Perspectives for Direct Interpretability in Multi-Agent Deep Reinforcement Learning

Yoann Poupart, Aurélie Beynier, Nicolas Maudet

Perspectives for Direct Interpretability in Multi-Agent Deep Reinforcement Learning

Yoann Poupart
LIP6, Sorbonne University
Paris, France
yoann.poupart@lip6.fr

Aurélie Beynier
LIP6, Sorbonne University
Paris, France
aurelie.beynier@lip6.fr

Nicolas Maudet
LIP6, Sorbonne University
Paris, France
nicolas.maudet@lip6.fr

ABSTRACT

Multi-Agent Deep Reinforcement Learning (MADRL) was proven efficient in solving complex problems in robotics or games, yet most of the trained models are hard to interpret. While learning intrinsically interpretable models remains a prominent approach, its scalability and flexibility are limited in handling complex tasks or multi-agent dynamics. This paper advocates for direct interpretability, generating post hoc explanations directly from trained models, as a versatile and scalable alternative, offering insights into agents' behaviour, emergent phenomena, and biases without altering models' architectures. We explore modern methods, including relevance backpropagation, knowledge edition, model steering, activation patching, sparse autoencoders and circuit discovery, to highlight their applicability to single-agent, multi-agent, and training process challenges. By addressing MADRL interpretability, we propose directions aiming to advance active topics such as team identification, swarm coordination and sample efficiency.

KEYWORDS

Interpretability, Multi-Agent Systems, Reinforcement Learning, Deep Neural Networks

1 INTRODUCTION

The increasing complexity of agents trained by Reinforcement Learning (RL) has raised significant safety and ethical concerns [82, 114, 125]. These considerations are even more crucial when training multiple agents based on Deep Neural Networks (DNNs), commonly referred to as black boxes, i.e., in Multi-Agent Deep Reinforcement Learning [19]. MADRL enables solving more complex problems through cooperation or opponent modelling [38, 46, 132], and finds applications in robotics [87], video games [124] or even health [112]. Recent advancements, such as pre-trained world models [6, 16, 99, 134] and the integration of Large Language Models (LLMs), as standalone agents [128] or within Multi-Agent Systems (MAS) [43, 68, 133], further exacerbate the interpretability challenge. While the field of eXplainable RL (XRL) is growing by the year [10, 47, 49, 79, 93], with one of the first dedicated workshops organised at the first RL Conference edition [61], interpretability is anecdotal in MADRL [48, 59, 75, 80, 127, 138]. Yet, as we expose in Section 3, interpretability could help advance specific challenges in MADRL, such as team identification, swarm coordination and sample efficiency.

Existing efforts in agent interpretability predominantly focus on intrinsically interpretable models [10, 47, 49, 79, 93], emphasising simplicity in architecture to make systems inherently understandable [17, 103]. However, these approaches often need to be revised for large and performant systems where expressiveness, scalability and flexibility are essential [105]. We thus propose to focus on direct interpretability, i.e., methods that are post-hoc, applicable after training, and generate explanations directly from DNNs. This class of methods enables probing complex systems without constraining their design or needing to extract an interpretable model. Inspired by modern interpretability methods [28, 31, 58, 141], and new XRL approaches [64, 67, 110], we decided to anticipate the adoption of explainability in the expanding field of MADRL and encourage the AAMAS community to use and engage more systematically with modern direct interpretability methods.

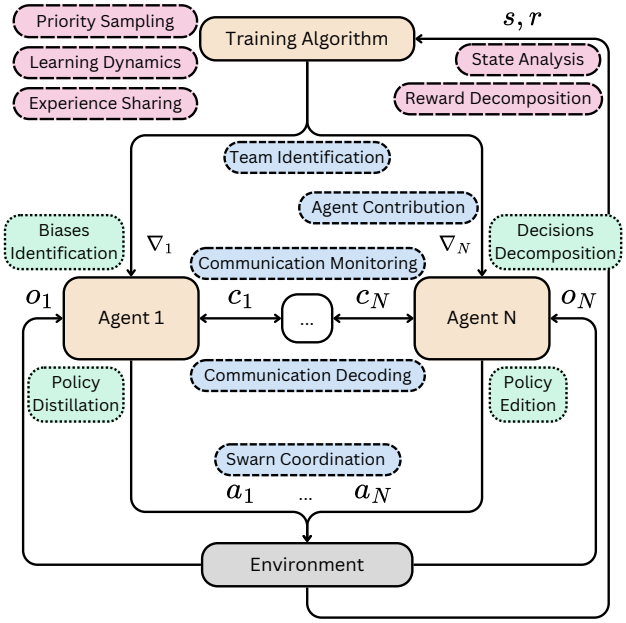

Figure 1: Visual taxonomy of MADRL challenges that could benefit from direct interpretability methods. In green (dots) challenges related to a single agent, in blue (short dashes) to multiple agents and in red (long dashes) to the training process.

We list our contributions as follows:

- Arguments to engage with direct interpretability methods.
- A taxonomy to position direct interpretability in MADRL.
- Potential applications of direct interpretability to solve modern MADRL challenges.

In this article, we first present an initial background about the systems of study and the methods advocated. Then, we propose a simple taxonomy to position modern interpretability methods in

the MADRL framework. Finally, we outline the limitations of some current works while proposing alternative ideas tracks.

