Explanations for Sequential Decision-Making - An Overview

Hendrik Baier[1*], Mark T. Keane[2], Sarath Sreedharan[3], Silvia Tulli[4], Abhinav Verma[5]

[1]Eindhoven University of Technology, NL.
[2]University College Dublin, IE.
[3]Colorado State University, US.
[4]Sorbonne Université, FR.
[5]Pennsylvania State University, US.

h.j.s.baier@tue.nl, mark.keane@ucd.ie, sarath.sreedharan@colostate.edu, silvia.tulli@sorbonne-universite.fr, verma@psu.edu

Abstract

In this paper, we highlight the field of explainable sequential decision making. We discuss how the problem of explaining sequential decisions gives rise to problems and challenges that are absent from scenarios that focus on explaining single-shot decision making. We provide a short survey of some of the more prominent subareas within explainable sequential decision-making and their unique focuses and blind spots. Here, we argue that we need to go beyond simply focusing on individual subareas like explainable planning, reinforcement learning, or robotics, and move towards studying and tackling the more general problem of explainable sequential decision-making. Such a holistic approach will not only allow us to identify previously ignored problems, but also provide us with the ability to transfer ideas and intuitions from one subarea of explainable sequential decision-making to another. We end the paper with a discussion on future directions and some of the most pressing open questions.

Introduction

The modern field of explainable AI (XAI) traces much of its identity back to the DARPA XAI program (Gunning and Aha 2019). Efforts to generate explanations for AI system decisions had existed before; in fact, within the context of expert systems, the capacity to explain and justify decisions was long regarded as essential for successful deployment (Swartout, Paris, and Moore 1991). Nonetheless, the priorities and aims of XAI as a field were strongly shaped by the DARPA initiative and by the types of AI systems it emphasized. In particular, many early advances in XAI focused on single-shot decision-making systems such as neural-network-based classifiers. While this emphasis yielded valuable methods, it also meant that research often overlooked a broader and arguably more compelling class of problems: namely. those involving sequential decision-making.

Sequential decision-making is not only significant because it generalizes single-shot decision-making; it also introduces challenges that are entirely absent from the latter setting. This becomes evident when trying to apply single-shot explanation methods to sequential decision-making tasks. For instance, applying feature attribution techniques to a reinforcement learning policy network (cf. (Huber, Schiller, and André 2019)) may reveal the features influencing the selection of a specific action. However, such methods do not clarify how that action contributes to the agent's long-term objective, or how the agent expects to handle its possible consequences. For researchers in this area, this limitation is unsurprising, given the complexity inherent in both the decisions themselves and the underlying decision-making processes. Indeed, a rich body of work already addresses the problem of explaining the validity of action sequences, employing approaches such as validation structures (Kambhampati and Hendler 1992) and causal link explanations (Bercher et al. 2014) to justify the role of individual actions.

Therefore, one of the central goals of this paper is to advocate for sequential decision-making as a productive level of abstraction at which to frame a large class of XAI problems. This perspective stands in contrast to the current tendency to divide efforts into narrower domains, such as explainable planning (Chakraborti, Sreedharan, and Kambhampati 2020), explainable reinforcement learning (Milani et al. 2024), or explainable robotics (Sakai and Nagai 2022). While such fragmentation has driven progress within individual subfields, it has also hindered the transfer of insights and methods across contexts. By adopting a more abstract stance, we aim to highlight commonalities across problems and explore how techniques can be adapted to the distinct needs of specific settings.

In line with this objective, this paper is structured as follows. We first provide a general definition of sequential decision-making problems. We then describe key explanation-generation challenges in this space, focusing on the particular components of the decision-making process that existing algorithms address. After that, we examine important subcategories of sequential decision-making problems and outline challenges unique to each. Finally, we conclude by identifying promising directions for future research and offering broader reflections. Many of the insights presented here are based on discussions from the Dagstuhl seminar on "Explainable AI for Sequential Decision Making", held in September 2024 (Baier et al. 2025). Readers interested in further detail are encouraged to consult the full sem-

[*]Authors are listed in alphabetical order.

inar report.

