Domain-Level Explainability – A Challenge for Creating Trust in Superhuman AI Strategies

Jonas Andrulis[1], **Ole Meyer**[2], **Grégory Schott**[1], **Samuel Weinbach**[1], **Volker Gruhn**[2]
[1]Aleph Alpha GmbH, Heidelberg, Germany
{firstname.lastname}@aleph-alpha.de
[2]University of Duisburg-Essen, Schützenbahn 70, 45127 Essen, Germany
{firstname.lastname}@uni-due.de

Abstract

For complex problems, intelligent systems based on Deep Reinforcement Learning (DRL) have demonstrated an impressive ability to learn solutions that can go beyond human capabilities. While this creates new opportunities for the development of assistance systems with ground-breaking functionalities, applying this technology to real-world problems carries significant risks and need therefore to be transparent and trustworthy. Superhuman strategies are non-intuitive and it is difficult to achieve trust in them, as reliable performance evaluation in real-world scenarios can be prohibitively complex. Explainable AI (XAI) has improved the transparency of modern AI systems through a variety of measures, however, research has not yet provided solutions enabling domain-level insights for expert users of DRL systems in strategic situations. In this position paper, we discuss the existence of superhuman DRL-based strategies, their properties, the requirements and challenges for applying them to real-world environments, and the need for explainability at the domain-level as a key ingredient to enable trust.

Introduction

Reinforcement Learning (RL) is a promising area of machine learning where the system learns from interacting with the environment and actions are reinforced based on reward values. The algorithms optimize the expected long-term reward and continuously improve their policies. Reward signals are available in many scenarios and come with lower costs than labels required for supervised methods or can be used when no clear labeling is possible (Feng et al. 2018). Because RL only specifies the problem to solve and not the solution, RL systems have the potential to achieve performance beyond that of domain experts. This superhuman capability makes them interesting in real-world applications.

In strategic problems, an agent has to achieve a long-term objective through a complex set of highly-significant actions. Such problems have been described as "dynamic, hostile, and smart" (Buro 2003) and share aspects of complexity with video games, such as: decision under uncertainty, spatial and temporal reasoning, and agent collaboration. Strategy games have been used to support real-world (military) training (Herz and Macedonia 2002) and have also proven ideal for the development of methods in the AI field.

In 2016, AlphaGo (Silver et al. 2016), a Deep Reinforcement Learning (DRL) algorithm, demonstrated a performance that surpassed that of the best human players of Go, a strategy game considered beyond the reach of traditional AI due to its prohibitively large branching factor. One year later, AlphaZero (Silver et al. 2017) proved that an AI can learn a superior Go strategy from the game's rules alone, without any expert input. Since then, DeepMind's AlphaStar (Vinyals et al. 2019) and OpenAI Five (Berner et al. 2019) have further advanced the capabilities of RL systems. This research did not only focus on existing games but introduced flexible game-like environments for general strategic AI research (Tian et al. 2017).

While the research on DRL-based strategies is steadily making progress for complex video games, its transfer to practical real-world applications, where DRL may have a potential for superhuman disruption, is lacking. Because a DRL agent's reasoning is not explicit but implicitly learned and evaluated – based on the current state and past experiences – and encoded in a Deep Neural Network, the explainability of actions an assistant system formulates is important for a human to understand and accept those actions. Strategies obtained this way cannot be represented with regular planning techniques and we argue new forms of explainability need to be developed for that purpose.

In this paper, we discuss the potential for superhuman strategies and possible challenges based on our combined experience from industrial and practical uses. From a combination of machine learning, trust, and innovation research we identify domain-level transparency as one of the core difficulty facing AI applications in order to leverage successful superhuman DRL solutions in strategic real-world scenarios and specify approaches to provide explainability.

Potential barriers and challenges

Strategies adopted by DRL agents are similar to disruptive innovation processes in business models (Christensen 2013), as they are able to provide superhuman solutions that challenge established structures and theories. Following the description of innovative business models (Chesbrough 2010),

