Agent Strategy Summarization

Blue Sky Ideas Track

Ofra Amir
Technion IE&M
oamir@technion.ac.il

Finale Doshi-Velez
Harvard University
finale@seas.harvard.edu

David Sarne
Bar-Ilan University
sarned@cs.biu.ac.il

ABSTRACT

Intelligent agents and AI-based systems are becoming increasingly prevalent. They support people in different ways, such as providing users with advice, working with them to achieve goals or acting on users' behalf. One key capability missing in such systems is the ability to present their users with an effective summary of their strategy and expected behaviors under different conditions and scenarios. This capability, which we see as complimentary to those currently under development in the context of "interpretable machine learning" and "explainable AI", is critical in various settings. In particular, it is likely to play a key role whenever a user needs to understand the strategy of an agent she is working along with, when having to choose between different available agents to act on her behalf, or when requested to determine the level of autonomy to be granted to the agent or approve its strategy. In this paper, we pose the challenge of developing capabilities for strategy summarization, which is not addressed by current theories and methods in the field. We propose a conceptual framework for strategy summarization, which we envision as a collaborative process that involves both agents and people. Last, we suggest possible testbeds that could be used to evaluate progress in research on strategy summarization.

KEYWORDS

Strategy summarization; Explainable AI

ACM Reference Format:
Ofra Amir, Finale Doshi-Velez, and David Sarne. 2018. Agent Strategy Summarization. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 5 pages.

1 INTRODUCTION

Intelligent systems play a growing role in our daily lives, from voice-controlled assistants and online ordering aides to tools for recognizing cancerous tumors and computer-assisted driving [38]. Many of these systems go even further, autonomously carrying out tasks on behalf of their user rather than simply providing advice. These can either take the form of virtual agents (e.g., Poker bots that play on real money, automatic news feed generators) or physical ones (e.g., autonomous cars, vacuum robots). The behavior of these systems is often opaque to human users. For example, a robotic vacuum may be equipped with several coverage algorithms and the choice of which to be used at a time may depend on environment conditions, which might be beyond the user's reach.

Users' familiarity with the strategies and expected behaviors of agents under different conditions and scenarios is essential for many purposes. First, an understanding of agents' behaviors can facilitate choosing between interchangeable systems (e.g., Siri, Cortana, Alexa). Second, knowing the agent's strength and weaknesses can improve the ability of people to collaborate with agents (e.g., with a surgical robot). Last, users may need to determine how much autonomy to grant to an agent, and knowing its expected behavior can help them make more informed decisions, and trust that the agent could perform its designated role.

However, explaining complex system behavior to users is challenging because the behavior of these systems is often determined using sophisticated computational techniques (e.g., machine-learning models, deep learning, Markov decision processes and in many cases an ensemble of methods). People are inherently bounded-rational and find it difficult to map from a design and logic to actual behavior of the system. It has been shown that users' mental models of system behaviors are incomplete, parsimonious and unstable; that people are limited in their ability to "run" these models to predict expected behavior; and that people often confuse different mental models [33]. Moreover, attempting to specify the system's actual behavior in each world state is typically infeasible because the space of possible world states the system may run into is often far more immense than what a human can manage. For example, the state space in which autonomous vehicles make decisions is based on speed, weather, road conditions, distance and much other data gathered by a variety of sensors, including cameras, LiDARS (Light Detection and Ranging), and radars. While the user may be a very experienced driver, it is very likely that she has never experienced or ever considered many of the possible states in this space (e.g., spotting a child running out to retrieve a bouncing ball or spotting a tire of the car in front of you exploding).

Supporting users in understanding and anticipating agents' behavior thus poses significant computational problems. While there has been growing recent discussion of the need for "explainable AI" and "human-aware AI" — which all relate to the problems specified above — these concepts are very broad and are not immediately translatable to concrete research problems, making it hard to measure progress in these areas. Therefore, we pose a more specific challenge: the development of core capabilities for *agent strategy summarization*. This challenge is not addressed by current state-of-the-art theories and methods. However, it is more concrete than the high-level goal of "explainable AI" and thus allows us to identify specific open problems that need to be solved in order to meet this challenge. These include the development of algorithms that generate summaries of agent behaviors, as well as the design of collaborative interfaces through which users can review and explore these summaries. We further consider methodologies and testbeds

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden. © 2018 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

for proper experimental evaluation of the effectiveness of such methods in improving people's understanding of agent behavior.