2 BACKGROUND

2.1 Multi-Agent Deep Reinforcement Learning

A typical system consists of the following components: agents, an environment, and a training algorithm, as depicted in Figure 2. Formally, we consider a system with N agents, each indexed by $i \in \{1, \ldots, N\}$. At each time step, the agent i is presented with an observation o_i and produces an action a_i. For the sake of generality, we included a possible communication channel c_i, seeing that it is increasingly used [140]. In principle, we can extend the definition of communication to include the most common MADRL methods like parameter sharing [23, 40], which can be seen as a form of latent space communication. Finally, the training algorithm provides feedback ∇_i to each agent.

Training algorithms in MADRL can be centralized, decentralized, or hybrid. Centralized training uses the joint action $a = (a_1, ..., a_N)$ and the state s, which can be understood as an observation augmented by information at training time [63], and consists of applying classical RL to multi-agent problems like for AplhaStar [76]. While decentralized training restricts each agent to local observations o_i, possibly including a local reward r_i, see IDQN [120] or IPPO [137]. Hybrid approaches, such as centralized training with decentralized execution, leverage global information during training but allow agents to act independently using only local observations during execution, see VDN [119], QMIX [98], MADPG [70] or MAPPO [137]. Here, we consider agents based on DNNs; therefore, the feedbacks ∇_i are gradients of a loss ℓ. Depending on the training algorithm, this loss can be a function of the reward r, the state s, the actions a_i, the observations o_i and the communications c_i. For simplicity, we didn't include those dependencies in Figure 2.

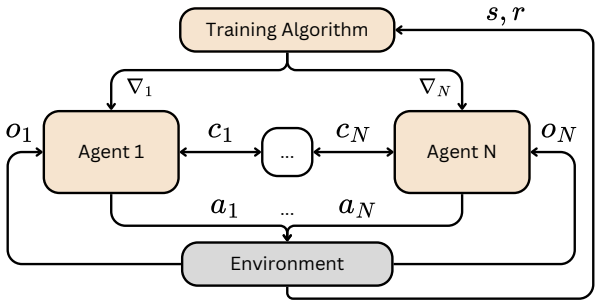

Figure 2: Schema of a simplified view of MADRL systems. At each time step, the agent i receives the initial observation o_i, complemented by potential communications c_i and produces an action a_i. The agent learns throughout training by the means of gradients ∇_i.

2.2 Direct Interpretability of DNNs

We now present an overview of the modern methods widely used to interpret DNNs in Computer Vision (CV) and Natural Language Processing (NLP). As these domains heavily relied on pre-trained models [44, 96, 116], direct post-hoc methods have dominated the research landscape, providing key hindsight without altering models' architectures.

Feature importance. Typical methods used in CV to understand convolutional networks involve visualising important pixels, i.e. saliency maps, [109, 139]. Other methods compute importance by perturbing the input [27], using the gradients [95, 109, 115, 117] or locally decomposing relevance [8, 83]. Recent works in NLP focus on the Transformer architecture and its attention mechanism [123], providing token-level insights [2, 131].

Prototypes: a class of methods that creates explanations based on characteristic samples. In CV, it is common to analyse neurons using activation maximisation to create pre-images [74], or find related images [20]. Prototypes can be of various forms like perturbed images [100], cropped images [29] or latent space vector [4, 60]. Recent works based on sparse autoencoders were able to elicit interpretable features in LLMs, i.e., prototypes [28].

Latent manipulation: techniques that further extend the interpretability of concepts and features by exploring the internal representations learned by models. These methods were introduced in CV with [60], later derived as the field of representation engineering [141]. Such latent features enable locating, editing, erasing or decoding models' knowledge [12, 35, 78], but causally modify or analyse the produced outputs [62, 101].

Circuit analysis: provides a more granular understanding of model internals by examining pathways and dependencies between models' components, usually neurons or attention heads. Circuits were first discovered in CNNs [85] before being formalised for Transformers [32]. These circuits revealed peculiar models' components that learned precise mechanisms like induction [86]. Using specific datasets, relevant circuits can be automatically discovered [26]. More recent works focus on larger models' components at the layer scale [31].

3 ADVOCATING DIRECT INTERPRETABILITY

Direct methods offer a significant advantage in their applicability to models during and after training, enabling developers to analyse and interpret complex systems without requiring architectural changes. This flexibility makes them particularly suitable for MADRL systems compared to intrinsic methods that might be challenging to scale with several agents. Figure 1 outlines speculative research directions and methodologies that can enhance systems understanding at different levels, from individual agents to the overall training process.

3.1 Single-Agent Challenges

To understand agents trained using MADRL, we can study each agent independently. Methods drawn from XRL and general interpretability are thus directly applicable to tackle single-agent challenges.

Biases identification: eliciting models' biases learned during training. In order to debug those "Clever Hans"[1], it is possible to use

[1]Cognitive bias that was learned due to spurious correlations, see [65].

feature importance techniques, described in Section 2.2. Previous work [65] showed that this debugging could be semi-automated by combining LRP [8] with spectral clustering [126]. While these methods are relatively established in XRL, further improvements tailored to MADRL could offer more context-specific explanations, e.g., by comparing different agents' perceptions.

Policy distillation, converting a model into a simpler one, can be achieved by training a new smaller model [106], or by extracting intrinsically interpretable models [9, 104], even for MADRL [80]. Yet, these distillation methods are computationally expensive. Recent works proposed network compression based on interpretability, using weights relevance [136] or circuit analysis [90].

Decision decomposition: could be achieved by internally decomposing an agent's decision into functional modules or representations. This methodology was proven efficient to elicit the algorithms behind certain capabilities, like addition or modular addition [84, 94]. Future work could focus on extracting different circuits using ACDC [26] to analyse simple shared actor-critic architectures, e.g., to extract the actor subnetwork.

