Explainable Search: An Exploratory Study in SameGame

Explainable Search: An Exploratory Study in SameGame

Chiara F. Sironi, Anna Wilbik, Mark H. M. Winands
Department of Advanced Computing Sciences, Maastricht University
Maastricht, The Netherlands
{c.sironi, a.wilbik, m.winands}@maastrichtuniversity.nl

***Abstract*—The field of Explainable Artificial Intelligence has gained popularity in recent years, due to the need for users to understand AI-made decisions, in order to increase their trust in the AI system. However, not much work has been performed on explaining recommendations made by search algorithms, which do not focus on single decisions, but on complex plans of action. This paper investigates promising directions for research in Explainable Search (XS), by evaluating with a user study different types of explanations for a search-based algorithm. Preliminary results suggest that users prefer explanations generated using context-based features, which are not only based on the current state of the problem, but are extracted from different parts of the tree generated by the search algorithm.**

***Index Terms*—Explainable Search, search algorithms, games**

I. Introduction

Recently, Artificial Intelligence (AI) has found its way in many practical applications and society increasingly relies on it to address a wide variety of decision-making problems. As a consequence, the need for users to understand decisions made by an AI system has gained more attention. This has prompted the development of the research field of Explainable Artificial Intelligence (XAI) [1], which aims at creating intelligent agents that are able to explain their decisions to a user.

Most of the research efforts made so far in XAI have focused on explaining (black-box) machine-learning models, mainly addressing explanations of single decisions [2], [3]. Some work on explaining entire decision policies has been performed in the field of explainable reinforcement learning (XRL) [4]. However, not much work has been performed to explain decisions made by intelligent search techniques, except attempts at visualizing trees generated by the algorithms [5]. Search techniques do not focus on single, independent decisions, but on generating plans of action, taking into account complex trees of expected contingencies and eventualities.

Baier and Kaisers [6] have already highlighted the relevance of the sub-field of Explainable Search (XS). Currently, we are facing the need to plan ahead in increasingly more complex settings where more options can or should be investigated, for instance in logistics [7], healthcare [8] and structural engineering [9]. Thus, we need to fill the gap between results of intelligent search techniques and human understanding.

979-8-3503-2277-4/23/$31.00 ©2023 IEEE

Research in XS comes with many challenges, such as how to generate proper explanations without hindering the performance of the search algorithms, which features are relevant to generate such explanations and how should information be combined into the explanations. The work presented in this paper serves as an exploratory study that investigates promising directions in the field of XS. More precisely, we evaluate which characteristics and what kind of information are more suitable to generate understandable and useful explanation of search-based sequential decision making. For this study, we use the SameGame puzzle [10] as test domain, and we generate explanations for the decisions of the Monte-Carlo tree search (MCTS) algorithm [11].

The remainder of the paper is structured as follows. Section II describes the considered types of explanations and Section III covers the methodology followed in this study to evaluate them by means of a user survey. Results of the user survey are discussed in Section IV and Section V gives the conclusions of the study and recommendations for future work.

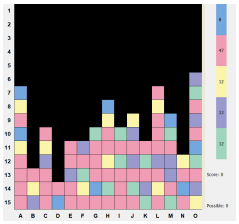
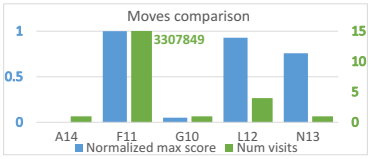
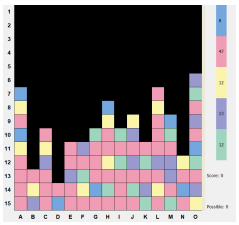
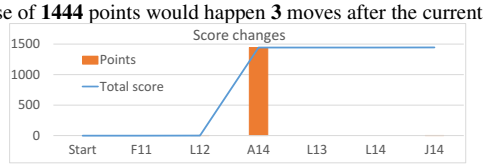
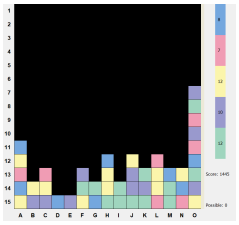
II. Envisioned types of explanations

This work focuses on evaluating post-hoc explanations of decisions made by a search algorithm for specific instances of the problem (i.e. states of the game). We do not aim at explaining the entire policy used by the search algorithm. Moreover, for each problem instance, two types of explanations are considered: *"Why?"-explanations* provide the reason why an action is recommended by the algorithm, and *"Why not"-explanations* provide the reason why an action is not recommended.