Sequential Decision Making

Before we start describing the exact problem of generating explanations, we will start by providing a sketch of what we mean by sequential decision-making problems. Specifically, in this paper, when we use the term sequential decision making or SDM, it always involves an agent making a series of related decision or actions over time, where each decision may be influenced by earlier decisions, current observations and/or projected future states, while trying to achieve some goal, or optimizing for an objective, a total reward or a cost.

Trivially, any single-shot decision-making setting is an instance of the more general sequential decision-making problem. However, we would argue that it is a less interesting one from an explanation point of view. To start with, the fact that SDMs deal with temporally-extended sequences of decisions brings with it a number of novel challenges. Here, each choice needs to account not only for the current state, or even the history of observations it receives, but all future states that action could lead to and how that affects its ability to achieve its objectives. This decision-making context may thus require new concepts, like trajectories and key events (Dinu et al. 2020), to generate satisfying explanations. Moving on to objectives, we also see that the ones considered in the SDM literature tend to be more diverse and varied than those considered in single-shot decision-making. Just to cite a few, popular objective mechanisms include goal state specification, reward functions, and even temporally-extended goals, which require the agent to generate particular types of trajectories. Explaining the objective of the current task to an end user in itself could be a challenging problem. Finally, it is worth keeping in mind that for any single-shot decision, one could always conceive of a more complex sequential decision-making problem, where the original single-shot problems are simply individual actions performed in the context of this overall problem.

Furthermore, it is also worth keeping in mind how closely related the notion of autonomy and agency is to sequential decision-making. While all AI systems could in theory be conceived of as agents (Russell and Norvig 2021), everyday users are usually more open to assigning such attributes to systems performing sequential decision-making. Additionally, there are aspects of autonomy that are most evidently seen in the sequential decision-making setting. For instance, agents that have incomplete knowledge about the task and environment, may have to autonomously explore the context to make up for this missing knowledge. This closer association between autonomy and agency, also brings with it, additional explanatory challenges. For one, users may be more open to anthropomorphizing such systems, and in turn the explanation may need to be given in terms of the agent intent, beliefs and capabilities.

Objects of Explanation in SDM

SDM is markedly more complex than traditional settings of explainable machine learning in terms of the diversity and complexity of the objects that may need to be explained. We focus on explanations for three major facets of SDM, that are elaborated in turn below:

- The decision sequence and the factors leading an agent to this sequence;
- the actual operation of the decision-making algorithm within the agent;
- and the environment in which the agent is making decisions and executing actions.

Explaining the decision sequence and its influences Every aspect of a decision sequence (or abstractions of those aspects) can be the object of an explanation, whether it be the decision outcomes, the actions taken or to be taken, the states traversed or to be traversed, or the agent's choices and objectives with respect to the past, present, or future. Decision sequences may also need to be explained with respect to beliefs about or comparisons to any aspects of *other* decision sequences. An explanation of a decision sequence may also address factors that influenced its *generation*, such as human inputs, the agent's history, its task, goals etc. Within each of the relevant fields using sequential decision-making these explanations can focus on different specific aspects: for example, in the planning field, states, actions, plans, goals, and failures may need to be explained, whereas in reinforcement learning we may want to explain local decisions, global decisions (e.g., involving policies) and various influences on these decisions (e.g., influential training data).

Explaining the algorithm An explanation could also focus on the agent or learning method or decision-making algorithm itself, to illuminate its operation to more technical end-users. For instance, in planning we may require an explanation of the the actual planning algorithm, and in reinforcement learning we may need insight into the algorithm that generated the agent's policy and behaviors, its methods for exploration, credit assignment, etc.

Explaining the environment An explanation may also consider the environment as the object to be explained, as captured in the agent's world model; it could explain the agent's perception and comprehension of the real world, the process of forming its world model, and any discrepancies that may arise between the world model and reality. Again, in different research fields, this environment focus will have somewhat different flavors. In the robotics field for example, one may need to explain environmental constraints, failures that arise from the robots' own physical limitations, and other constraints of computation, memory, or power. In reinforcement learning, on the other hand, the focus might be on explaining the formulation of the task as a Markov Decision Problem (MDP), including its state and action spaces, transitions, rewards, and discount factor.

