Explainable AI Planning (XAIP): Overview and the Case of Contrastive Explanation (Extended Abstract)

Jörg Hoffmann[1(✉)] and Daniele Magazzeni[2]

[1] Saarland Informatics Campus, Saarland University, Saarbrücken, Germany
hoffmann@cs.uni-saarland.de
[2] King's College London, London, UK
daniele.magazzeni@kcl.ac.uk

Abstract. Model-based approaches to AI are well suited to explainability in principle, given the explicit nature of their world knowledge and of the reasoning performed to take decisions. AI Planning in particular is relevant in this context as a generic approach to action-decision problems. Indeed, explainable AI Planning (XAIP) has received interest since more than a decade, and has been taking up speed recently along with the general trend to explainable AI. In the lecture, we provide an overview, categorizing and illustrating the different kinds of explanation relevant in AI Planning; and we outline recent works on one particular kind of XAIP, contrastive explanation. This extended abstract gives a brief summary of the lecture, with some literature pointers. We emphasize that completeness is neither claimed nor intended; the abstract may serve as a brief primer with literature entry points.

1 Explainable AI Planning: Overview

The need for explainable AI (XAI) first became prominent in Machine Learning, where the lack of understandable decision rationales is particularly daunting. Model-based techniques are fundamentally better suited to providing explanations, yet their explainability has traditionally not received much interest. This has changed with the XAI trend. In particular, research on explainable AI planning (XAIP) has received increasing interest in recent years. One culminating point of this trend is the nascent series of XAIP workshops[1] at the International Conference on Automated Planning and Scheduling (ICAPS).

As is natural for a nascent area, at this time the XAIP landscape is still in the making. XAIP has attracted interest from researchers with widely different backgrounds and points of view, and it is too early to give a conclusive systematization into sub-topics and issues of interest. A roadmap for XAIP was

[1] See the 2019 edition at https://kcl-planning.github.io/XAIP-Workshops/ICAPS_2019.

M. Krötzsch and D. Stepanova (Eds.): Reasoning Web 2019, LNCS 11810, pp. 277–282, 2019.
https://doi.org/10.1007/978-3-030-31423-1_9

proposed by Fox et al. [10], categorizations have been attempted [16], and a systematization of possible objectives has just been published [6]. XAIP includes topics ranging from epistemic logic to machine learning, and techniques ranging from domain analysis to plan generation and goal recognition. Nevertheless, some major themes have emerged, that we refer to here as *plan explanation*, *contrastive explanation*, *human factors*, and *model reconciliation*.

Plan explanation is the oldest branch of XAIP. It aims at helping humans to understand the inner workings of a plan suggested by the system (e. g., [1,2,13,17,20,24,27]). This involves, in particular, the transformation of planner output into forms that humans can easily understand; the description of causal and temporal relations between individual plan steps; and the design of interfaces, in particular suitable dialogue systems, supporting human interaction and understanding.

In contrastive explanation, the aim is to answer user questions of the kind "Why do you suggest to do A here? (rather than B which seems more appropriate to me)". This is a frequent form of question as highlighted by a recent analysis [19] of lessons to be learned for XAI from social sciences. Answers to such questions take the form of reasons why A is preferable over B. Contrastive explanation is the major focus of this lecture, so we discuss it in more detail in Sects. 2 and 3.

Human factors research naturally has to be a major component of XAIP, whose ultimate aim is to communicate with human users. Manuela Veloso and her team investigate verbalizations describing the robot experience and intentions to human users [22]. Other work [29] focuses on a human's interpretation of plans. Learning is used to create a model of the interpretation, which is then used to measure the explicability and predictability of plans. A recent proposal is to combine cognitive measures with epistemic planning [21]. Many works, also ones cited here as belonging to other themes, include human factors research to varying degrees.

In *model reconciliation*, the focus is on the agent vs. human having different world models. The explanation must then identify and reconcile the relevant model differences. This has been intensively investigated in the last years [4, 15,23,28], with mature results and outreach to the robotics [7] and multi-agent communities [12].

There are of course various works on XAIP, or relating to XAIP, that do not fit into this categorization. To name but a few examples: Göbelbecker et al. [11] proposed a framework for "excuses", which can be viewed as explanations why a planning task is unsolvable; Smith [25] put forward the challenge of *planning as an iterative process*, which among others requires explanation facilities; and some work has considered particular forms of communication like lying [5].

2 Contrastive Explanations

As mentioned, an important type of questions in Explainable Planning are *contrastive* questions, of the form "Why action A instead of action B?". These questions arise when the planner is suggesting something different from what

the user would expect. In such a scenario, one way to address this type of question is to allow the user to compare the plan suggested by the planner with what she/he was expecting. These are contrastive explanations that can highlight the differences between the decisions that have been made by the planner and what the user would expect, as well as to provide further insight into the model and the planning process. A detailed analysis of contrastive explanations in AI has been proposed by Tim Miller in [18].

