ICGA Journal 46 (2024) 67–75
DOI 10.3233/ICG-240256
IOS Press

Chess and explainable AI

Yngvi Björnsson
Department of Computer Science, Reykjavik University

Abstract. Chess, once famously referred to as the *drosophila* of artificial intelligence (AI) research, has been a significant domain for developing intelligent AI agents capable of achieving super-human performance in domains previously dominated by humans. However, the emphasis on unceasingly improved playing strength has come at the cost of neglecting other fundamental aspects of intelligent agents, such as being capable of explaining the rationality behind their decisions in human-understandable terms. The need for such capabilities may be even more profound now than before, partly because such agents may be capable of learning novel concepts of interest to us humans, for example, as recently demonstrated in the game of chess. In this paper, we survey the state of explainable AI in chess-playing agents, arguing that chess may indeed hold a promise as an admissible domain for explainable AI.

Keywords: Explainable AI, chess, game playing

1. INTRODUCTION

Today's chess programs have reached a super-human level of play (Romstad et al. (2023); Pascutto et al. (2023); Silver et al. (2018)). The research effort into computer chess over the past few decades has first and foremost concentrated on developing new techniques and algorithms for improved gameplay while mostly neglecting other important artificial intelligence aspects, such as how to build intelligent chess tutoring systems. Contemporary chess engines are thus – albeit of a value for master-level players for assisting with their analysis and opening repository preparation – of a limited use for the regular club players that want to use them to improve their understanding of the game. Moreover, recent advancements in the field where deep neural-networks (DNNs) are used both for evaluating board positions and for action selection in the think-ahead process (e.g., in MCTS), such as in AlphaZero (Silver et al. (2018)), do seem to make the decision-making process even further non-transparent to humans, for example, by making it almost impossible in many situations to understand the rationality for why a particular game position is evaluated in favor of one player over the other, even for expert-level players.

The above-mentioned interpretability problem is not specific to chess engines, or other game-playing programs for that matter. As AI systems in diverse fields, such as healthcare and banking, become increasingly convoluted and ubiquitous, the need for humans to understand their decisions becomes increasingly crucial – not only to learn from the systems, but also due to concerns of correctness, ethics, and trust. This need has spurred a renewed interest in the field of *explainable AI*, that is, in designing of AI whose decisions can be better understood by humans.

In this paper,[1] we survey the state of explainable AI in contemporary chess agents, using recently proposed obligatory abilities of explainable agency as a yardstick. We also inspect the potentials of chess becoming an admissible domain for explainable AI, arguing that chess, once famously labeled

[1] A previous version of the paper was presented at the (non-archival) *IJCAI 2023 Workshop on Explainable Artificial Intelligence (XAI)*.

as the *drosophila* of AI research (but possibly side-tracked by over-emphasis on performance over understanding) may indeed hold such promise.

The rest of the paper is structured as follows: First, we give an up-to-date overview of desirable properties of explainable AI systems, then we review intelligent chess-playing agents from the perspective of explainable agency, both its current state and the potentials the domain holds for becoming an admissible test-bed, next we survey recent work on explainable game-playing agents and, finally, we conclude.

2. EXPLAINABLE AI

Explainable AI (Schwalbe and Finzel (2023)), also referred to as *interpretable* or *transparent AI*, concentrates on developing AI techniques that can explain the reasons behind their decisions in a way easily understandable to humans. This is in contrast to so-called *black-box AI* techniques, most notable artificial neural-networks (ANNs), that give answers without providing much insights into how a particular decision was reached. A related problem is when an AI system provides a detailed trace of its internal sub-decisions as an explanation to the human, but from a practical standpoint the amount of information is simply too excessive for the human to process and analyze, thus effectively rendering the explanation useless. This problem is typically referred to as *information overload*, *information explosion* or, more informally, *infobesity*.

Saliency maps (Kadir and Brady (2001)) visually represent how much a given input affects a model's output. They are commonly used in image classification to visualize relevant aspects of a given input image to a given classification, for example, highlighting areas of interest (Simonyan et al. (2014)).

The *Local Interpretable Model-Agnostic Explanations (LIME)* is a well-known framework for model interpretation (Ribeiro et al. (2016)). It interprets individual model predictions based on approximating the model around a given prediction using a local linear explanation model. More recently, *SHAP (SHapley Additive exPlanations)* was proposed as a unified framework for interpreting predictions, unifying several existing methods (including LIME) as well as presenting new ones (Lundberg and Lee (2017)).