The idea behind strategy summarization is to provide users with some form of a summary (either textual or visual, through an interactive interface) that demonstrates system behavior in carefully selected world states. With this new paradigm, users gain a better understanding of the system in a range of diverse world states, in a relatively short time. While recent efforts developed methods for explaining one-shot decisions made by autonomous agents (e.g., Khan et al. [23]) or machine learning algorithms (e.g., [34]) in retrospect, or *ex-post*, strategy summarization offers a complementary important capability, which is the cohesive description of the behavior an agent is likely to exhibit, *ex-ante*. Few recent works propose user interface designs enabling users to query an agent's policy (e.g., Hayes & Shah [20]), yet these require much sophistication, knowledge and effort from users in order to properly understand the system, and do not generate automated summaries. In contrast, the strategy summarization approach aims to communicates the actual agent behavior in a scenario-based manner, rather than conveying its underlying decision-making model (e.g., a decision tree, the coefficients of a logistics regression model). It reduces user effort in that it presents agent behavior in a range of scenarios rather than requiring users to specify many queries on their own.

The study of strategy summarization will contribute to theories and methods in the areas of decision making under uncertainty, human-agent interaction, explainable AI and multi-agent literature. It requires expertise in diverse fields, in particular human-agent interaction, planning and learning algorithms and representations, interpretability, and machine learning. Still, we argue it is both feasible and worth pursuing, as methods that help people better understand agents are expected to have an impact in many areas, including domains of societal importance such as healthcare, education and transportation.

2 RELATED WORK

While there is relatively little work on ex-ante descriptions of agent strategy, there is a broader literature of summarizing hierarchical plans, explaining robot behavior, explaining decisions in Markov Decision Processes, interpretable machine learning, as well studies concerned with users' mental models of systems. We review works in these areas, which strategy summarization research can and should draw on.

Summarizing hierarchical plans. Recent research has considered several new approaches for enabling users understand the strategy of the systems they use. For example, Myers [31] proposed a method for summarizing plans represented as Hierarchical task networks (HTNs) to help people in reviewing and comparing them, emphasizing features such as the allocation of roles to agents, tasks included in the plan and tasks absent. However, this approach is limited in the sense that it is only appropriate for short-term plans toward achieving a specific goal and relies on an HTN model.

Explaining robot and agent behavior. In the context of human-robot interaction, prior work has developed methods for supporting users in debugging a robot or to improve the ability of the human and the robot to collaborate effectively. For example, Nikolaidis et al. [32] proposed a cross-training approach to help parties develop a better understanding of their teammate. Lomas et al. [27] developed a system that enables a user to ask robots questions. Brooks et al. [5] developed a system that visualizes all the past actions of a robot and includes explanations for the actions. Hayes & Shah [20] proposed several methods for explaining robot policies to people using past execution traces, enabling users to query the agent's behavior in different states and request explanations. In other work, animation techniques depicting anticipation and reaction of robots were used to help people predict what a robot will do next [40]. Recently, Zhang et al. [51] proposed methods for plan explicability and predictability which can be used to generate plans that are more understandable to people. Other works have used argumentation approaches to explain agent behavior [1, 37]. These approaches are complementary to the strategy summarization approach, in that they provide different ways of examining the behaviors of agents, yet they do not attempt to generate summaries of the agent's behavior. Instead, they present complete plans or attempt to explain the logic underlying specific decisions.

Several prior works suggested methods for explaining recommendations given by MDP-based intelligent assistants [9, 11, 12, 22, 23] or explaining plans [36]. Wang et al. generated explanations of robot reasoning based on Partially Observable Markov Decision Problems (POMDPs) [47] and similar approaches have been developed for explaining decisions in the context of HTN planning, explaining an agent's actions based on its task model [29, 30]. The problem we address differs from the problem of generating explanations for specific decisions, as rather than explaining an action taken (or a suggested action), we aim to describe *which* actions would be taken in information-critical states, before the agent actually acts.