Explanations are generated using different types of features and are categorized based on their *domain dependence*:
Statistics: features based on the information collected in the search tree generated by the search algorithm.
Context: features based on the characteristics of the state.
Statistics + context: all features from the previous two categories.

Next, we distinguish features depending on the *tree scope* from where they are extracted, namely:
Flat: features extracted only from the state in the tree that corresponds to the action (decision) that we want to explain.
Siblings: features that are extracted from the states in the tree that correspond to the action (decision) that we want to explain and to its siblings (i.e. its alternatives in the state).

TABLE I: Example of "Why?"-explanations according to their dimensions

Tree scope		Feature domain: Statistics	Feature domain: Context
	Flat	The recommended move is **F11** because it has the highest maximum score of **1481**. The algorithm has evaluated this move **3307849** times.	The recommended move is **F11**. The figure shows the state reached by playing this move. This state is good because of the following reasons: (i) the number of smaller clusters of size **10** is decreased from **1** to **0**, (ii) the number of larger clusters of size **37** is increased from **0** to **1**, (iii) in this state, the number of new connections created between blocks of the same color is higher than the number of connections that have been destroyed (**2** newly created connections vs **1** destroyed connections). 
	Siblings	The recommended move is **F11** because the maximum score of **1481** that can be achieved with this move is better than the ones of its siblings. See the figure for a comparison of the statistics of all the moves in this state. 	The recommended move is **F11**. The figure shows the state reached by playing this move. This move is better than its siblings because of the following reasons: (i) the resulting board has one of the biggest clusters, (ii) the resulting board has bigger clusters on average, (iii) the resulting board has higher columns on average. 
	Sequence	The recommended move is **F11** because you can reach the end of the game with the best score found during the search, which will give you a difference of **1481** points with respect to the current state. The figure shows how the score would change over the next **6** moves, if you would follow the path recommended by the algorithm after playing move **F11**. In particular, the highest score increase of **1444** points would happen **3** moves after the current state. 	The recommended move is **F11** because you can reach the good future state in the figure by performing the following sequence of moves: **[F11, L12, A14]**. This state is good when compared to the current state for the following reasons: (i) the latest removed cluster, which has size **40** is one of the largest in the state, (ii) in this state, the number of new connections created between blocks of the same color is higher than the number of connections that have been destroyed (**12** newly created connections vs **0** destroyed connections). 

Sequence: features that are extracted from the states on the principal variation of the sub-tree of the considered move.

Combining the categories from the feature domain and from the tree scope, we obtain 9 possible types of explanations, one for each combination of categories. Consider the game board in Fig. 1 for *SameGame*, a puzzle where the aim is to empty a board full of colored blocks by eliminating groups of blocks of the same color. Table I gives an example of a "Why?"-explanation for the search algorithm recommending move F11 in this game position for all the categories. Note that explanations that use features in the "Statistics + context" category are obtained by simply combining the explanations in the "Statistics" and "Context" columns of the table. Similar explanations can be generated for moves that are not recommended (i.e. "Why not?"-explanations).