Policy edition, an essential aspect to regain control over DNNs. Indeed, being able to edit a trained policy is essential to remove biases, unwanted associations or dangerous behaviours without needing to retrain the model. In this respect, direct interpretability is perfectly suited for the task with methods leveraging CAVs [30] or causal tracing [78].

3.2 Multi-Agent Challenges

Interpretability could be a powerful tool for automating the oversight of systems involving multiple agents. Indeed, such systems become more complex through inter-agent interactions, coordination strategies, and emergent behaviours.

Team identification: grouping together agents with similar roles or policies. This is particularly interesting to reduce the complexity of MAS by having fewer agents to train or could be an avenue to extend the mean-field framework [135]. Previous work showed that selective parameter-sharing can be based on latent spaces [21]. Further improvements could consider dynamic teams throughout learning by analysing mixing networks [98], e.g., by partitioning the positive weights using NMF [88], or using other prototype methods like SAE [28].

Agent contribution, or agent credit assignment, is a well-known challenge introduced by MAS. Shapley values theoretically give the individual agent contributions [113], and thus can be computed using SHAP or equivalent methods [48, 71, 127]. Yet, as it can be expensive to compute, it might be beneficial to explore other versatile methods like LRP [8], e.g., by designing specific relevance propagation rules.

Communication monitoring. In settings with natural language communication between agents, leveraging LLMs or pre-trained models can enable a seamless integration [140]. Yet, these models are highly opaque and would benefit from interpretability, offering an avenue to supervise and interpret conversations. Applications could make use of feature importance methods, like AttnLRP [2], to spot key information used in the agent prediction.

Communication decoding. For learned communication analyses, it becomes harder and might be reduced to finding patterns or comparing and aligning latent spaces to spot similar messages between agents. In order to uncover how agents derive meaning from these interactions, causal interventions might yields interesting hindsights [62].

Swarm coordination: an inherent challenge of MAS that becomes increasingly complex as the number of agents scales. Fortunately, modern direct interpretability offers means to control models using methods from representation engineering [141], like activation steering [101]. The latter method has proven useful to control an agent's policy by favouring different goals [81]. Further application to MADRL could improve swarm coordination by enhancing traits like cooperativeness or better distributing goals among agents, e.g., by alternating resource collection among sites and agents.

3.3 Training Process Challenges

Training multiple agents simultaneously demands more computing power and can lead to learning instabilities. Therefore, it is crucial to better understand the training process of MADRL at different levels by improving learning efficiency and ensuring robustness.

State analysis. In order to model complex environments, one can train world models [16], later used by an agent [41]. The condensed latent representation obtained can be analysed [52] with tools like the tuned lens [11]. This framework offers a better view of the transition function, which could help guide the agents towards unbiased training if analysed thoroughly.

Reward decomposition: often achieved by learning separate value functions aggregated afterwards [57, 122]. To avoid arbitrary decompositions, one could rely on local backpropagation methods like LRP or CRP [1, 8], enabling the discovery of concepts that can later clarify the influence of the reward on the learning process of a policy. Further improvements could consider generating an adaptative curriculum [56], prioritizing the concepts to learn.

Priority sampling, a staple method in RL that improves sample efficiency [107]. Also, in RL, interpretability was proven efficient to prioritize the important pixels for a visual policy by means of a consistency loss [13]. Such a framework could be extended to compute importance over multiple inputs, creating a metric for better eliciting shared critical training samples.

Learning dynamics: trying to understand the agents throughout training, e.g., by observing the trained policies. Yet, it becomes more complicated as the number of agents scales and requires automated methods beyond observing policies. A widely used method to detect learned concepts in a model is to train linear probes [4], which gave valuable insights for the analysis of AlphaZero networks [77]. By monitoring each agent, it would be possible to gain a more nuanced understanding of the swarm development and track the emergence or disappearance of certain capabilities.

Experience sharing: a method introduced to scale MADRL by improving sample efficiency [22]. Further improvements shared

the data selectively according to exploration metrics [34]. Yet, this framework is missing a key point: you might want to select agents that share their experience similarly to parameter sharing [21]. A naive method could be to cluster experiences based on some latent representation of the different agents, enabling efficient knowledge sharing [141].

4 DISCUSSION

4.1 Post-Hoc Interpretability in Deep RL

Post hoc interpretability in Deep Reinforcement Learning (DRL) is an increasingly important field, with methods such as saliency maps already being used to visualize agent behaviour [37], debug learned concepts [50, 53], and inform sampling strategies to improve efficiency [13]. Other approaches analyse agent behaviour by querying interaction data [111] or by visualising pattern prototypes [5, 97]. More extensive efforts have focused on interpreting well-known chess engines like AlphaZero [42, 54, 69, 77, 92, 108] and Stockfish [89], providing valuable insights into learned strategies. Ongoing efforts are also focused on exposing the key mechanisms behind planning, especially with games as a testbed [24, 39, 55, 121].

Other post hoc methods, like policy distillation into interpretable models, often referred to as model extraction, have also been a central focus. Techniques such as DAGGER [104] and VIPER [9] leverage imitation learning to simplify policies. However, these methods struggle to scale effectively when applied globally to complex models, limiting their applicability to large-scale systems.

4.2 Interpretability in MADRL

Interpretability in MADRL is an evolving field with several promising approaches. Shapley values have been widely applied to analyse individual agent contributions, providing a robust theoretical framework for evaluating each agent's influence on team performance [48, 75, 127]. Diversity measures of agent policies have also emerged as a valuable tool for understanding agent behaviour, revealing distinctions between individual strategies and their roles in collective dynamics [59].