One drawback of the methods mentioned above is that they mainly work on detecting the sensitivity of the network's predictions to the individual input parameters. In the case of DNNs, the input parameters are typically low-level features that are not necessarily meaningful for human-based interpretation (e.g., a single pixel in an image). Recently, concept-based explanation methods have shown promise, but they go beyond per-sample input features to explain higher-level human-friendly concepts across entire datasets. Testing with *Concept Activation Vectors (TCAV)* (Kim et al. (2018)) allows one to quantify how vital arbitrary user-defined (binary and non-binary) concepts are to neural-network predictions via model-probing (Alain and Bengio (2018)). The *Automatic Concept-Based Explanations (ACE)* method (Ghorbani et al. (2019)) further extends this line of concept-based explanations (in image recognition) by automatically discovering concepts, as opposed to them being human-provided.

In the context of intelligent autonomous agents, particularly in planning (Chakraborti et al. (2018)), the ability of an agent to explain the reasoning behind its decisions has been labeled *explainable agency* (Langley et al. (2017)), and which requires four distinct abilities: *(i)* the agent must be able to explain decisions made during plan generation, *(ii)* report which actions it executed at different levels of abstraction, *(iii)* show how actual events diverged from planned ones and what adaptions were necessary, and, finally, *(iv)* communicate its decisions and reasoning effectively in a formalism natural to humans. Furthermore, work on explainable agency in systems based on heuristic search

tends to distinguish between two types of self-explanations: *process* vs. *preference* oriented. The former emphasizes the (thought) process leading to finding the solutions, whereas the latter focuses on the solutions themselves without concerns about how they were found (Langley (2019)).

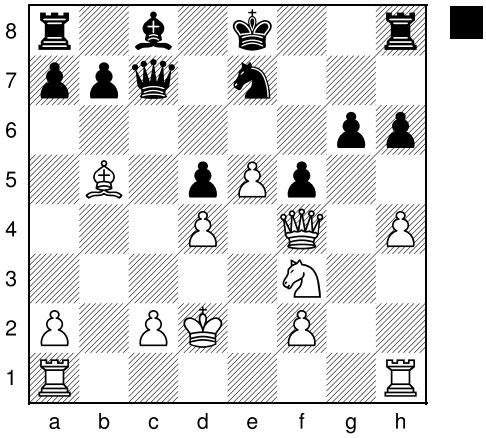

Fig. 1. From the game *Fischer vs. Rossolimo, U.S.A. Championship 1955-6*. Fisher's commentary after the move **17...♔f8?** (from (Fischer (2012))): on **17...♘c6** (if **17...♗d7 18 ♗×d7+ ♕×d7 19 e6!**) **18 ♗×c6+ b×c6** (**18...♕×c6** is again met by **19 e6! ♗×e6 20 ♘e5 ♕d6 21 ♘×g6 ♕×f4+ 22 ♘×f4** and the Knight beats the Bishop in the ending) **19 ♖hg1**, etc. Black's best change, however, is to try and reach sanctuary with **17...♔d8! 18 ♗d3 ♗e6**. White undoubtedly has the initiative, but it's hard to get at the King.

3. EXPLAINABLE CHESS AI

The state of explainable agency in contemporary chess-playing agents is weak, which may not be surprising given the lack of research focus. Although there exists some research literature on intelligent chess-tutoring systems (Guid et al. (2013); Sadikov et al. (2006)), it is sparse, and the reported work either restricted to simple chess endgames or is somewhat preliminary. When it comes to strong chess-playing agents (engines), their explanation capability is embarrassingly poor.

The sole explanations chess engines typically provide in a given board position are limited to the projected best continuation of play – the so-called *principal-variation (PV)* – and a single numeric score evaluating the merit of the board position at the tip of the PV. Although some engines can provide a somewhat more detailed explanation upon request, e.g., how different components such as material, mobility, pawn-structure, and king-safety contribute to the score, most engines do not offer such an option (and ANNs-based engines are inherently incapable of doing so).