Interpretable machine learning. Recently, many approaches have been proposed for developing interpretable machine learning models, that is, models that are understandable to people [10, 26, 35, 44]. These approaches typically seek to explain a decision made by a machine learning model (e.g., by showing a simpler locally correct model [34] that explains the classification of a sample). Similar to the methods for explaining MDP decisions, these approaches explain a single decision after the fact, rather than provide a description of a strategy or behavior of an agent in different situations. While some methods suggest complete decision-making models that can be more intuitive to people [24, 50], they do not account for sequential decision-making settings.

Users' understanding of system behavior. The strategy summarization idea relates to the literature on users' mental models of systems (software, robots). In a study of users' trust in personal assistants, Glass et al [16] found that system transparency is important to users and suggested that explanations of system behavior can facilitate trust. More accurate mental models of users about robots' behavior can result in improved performance when collaborating with robots [8]. However, research in HCI has shown that people face difficulties in forming accurate mental models of systems and in practice their models of system behaviors are incomplete, parsimonious and unstable; and they are limited in their ability to predict a system's behavior [33]. Letting users interact with a robot's behavior has been shown to help them establish appropriate mental model[39]. We hypothesize that presenting users with summaries of agent strategies will also help them establish mental models of these agents, facilitating trust and collaboration.

3 A CONCEPTUAL FRAMEWORK FOR STRATEGY SUMMARIZATION

While strategy summarization is a complex task, we argue that it can be broken down into manageable subcomponents. We suggest a conceptual framework for the process of strategy summarization, illustrated in Figure 1, which we envision as a collaborative effort that involves both agents and people. We next describe the main components required for generating and presenting summaries, as well as how they interact.

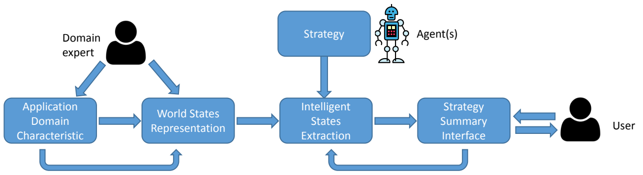
Domain expert
Strategy
Agent(s)
Application Domain Characteristic
World States Representation
Intelligent States Extraction
Strategy Summary Interface
User

Figure 1: A conceptual model of the summarization process.

A key capability required for strategy summarization is identifying states that are likely to be of interest to people interacting with the agent, such that the behavior of the agent in this states can be conveyed to users. To this end, we suggest developing **intelligent state extraction methods**. These methods should take as input the strategy of the agent (or agents) that needs to be summarized and identify and prioritize a subset of states to include in the summary. The implementation of such methods requires a specification of the desired properties of a set of world-states to present in a summary. We discuss these in Section 3.1.

Another important capability required for the summarization process is the ability to properly **represent world states** (Section 3.2), meaning, how to encode states in a given application domain. We expect that internally, the agent may have a highly complex representation of the world that it uses for decision-making. For example, a decision to invest in a stock may depend not only on current stock prices but also on long-term and short-term trends as well as a variety of different types of external financial and political events. This representation likely does not match the human user's notion of essential decision-making factors. A good world-state representation will substantially reduce the potentially immense and inscrutable space of agent internal representations to those that are meaningful to the user.

Finally, users will be the ultimate consumers of the strategy summary, and thus also essential in the process of determining what is relevant to them. We envision **strategy summary interface**, facilitating mixed-initiative interaction, where the exploration of the summary is guided by both the user and the system. We discuss key design and computational problems toward supporting this collaborative exploration in Section 3.3.