Flavors of SDM and Their Explanations

Next, we will look at some of the more important subfields within SDM and some prominent explanation generation methods related to those subfields. Please note that this section isn't meant to be an exhaustive summary of all the methods that exist in each field.

Planning With planning, we refer to methods that are generally focused on techniques that generate decisions from models of the task. This includes methods designed for classical planning problems (Geffner and Bonet 2013) and more expressive formulations like those represented as Markov Decision Processes (MDP) (Puterman 1990), and Partially Observable Markov Decision Processes (POMDP) (Astrom 1965). We will also consider methods that support online planning (Sutton and Barto 2018). Starting with classical planning, there exists a pretty rich literature on generating explanations. Many of them could be traced back to early interests in case-based planning and plan adaptation (Kambhampati 1990). As part of these systems, many of the plans were associated with explanations as to why they constituted a valid plan. Some prominent examples include validation structure (Kambhampati and Hendler 1992). Causal links, a mechanism used to represent the relationship between actions in a plan, have also been widely used as a way to explain or justify the contributions made by an individual action to the success of the overall plan (Bercher et al. 2014). Coming to more recent works, some of the more popular approaches used in explanations for planning include contrastive explanations against counterfactual plans (Cashmore et al. 2019), the use of abstractions to simplify model information (Sreedharan, Srivastava, and Kambhampati 2021; Vasileiou and Yeoh 2023), and model reconciliation explanations (Sreedharan, Chakraborti, and Kambhampati 2021; Vasileiou et al. 2022). While rare, there have also been works that have tried to present visualizations of the decision-making process itself (Magnaguagno et al. 2017) or of explanations (Kumar et al. 2022). There are also methods that are focused on generating explanations for decisions derived from specific planning methods like MCTS (Baier and Kaisers 2021). For the readers interested in learning more, we will point them to look at the position paper on explainable planning (Fox, Long, and Magazzeni 2017) and the survey by Chakraborti, Sreedharan, and Kambhampati (2020).

Reinforcement Learning Here, with Reinforcement Learning (RL), we refer to SDM methods that learn from experience, either by directly interacting with an environment or a simulator. We will consider both tabular RL and those that use function approximators for representing policies or value functions. The history of explanations for reinforcement learning is relatively new, with some exceptions for generating explanations for MDP policies that could in theory be adapted to RL systems (cf. (Khan, Poupart, and Black 2009)). A large focus on early RL explanations includes the goal of helping users interpret policies generated by RL algorithms. Here, the efforts have included methods that leverage methods from single-shot decision-making systems to make sense of action generated via policy networks (cf. (Huber, Schiller, and André 2019)), generating symbolic rules that summarize the overall policy (cf. (Hayes and Shah 2017)), and learning other post-hoc policy representations like finite-state machines (Danesh et al. 2021). Other related approaches include the use of programmatic RL, which aims to learn inherently interpretable policies (cf. (Verma et al. 2018)), communicating the overall policy by the use of demonstration of the policy at some pre-identified states (cf. (Amir and Amir 2018)), and even the use of state-abstraction to simplify the overall policy (cf. (Topin and Veloso 2019)). Other prominent methods include decomposition of the value function into more interpretable reward sources (Erwig et al. 2018), and the use of learned symbolic or causal models to generate explanations (Madumal et al. 2020). More recently, there have also been efforts to generate counterfactual explanations for actions identified by RL methods (Sreedharan et al. 2020). We point the readers to the survey by Milani et al. (2024) for a more comprehensive treatment of the existing methods within the space.

Robotics Here, we focus on methods that are designed to generate explanations for behavior generated by physically embodied agents. Explanations for robotics have a large overlap with those meant for planning and reinforcement learning. This is not particularly surprising given the fact that robotics relies on the earlier methods to generate decisions. However, the specifics of the application setting, including the fact that it is an embodied agent, and the interplay of task-level and motion-level constraints, introduce new opportunities and challenges not usually considered in the more general settings. A significant outcome of the use of an embodied agent is the possibility of using agent behavior to communicate information about the decision-making with a human observer implicitly. This includes the use of legible behavior to communicate intent (Dragan, Lee, and Srinivasa 2013), the use of predictable behavior to generate behavior that the user can predict (Fisac et al. 2020), the use of explicable behavior to avoid human confusion (Zhang et al. 2017), and even communicating robot incapability through behavior (Kwon, Huang, and Dragan 2018). It has been shown that even robot morphology can be used to communicate information about the robot's intent and capabilities (Haring, Matsumoto, and Watanabe 2013). New explanation communication paradigms, like augmented reality, have also been considered in the context of explaining robotics (Chakraborti et al. 2018). Other robotics-specific explanation generation methods include explanation methods that can switch between task-level and motion-level explanations (Sreedharan, Srivastava, and Kambhampati 2021) and methods that are designed to verbalize robot experience (Rosenthal, Selvaraj, and Veloso 2016). We point the readers to the survey by Sakai and Nagai (2022) for more details on existing works in this space.