Specifically, when measuring engines on the four properties of explainable agency, they fail on all accounts. For example, albeit showing a plan (the PV), no explanations are provided for why a given line of play is desirable, nor why it was chosen over other possible continuations. Also, if initially an alternative more promising plan is followed, but refuted by the opponent, the observer will never know. One could in some cases have the chess engines (programmatically) output more detailed search information, but without a proper human-like level of abstraction for reasoning and communicating one would soon fall prey to infobesity. This is in clear contrast to how an expert human would explain the rationality for her play, i.e., combining variations and higher-level goals, showing potential alternatives and the reasons for why they do or do not work, etc. An example thereof is illustrated in Fig. 1, for example, giving an example of why an initially promising alternative does not work (blocking

the check by moving a Knight to c6), and a long-term strategic evaluation (Knight is superiour to a Bishop in a given ending potentially arising).

It is also worth noting that in chess (and other abstract games), both the process and preference types of self-explanation are relevant, that is, a human player looking to improve would be interested in a chess-tutoring system explaining (at a human-friendly level of abstraction of course) both how a desirable game position (solution) could be reached and why it is so preferable. It is a non-trivial task to explain the reasoning process as the agent needs to show, in addition to the preferred continuation, what other candidate lines of play it considered, explain why they were consideration worthy, and why they ultimately proved inferior to the chosen line of play. This is different from some other problem domains, where only one or the other (process or preference) is relevant, as hypothesised in the original paper on the topic (Langley (2019)).

Although many abstract board games provide a good testbed for AI and XAI research because of their simple rules yet require non-trivial strategies to play well, chess offers additional benefits. Chess is a popular game in Western culture, and its following is increasing, making it more relatable as a problem domain than most other abstract board games. Most importantly, the vast amount of literature on chess provides prime examples of how chess decisions are best communicated and explained to humans, from beginners to experts, at different levels of detail. Not many problem domains offer the same benefit – in games or otherwise. Furthermore, the game is easy to scale in complexity, ranging from studying trivial endgames to complex middle-game patterns. For research, the ability to scale the domain difficulty is instrumental (some other abstract board games offer similar benefits but few to the same level). From a technical standpoint, there are also good arguments for using chess as a problem domain. First, the reasoning approaches used in top chess-playing agents are either minimax- or simulation-based (e.g., alpha-beta and MCTS search, respectively); this allows researchers to experiment with techniques in both those dominant heuristic-search paradigms. Second, many freely available open-source chess-playing agents are available, which provides researchers with a valuable head-start and a more objective way of evaluating their results (i.e., as opposed to using custom-built software).

Some earlier work defining desirable properties of machine learning systems took a holistic look, where learning performance is considered only one of several desirable criteria for evaluating the capabilities of such systems. For example, Michie (1988) defined three criteria for machine learning: *(i)* weak: the learning system improves its performance through experience; *(ii)* strong criterion: additionally, it can describe what it has learned in explicit symbolic form; *(iii)* ultra-strong criterion: additionally, the symbolic description is human-understandable and suited for improving the human's performance at the task. AlphaZero-style agents still have a long way to go to fulfill the last two criteria and, by this definition, are considered weak learners, as argued in (Bratko (2018)).

Finally, research into explainable AI is not orthogonal to other relevant AI research directions. An important aspect of explainable AI, as we have seen, is to be able to reason and communicate at an abstract level that is natural to humans. Internal representations used for that purpose can provide synergies with other types of reasoning and learning approaches. For example, when applying reinforcement learning in chess, sparse rewards and the absence of tangible sub-goals can make learning convergence slow. The aforementioned internal representations could potentially also be useful for expediting reinforcement learning, e.g., by introducing sub-goals (e.g., to mate the opponent's king in a given endgame, one must first restrain it to any side of the board, then fix it to an adjacent corner, and only then go for the mate). Active learning, where a learning system may ask humans only a limited number of queries to help expedite its learning process and require fewer labeled examples, is another example of where synergies may occur.

4. SURVEY OF RECENT WORK

Chess provides many exciting challenges for explainable AI. First, neural-network-based evaluation functions are gaining ground, often learning intricate concepts not immediately visible to humans (McGrath et al. (2022)). Thus, developing computational explainability methods for gauging into those networks to assist with analyzing the knowledge encoded there is valuable. Second, explainability research has focussed on model interpretability with little attention to the think-ahead process until recently, with the so-called explainable search (Baier and Kaisers (2020)). Third, the development of general chess-tutoring systems has been hampered, among other things, by a lack of research attention. Hopefully, advances in explainable AI will pave the road for more capable tutoring systems. We now survey recent work in chess (and relevant work in a few other abstract board games) along those three dimensions.