3.1 Intelligent State Extraction

While the specific requirements for an effective strategy summary will likely vary across problem domains, we hypothesize that there is a basic set of requirements that are common across different settings: the summary should provide a high coverage of states that are likely to be of interest to users and should be of reasonable length (such that it does not require too much time/cognitive effort to review). In addition, the summary should be relevant to the user's goal (e.g., whether it is to choose between agent or work together with an agent), and should provide information that would help the user to contextualize the information presented (e.g., the likelihood of encountering different states).

We suggest two potential directions for the development of methods that determine which states (and the corresponding actions taken in those states) to include in the summary. The first approach draws on the agent's own reasoning to identify states of interest. For example, states in which the agent could take an action that leads to a catastrophic outcome, or states where the agent is less confident might be considered as potentially interesting for the user. These can be identified by reasoning about the agent's decision making model. For instance, the distribution of Q-values in a certain state could be used to determine the importance of the choice of action in that state, as well as the uncertainty of the agent regarding its decision. Similar ideas have been used in the past in the context of student-teacher reinforcement learning to determine effective teaching opportunities for agents [3, 7, 41]. In preliminary work on strategy summarization, we used this approach to generate summaries and showed that these summaries helped people choose between different agents [2]. Distance between states can also be computed to avoid redundancy in the summary, and the likelihood of encountering a state can be computed to account for both common and rare states in a summary. If the goal is to distinguish between different users, their summaries can focus on states where their policies diverge.

The second approach utilizes people's judgment to identify states that will likely be of importance to users. An example for possible methods that take this approach is the use of Peer-Designed Agents (PDAs) [6, 13, 14, 28]. These agents have been used in recent years for generating realistic behaviors for the purpose of predicting human behavior and their reaction to changes in their environment. The idea is to provide people with a skeleton agent equipped with all the required functionality except for its behavioral layer and have people (either directly or through the mediation of a programmer) design and program into them the strategy they would have used in similar decision situations. The PDAs' logic can then be used as a reflection of what their developers considered to be important (or worth distinguishing) when reasoning about the strategy to be used. This could enable synthesizing the set of world states that are most relevant for demonstrating the system behavior, either through a manual code review, seeking for points of code divergence, or by applying standard clustering algorithms for identifying those states that are highly distinguishable in terms of the codes used in large by the population (when using several PDAs).

3.2 World-State Representation

Deciding how to encode the state representation such that states could be effectively conveyed to people and such that the space of states to consider for the summary will be reduced is a hard problem. We expect that this would require either analysis of large sets of strategies (e.g., a set of strategies programmed into PDAs for that application domain) or the design of effective processes for eliciting such state representation from domain experts. The idea in both approaches is to reason about what types of raw states can be logically aggregated to a single state, as far as the user is concerned, and to what extent prior events as well as various measurable factors should be considered for distinguishing between states of interest. State representation encoding using experts is inherently a manual process, and designing a process for querying experts in an efficient way (e.g., using active learning approaches) will be key to making this process feasible. Extraction based on strategies could be made either based on manual code analysis or using unsupervised clustering over the raw world states to join states for which a similar action is used by a large subset of strategies.

3.3 Strategy Summary Interface

Naturally, different forms of presenting summaries would be appropriate in different settings. For example, for physical agents such as self-driving cars or home robots, it might be more helpful to visually show their actual behavior (e.g., a video of their execution) than showing a textual summary of their expected behavior. However, for virtual agents such as a finance investment advising agent some form of a textual summary, typically in the form of rules to be applied over sets of world-states may be preferred. A key question is thus how to present summaries to people. There are additional important questions such as how to provide people with sufficient context about the states shown in the summary without overwhelming them with non-important low-level details.

As discussed in Section 3.1, there could be different criteria for deciding what information to include in the summary, and these criteria might be in conflict. Therefore, there is a need to design collaborative interfaces which will allow users to guide the generation of summaries by stating preferences. Such interfaces could also allow users to directly query the agent's strategy and explore its behavior in situations that were not described in the summary. A key design challenge in developing these interaction methods is to ensure that users can express their needs in their own language, rather than being asked to specify low-level system parameters. Moreover, exploring the behavior of an agent can be tedious. To make the process more efficient, the design of mixed-initiative interactions [21] where the system tries to help the user in exploring the agent's strategy will likely be required.