Similarly to XRL, policy extraction techniques, such as VIPER [9], have been extended to leverage MADRL training to distil interpretable policies from complex models [80]. Furthermore, predicting high-level concepts instead of actions offers a novel pathway to intrinsically interpretable models, aligning model outputs with human-understandable abstractions [138]. These advancements highlight the growing potential of interpretability methods in uncovering insights into multi-agent behaviour and learning processes.

4.3 Limits of Intrinsically Interpretable Models

Intrinsically interpretable models, whether obtained by design or post hoc extraction, have long been a dominant paradigm in agent interpretability research, relying on predefined, transparent model architectures. Design frameworks like XAg [102], concept bottlenecks [91], learning skills with decision trees [130], or learning modularised agents [25], aim to embed interpretability directly into model structures. However, such approaches face challenges in scalability and flexibility, particularly in multi-agent settings or with complex DRL models like the latest pre-trained world models [6, 16, 99, 134]. The rigidity of design-based interpretability often compromises performance and fails to capture emergent behaviours, highlighting the need for alternative approaches that can adapt to the complexity and scale of modern systems [72]. New hybrid paradigms like Wrapper Boxes [118], might be required to overcome those limitations.

5 PERSPECTIVES

5.1 MADRL Should Leverage Direct Interpretability

Engaging and expanding interpretability is an opportunity to address existing challenges in MADRL. Direct approaches are particularly well-suited for analysing communication dynamics, coordination strategies, and emergent behaviours in MAS. Graph-based analysis, for instance, could provide insights into inter-agent interactions, while feature importance techniques can identify biases and ensure fairness in decision-making. By systematically exploring and applying scalable direct methods to trained models, researchers can better address the inherent complexities of MADRL, enabling the development of more transparent, robust, and accountable systems for real-world applications.

Although previous calls to action are prone to integrate interpretability beforehand [102], this paper claims that the interpretation of models post hoc is highly valuable. Direct interpretability offers greater flexibility, particularly for existing models where architectural modifications are impractical.

5.2 Robust Evaluation Protocols

As repeatedly outlined, direct post-hoc methods are easily actionable and scalable. However, their adoption requires acknowledging and addressing limitations such as the inherent shortcomings of saliency maps [3, 14], counterfactual explanations [66], or other interpretability illusions [15, 33, 33]. In fact, these methods often generate metrics with limited predictive power, and thus, claims should be reasonable.

A key priority is the development of robust evaluation protocols for direct methods. Given the absence of ground-truth explanations, reliable metrics and standardized evaluation frameworks must be established to assess the quality and utility of these methods [7, 18, 36, 45, 51, 73, 129]. Advancing evaluation thoroughly, e.g., by evaluating out of distribution, is especially important to develop scalable, effective, and actionable interpretability solutions.

6 CONCLUSION

We outlined that direct interpretability might be vital for addressing the challenges of scalability and complexity in modern MADRL. It enables the analysis of trained models without imposing architectural constraints, providing critical insights into agent behaviour, emergent dynamics, and biases. Advancing these methods will ensure scalable oversight of these systems, which is a precious desideratum for real-world applications. However, challenges such as explanation illusions, lack of robust evaluation metrics, and difficulty disentangling causal effects should be considered and tackled.

REFERENCES

[1] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. 2022. From attribution maps to human-understandable explanations through Concept Relevance Propagation. Nature Machine Intelligence 5 (2022), 1006 – 1019.

[2] Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2024. AttnLRP: Attention-Aware Layer-wise Relevance Propagation for Transformers. ArXiv abs/2402.05602 (2024).

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In Neural Information Processing Systems.

[4] Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. arXiv:1610.01644 [stat.ML]

[5] Gülsüm Alicioğlu and Bo Sun. 2024. Use Bag-of-Patterns Approach to Explore Learned Behaviors of Reinforcement Learning. In xAI.

[6] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. [n.d.]. Diffusion for World Modeling: Visual Details Matter in Atari. In Thirty-eighth Conference on Neural Information Processing Systems.

[7] José Pereira Amorim, Pedro Henriques Abreu, João A. M. Santos, and Henning Müller. 2023. Evaluating Post-hoc Interpretability with Intrinsic Interpretability. ArXiv abs/2305.03002 (2023).

[8] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS ONE 10 (2015).

[9] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. 2018. Verifiable Reinforcement Learning via Policy Extraction. In Neural Information Processing Systems.

[10] Yanzhe Bekkemoen. 2023. Explainable reinforcement learning (XRL): a systematic literature review and taxonomy. Machine Learning 113 (2023), 355–441.

[11] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor V. Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting Latent Predictions from Transformers with the Tuned Lens. ArXiv abs/2303.08112 (2023).

[12] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. arXiv:2306.03819 [cs.LG]

[13] David Bertoin, Adil Zouitine, Mehdi Zouitine, and Emmanuel Rachelson. 2022. Look where you look! Saliency-guided Q-networks for generalization in visual Reinforcement Learning. In Neural Information Processing Systems.

[14] Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. 2022. Impossibility theorems for feature attribution. Proceedings of the National Academy of Sciences of the United States of America 121 (2022).

[15] Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Vi'egas, and Martin Wattenberg. 2021. An Interpretability Illusion for BERT. ArXiv abs/2104.07143 (2021).

[16] Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal M. P. Behbahani, Stephanie Chan, Nicolas Manfred Otto Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktaschel. 2024. Genie: Generative Interactive Environments. ArXiv abs/2402.15391 (2024).

[17] Aditya Chattopadhyay, Stewart Slocum, Benjamin David Haeffele, René Vidal, and Donald Geman. 2022. Interpretable by Design: Learning Predictors by Composing Interpretable Queries. IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2022), 7430–7443.