Some recent work introduced contrastive explanations for Explainable Planning. In particular, in [14] contrastive questions are compiled into constraints that form a hypothetical model. Such a hypothetical model can be used to generate the hypothetical plan that the user would expect and from here the contrastive explanation can be presented to the user. The work focuses on temporal planning and presents domain-independent compilations.

Another related line of work focuses on providing contrastive explanations *as a service* [3]. Here the idea is to create a wrapper around an existing planner and use automatic compilations of user questions into models. In this way, the explanations are generated using the same planner already used by the user, and this increases the user confidence in the explanations provided.

In the lecture we give an overview of recent progress on using contrastive explanations for Explainable Planning.

3 Contrastive Explanation of Plan Space Through Plan-Space Dependencies

We finally consider a line of work, conducted by the authors, starting from the idea to answer questions "Why does the plan π start with action A rather than B?" by generating a new plan π' starting with B and highlighting undesirable properties of π'. A weakness of this approach is that there may be differences between π and π' unrelated to the use of A vs. B. Many comparison aspects (e. g. which other actions are used, or which "soft" objectives are satisfied) may be affected by arbitrary decisions in the planner's search. Therefore, the idea is to replace the *existential* answer generating a single alternative plan π' with a *universal* answer pertaining to *all* possible such alternatives.

This can be done at the level of *plan properties*: Boolean functions on plans that capture aspects of plans the user cares about (whether or not the plan starts with a particular action, whether or not a particular soft objective is satisfied, etc). Given a set of plan properties, one can determine dependencies across these properties, i. e., plan-space entailments: a plan property p entails another property p' if every plan that satisfies p also satisfies p'. A user question "Why does the current plan π satisfy p rather than q?" can then be answered in terms of the properties q' not true in π but entailed by q: things that will necessarily change when satisfying q.

We put forward, and explain in the lecture, a generic framework for this kind of analysis, as well as an instantiation and experiments in the context of oversubscription planning [8,26] where resources are insufficient to achieve all

goals, and plan properties of obvious interest are those goals achieved by a plan. A first paper on this approach is published at XAIP'19 and serves as a reference for the reader interested in details [9].

Acknowledgments. This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-18-1-0245. Jörg Hoffmann's research group has received support by DFG grant 389792660 as part of TRR 248 (see https://perspicuous-computing.science). Daniele Magazzeni's research group has received support by EPSRC grant EP/R033722/1: Trust in Human-Machine Partnerships.

References

1. Bercher, P., et al.: Plan, repair, execute, explain - how planning helps to assemble your home theater. In: Chien, S., Do, M., Fern, A., Ruml, W. (eds.) Proceedings of the 24th International Conference on Automated Planning and Scheduling (ICAPS 2014). AAAI Press (2014)
2. Bidot, J., Biundo, S., Heinroth, T., Minker, W., Nothdurft, F., Schattenberg, B.: Verbal plan explanations for hybrid planning. In: Proceedings MKWI (2010)
3. Cashmore, M., Collins, A., Krarup, B., Krivic, S., Magazzeni, D., Smith, D.: Explainable planning as a service. In: ICAPS-19 Workshop on Explainable Planning (2019)
4. Chakraborti, T., Sreedharan, S., Zhang, Y., Kambhampati, S.: Plan explanations as model reconciliation: moving beyond explanation as soliloquy. In: Sierra, C. (ed.) Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017). AAAI Press/IJCAI (2017)
5. Chakraborti, T., Kambhampati, S.: (how) can ai bots lie? In: Proceedings of the 2nd Workshop on Explainable Planning (XAIP 2019) (2019)
6. Chakraborti, T., Kulkarni, A., Sreedharan, S., Smith, D.E., Kambhampati, S.: Explicability? Legibility? Predictability? Transparency? Privacy? Security? The emerging landscape of interpretable agent behavior. In: Proceedings of the 29th International Conference on Automated Planning and Scheduling (ICAPS 2019). AAAI Press (2019)
7. Chakraborti, T., Sreedharan, S., Grover, S., Kambhampati, S.: Plan explanations as model reconciliation. In: Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2019), pp. 258–266 (2019)
8. Domshlak, C., Mirkis, V.: Deterministic oversubscription planning as heuristic search: abstractions and reformulations. J. Artif. Intell. Res. **52**, 97–169 (2015)
9. Eifler, R., Cashmore, M., Hoffmann, J., Magazzeni, D., Steinmetz, M.: Explaining the space of plans through plan-property dependencies. In: Proceedings of the 2nd Workshop on Explainable Planning (XAIP 2019) (2019)
10. Fox, M., Long, D., Magazzeni, D.: Explainable planning. In: Proceedings of IJCAI 2017 Workshop on Explainable AI (2017)
11. Göbelbecker, M., Keller, T., Eyerich, P., Brenner, M., Nebel, B.: Coming up with good excuses: what to do when no plan can be found. In: Brafman, R.I., Geffner, H., Hoffmann, J., Kautz, H.A. (eds.) Proceedings of the 20th International Conference on Automated Planning and Scheduling (ICAPS 2010), pp. 81–88. AAAI Press (2010)