4 EVALUATION METHODOLOGIES

To assess progress in the area of strategy summarization and ensure that generated summaries are helpful for users, it is essential to thoughtfully consider means of evaluation. To this end, there are numerous testbeds of common use in the AI-community which can be used for evaluation. Basic testbeds may include robotics applications, e.g., in exploration and search tasks [25, 43, 45, 46, 49]. More advanced testbeds include trading agents and automated negotiation infrastructures (e.g., TAC [17, 42, 48] or ANAC [4] simulators). Using such well-studied testbeds will enable drawing on already established agent repositories (with a variety of agent designs) and environments associated with these competitions. Route planning (and re-route planning as the congestion predictions change) is also an ideal testbed and there are various multi-layered multi-agent based implementations that can be used [18, 19].

Alongside experimentation with the above testbeds, it is important to have some real test-cases to validate. One ideal application on which strategy summarization can be tested is robotic vacuums. More complex domains include clinical decision-support (e.g., for treatment management [15]) and autonomous vehicles. The simpler domains will facilitate relatively fast cycles of testing and improvement of the developed methods, enabling gradual progress towards evaluation in the more complex domains, where substantial implementation and careful experimental design will be required.

Automated summarization should be assessed both in terms of computational aspects of the summarization methods and human-centered evaluation criteria. Metrics of the latter type can be both objective, e.g., people's understanding of strategies (which could be measured by assessing their ability to predict agents' actions or rank different agents in terms of the performance they are likely to achieve in a given environment) and their performance when collaborating with agents, as well as subjective, including the perceived usefulness of the methods and cognitive effort.

5 DISCUSSION

The study of strategy summarization will contribute much to theories and methods in the areas of decision making under uncertainty, human-agent interaction, explainable AI and multi-agent literature. However, more importantly, its deliverables will enable both novice and expert users to better understand the systems they use, in particular complex (AI-based) systems. With the growing use of AI-based agents and shift toward the design of intelligent agents that can collaborate effectively with people [38], we expect that improved user understanding of agents' capabilities and limitations will lead to improved outcomes in many areas.

One key area in which we expect the developed methods to make a substantial impact is the emerging use of autonomous and semi-autonomous vehicles. There is no doubt that we are on the verge of a quantum shift in the way vehicles, humans and the transportation infrastructures interact. The successful operation of autonomous transportation systems (e.g., the autonomous car or Amazon's UAVs) requires providing the highest level of assurance to legislators, authorities (e.g., highway authorities) and users (e.g., car buyers). With strategy summarization, both legislators and users will be able to better understand the way in which these systems work, potentially leading to shorter approval cycles and more effective use. For example, a better understanding of the expected behavior of an autonomous car will help drivers anticipate situations in which the car needs to transfer control to them.

Strategy summarization could also be beneficial in areas of vast societal importance such as education and healthcare. In education, the methods to be developed will enable parents and educators to make better choices when deciding on the educational systems kids will become engaged with, hence improving education development. In medical domains, patients will be able to better understand treatment protocols leading to a better state of mind while being treated, and professionals will be able to reason about the fit of such plans to patients, potentially improving patients' outcomes.

ACKNOWLEDGMENTS

This research was partially supported by the ISRAEL SCIENCE FOUNDATION (grant No. 1162/17) and the ISF-NSFC joint research program (grant No. 2240/15).

REFERENCES

[1] Leila Amgoud and Henri Prade. 2009. Using arguments for making and explaining decisions. *Artificial Intelligence* 173, 3-4 (2009), 413–436.

[2] Dan Amir and Ofra Amir. 2018. HIGHLIGHTS: Summarizing Agent Behavior to People. In *Proc. of the 17th International conference on Autonomous Agents and Multi-Agent Systems (AAMAS).*

[3] Ofra Amir, Ece Kamar, Andrey Kolobov, and Barbara J Grosz. 2016. Interactive teaching strategies for agent training. International Joint Conferences on Artificial Intelligence.