[18] Maheep Chaudhary and Atticus Geiger. 2024. Evaluating Open-Source Sparse Autoencoders on Disentangling Factual Knowledge in GPT-2 Small. ArXiv abs/2409.04478 (2024).

[19] Paul Constantin Chelarescu. 2021. Deception in Social Learning: A Multi-Agent Reinforcement Learning Perspective. ArXiv abs/2106.05402 (2021).

[20] Zhi Chen, Yijie Bei, and Cynthia Rudin. 2020. Concept whitening for interpretable image recognition. Nature Machine Intelligence 2 (2020), 772 – 782.

[21] Filippos Christianos, Georgios Papoudakis, Arrasy Rahman, and Stefano V. Albrecht. 2021. Scaling Multi-Agent Reinforcement Learning with Selective Parameter Sharing. ArXiv abs/2102.07475 (2021).

[22] Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. 2020. Shared Experience Actor-Critic for Multi-Agent Reinforcement Learning. ArXiv abs/2006.07169 (2020).

[23] Xiangxiang Chu and Hangjun Ye. 2017. Parameter Sharing Deep Deterministic Policy Gradient for Cooperative Multi-agent Reinforcement Learning. ArXiv abs/1710.00336 (2017).

[24] Stephen Chung, Scott Niekum, and David Krueger. 2024. Predicting Future Actions of Reinforcement Learning Agents. ArXiv abs/2410.22459 (2024).

[25] Alex Cloud, Jacob Goldman-Wetzler, Evzen Wybitul, Joseph Miller, and Alexander Matt Turner. 2024. Gradient Routing: Masking Gradients to Localize Computation in Neural Networks. ArXiv abs/2410.04332 (2024).

[26] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards Automated Circuit Discovery for Mechanistic Interpretability. arXiv:2304.14997 [cs.LG]

[27] Ian Covert, Scott M. Lundberg, and Su-In Lee. 2020. Explaining by Removing: A Unified Framework for Model Explanation. J. Mach. Learn. Res. 22 (2020), 209:1–209:90.

[28] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse Autoencoders Find Highly Interpretable Features in Language Models. ArXiv abs/2309.08600 (2023).

[29] Maximilian Dreyer, Reduan Achtibat, Wojciech Samek, and Sebastian Lapuschkin. 2023. Understanding the (Extra-)Ordinary: Validating Deep Model Decisions with Prototypical Concept-based Explanations. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2023), 3491–3501.

[30] Maximilian Dreyer, Frederik Pahde, Christopher J. Anders, Wojciech Samek, and Sebastian Lapuschkin. 2023. From Hope to Safety: Unlearning Biases of Deep Models via Gradient Penalization in Latent Space. In AAAI Conference on Artificial Intelligence.

[31] Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024. Transcoders Find Interpretable LLM Feature Circuits. ArXiv abs/2406.11944 (2024).

[32] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. Transformer Circuits Thread 1, 1 (2021), 12.

[33] Dan Friedman, Andrew K. Lampinen, Lucas Dixon, Danqi Chen, and Asma Ghandeharioun. 2023. Interpretability Illusions in the Generalization of Simplified Models. ArXiv abs/2312.03656 (2023).

[34] Matthias Gerstgrasser, Tom Danino, and Sarah Keren. 2023. Selectively Sharing Experiences Improves Multi-Agent Reinforcement Learning. ArXiv abs/2311.00865 (2023).

[35] Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A Unifying Framework for Inspecting Hidden Representations of Language Models. ArXiv abs/2401.06102 (2024).

[36] Navdeep Gill, Patrick Hall, Kim Montgomery, and Nicholas Schmidt. 2020. A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing. Inf. 11 (2020), 137.

[37] Sam Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. 2017. Visualizing and Understanding Atari Agents. ArXiv abs/1711.00138 (2017).

[38] Sven Gronauer and Klaus Diepold. 2021. Multi-agent deep reinforcement learning: a survey. Artificial Intelligence Review 55 (2021), 895 – 943.

[39] Hung Guei, Yan-Ru Ju, Wei-Yu Chen, and Ti-Rong Wu. 2024. Interpreting the Learned Model in MuZero Planning.

[40] Jayesh K. Gupta, Maxim Egorov, and Mykel J. Kochenderfer. 2017. Cooperative Multi-agent Control Using Deep Reinforcement Learning. In AAMAS Workshops.

[41] Danijar Hafner, J. Pazukonis, Jimmy Ba, and Timothy P. Lillicrap. 2023. Mastering Diverse Domains through World Models. ArXiv abs/2301.04104 (2023).

[42] Patrik Hammersborg and Inga Strümke. 2023. Information based explanation methods for deep learning agents–with applications on large open-source chess models. arXiv preprint arXiv:2309.09702 (2023).

[43] Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. 2024. LLM Multi-Agent Systems: Challenges and Open Problems. ArXiv abs/2402.03578 (2024).

[44] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015), 770–778.

[45] Anna Hedström, Leander Weber, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. 2022. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations. ArXiv abs/2202.06861 (2022).

[46] Pablo Hernandez-Leal, Bilal Kartal, and Matthew E. Taylor. 2018. A survey and critique of multiagent deep reinforcement learning. Autonomous Agents and Multi-Agent Systems 33 (2018), 750 – 797.

[47] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz Rodríguez. 2020. Explainability in Deep Reinforcement Learning. Knowl. Based Syst. 214 (2020), 106685.