Table 1: Explainability Challenges for Superhuman Strategies found by DRL Agents

Domain Level	**Strategic Complexity**	**C1**	**Spatial & temporal reasoning:** Actions are not only conditioned on the currently "visible" state but also on past and future states. Relevance of observations depends on both state and actions.
		C2	**Collaboration:** Interdependencies between actions of competitors and collaborators are essential in game-theory-like scenarios.
		C3	**Decision-making under uncertainty:** Incomplete and uncertain information plays a major role in detemining an optimal strategy.
		C4	**Resource management:** Short term resources must be allocated towards a long-term strategic goal. Measurable advantages may appear long after a decisive action, with results appearing disadvantageous in the mean time.
		C5	**Opponent modeling & learning:** Learning from experience and adapting to scenarios and opponents.
		C6	**Adversarial real-time planning:** Long-time planning may be required due to sparse reward signals.
		C7	**Huge action- & state-spaces:** Some environments have a number of variables, observations, possible actions, or rules that is much larger than in classical strategy games.
	Strategic Explainability	**E1**	**Future projections:** Analysis of (potential) future states, events, and competitor behavior.
		E2	**Hypotheticals scenarios:** Study of "what-if" scenarios on changes/hypothetical/potential future states.
		E3	**Risk, transparency & safety:** Risks due to "real" randomness or uncertain collaborator/competitor behavior.
		E4	**Uncertainty:** Simulation results have to indicate how well the model is likely to capture a given situation (e.g. detection of out-of-distribution cases).

we derive preconditions for potential disruptions based on DRL-AI:

- **Existence:** There may be no real-world strategic problem with challenges reaching the level of complexity of the games used in research (such as Starcraft and DOTA 2).
- **Potential:** Real-world strategic challenges may not provide rewards that warrant the use of DRL.
- **Trust:** Real-world applications may require a high level of trust that DRL methods do not currently provide.

In the rest of this paper, we examine each of these requirements based on existing use-cases, applications, and research examples. We discuss complexity and explainability challenges for superhuman strategies of DRL agents and provide an outlook on the critical questions to be addressed in order to enable the use of DRL applications in industrial practice, in particular by creating trust and enhancing explainability.

Existence of strategic real-world challenges

Strategic environments with high complexity and high uncertainty have been traditionally tackled with stochastic models or scenario planning (Schoemaker et al. 2004).

There is a long history of model development in game theory and operations research, which began during World War II and gained popularity during Cold War conflicts. Besides applications of game theory to conflict research (Slantchev 2017) and operations research (Forder 1998), market competition has gained a considerable importance (Moorthy 1993). The competitive nature of pricing strategy (Andrulis and Ender 2009) and logistics (Cachon and Netessine 2006) define problems in which strategy models have proven helpful in navigating these challenging environments.

The scope of the problems that can be addressed by these methods is, however, limited: in Table 1, we compile strategic complexity challenges that have been seen to apply to research and practical applications as will be discussed below. When met, those challenges prevent simplification without loosing the potential for superhuman discovery which require unique capabilities (labeled C1 to C7 in the table) of DRL agents that have been demonstrated in research publications.

For example, DOTA 2 (Berner et al. 2019) is a remarkable case that excels in all those criteria: in this incomplete information game (C3) teams of five compete against each other (C2) with limited information about the actions of the competitors (C5). Successful moves require coordination and planning (C1) forcing players to adapt and build ressources in a long-term resource building effort (C4). To win a match each agent plans up to 80.000 turns with each time up to 170.000 possible actions (C7) while there are no meaningful rewards until a match is either won or lost. OpenAI has been able to build agents that compete successfully against the world's best DOTA 2 players in real-time competition (C6).

Strategic problems with complexity beyond the limits of established strategy models and scenario planning can be found in real-world areas. There are criteria that suggest that DRL may be uniquely suited to address these challenges.

Potential for superhuman disruption

Modern DRL has demonstrated unexpected superhuman results, in several environments. This has even been the case in games that are well established and have received the massive attention of millions of players worldwide for decades or even centuries. DRL agents have been able to develop, against world's best human players, superior and previously

unknown approaches – which we classify as a superhuman disruption.

In Go, one great example for this is turn 37 of game 2 between AlphaGo (Silver et al. 2016) and Lee Sedol that resulted in the AI's victory (Holcomb et al. 2018).

> "During the games, AlphaGo played several inventive winning moves, several of which - including move 37 in game two - were so surprising that they upended hundreds of years of wisdom. Players of all levels have extensively examined these moves ever since."[1]

Similar results have been observed in DOTA 2 – a game that is massively more complex than Go – with one of the world's best players stating after losing against OpenAI Five:

> "It did things that we had never seen anybody else do and it has set a type of play style that we pretty much just copy now. When I see the bot make a play, it clicks in my head. I'm like, 'why aren't we doing that?'"[2]

Superhuman disruptive results are by definition difficult to plan for. However, we argue that experience with complex games has shown that we can expect a potential for superhuman disruption more often than not. The reason for the lack of superhuman DRL in real-world scenarios cannot reasonably be attributed to a lack of potential.