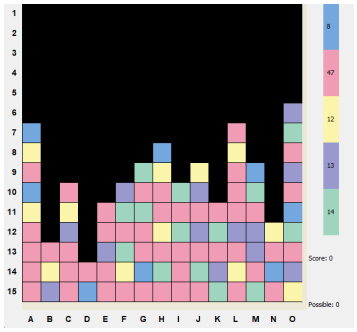


Fig. 1: Intermediate board of SameGame

III. Methodology

A. Experimental design

This work follows an experimental design methodology. The questionnaire designed for this study is guided by the following six key design decisions:

1) Target users: The users targeted in this study are university students and university employees, which are categorized according to two different criteria. First, we distinguish between *AI novices* and *AI experts*, and second, we distinguish between *domain novices* and *domain experts*. To identify which type of users responded to the survey, participants are asked to indicate (i) how familiar they are with AI methods and (ii) how familiar they are with the considered problem domain (SameGame) on a 5-point Likert scale.

2) Use case: As test domain for which explanations are generated, we chose to use the SameGame puzzle [10]. This game seems a suitable use case for this exploratory work because it is quite fast for a user to learn how to play, yet

TABLE II: Board position assignment per explanation type.

		Feature domain		
		Statistics	**Context**	**Statistics+Context**
Tree scope	**Flat**	Game1-initial	Game3-end	Game2-mid
	Siblings	Game2-end	Game1-mid	Game3-initial
	Sequence	Game3-mid	Game2-initial	Game1-end

it presents a certain degree of complexity and requires non-trivial play strategies. SameGame is played on a rectangular board (15×15 in this study), which is initially randomly filled with blocks of 5 colors. The goal of the game is to remove from the board as many clusters of blocks (of size ≥ 2) as possible. Blocks that are no longer supported will fall down, and columns will shift to the left when possible. For each removed group of size n, points are awarded according to the following formula: $(n-2)^2$.

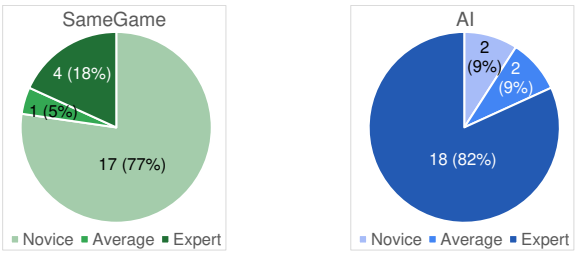
3) XS method: Although not much work has been performed on generating explanations for search-based models, we refer to the terminology in [1] to frame our approach. We generate *post-hoc*, *model-specific* explanations for search algorithms. They are based on pre-determined templates that are filled and completed using the values of the features extracted from the output (i.e. search tree and game states) obtained by applying the search algorithm on a given instance (board) of the problem. As an example, each explanation given in Table I is a different template, where the text in bold is filled using the features extracted from the output of the application of the search algorithm on the board in Fig. 1. Note that, for context-based explanations, the reasons listed to explain why a state is good are sub-templates that are activated depending on the values of the corresponding features [12].

4) Search algorithm: In this study, we chose to focus on explaining recommendations of the MCTS algorithm [11]. This algorithm has been chosen not only because it has been successfully applied to many problems, but also because it is a challenging algorithm to generate explanations for. MCTS is a simulation-based search algorithm that incrementally builds a tree representation of the state space of the problem. To handle large state spaces, it is selective and tends to focus on the parts of the tree that are most promising, while trying to balance exploitation of good moves with exploration of less visited ones. The MCTS implementation used in this study is the one of Schadd et al. [10], which is optimized for SameGame. More precisely, it has a high exploitation rate and the simulation strategy focuses on creating large groups of the same color, in order to achieve more points when they are removed.

5) Evaluation criteria: To evaluate the generated explanations we use two of the metrics that [13] identified as key attributes of a satisfying explanation:

- *Understandability*: measures if the user understands the explanation.
- *Usefulness*: measures if the user finds the explanation useful to make better decisions or to perform an action.

The perceived fulfilment of each criteria is measured by a user's agreement or disagreement indicated through his or her rating on a 5-point Likert scale.