[48] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz Rodríguez. 2021. Collective eXplainable AI: Explaining Cooperative Strategies and Agent Contribution in Multiagent Reinforcement Learning With Shapley Values. IEEE Computational Intelligence Magazine 17 (2021), 59–71.

[49] Tom Hickling, Abdelhafid Zenati, Nabil Aouf, and Phillippa Spencer. 2022. Explainability in Deep Reinforcement Learning: A Review into Current Methods and Applications. Comput. Surveys 56 (2022), 1 – 35.

[50] Jacob Hilton, Nick Cammarata, Shan Carter, Gabriel Goh, and Christopher Olah. 2020. Understanding RL vision.

[51] Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. 2024. RAVEL: Evaluating Interpretability Methods on Disentangling Language Model Representations. ArXiv abs/2402.17700 (2024).

[52] Michael I. Ivanitskiy, Alex F Spies, Tilman Rauker, Guillaume Corlouer, Chris Mathwin, Lucia Quirke, Can Rager, Rusheb Shah, Dan Valentine, Cecilia G. Diniz Behn, Katsumi Inoue, and Samy Wu Fung. 2023. Structured World Representations in Maze-Solving Transformers. ArXiv abs/2312.02566 (2023).

[53] Max Jaderberg, Wojciech M. Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio García Castañeda, Charlie Beattie, Neil C. Rabinowitz, Ari S. Morcos, Avraham Ruderman, Nicolas Sonnerat, Tim Green, Louise Deason, Joel Z. Leibo, David Silver, Demis Hassabis, Koray Kavukcuoglu, and Thore Graepel. 2018. Human-level performance in 3D multiplayer games with population-based reinforcement learning. Science 364 (2018), 859 – 865.

[54] Erik Jenner, Shreyas Kapur, Vasil Georgiev, Cameron Allen, Scott Emmons, and Stuart Russell. 2024. Evidence of Learned Look-Ahead in a Chess-Playing Neural Network. arXiv:2406.00877 [cs.LG]

[55] Erik Jenner, Shreyas Kapur, Vasil Georgiev, Cameron Allen, Scott Emmons, and Stuart Russell. 2024. Evidence of Learned Look-Ahead in a Chess-Playing Neural Network. ArXiv abs/2406.00877 (2024).

[56] Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. 2020. Prioritized Level Replay. In International Conference on Machine Learning.

[57] Zoe Juozapaitis, Anurag Koul, Alan Fern, Martin Erwig, and Finale Doshi-Velez. 2019. Explainable Reinforcement Learning via Reward Decomposition.

[58] Shahar Katz, Yonatan Belinkov, Mor Geva, and Lior Wolf. 2024. Backward Lens: Projecting Language Model Gradients into the Vocabulary Space. ArXiv abs/2402.12865 (2024).

[59] Wiem Khlifi, Siddarth Singh, Omayma Mahjoub, Ruan de Kock, Abidine Vall, R. Gorsane, and Arnu Pretorius. 2023. On Diagnostics for Understanding Agent Training Behaviour in Cooperative MARL. ArXiv abs/2312.08468 (2023).

[60] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). arXiv:1711.11279 [stat.ML]

[61] Hector Kohler, Quentin Delfosse, Paul Festor, and Philippe Preux. 2024. Towards a Research Community in Interpretable Reinforcement Learning: the InterpPol Workshop. ArXiv abs/2404.10906 (2024).

[62] J'anos Kram'ar, Tom Lieberum, Rohin Shah, and Neel Nanda. 2024. AtP*: An efficient and scalable method for localizing LLM behaviour to components. ArXiv abs/2403.00745 (2024).

[63] Gaspard Lambrechts, Adrien Bolland, and Damien Ernst. 2023. Informed POMDP: Leveraging Additional Information in Model-Based RL. In RLC.

[64] Moritz Lange, Raphael C. Engelhardt, Wolfgang Konen, and Laurenz Wiskott. 2024. Interpretable Brain-Inspired Representations Improve RL Performance on Visual Navigation Tasks. ArXiv abs/2402.12067 (2024).

[65] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. Nature Communications 10 (2019).

[66] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, X. Renard, and Marcin Detyniecki. 2019. The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations. In International Joint Conference on Artificial Intelligence.

[67] Mark Levin and Hana Chockler. 2023. Clustered Policy Decision Ranking. ArXiv abs/2311.12970 (2023).

[68] Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. Vicinagearth (2024).

[69] Charles Lovering, Jessica Forde, George Konidaris, Ellie Pavlick, and Michael Littman. 2022. Evaluation Beyond Task Performance: Analyzing Concepts in AlphaZero in Hex. Advances in Neural Information Processing Systems 35 (2022), 25992–26006.

[70] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, P. Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. ArXiv abs/1706.02275 (2017).

[71] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In Neural Information Processing Systems.

[72] Andreas Madsen, Himabindu Lakkaraju, Siva Reddy, and Sarath Chandar. 2024. Interpretability Needs a New Paradigm. ArXiv abs/2405.05386 (2024).

[73] Andreas Madsen, Siva Reddy, and A. P. Sarath Chandar. 2021. Post-hoc Interpretability for Neural NLP: A Survey. Comput. Surveys 55 (2021), 1 – 42.

[74] Aravindh Mahendran and Andrea Vedaldi. 2015. Visualizing Deep Convolutional Neural Networks Using Natural Pre-images. International Journal of Computer Vision 120 (2015), 233–255.

[75] Omayma Mahjoub, Ruan de Kock, Siddarth Singh, Wiem Khlifi, Abidine Vall, Kale ab Tessera, and Arnu Pretorius. 2023. Efficiently Quantifying Individual Agent Importance in Cooperative MARL. ArXiv abs/2312.08466 (2023).