Trust in superhuman AI-strategies

Even if technical challenges can be solved, trust is essential in practical AI applications (Ferrario, Loi, and Viganò 2019), as safety requirements and threats can lead to economic costs, risks, and even regulatory issues. A human expert has to make the decision to delegate to the AI some aspect of importance in achieving a goal without the possibility to completely verify the AI's suggestion and all its potential implications (Grodzinsky et al. 2011).

Four key components have been shown to build trust in AI: tangibility, transparency, reliability and task characteristrics (Glikson and Woolley 2020). Of those four, transparency and reliability do not depend on the specific systems and tasks and can be the basis for general machine learning requirements: they correspond to the technically researched areas of robustness and explainability and define a trusted zone shown on Figure1.

Robustness: Robustness is a concept developed in control theory (Sastry and Bodson 2011), which is intended for dealing with the effects of uncertainties. This idea has been applied to machine learning models by measuring the impact of fluctuating inputs or environments such as uncertainties coming from modeling errors (Reinelt, Garulli, and Ljung 2002), poor generalization due to overfitting, or intentional adversial attacks. Robustness is an ongoing challenge for DRL. It has been very difficult to build generalizing DRL agents (Cobbe et al. 2019) with recent success only for simple environments (Badia et al. 2020) and, for some DRL models, even naively executed adversial attacks can have a significant impact on performance (Pattanaik et al. 2017). Seemingly unimportant changes in hyperparameters or the random seed can also produce drastically different results (Henderson et al. 2017). Given these issues with robustness, one way to satisfy safety concerns is through the integration of safety constraints (Junges et al. 2016; Cheng et al. 2019). Overall, research in this area is still active and currently represents a major challenge in the development of DRL systems.

Explainability: Current AI explainability methods can be divided into three categories: model explainability, outcome explainability and model inspection (Guidotti et al. 2018).

Model explainability approaches XAI by approximating the results of one model with a second model that is by design easier to understand – for example, by using decision trees (van der Waa et al. 2018; Johansson and Niklasson 2009; Craven and Shavlik 1996). However, even human-readable rules quickly become complex and incomprehensible, especially when they occur in large numbers (Lage et al. 2019).

Outcome explainability approaches XAI by illustrating the effects of different inputs on the outputs of the model while mainly ignoring what is going on within the model itself. The most common technique masks the actual input spotting the input-subset primarily responsible for the model result. The layer-wise relevance propagation approach (Bach et al. 2015) uses backpropagation. Creating attention maps over the input is another possibility (Xu et al. 2015; Fong and Vedaldi 2017). There are also experiments of model agnostic methods, such as Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh, and Guestrin 2016). This approach is particularly well suited for images, because in this case humans are able to process the large amount of information (pixel by pixel) quickly. In other areas, however, this becomes difficult because the ability to capture the information is not sufficiently available.

Model inspection directly analyzes on a technical level how model results are generated. Examples are sensitivity analysis (Saltelli 2002), the representation of dependencies between features and outputs (Friedman 2001) and Activation Maximization (AM) (Yosinski et al. 2015).

Some XAI methods have been specifically designed for DRL systems. Those can be divided into the dimensions time (intrinsic or post-hoc) and scope (local or global) (Puiutta and Veith 2020). Both approaches belong to the model explainability category. The instrinsic approach try to use interpretable function approximators for the policy function during training (Hein, Udluft, and Runkler 2018; Verma et al. 2018) and thereby likely limiting the potential for the DRL system to learn complex superhuman strategies in the first place. Post-hoc approaches use a simpler and more explainable model to analyze and provide explanations for the original successfully-trained complex model (Du, Liu, and Hu 2019). These are more common than intrinsic approaches (Liu et al. 2018; Madumal et al. 2019) however

[1] https://deepmind.com/research/case-studies/alphago-the-story-so-far

[2] William "BlitzDota" Lee on OpenAI Five playing DOTA 2.

it seems unlikely that superhuman strategies would be meaningfully preserved in a simplified explanation model.

In their review, the authors of (Puiutta and Veith 2020) find that XAI for reinforcement learning needs to exhibit context awareness by adapting to environment and user. One of the established and easily understandable ways to do this is to offer contrastive explanations comparing different strategy options. This kind of XAI output - which is especially useful for domain experts without any further AI knowledge - can be found in three of thirteen papers they reviewed: (Madumal et al. 2019; Sequeira and Gervasio 2019; van der Waa et al. 2018).

