YU Chao. Coactive design of explainable agent-based task planning and deep reinforcement learning for human-uavs teamwork. *Chinese Journal of Aeronautics*, 33(11):2930–2945, 2020.
- [6] Antoni Climent, Dmitry Gnatyshak, and Sergio Alvarez-Napagao. Applying and verifying an explainability method based on policy graphs in the context of reinforcement learning. In *Artificial Intelligence Research and Development*, pages 455–464. IOS Press, 2021.
- [7] Francisco Cruz, Richard Dazeley, and Peter Vamplew. Memory-based explainable reinforcement learning. In *Australasian Joint Conference on Artificial Intelligence*, pages 66–77. Springer, 2019.
- [8] Francisco Cruz, Richard Dazeley, Peter Vamplew, and Ithan Moreira. Explainable robotic systems: Understanding goal-driven actions in a reinforcement learning scenario. *Neural Computing and Applications*, pages 1–18, 2021.
- [9] Michael Dammann. Deep q-learning and explainable ai: Elucidating training behaviour, transfer learning and modularity. 2021.
- [10] Giang Dao, Wesley Houston Huff, and Minwoo Lee. Learning sparse evidence-driven interpretation to understand deep reinforcement learning agents. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7. IEEE, 2021.
- [11] Yashesh Dhebar, Kalyanmoy Deb, Subramanya Nagesh Rao, Ling Zhu, and Dimitar Filev. Interpretable-ai policies using evolutionary nonlinear decision trees for discrete action systems. *arXiv preprint arXiv:2009.09521*, 2020.
- [12] Zihan Ding, Pablo Hernandez-Leal, Gavin Weiguang Ding, Changjian Li, and Ruitong Huang. Cdt: Cascading decision trees for explainable reinforcement learning. *arXiv preprint arXiv:2011.07553*, 2020.
- [13] Nathan Douglas, Dianna Yim, Bilal Kartal, Pablo Hernandez-Leal, Frank Maurer, and Matthew E Taylor. Towers of saliency: A reinforcement learning visualization using immersive environments. In *Proceedings of the 2019 acm international conference on interactive surfaces and spaces*, pages 339–342, 2019.
- [14] Jeff Druce, Michael Harradon, and James Tittle. Explainable artificial intelligence (xai) for increasing user trust in deep reinforcement learning driven autonomous systems. *arXiv preprint arXiv:2106.03775*, 2021.
- [15] Jeff Druce, James Niehaus, Vanessa Moody, David Jensen, and Michael L Littman. Brittle ai, causal confusion, and bad mental models: Challenges and successes in the xai program. *arXiv preprint arXiv:2106.05506*, 2021.

- [16] Mira Finkelstein, Nitsan Levy Schlot, Lucy Liu, Yoav Kolumbus, Jeffrey Rosenschein, David C Parkes, and Sarah Keren. Deep reinforcement learning explanation via model transforms. In *Deep RL Workshop NeurIPS 2021*, 2021.
- [17] Jasmina Gajcin, Rahul Nair, Tejaswini Pedapati, Radu Marinescu, Elizabeth Daly, and Ivana Dusparic. Contrastive explanations for comparing preferences of reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2022.
- [18] Suna Sihang Guo, Ruohan Zhang, Bo Liu, Yifeng Zhu, Dana Ballard, Mary Hayhoe, and Peter Stone. Machine versus human attention in deep reinforcement learning tasks. *Advances in Neural Information Processing Systems*, 34:25370–25385, 2021.
- [19] Weisi Guo. Partially explainable big data driven deep reinforcement learning for green 5g uav. In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pages 1–7. IEEE, 2020.
- [20] Wenbo Guo, Xian Wu, Usman Khan, and Xinyu Xing. Edge: Explaining deep reinforcement learning policies. *Advances in Neural Information Processing Systems*, 34:12222–12236, 2021.
- [21] Lei He, Nabil Aouf, and Bifeng Song. Explainable deep reinforcement learning for uav autonomous path planning. *Aerospace science and technology*, 118:107052, 2021.
- [22] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. Collective explainable ai: Explaining cooperative strategies and agent contribution in multiagent reinforcement learning with shapley values. *IEEE Computational Intelligence Magazine*, 17(1):59–71, 2022.