[76] Michaël Mathieu, Sherjil Ozair, Srivatsan Srinivasan, Caglar Gulcehre, Shangtong Zhang, Ray Jiang, Tom Le Paine, Richard Powell, Konrad Zolna, Julian Schrittwieser, David Choi, Petko Georgiev, Daniel Toyama, Aja Huang, Roman Ring, Igor Babuschkin, Timo Ewalds, Mahyar Bordbar, Sarah Henderson, Sergio Gomez Colmenarejo, Aäron van den Oord, Wojciech M. Czarnecki, Nando de Freitas, and Oriol Vinyals. 2023. AlphaStar Unplugged: Large-Scale Offline Reinforcement Learning. ArXiv abs/2308.03526 (2023).

[77] Thomas McGrath, Andrei Kapishnikov, Nenad Tomaš ev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. 2022. Acquisition of chess knowledge in AlphaZero. Proceedings of the National Academy of Sciences 119, 47 (nov 2022). https://doi.org/10.1073/pnas.2206625119

[78] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. In Neural Information Processing Systems.

[79] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. 2023. Explainable Reinforcement Learning: A Survey and Comparative Review. Comput. Surveys 56 (2023), 1 – 36.

[80] Stephanie Milani, Zhicheng Zhang, Nicholay Topin, Zheyuan Ryan Shi, Charles A. Kamhoua, Evangelos E. Papalexakis, and Fei Fang. 2022. MAVIPER: Learning Decision Tree Policies for Interpretable Multi-Agent Reinforcement Learning. In ECML/PKDD.

[81] Ulisse Mini, Peli Grietzer, Mrinank Sharma, Austin Meek, Monte Stuart MacDiarmid, and Alexander Matt Turner. 2023. Understanding and Controlling a Maze-Solving Policy Network. ArXiv abs/2310.08043 (2023).

[82] Catalin Mitelut, Ben Smith, and Peter Vamplew. 2023. Intent-aligned AI systems deplete human agency: the need for agency foundations research in AI safety. arXiv:2305.19223 [cs.AI]

[83] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2015. Explaining nonlinear classification decisions with deep Taylor decomposition. ArXiv abs/1512.02479 (2015).

[84] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. ArXiv abs/2301.05217 (2023).

[85] Christopher Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom In: An Introduction to Circuits.

[86] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova Dassarma, Tom Henighan, Benjamin Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, John Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom B. Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Christopher Olah. 2022. In-context Learning and Induction Heads. ArXiv abs/2209.11895 (2022).

[87] James Orr and Ayan Dutta. 2023. Multi-Agent Deep Reinforcement Learning for Multi-Robot Applications: A Survey. Sensors (Basel, Switzerland) 23 (2023).

[88] Pentti Paatero and Unto Tapper. 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values†. Environmetrics 5 (1994), 111–126.

[89] Aðalsteinn Pálsson and Yngvi Björnsson. [n.d.]. Unveiling concepts learned by a world-class chess-playing agent.

[90] Nicholas Pochinkov and Nandi Schoots. 2024. Dissecting Language Models: Machine Unlearning via Selective Pruning. ArXiv abs/2403.01267 (2024).

[91] Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. 2023. Concept-based Explainable Artificial Intelligence: A Survey. ArXiv abs/2312.12936 (2023).

[92] Yoann Poupart. 2024. Contrastive Sparse Autoencoders for Interpreting Planning of Chess-Playing Agents. ArXiv abs/2406.04028 (2024).

[93] Yunpeng Qing, Shunyu Liu, Jie Song, and Mingli Song. 2022. A Survey on Explainable Reinforcement Learning: Concepts, Algorithms, Challenges. ArXiv abs/2211.06665 (2022).

[94] Philip Quirke and Fazl Barez. 2023. Understanding Addition in Transformers. ArXiv abs/2310.13121 (2023).

[95] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. CoRR abs/1511.06434 (2015).

[96] Alec Radford and Karthik Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training.

[97] Ronilo J. Ragodos, Tong Wang, Qihang Lin, and Xun Zhou. 2022. ProtoX: Explaining a Reinforcement Learning Agent via Prototyping. ArXiv abs/2211.03162 (2022).

[98] Tabish Rashid, Mikayel Samvelyan, C. S. D. Witt, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. ArXiv abs/1803.11485 (2018).

[99] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley D. Edwards, Nicolas Manfred Otto Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. 2022. A Generalist Agent. ArXiv abs/2205.06175 (2022).

[100] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In AAAI Conference on Artificial Intelligence.

[101] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering Llama 2 via Contrastive Activation Addition. arXiv:2312.06681 [cs.CL]

[102] Sebastian Rodriguez and John Thangarajah. 2024. Explainable Agents (XAg) by Design. In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (Auckland, New Zealand) (AAMAS '24). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2712–2716.

[103] Sebastian Rodriguez, John Thangarajah, and Andrew Davey. 2024. Design Patterns for Explainable Agents (XAg). In Adaptive Agents and Multi-Agent Systems.

[104] Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. 2010. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In International Conference on Artificial Intelligence and Statistics.

[105] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2021. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. ArXiv abs/2103.11251 (2021).

[106] Andrei A. Rusu, Sergio Gomez Colmenarejo, Çaglar Gülçehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. 2015. Policy Distillation. CoRR abs/1511.06295 (2015).

[107] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2015. Prioritized Experience Replay. CoRR abs/1511.05952 (2015).