As key finding, they conclude that the ability to not only extract or generate explanations for the decisions of the model, but also to present this information in a way that is understandable by human (non-expert) users, makes it possible to predict the behaviour of a model. This definition of XAI implicitly assumes that the expert perfectly understands the strategic problem and can easily judge the right action. In a domain of high complexity and uncertainty, where intuitive judgment should not be trusted (Hogarth 2001), this cannot be easily expected and strategic explanations become a key challenge – making AlphaGo's turn 37 in game 2 predictable through XAI is a far greater challenge than the examples the authors had in mind.

Summary: Trust is an essential factor for the deployment and leveraging of DRL systems in real-world scenarios. Both current robustness and explainability methods are not suited for the requirements of complex strategic environments and constitute active areas where further research is required.

Domain-Level XAI

Considering the limits for XAI in strategy contexts, discussed in the last section, one may ask: What are the key ingredients for a 'strategic' XAI that will help human experts learn from superhuman AIs? This question cannot be answered on a technical level alone but must also address the strategic complexity (C1-C7) in a way that a domain expert with no machine learning mastery can use this information.

To achieve this, established tools for scenario planning (Schoemaker et al. 2004) can be adopted and translated into requirements for 'strategic explainability' in a DRL context (E1-E4 in Table 1). While scenario planning is limited and mostly qualitative, its approach to uncertainty (Courtney, Kirkland, and Viguerie 1997) and the criteria for scenario selection provide solid foundations for strategic explainability of DRL-AI results.

The main questions and drivers for scenarios are predictions of the future (E1) and the impact of changes in hypothetical scenarios (E2). While classical scenario planning has no way of quantify probability distributions, DRL adds this quantitative dimension with the potential to add a measure for risk (E3) and model uncertainty (E4).

Transparency of superhuman AI-strategies is essential and future research must further focus on domain-level explainability (E1-E4) in strategically complex environments

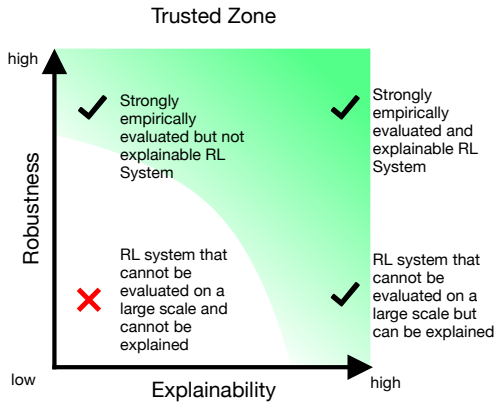

Figure 1: Trusted Zone for Strategic Real-World AI Systems

(C1-C7). Established approaches to strategy modeling, such as scenario planning, may be a great resource for building strategic AI-assistant systems and gaining the trust of experts in the potential for superhuman disruptions.

Conclusion & Future Work

There are real-world strategic use-cases that offer great potential for superhuman innovation through the use of DRL-AIs. Because the robustness of complex real-world strategies often cannot be empirically validated, trust in these systems must be build through transparency and explainability. Current XAI methods cannot offer the domain-level strategy explanations that are necessary for an expert to understand counter-intuitive superhuman strategies (Hogarth 2001). In order to build trust-enabling transparency into strategy AIs, the current concepts of explainability need to be enhanced. The implicitly learned strategic complexity requires an explainability that can address concepts beyond the relationship of individual input and output combinations for users without technical machine learning knowledge. Those need to be implemented as readily available tools that can be applied to AI agents. Finally, future studies will have to show that domain-level strategic explainability is possible so that human experts can trust and benefit from superhuman strategies issued by a DRL-AI in real-world applications.

References

Andrulis, J., and Ender, M. 2009. Strategic retail banking competition in distributed markets with varying switching costs. In *Quantitative Methods in Finance Conference*.

Bach et al. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10(7):e0130140.

Badia et al. 2020. Agent57: Outperforming the Atari Human Benchmark. *arXiv:2003.13350*.

Berner et al. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. *arXiv:1912.06680*.

Buro, M. 2003. Real-time strategy games: A new AI research challenge. In *IJCAI*, volume 2003, 1534–1535.

Cachon, G. P., and Netessine, S. 2006. Game theory in supply

chain analysis. In *Models, methods, and applications for innovative decision making*. INFORMS. 200–233.

Cheng et al. 2019. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3387–3395.

Chesbrough, H. 2010. Business model innovation: opportunities and barriers. *Long range planning* 43(2-3):354–363.

Christensen, C. M. 2013. *The innovator's dilemma: when new technologies cause great firms to fail*. Harvard Business Review Press.