- [23] Tobias Huber, Dominik Schiller, and Elisabeth André. Enhancing explainability of deep reinforcement learning through selective layer-wise relevance propagation. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 188–202. Springer, 2019.
- [24] Alessandro Iucci, Alberto Hata, Ahmad Terra, Rafia Inam, and Iolanda Leite. Explainable reinforcement learning for human-robot collaboration. In *2021 20th International Conference on Advanced Robotics (ICAR)*, pages 927–934. IEEE, 2021.
- [25] Theo Jaunet, Romain Vuillemot, and Christian Wolf. Drlviz: Understanding decisions and memory in deep reinforcement learning. In *Computer Graphics Forum*, volume 39, pages 49–61. Wiley Online Library, 2020.
- [26] Ho-Taek Joo and Kyung-Joong Kim. Visualization of deep reinforcement learning using gradcam: how ai plays atari games? In *2019 IEEE Conference on Games (CoG)*, pages 1–2. IEEE, 2019.
- [27] Shirel Josef and Amir Degani. Deep reinforcement learning for safe local planning of a ground vehicle in unknown rough terrain. *IEEE Robotics and Automation Letters*, 5(4):6748–6755, 2020.
- [28] Jakob Karalus and Felix Lindner. Accelerating the convergence of human-in-the-loop reinforcement learning with counterfactual explanations. *arXiv preprint arXiv:2108.01358*, 2021.
- [29] Olivera Kotevska, Jeffrey Munk, Kuldeep Kurte, Yan Du, Kadir Amasyali, Robert W Smith, and Helia Zandi. Methodology for interpretable reinforcement learning model for hvac energy control. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1555–1564. IEEE, 2020.
- [30] Mikel Landajuela, Brenden K Petersen, Sookyoung Kim, Claudio P Santiago, Ruben Glatt, Nathan Mundhenk, Jacob F Pettit, and Daniel Faissol. Discovering symbolic policies with deep reinforcement learning. In *International Conference on Machine Learning*, pages 5979–5989. PMLR, 2021.
- [31] Huiling Li, Jun Wu, Hansong Xu, Gaolei Li, and Mohsen Guizani. Explainable intelligence-driven defense mechanism against advanced persistent threats: A joint edge game and ai approach. *IEEE Transactions on Dependable and Secure Computing*, 19(2):757–775, 2021.
- [32] Roman Liessner, Jan Dohmen, and Marco A Wiering. Explainable reinforcement learning for longitudinal control. In *ICAART (2)*, pages 874–881, 2021.

- [33] Guiliang Liu, Oliver Schulte, Wang Zhu, and Qingcan Li. Toward interpretable deep reinforcement learning with linear model u-trees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 414–429. Springer, 2018.
- [34] Guiliang Liu, Xiangyu Sun, Oliver Schulte, and Pascal Poupart. Learning tree interpretation from object representation for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 34:19622–19636, 2021.
- [35] Jakob Løver, Vilde B Gjørsum, and Anastasios M Lekkas. Explainable ai methods on a deep reinforcement learning agent for automatic docking. *IFAC-PapersOnLine*, 54(16):146–152, 2021.
- [36] Daoming Lyu, Fangkai Yang, Bo Liu, and Steven Gustafson. Sdrl: interpretable and data-efficient deep reinforcement learning leveraging symbolic planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2970–2977, 2019.
- [37] Aditi Mishra, Utkarsh Soni, Jinbin Huang, and Chris Bryan. Why? why not? when? visual explanations of agent behaviour in reinforcement learning. In *2022 IEEE 15th Pacific Visualization Symposium (PacificVis)*, pages 111–120. IEEE, 2022.
- [38] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.
- [39] Xiaotong Nie, Motoaki Hiraga, and Kazuhiro Ohkura. Visualizing deep q-learning to understanding behavior of swarm robotic system. In *Symposium on Intelligent and Evolutionary Systems*, pages 118–129. Springer, 2019.
- [40] Tiago Miguel Monteiro Nunes, Clark Borst, Erik-Jan van Kampen, Brian Hilburn, and Carl Westin. Human-interpretable input for machine learning in tactical air traffic control.