4.1. Explaining evaluations

Saliency maps are commonly used in image classification to visualize relevant aspects of a given input to the produced output. They have recently been adapted to help visualize and better understand game board evaluations. Fritz and Fürnkrans (2021) analyze the use of the *Specific and Relevant Feature Attribution (SARFA)* method in chess using different chess engines and pinpoint some of the pros and cons of such an approach, and propose some improvements to address identified shortcomings. Along similar lines, Pálsson and Björnsson (2022) evaluate the applicability and effectiveness of several saliency-map-based methods for explaining the evaluation of positions in the game of Breakthrough, demonstrating that the more applicable methods (like Shapley Value Sampling and LIME) provide valuable insights into the importance of game pieces and other domain-dependent knowledge learned by the model.

McGrath et al. (2022) analyze the knowledge acquired by AlphaZero in chess, drawing an analogy to chess concepts learned by humans, applying (linear) concept probes to the neural network and behavioral analysis of the agents opening play. The probing examination showed that many human-understandable chess concepts could be accurately regressed from the AlphaZero neural network both after and during training. Furthermore, a qualitative analysis of AlphaZero's play by GM Vladimir Kramnik shed light on how its chess knowledge (as judged by an expert observer) developed alongside its training. The work shows, for example, that the agent's opening knowledge undergoes a period of rapid development around the same time that many human-like concepts become predictable from network activations.

Using Stockfish (Romstad et al. (2023)), a world-class superhuman-strength chess-playing engine, as a testbed, Pálsson and Björnsson (2023) show how recent interpretability techniques, including surrogate models and concept probing, can illuminate human-understandable chess concepts learned by the engine's neural network. Furthermore, the work contrasts the state evaluations of the learned neural network to that of its counterpart hand-crafted evaluation model. They identify and explain critical differences in the game state assessments by doing so. For example, the neural network could statically detect threats such as forks, promotions, and attacking potentials, which would require a look-ahead search in the classical version of Stockfish. Also of interest was the low agreement on king-safety evaluation between hand-crafted and neural-network models – the neural network had seemingly discovered an alternative and more effective way of evaluating king safety.

Lovering et al. (2022) present a highly related work, albeit in the game of Hex. It uses model probing and behavioral tests to investigate how and what information is encoded in an AlphaZero-style agent

trained to play Hex. Their analyses suggest that the model neural network learned to represent and use concepts humans consider important for the game. However, they found gaps in embodied knowledge, such as dead cells and the lack of urgency to go for an imminent win. They also show that the training encodes short-term end-game-related concepts in the final layers of the network, whereas it encodes concepts related to long-term planning in the middle layers. They also show that MCTS typically discovers relevant concepts before the neural network learns to encode them. This partially resembles how expert human players learn: dynamic tactical motives, over time, become seen as statically detectable patterns.

Whereas the abovementioned work on concept probing checks only for the presence of pre-defined human-constructed concepts, Schut et al. (2023) take that approach a step further using the AlphaZero agent to discover new chess concepts, thus hopefully extending the scope of existing human chess knowledge. They first employ convex optimization to find suitable concept candidates, then filter out non-novel or non-teachable candidates based on spectral analysis and the notion of informativeness, respectively; finally, they validate the remaining candidates by presenting them to strong chess players and see if the new knowledge results in improved play (i.e., play better aligned with AlphaZero's move choices).

4.2. Explainable search

Explainable search is a recently emerging research direction targeted toward explaining the decisions of search-based agents in sequential decision-making domains, such as chess.

As of today, the most mature research work towards that goal is (arguably) in the field of autonomous planning, with some recent work making noteworthy headway towards explainable planning. Chakaborti et al. (Chakraborti et al. (2020)) provide a comprehensive survey of Explainable AI Planning (XAIP) and compare that to earlier efforts in the field in terms of techniques, target users, and delivery mechanisms, with a particular focus on the role of explanations in the design of an effective human-in-the-loop planning systems.

On more general notes, Baier and Kaisers (2020) highlight six research challenges relating to explainable search: *(i)* explanations as conversations; *(ii)* explanations as a two-way street; *(iii)* explanations in long-term interactions with users; *(iv)* explanation-aware search; *(v)* counterfactual explanations of search; and, finally, *(vi)* integrated explanations of search and evaluation. Whereas all the above challenges are admissible to explainable AI for chess, some seem more relevant than others, particularly integrating explanations of search and evaluation. Explanation-aware search is also exciting and may become essential to future intelligent chess-tutoring systems.