(a) Expertise in SameGame (b) Expertise in AI

Fig. 2: User characteristics

6) Questionnaire design: The designed questionnaire is composed of eleven individual sections. The first section asks users to indicate their familiarity with SameGame and AI. Nine sections focus each on one of the 9 types of explanations described in Section II. To generate these explanations, we considered the first 3 games in a standardized test set of 20 games.[1] For each of these 3 games, we extracted a board position at the start of the game (*initial* board), one after 25 moves (about *mid*-game) and one after 50 moves (about *end*-game). On each of these 9 board positions, we performed a run of MCTS for 20 seconds and used the resulting tree to extract the features and generate the 9 explanations types with the assignment shown in Table II. For each type, we generated a "Why?"-explanation for the recommended move and a "Why not?"-explanation for one of the other moves that are available in the board position, but not recommended by MCTS. The remaining section focuses on a deceiving explanation generated manually, in which the recommended move is motivated by negative reasons, while the non-recommended move is motivated by positive reasons. The targeted time to complete the questionnaire was 15 minutes.

B. Survey process

The survey was implemented in Google Forms and distributed through multiple channels targeting university employees and students mainly, but not exclusively, with a background in AI and/or games. The survey was conducted within one week during which 22 target users participated.

IV. Results

To analyze the background of the respondents to the survey, we categorized them depending on whether they are *experts* (> 3), *average* (3) or *novices* (< 3) in SameGame and in AI. Fig. 2 shows the percentages for each category of users that responded to the survey, from which we can see that we have a majority of SameGame novices and AI experts.

Tables III and IV show the results of the survey. For each of the 9 explanation types, the tables report the mean score and standard deviation over all the respondents of the understandability and usefulness criterion, respectively. Moreover, the mean scores and standard deviations are given in both tables also for all the categories of each dimension separately.

[1]The games can be found at http://www.js-games.de/eng/games/samegame

TABLE III: Results for the understandability criterion

		Feature domain			
		Statistics	**Context**	**Statistics+Context**	**Total**
Tree scope	**Flat**	3.55 ± 1.10	3.82 ± 0.66	3.59 ± 1.01	3.65 ± 0.94
	Siblings	3.64 ± 1.09	$\mathbf{4.23} \pm 0.53$	4.00 ± 0.87	$\mathbf{3.95} \pm 0.88$
	Sequence	3.27 ± 0.88	4.14 ± 0.77	3.50 ± 1.06	3.64 ± 0.97
Total		3.48 ± 1.03	$\mathbf{4.06} \pm 0.68$	3.70 ± 0.99	

TABLE IV: Results for the usefulness criterion

		Feature domain			
		Statistics	**Context**	**Statistics+Context**	**Total**
Tree scope	**Flat**	2.05 ± 0.84	3.59 ± 0.67	2.95 ± 0.84	2.86 ± 1.01
	Siblings	2.77 ± 1.23	3.86 ± 0.77	$\mathbf{4.00} \pm 1.02$	$\mathbf{3.55} \pm 1.15$
	Sequence	2.59 ± 0.91	3.91 ± 0.81	3.55 ± 1.18	3.35 ± 1.12
Total		2.47 ± 1.04	$\mathbf{3.79} \pm 0.75$	3.50 ± 1.10	

First, we can see positive results for the understandability metric, for which each explanation type has scored quite high on average (around at least 3.5). Moreover, there seems to be no big difference between all the scores, indicating that the designed templates might be a suitable presentation of the considered features as explanations to the user.

More difference in the mean scores can be seen for the usefulness metric. The explanation type that scored the highest on average is based on statistics and context features extracted from all the siblings of the considered move, but also explanations based on context features extracted from the sequence or from the siblings have a quite close score. In general, it seems that the explanations that are perceived as the least useful are the ones with either a flat tree scope, or statistic-based features. This seems to be an indication that these types of explanations do not provide sufficient information to the users.

Explanations that the users perceived as most useful seem to be the ones with context-based features. This could be because they are more closely related to how humans would analyze the game (i.e. reasoning about structures in the state, like cluster size, connections, and colors, rather than calculating statistics on the expected maximum score for each available move).