[108] Lisa Schut, Nenad Tomasev, Tom McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. 2023. Bridging the Human-AI Knowledge Gap: Concept Discovery and Transfer in AlphaZero. arXiv:2310.16410 [cs.AI]

[109] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision 128 (2016), 336 – 359.

[110] Hyunki Seong and David Hyunchul Shim. 2024. Self-Supervised Interpretable End-to-End Learning via Latent Functional Modularity. In International Conference on Machine Learning.

[111] Pedro Sequeira, Eric Yeh, and Melinda T. Gervasio. 2019. Interestingness Elements for Explainable Reinforcement Learning through Introspection. In IUI Workshops.

[112] Thanveer Basha Shaik, Xiaohui Tao, Haoran Xie, Lin Li, Jianming Yong, and Hongning Dai. 2023. Adaptive Multi-Agent Deep Reinforcement Learning for Timely Healthcare Interventions.

[113] Lloyd S. Shapley. 1988. A Value for n-person Games.

[114] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Contributing authors, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, Katarina Slama, Lama Ahmad, Paul McMillan, Alex Beutel, Alexandre Passos, and David G. Robinson. 2023. Practices for Governing Agentic AI Systems.

[115] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. ArXiv abs/1605.01713 (2016).

[116] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR abs/1409.1556 (2014).

[117] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise. ArXiv abs/1706.03825 (2017).

[118] Yiheng Su, Juni Jessy Li, and Matthew Lease. 2023. Interpretable by Design: Wrapper Boxes Combine Neural Performance with Faithful Explanations. ArXiv abs/2311.08644 (2023).

[119] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech M. Czarnecki, Vinícius Flores Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. 2017. Value-Decomposition Networks For Cooperative Multi-Agent Learning. ArXiv abs/1706.05296 (2017).

[120] Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. 2015. Multiagent cooperation and competition with deep reinforcement learning. PLoS ONE 12 (2015).

[121] Mohammad Taufeeque, Philip Quirke, Maximilian Li, Chris Cundy, Aaron David Tucker, Adam Gleave, and Adrià Garriga-Alonso. 2024. Planning in a recurrent neural network that plays Sokoban.

[122] Harm van Seijen, Mehdi Fatemi, Romain Laroche, Joshua Romoff, Tavian Barnes, and Jeffrey Tsang. 2017. Hybrid Reward Architecture for Reinforcement Learning. ArXiv abs/1706.04208 (2017).

[123] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In Neural Information Processing Systems.

[124] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, L. Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom Le Paine, Caglar Gulcehre, Ziyun Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature 575 (2019), 350 – 354.

[125] Ajay Vishwanath, Louise A. Dennis, and Marija Slavkovik. 2024. Reinforcement Learning and Machine ethics:a systematic review. ArXiv abs/2407.02425 (2024).

[126] Ulrike von Luxburg. 2007. A tutorial on spectral clustering. Statistics and Computing 17 (2007), 395–416.

[127] Jianhong Wang, Yuan Zhang, Yunjie Gu, and Tae-Kyun Kim. 2021. SHAQ: Incorporating Shapley Value Theory into Multi-Agent Q-Learning. In Neural Information Processing Systems.

[128] Lei Wang, Chengbang Ma, Xueyang Feng, Zeyu Zhang, Hao ran Yang, Jingsen Zhang, Zhi-Yang Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji rong Wen. 2023. A Survey on Large Language Model based Autonomous Agents. ArXiv abs/2308.11432 (2023).

[129] Jiawen Wei, Hugues Turb'e, and Gianmarco Mengaldo. 2024. Revisiting the robustness of post-hoc interpretability methods. ArXiv abs/2407.19683 (2024).

[130] Yongyan Wen, Siyuan Li, Rongchang Zuo, Lei Yuan, Hangyu Mao, and Peng Liu. 2024. SkillTree: Explainable Skill-Based Deep Reinforcement Learning for Long-Horizon Control Tasks.

[131] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. In Conference on Empirical Methods in Natural Language Processing.

[132] Annie Wong, Thomas Bäck, Anna V. Kononova, and Aske Plaat. 2021. Multiagent Deep Reinforcement Learning: Challenges and Directions Towards Human-Like Approaches. ArXiv abs/2106.15691 (2021).

[133] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv:2308.08155 [cs.AI]

[134] Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, P. Abbeel, and Dale Schuurmans. 2023. Foundation Models for Decision Making: Problems, Methods, and Opportunities. ArXiv abs/2303.04129 (2023).

[135] Yaodong Yang, Rui Luo, Minne Li, M. Zhou, Weinan Zhang, and Jun Wang. 2018. Mean Field Multi-Agent Reinforcement Learning. ArXiv abs/1802.05438 (2018).

[136] Seul-Ki Yeom, Philipp Seegerer, Sebastian Lapuschkin, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. 2019. Pruning by Explaining: A Novel Criterion for Deep Neural Network Pruning. ArXiv abs/1912.08881 (2019).

[137] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre M. Bayen, and Yi Wu. 2021. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In Neural Information Processing Systems.

[138] Renos Zabounidis, Joseph Campbell, Simon Stepputtis, Dana Hughes, and Katia P. Sycara. 2023. Concept Learning for Interpretable Multi-Agent Reinforcement Learning. ArXiv abs/2302.12232 (2023).

[139] Matthew D. Zeiler and Rob Fergus. 2013. Visualizing and Understanding Convolutional Networks. ArXiv abs/1311.2901 (2013).

[140] Changxi Zhu, Mehdi M. Dastani, and Shihan Wang. 2022. A survey of multi-agent deep reinforcement learning with communication. Autonomous Agents and Multi-Agent Systems 38 (2022), 1–48.

[141] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405 [cs.LG]