Finally, Baier and Kaisers (2021) make some initial steps towards addressing some of the abovementioned challenges as applied to MCTS-based agents. They also rightfully acknowledge the need for a more robust and flexible formalization of explainable search and the need for carefully constructed user studies to get informative feedback on how preferred and practical explanations of search should look in practice. In chess, we believe this need is already partially met by the existing rich chess literature on compelling explanations, thus further supporting the case of chess being an ideal and fitting domain for explainable AI research.

4.3. Chess tutoring systems

There is little to no recent work on building intelligent chess-tutoring systems, except for possibly *DecodeChess* (DecodeChess), a commercial online platform where one can submit chess games for

annotations. Their explanation approach is proprietary, and thus, little is known; however, they state that they use cutting-edge cognitive computing approaches to emulate abstract human thinking and look for and match relevant chess concepts to the positions at hand. The provided text explanations refer to known chess concepts. However, they often come across as elementary and superficial and far from being as insightful as annotations one is accustomed to in the chess literature. Although some older research on intelligent chess-tutoring systems exists, e.g. (Guid et al. (2013); Sadikov et al. (2006)), it is sparse, and the reported work is either restricted to simple chess endgames or is somewhat preliminary.

Another recent development worth mentioning uses natural-language generation techniques, including large language models, to provide online chess commentary, some in conjunction with symbolic reasoning (see, e.g., (Zang et al. (2019); Lee et al. (2022)) for such approaches as well as an overview of this recently emerging field).

Developing a fully general and robust tutoring system for chess is a lofty research goal, needing to address more or less all the abovelisted research challenges of explainable search. A more attainable yet challenging research goal would be to develop methods to fully annotate chess games with insightful human-like comments (like those in Fig. 1). For that, one would not need to be concerned with the interactive or two-way aspect of the explanation. Instead, one could concentrate on the tasks of integrating evaluation and think-ahead explanations, even altering them based on the expert level of the intended audience. [2]

5. CONCLUSION

Chess-playing agents are now playing at a level far exceeding even the strongest human grandmasters. However, there are still ample opportunities to further improve the intelligence of the chess agents – not by further improving their playing strength, but by improving their explainable agency.

We believe that, in this respect, chess will continue to serve as an important domain for AI research. For example, it offers exciting challenges and future research directions into XAI, including: (i) explaining evaluations in human-friendly manners for audiences of different levels of expertise (building on the rich existing chess literature); (ii) explaining the reasoning process in combination with the evaluations (in games the think-a-head process is also of importance when explaining decisions); (iii) discovering and explaining concepts yet not fully appreciated by humans (some of the best chess programs play at a super-human level), to name a few.

REFERENCES

Alain, G. & Bengio, Y. (2018). Understanding intermediate layers using linear classifier probes.

Baier, H. & Kaisers, M. (2020). Explainable search. In *IJCAI-PRICAI Workshop on Explainable Artificial Intelligence*.

Baier, H. & Kaisers, M. (2021). Towards explainable MCTS. In *AAAI Workshop on Explainable Agency in AI*.

[2]The International Computer Games Association (ICGA), formerly named the International Computer Chess Association (ICCA), used to hand out prizes to recognize chess systems providing good annotations; to further encourage such research, maybe it is time to re-instantiate that practice.

Bratko, I. (2018). AlphaZero – What's Missing? *Informatica (Slovenia)*, *42*.

Chakraborti, T., Kulkarni, A., Sreedharan, S., Smith, D.E. & Kambhampati, S. (2018). Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior. *CoRR*. arXiv:1811.09722.

Chakraborti, T., Sreedharan, S. & Kambhampati, S. (2020). *The Emerging Landscape of Explainable AI Planning and Decision Making*.

DecodeChess. Chess Analysis, Powered by AI. https://decodechess.com/.

Fischer, B. (2012). *My 60 Memorable Games: chess tactics, chess strategies with Bobby Fischer*. Batsford Chess. Pavilion Books.

Fritz, J. & Fürnkranz, J. (2021). Some chess-specific improvements for perturbation-based saliency maps. In *2021 IEEE Conference on Games (CoG)* (pp. 01–08). doi:10.1109/CoG52621.2021.9619015.

Ghorbani, A., Wexler, J., Zou, J.Y. & Kim, B. (2019). Towards automatic concept-based explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/77d2afcb31f6493e350fca61764efb9a-Paper.pdf.