Moving Towards Explainable SDM Finally, let's go from individual pieces to the larger picture of explanations for sequential decision-making. Going through each individual explanation method discussed in earlier techniques, there exists no method that, in principle, doesn't apply to other sub-areas. The choice for each subarea to focus on certain subproblems results from the information available to the system and what problems seem more pressing to the practitioner. For example, the prevalence of methods to effectively communicate decisions in RL isn't shared by works in explainable planning. Even though the solution concepts used for MDPs and POMDPs are as complex, if not more

complex, than the ones leveraged in traditional RL settings. Even within classical planning, where the solutions are simply sequences of action called plans, communicating a plan that contains hundreds, if not thousands, of steps is not trivial. On the other hand, methods like model reconciliation, where explanations that try to resolve the user's misunderstanding about the task and robot model, still remain relevant in reinforcement learning settings. However, very little work has been done to port those works over to RL settings. Finally, apart from works that leverage simple visualization, very little work has been done on relying on accounts as a communication medium during explanation generation in both planning and RL contexts. For example, methods like augmented reality could be used for non-embodied decision-making too. As mentioned, one of the problems we believe is that very few people look at the entire landscape of explanation for sequential decision-making.

Open Questions and Next Steps

We would like to close out this paper by considering important open questions that arise in explainable sequential decision-making, ones that may not arise at all in single-shot decision-making settings. Indeed, many of the questions that one would imagine to be fundamental to explanations, in general, take on new dimensions in the context of sequential decision making.

For instance, in single-shot decisions, the assumed explanatory goals are often monolithic, unchanging and unquestioned by system developers (e.g., the goal of explaining a classification is to communicate the feature importances). Yet, in sequential decision-making, an agent can have many different explanatory goals (e.g., challenging an individual user or promoting team cooperation), resulting in different explanations to achieve some overall strategy. Similarly, when we revisit a question like how to incorporate user knowledge and background into generated explanations, we see more examples of methods that leverage such consideration within sequential decision-making than in single-shot decision-making settings (cf. (Sreedharan, Chakraborti, and Kambhampati 2021)).

As we move to questions that are more unique to sequential decision-making, a good starting point would be that of how to communicate the identified solution itself. In many cases, the policy or plan generated could be quite complex, and communicating them to the user might require additional mechanisms (cf. (Sreedharan, Srivastava, and Kambhampati 2020)). Such communication could be further complicated by many factors, such as multiple objectives, constraints, partial observability, and model uncertainty. In many cases, it might even be hard to communicate what the solution is meant to accomplish. Furthermore, apart from the issue of *how* to explain the decision, there is also the question of *when* to communicate that explanation. The very fact that the agent could treat the communication of explanation as yet another action opens up many possibilities that simply do not arise in single-shot decision-making problems.

A core characteristic of many sequential decision-making problems is the usual separation between the decision-making stage and the execution phase. While one could argue that such considerations are present in single-shot decision-making, in such settings, it would not make sense to talk about the agent reasoning about the consequences of its actions in the environment. In relation to this separation, one important question would be how one provides an account for events that the system's model did not capture. Here, the possibilities could range from events that arise due to unmodeled side-effects of the agent's actions or acting in a dynamic environment. Another consequence of this separation between decision-making and execution is the possibility that the agent might choose to change its plan or policy. A relevant question, then, is how one explains such changes in behavior.