It would also be reasonable to think that combining both statistics and context features into an explanation could provide more useful information about the inner workings of the algorithm. However, users have rated such explanations generally lower, especially when features are extracted from a sequence of moves in the tree. This could be due to an information overload, which has also been indicated by some participants as open feedback at the end of the survey.

Information overload seems to also be the reason for the scores of the deceiving explanation being higher than expected (i.e. 3.55 ± 0.91 for understandability and 3.45 ± 0.96 for usefulness). Being presented with a long explanation based both on statistics and context features might have given users the idea of a complete and rich explanation, while distracting them from checking whether the presented reasons were consistent with what they were observing.

V. Conclusion and Future Work

This paper presented an exploratory study in the newly developed field of Explainable Search. Different types of explanations have been evaluated for Monte-Carlo tree search decisions in the SameGame puzzle. Results show that a template-based approach is a suitable starting point to generate understandable explanations. Moreover, to generate useful explanations we may recommend to use context-based features and extract features from different parts of the tree generated by the search algorithm, instead of limiting the scope only to features of the current state.

Results presented in this study are encouraging and support the need for further research on XS. First of all, a more extensive study with more users with varying levels of expertise should be performed. In addition, we should evaluated "Why?"- and "Why not?"-explanations separately. As remarked by [6], selective algorithms like MCTS might not visit sub-optimal moves enough to generate reliable explanations for why they are not recommended. The trade-off between quality of the search and quality of "Why not?"-explanations should be object of future research.

References

[1] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[2] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[3] C. Wrede, M. H. M. Winands, and A. Wilbik, "Linguistic summaries as explanation mechanism for classification problems," in *The 34th Benelux Conference on Artificial Intelligence and the 31th Belgian Dutch Conference on Machine Learning*, 2022.

[4] E. Puiutta and E. Veith, "Explainable Reinforcement Learning: A survey," in *Machine Learning and Knowledge Extraction. CD-MAKE 2020*, ser. LNCS, vol. 1227. Springer, 2020, pp. 77–95.

[5] R. Coulom, "Treemaps for search-tree visualization," in *The Seventh Computer Olympiad Computer-Games Workshop Proceedings*, J. W. H. M. Uiterwijk, Ed. Maastricht, The Netherlands: Maastricht University, 2002.

[6] H. Baier and M. Kaisers, "Explainable search," in *2020 IJCAI-PRICAI Workshop on Explainable Artificial Intelligence*, 2020.

[7] S. Edelkamp, M. Gath, C. Greulich, M. Humann, O. Herzog, and M. Lawo, "Monte-Carlo tree search for logistics," in *Commercial Transport*. Springer International Publishing, 2016, pp. 427–440.

[8] G. Zhu, D. Lizotte, and J. Hoey, "Scalable approximate policies for Markov decision process models of hospital elective admissions," *Artificial Intelligence in Medicine*, vol. 61, no. 1, pp. 21—34, 2014.

[9] L. Rossi, M. H. M. Winands, and C. Butenweg, "Monte Carlo tree search as an intelligent search tool in structural design problems," *Engineering with Computers*, vol. 38, p. 3219–3236, 2022.

[10] M. P. D. Schadd, M. H. M. Winands, M. J. W. Tak, and J. W. H. M. Uiterwijk, "Single-player Monte-Carlo Tree Search for SameGame," *Knowledge-Based Systems*, vol. 34, pp. 3–11, 2012.

[11] R. Coulom, "Efficient selectivity and backup operators in Monte-Carlo Tree Search," in *Computers and Games (CG 2006)*, ser. LNCS, H. J. van den Herik, P. Ciancarini, and H. H. L. M. Donkers, Eds., vol. 4630. Berlin Heidelberg, Germany: Springer-Verlag, 2007, pp. 72–83.

[12] B. Hornig, E. Doe, L. Padolevicius, and X. Wigman, "Explainable search for SameGame," Maastricht University, Tech. Rep., 2022.

[13] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.