Guid, M., Mozina, M., Bohak, C., Sadikov, A. & Bratko, I. (2013). Building an intelligent tutoring system for chess endgames. In O. Foley, M.T. Restivo, J.O. Uhomoibhi and M. Helfert (Eds.), *CSEDU 2013 – Proceedings of the 5th International Conference on Computer Supported Education*, Aachen, Germany, 6–8 May, 2013 (pp. 263–266). SciTePress.

Kadir, T. & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision*, *45*(2), 83–105. doi:10.1023/A:1012460413855.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F. & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In J. Dy and A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning*. Proceedings of Machine Learning Research (Vol. 80, pp. 2668–2677). PMLR. https://proceedings.mlr.press/v80/kim18d.html.

Langley, P. (2019). Varieties of explainable agency. In *ICAPS 2019 Workshop on Explainable AI Planning (XAIP)*.

Langley, P., Meadows, B., Sridharan, M. & Choi, D. (2017). Explainable agency for intelligent autonomous systems. In S.P. Singh and S. Markovitch (Eds.), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, California, USA, February 4–9, 2017 (pp. 4762–4764). AAAI Press.

Lee, A., Wu, D., Dinan, E. & Lewis, M. (2022). *Improving Chess Commentaries by Combining Language Models with Symbolic Reasoning Engines*.

Lovering, C., Forde, J., Konidaris, G., Pavlick, E. & Littman, M. (2022). Evaluation beyond task performance: Analyzing concepts in AlphaZero in Hex. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 25992–26006). Curran Associates, Inc, https://proceedings.neurips.cc/paper_files/paper/2022/file/a705747417d32ebf1916169e1a442274-Paper-Conference.pdf.

Lundberg, S.M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (Eds.), *Advances in*

Neural Information Processing Systems (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

McGrath, T., Kapishnikov, A., Tomašev, N., Pearce, A., Wattenberg, M., Hassabis, D., Kim, B., Paquet, U. & Kramnik, V. (2022). Acquisition of chess knowledge in AlphaZero. In *Proceedings of the National Academy of Sciences* (Vol. 119).

Michie, D. (1988). Machine learning in the next five years. In *Proc. Third European Working Session on Learning (EWSL)* (pp. 107–122).

Pálsson, A. & Björnsson, Y. (2022). Evaluating interpretability methods for DNNs in game-playing agents. In C. Browne, A. Kishimoto and J. Schaeffer (Eds.), *Advances in Computer Games* (pp. 71–81). Cham: Springer. doi:10.1007/978-3-031-11488-5_7.

Pálsson, A. & Björnsson, Y. (2023). Unveiling Concepts Learned by a World-Class Chess-Playing Agent. In *International Joint Conference on Artificial Intelligence (IJCAI23)*.

Pascutto, G.-C., Linscott, G., Lyaskuk, A. & Huizinga, F. (2023). Leela Chess Zero. http://lczero.org.

Ribeiro, M., Singh, S. & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 97–101). San Diego, California: Association for Computational Linguistics. https://aclanthology.org/N16-3020. doi:10.18653/v1/N16-3020.

Romstad, T., Costalba, M., Kiiski, J., Linscott, G., Nicolet, S., Geschwentner, S. & VandeVondele, J. (2023). Stockfish. https://stockfishchess.org.

Sadikov, A., Mozina, M., Guid, M., Krivec, J. & Bratko, I. (2006). Automated chess tutor. In H.J. van den Herik, P. Ciancarini and H.H.L.M. Donkers (Eds.), *Computers and Games*. Lecture Notes in Computer Science (Vol. 4630, pp. 13–25). Springer. doi:10.1007/978-3-540-75538-8_2.

Schut, L., Tomasev, N., McGrath, T., Hassabis, D., Paquet, U. & Kim, B. (2023). Bridging the Human-AI Knowledge Gap: Concept Discovery and Transfer in AlphaZero. *CoRR*. arXiv:2310.16410. doi:10.48550/ARXIV.2310.16410.

Schwalbe, G. & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*. doi:10.1007/s10618-022-00867-8.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, *362*(6419), 1140–1144. doi:10.1126/science.aar6404.

Simonyan, K., Vedaldi, A. & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*.

Zang, H., Yu, Z. & Wan, X. (2019). Automated chess commentator powered by neural chess engine. In A. Korhonen, D.R. Traum and L. Màrquez (Eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, Florence, Italy, July 28–August 2, 2019, Volume 1: Long Papers (pp. 5952–5961). Association for Computational Linguistics. doi:10.18653/v1/p19-1597.