The main message we would like to send through this paper is the need to look beyond just generating explanations for specific instances of sequential decision-making and look at the problem in a more holistic manner. One of the critical next steps towards supporting such efforts would be to create a unified taxonomy for explanations within sequential decision-making that can encompass all existing works within the individual subfields. We hope that this write-up will encourage more researchers to take this viewpoint and work towards creating such general frameworks.

Acknowledgments

This work was initiated by Dagstuhl Seminar 24372 "Explainable AI for Sequential Decision Making". We are indebted to all seminar participants for their valuable insights.

This work has received funding from NSF grant 2303019, the project ALIGN4Energy (NWA.1389.20.251) of the Dutch research programme NWA ORC 2020, and from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement number 101120406. The paper reflects only the authors' view, and the EC is not responsible for any use that may be made of the information it contains.

References

Amir, D.; and Amir, O. 2018. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems*, 1168–1176.

Astrom, K. J. 1965. Optimal control of Markov decision processes with incomplete state estimation. *J. Math. Anal. Applic.*, 10: 174–205.

Baier, H.; and Kaisers, M. 2021. Towards explainable MCTS. In *2021 AAAI Workshop on Explainable Agency in AI*, volume 178.

Baier, H.; Keane, M. T.; Sreedharan, S.; Tulli, S.; Verma, A.; and Vasileiou, S. L. 2025. Explainable AI for Sequential Decision Making (Dagstuhl Seminar 24372). *Dagstuhl Reports*, 14(9): 67–103.

Bercher, P.; Biundo, S.; Geier, T.; Hoernle, T.; Nothdurft, F.; Richter, F.; and Schattenberg, B. 2014. Plan, repair, execute, explain—how planning helps to assemble your home theater. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 24, 386–394.

Cashmore, M.; Collins, A.; Krarup, B.; Krivic, S.; Magazzeni, D.; and Smith, D. 2019. Towards explainable AI planning as a service. *arXiv preprint arXiv:1908.05059*.

Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2020. The emerging landscape of explainable ai planning and decision making. *arXiv preprint arXiv:2002.11697*.

Chakraborti, T.; Sreedharan, S.; Kulkarni, A.; and Kambhampati, S. 2018. Projection-aware task planning and execution for human-in-the-loop operation of robots in a mixed-reality workspace. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4476–4482. IEEE.

Danesh, M. H.; Koul, A.; Fern, A.; and Khorram, S. 2021. Re-understanding finite-state representations of recurrent policy networks. In *International conference on machine learning*, 2388–2397. PMLR.

Dinu, M.-C.; Hofmarcher, M.; Patil, V. P.; Dorfer, M.; Blies, P. M.; Brandstetter, J.; Arjona-Medina, J. A.; and Hochreiter, S. 2020. XAI and strategy extraction via reward redistribution. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, 177–205. Springer.

Dragan, A. D.; Lee, K. C.; and Srinivasa, S. S. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 301–308. IEEE.

Erwig, M.; Fern, A.; Murali, M.; and Koul, A. 2018. Explaining deep adaptive programs via reward decomposition. In *IJCAI/ECAI workshop on explainable artificial intelligence*.

Fisac, J. F.; Liu, C.; Hamrick, J. B.; Sastry, S.; Hedrick, J. K.; Griffiths, T. L.; and Dragan, A. D. 2020. Generating plans that predict themselves. In *Algorithmic Foundations of Robotics XII: Proceedings of the Twelfth Workshop on the Algorithmic Foundations of Robotics*, 144–159. Springer.

Fox, M.; Long, D.; and Magazzeni, D. 2017. Explainable planning. *arXiv preprint arXiv:1709.10256*.

Geffner, H.; and Bonet, B. 2013. *A concise introduction to models and methods for automated planning*. Morgan & Claypool Publishers.

Gunning, D.; and Aha, D. W. 2019. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Mag.*, 40: 44–58.

Haring, K. S.; Matsumoto, Y.; and Watanabe, K. 2013. How do people perceive and trust a lifelike robot. In *Proceedings of the world congress on engineering and computer science*, volume 1, 425–430.

Hayes, B.; and Shah, J. A. 2017. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, 303–312.

Huber, T.; Schiller, D.; and André, E. 2019. Enhancing explainability of deep reinforcement learning through selective layer-wise relevance propagation. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, 188–202. Springer.