- [41] Juan Marcelo Parra-Ullauri, Antonio García-Domínguez, Nelly Bencomo, Changgang Zheng, Chen Zhen, Juan Boubeta-Puig, Guadalupe Ortiz, and Shufan Yang. Event-driven temporal models for explanations-etemox: explaining reinforcement learning. *Software and Systems Modeling*, 21(3):1091–1113, 2022.
- [42] Sindre Benjamin Remman and Anastasios M Lekkas. Robotic lever manipulation using hind-sight experience replay and shapley additive explanations. In *2021 European Control Conference (ECC)*, pages 586–593. IEEE, 2021.
- [43] Sindre Benjamin Remman, Inga Strümke, and Anastasios M Lekkas. Causal versus marginal shapley values for robotic lever manipulation controlled using deep reinforcement learning. *arXiv preprint arXiv:2111.02936*, 2021.
- [44] Stefano Giovanni Rizzo, Giovanna Vantini, and Sanjay Chawla. Reinforcement learning with explainability for traffic signal control. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3567–3572. IEEE, 2019.
- [45] Aaron M Roth, Jing Liang, and Dinesh Manocha. Xai-n: Sensor-based robot navigation using expert policies and decision trees. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2053–2060. IEEE, 2021.
- [46] Lincoln V Schreiber, Gabriel de O Ramos, and Ana LC Bazzan. Towards explainable deep reinforcement learning for traffic signal control.
- [47] Simon Schwaiger, Mohamed Aburaia, Ali Aburaia, and Wilfried Woeber. Explainable artificial intelligence for robot arm control. *Annals of DAAAM & Proceedings*, pages 640–648, 2021.
- [48] Alexander Sieusahai and Matthew Guzdial. Explaining deep reinforcement learning agents in the atari domain through a surrogate model. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 17, pages 82–90, 2021.
- [49] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharmashan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

- [50] Eduardo Soares, Plamen P Angelov, Bruno Costa, Marcos P Gerardo Castro, Subramanya Nagesh Rao, and Dimitar Filev. Explaining deep learning models through rule-based approximation and visualization. *IEEE Transactions on Fuzzy Systems*, 29(8):2399–2407, 2020.
- [51] Mark Stefik, Michael Youngblood, Peter Piroli, Christian Lebiere, Robert Thomson, Robert Price, Lester D Nelson, Robert Krivacic, Jacob Le, Konstantinos Mitsopoulos, et al. Explaining autonomous drones: An xai journey. *Applied AI Letters*, 2(4):e54, 2021.
- [52] Connor van Rossum, Candice Feinberg, Adam Abu Shumays, Kyle Baxter, and Benedek Bartha. A novel approach to curiosity and explainable reinforcement learning via interpretable sub-goals. *arXiv preprint arXiv:2104.06630*, 2021.
- [53] Marko Vasić, Andrija Petrović, Kaiyuan Wang, Mladen Nikolić, Rishabh Singh, and Sarfraz Khurshid. Moët: Mixture of expert trees and its application to verifiable reinforcement learning. *Neural Networks*, 151:34–47, 2022.
- [54] A Vijay and K Umadevi. Secured ai guided architecture for d2d systems of massive mimo deployed in 5g networks. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 468–472. IEEE, 2019.
- [55] George A Vouros. Explainable deep reinforcement learning: State of the art and challenges. *ACM Computing Surveys (CSUR)*, 2022.
- [56] Yuyao Wang, Masayoshi Mase, and Masashi Egi. Attribution-based salience method towards interpretable reinforcement learning. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (1)*, 2020.
- [57] Lindsay Wells and Tomasz Bednarz. Explainable ai and reinforcement learning—a systematic review of current approaches and trends. *Frontiers in artificial intelligence*, 4:550030, 2021.
- [58] Salomón Wollenstein-Betech, Christian Muise, Christos G Cassandras, Ioannis Ch Paschalidis, and Yasaman Khazaeni. Explainability of intelligent transportation systems using knowledge compilation: a traffic light controller case. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6. IEEE, 2020.
- [59] Duo Xu and Faramarz Fekri. Interpretable model-based hierarchical reinforcement learning using inductive logic programming. *arXiv preprint arXiv:2106.11417*, 2021.
- [60] Ke Zhang, Peidong Xu, and Jun Zhang. Explainable ai in deep reinforcement learning models: A shap method applied in power system emergency control. In *2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2)*, pages 711–716. IEEE, 2020.