Cobbe et al. 2019. Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning*, 1282–1289.

Courtney, H.; Kirkland, J.; and Viguerie, P. 1997. Strategy under uncertainty. *Harvard business review* 75(6):67–79.

Craven, M., and Shavlik, J. W. 1996. Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems*, 24–30.

Du, M.; Liu, N.; and Hu, X. 2019. Techniques for interpretable machine learning. *Communications of the ACM* 63(1):68–77.

Feng et al. 2018. Reinforcement learning for relation classification from noisy data. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ferrario, A.; Loi, M.; and Viganò, E. 2019. In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. *Philosophy and Technology*.

Fong, R. C., and Vedaldi, A. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, 3429–3437.

Forder, R. A. 1998. Military Operations Research: Quantitative Decision Making. *Journal of the Operational Research Society* 49(11):1227–1228.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232.

Glikson, E., and Woolley, A. W. 2020. Human trust in Artificial Intelligence: Review of empirical research. *Academy of Management Annals* (ja).

Grodzinsky et al. 2011. Developing artificial agents worthy of trust:"would you buy a used car from this artificial agent?". *Ethics and information technology* 13(1):17–27.

Guidotti et al. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51(5):1–42.

Hein, D.; Udluft, S.; and Runkler, T. A. 2018. Interpretable policies for reinforcement learning by genetic programming. *Engineering Applications of Artificial Intelligence* 76:158–169.

Henderson et al. 2017. Deep Reinforcement Learning that Matters. *arXiv:1709.06560*.

Herz, J., and Macedonia, M. R. 2002. Computer Games and the Military: Two Views. Center for Technology and National Security Policy, National Defense University.

Hogarth, R. M. 2001. *Educating intuition*. University of Chicago Press.

Holcomb et al. 2018. Overview on DeepMind and Its AlphaGo Zero AI. In *Proceedings of the 2018 International Conference on Big Data and Education*, ICBDE '18, 67–71. New York, NY, USA: Association for Computing Machinery.

Johansson, U., and Niklasson, L. 2009. Evolving decision trees using oracle guides. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*, 238–244. IEEE.

Junges et al. 2016. Safety-constrained reinforcement learning for MDPs. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, 130–146. Springer.

Lage et al. 2019. An evaluation of the human-interpretability of explanation. *arXiv:1902.00006*.

Liu et al. 2018. Toward interpretable deep reinforcement learning with linear model u-trees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 414–429. Springer.

Madumal et al. 2019. Explainable reinforcement learning through a causal lens. *arXiv:1905.10958*.

Moorthy, K. S. 1993. Competitive marketing strategies: Game-theoretic models. *Handbooks in operations research and management science* 5:143–190.

Pattanaik et al. 2017. Robust deep reinforcement learning with adversarial attacks. *arXiv:1712.03632*.

Puiutta, E., and Veith, E. 2020. Explainable Reinforcement Learning: A Survey. *arXiv:2005.06247*.

Reinelt, W.; Garulli, A.; and Ljung, L. 2002. Comparing different approaches to model error modeling in robust identification. *Automatica* 38(5):787–803.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144. ACM.

Saltelli, A. 2002. Sensitivity analysis for importance assessment. *Risk analysis* 22(3):579–590.

Sastry, S., and Bodson, M. 2011. *Adaptive Control: Stability, Convergence and Robustness*. Courier Corporation.

Schoemaker et al. 2004. Forecasting and scenario planning: the challenges of uncertainty and complexity. *Handbook of judgment and decision-making, eds., DJ Koehler and N. Harvey. Oxford, UK: Blackwell* 274–296.

Sequeira, P., and Gervasio, M. 2019. Interestingness Elements for Explainable Reinforcement Learning: Understanding Agents' Capabilities and Limitations. *arXiv:1912.09007*.

Silver et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587):484–489.

Silver et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550(7676):354–359.

Slantchev, B. L. 2017. On the Proper Use of Game-Theoretic Models in Conflict Studies. *Peace Economics, Peace Science and Public Policy* 23(4).

Tian et al. 2017. Elf: An extensive, lightweight and flexible research platform for real-time strategy games. In *Advances in Neural Information Processing Systems*, 2659–2669.

van der Waa et al. 2018. Contrastive explanations with local foil trees. *arXiv:1806.07470*.

Verma et al. 2018. Programmatically interpretable reinforcement learning. *arXiv:1804.02477*.

Vinyals et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575(7782):350–354.

Xu et al. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.

Yosinski et al. 2015. Understanding neural networks through deep visualization. *arXiv:1506.06579*.