Kambhampati, S. 1990. A classification of plan modification strategies based on coverage and information requirements. In *AAAI 1990 Spring Symposium on Case Based Reasoning*. Citeseer.

Kambhampati, S.; and Hendler, J. A. 1992. A validation-structure-based theory of plan modification and reuse. *Artificial Intelligence*, 55(2-3): 193–258.

Khan, O.; Poupart, P.; and Black, J. 2009. Minimal sufficient explanations for factored markov decision processes. In *Proceedings of the international conference on automated planning and scheduling*, volume 19, 194–200.

Kumar, A.; Vasileiou, S. L.; Bancilhon, M.; Ottley, A.; and Yeoh, W. 2022. Vizxp: A visualization framework for conveying explanations to users in model reconciliation problems. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 32, 701–709.

Kwon, M.; Huang, S. H.; and Dragan, A. D. 2018. Expressing robot incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 87–95.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2020. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 2493–2500.

Magnaguagno, M. C.; FRAGA PEREIRA, R.; Móre, M. D.; and Meneguzzi, F. R. 2017. Web planner: A tool to develop classical planning domains and visualize heuristic state-space search. In *2017 Workshop on User Interfaces and Scheduling and Planning (UISP@ ICAPS), 2017, Estados Unidos.*

Milani, S.; Topin, N.; Veloso, M.; and Fang, F. 2024. Explainable Reinforcement Learning: A Survey and Comparative Review. *ACM Comput. Surv.*, 56(7).

Puterman, M. L. 1990. Markov decision processes. *Handbooks in operations research and management science*, 2: 331–434.

Rosenthal, S.; Selvaraj, S. P.; and Veloso, M. M. 2016. Verbalization: Narration of Autonomous Robot Experience. In *IJCAI*, volume 16, 862–868.

Russell, S.; and Norvig, P. 2021. *Artificial Intelligence: A Modern Approach*. Pearson, 4 edition.

Sakai, T.; and Nagai, T. 2022. Explainable autonomous robots: a survey and perspective. *Advanced Robotics*, 36(5-6): 219–238.

Sreedharan, S.; Chakraborti, T.; and Kambhampati, S. 2021. Foundations of explanations as model reconciliation. *Artificial Intelligence*, 301: 103558.

Sreedharan, S.; Soni, U.; Verma, M.; Srivastava, S.; and Kambhampati, S. 2020. Bridging the gap: Providing post-hoc symbolic explanations for sequential decision-making problems with inscrutable representations. *arXiv preprint arXiv:2002.01080*.

Sreedharan, S.; Srivastava, S.; and Kambhampati, S. 2020. Tldr: Policy summarization for factored ssp problems using temporal abstractions. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 30, 272–280.

Sreedharan, S.; Srivastava, S.; and Kambhampati, S. 2021. Using state abstractions to compute personalized contrastive explanations for AI agent behavior. *Artificial Intelligence*, 301: 103570.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 2nd edition.

Swartout, W.; Paris, C.; and Moore, J. D. 1991. Explanations in knowledge systems: design for explainable expert systems. *IEEE Expert*, 6: 58–64.

Topin, N.; and Veloso, M. 2019. Generation of policy-level explanations for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2514–2521.

Vasileiou, S. L.; and Yeoh, W. 2023. PLEASE: Generating personalized explanations in human-aware planning. In *ECAI 2023*, 2411–2418. IOS Press.

Vasileiou, S. L.; Yeoh, W.; Son, T. C.; Kumar, A.; Cashmore, M.; and Magazzeni, D. 2022. A logic-based explanation generation framework for classical and hybrid planning problems. *Journal of Artificial Intelligence Research*, 73: 1473–1534.

Verma, A.; Murali, V.; Singh, R.; Kohli, P.; and Chaudhuri, S. 2018. Programmatically interpretable reinforcement learning. In *International conference on machine learning*, 5045–5054. PMLR.

Zhang, Y.; Sreedharan, S.; Kulkarni, A.; Chakraborti, T.; Zhuo, H. H.; and Kambhampati, S. 2017. Plan explicability and predictability for robot task planning. In *2017 IEEE international conference on robotics and automation (ICRA)*, 1313–1320. IEEE.