12. Kambhampati, S.: Synthesizing explainable behavior for human-AI collaboration. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2019), pp. 1–2 (2019)
13. Khan, O.Z., Poupart, P., Black, J.P.: Minimal sufficient explanations for factored Markov decision processes. In: Gerevini, A., Howe, A., Cesta, A., Refanidis, I. (eds.) Proceedings of the 19th International Conference on Automated Planning and Scheduling (ICAPS 2009). AAAI Press (2009)
14. Krarup, B., Cashmore, M., Magazzeni, D., Miller, T.: Model-based contrastive explanations for explainable planning. In: ICAPS 2019 Workshop on Explainable Planning (2019)
15. Kulkarni, A., Zha, Y., Chakraborti, T., Vadlamudi, S.G., Zhang, Y., Kambhampati, S.: Explicable planning as minimizing distance from expected behavior. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2019, Montreal, QC, Canada, 13–17 May 2019, pp. 2075–2077 (2019)
16. Langley, P., Meadows, B., Sridharan, M., Choi, D.: Explainable agency for intelligent autonomous systems. In: Singh, S., Markovitch, S. (eds.) Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI 2017). AAAI Press, February 2017
17. McGuinness, D.L., Glass, A., Wolverton, M., da Silva, P.P.: Explaining task processing in cognitive assistants that learn. In: Proceedings of the 20th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2007), pp. 284–289 (2007)
18. Miller, T.: Contrastive explanation: a structural-model approach. CoRR, abs/1811.03163 (2018)
19. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
20. Nothdurft, F., Behnke, G., Bercher, P., Biundo, S., Minker, W.: The interplay of user-centered dialog systems and AI planning. In: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDAL2015), pp. 344–353 (2015)
21. Petrick, R., Dalzel-Job, S., Hill, R.: Combining cognitive and affective measures with epistemic planning for explanation generation. In: Proceedings of the 2nd Workshop on Explainable Planning (XAIP 2019) (2019)
22. Rosenthal, S., Selvaraj, S.P., Veloso, M.M.: Verbalization: narration of autonomous robot experience. In: Kambhampati, S. (ed.) Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016). AAAI Press/IJCAI (2016)
23. Sarath, S., Alberto, O., Prasad, M., Subbarao, K.: Model-free model reconciliation. In: ICAPS-19 Workshop on Explainable Planning (2019)
24. Seegebarth, B., Müller, F., Schattenberg, B., Biundo, S.: Making hybrid plans more clear to human users - A formal approach for generating sound explanations. In: Bonet, B., McCluskey, L., Silva, J.R., Williams, B. (eds.) Proceedings of the 22nd International Conference on Automated Planning and Scheduling (ICAPS 2012). AAAI Press (2012)
25. Smith, D.: Planning as an iterative process. In: Hoffmann, J., Selman, B. (eds.) Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI 2012), Toronto, ON, Canada, pp. 2180–2185. AAAI Press, July 2012
26. Smith, D.E.: Choosing objectives in over-subscription planning. In: Koenig, S., Zilberstein, S., Koehler, J. (eds.) Proceedings of the 14th International Conference on Automated Planning and Scheduling (ICAPS 2004), Whistler, Canada, pp. 393–401. Morgan Kaufmann (2004)

27. Sohrabi, S., Baier, J.A., McIlraith, S.A.: Preferred explanations: theory and generation via planning. In: Burgard, W., Roth, D. (eds.) Proceedings of the 25th National Conference of the American Association for Artificial Intelligence (AAAI 2011), San Francisco, CA, USA. AAAI Press, July 2011
28. Sreedharan, S., Chakraborti, T., Kambhampati, S.: Handling model uncertainty and multiplicity in explanations via model reconciliation. In: Proceedings of the Twenty-Eighth International Conference on Automated Planning and Scheduling, ICAPS 2018, Delft, The Netherlands, 24–29 June 2018, pp. 518–526 (2018)
29. Zhang, Y., Sreedharan, S., Kulkarni, A., Chakraborti, T., Zhuo, H., Kambhampati, S.: Plan explicability and predictability for robot task planning. In: Proceedings of ICRA (2017)