[4] Tim Baarslag, Koen Hindriks, Catholijn M. Jonker, Sarit Kraus, and Raz Lin. 2012. The first automated negotiating agents competition (ANAC 2010). In *New Trends in Agent-based Complex Automated Negotiations*, Takayuki Ito, Minjie Zhang, Valentin Robu, Shaheen Fatima, and Tokuro Matsuo (Eds.). Springer, Berlin, Heidelberg, 113–135.

[5] Daniel J Brooks, Abraham Shultz, Munjal Desai, Philip Kovac, and Holly A Yanco. 2010. Towards State Summarization for Autonomous Robots.. In *AAAI Fall Symposium: Dialog with Robots*, Vol. 61. 62.

[6] Michal Chalamish, David Sarne, and Raz Lin. 2012. The effectiveness of peer-designed agents in agent-based simulations. *Multiagent and Grid Systems* 8, 4 (2012), 349–372.

[7] Jeffery Allen Clouse. 1996. On integrating apprentice learning and reinforcement learning. (1996).

[8] Sandra Devin and Rachid Alami. 2016. An implemented theory of mind to improve human-robot shared plans execution. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on.* IEEE, 319–326.

[9] Thomas Dodson, Nicholas Mattei, and Judy Goldsmith. 2011. A natural language argumentation interface for explanation generation in Markov decision processes. *Algorithmic Decision Theory* (2011), 42–55.

[10] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. (2017).

[11] Francisco Elizalde. 2008. Policy explanation in factored Markov decision processes. In *Proceedings of the 4th European Workshop on Probabilistic Graphical Models (PGM 2008).* 97âĂŞ–104.

[12] Francisco Elizalde, Luis Enrique Sucar, Alberto Reyes, and Pablo deBuen. 2007. An MDP Approach for Explanation Generation.. In *ExaCt.* 28–33.

[13] Avshalom Elmalech and David Sarne. 2014. Evaluating the applicability of peer-designed agents for mechanism evaluation. *Web Intelligence and Agent Systems* 12, 2 (2014), 171–191.

[14] Avshalom Elmalech, David Sarne, and Noa Agmon. 2016. Agent development as a strategy shaper. *Autonomous Agents and Multi-Agent Systems* 30, 3 (2016), 506–525.

[15] Amit X Garg, Neill KJ Adhikari, Heather McDonald, M Patricia Rosas-Arellano, PJ Devereaux, Joseph Beyene, Justina Sam, and R Brian Haynes. 2005. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *Jama* 293, 10 (2005), 1223–1238.

[16] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. 2008. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces.* ACM, 227–236.

[17] Amy Greenwald and Peter Stone. 2001. Autonomous Bidding Agents in the Trading Agent Competition. *IEEE Internet Computing* 5, 2 (2001), 52–60. https://doi.org/10.1109/4236.914648

[18] Rafik Hadfi and Takayuki Ito. 2016. Holonic Multiagent Simulation of Complex Adaptive Systems. In *Workshop on MAS for Complex Networks and Social Computation (CNSC).*

[19] Rafik Hadfi and Takayuki Ito. 2016. Multilayered Multiagent System for Traffic Simulation. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), Singapore, May 9-13, 2016.*

[20] Bradley Hayes and Julie A Shah. 2017. Improving Robot Controller Transparency Through Autonomous Policy Explanation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction.* ACM, 303–312.

[21] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems.* ACM, 159–166.

[22] O Khan, Pascal Poupart, J Black, LE Sucar, EF Morales, and J Hoey. 2011. Automatically generated explanations for Markov decision processes. *Decision Theory Models for Applications in AI: Concepts and Solutions* (2011), 144–163.

[23] Omar Zia Khan, Pascal Poupart, and James P Black. 2009. Minimal Sufficient Explanations for Factored Markov Decision Processes.. In *ICAPS.*

[24] Been Kim, Julie Shah, and Finale Doshi-Velez. 2015. Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction. In *Advances in Neural Information Processing Systems.*

[25] Shahar Kosti, David Sarne, and Gal A. Kaminka. 2014. A novel user-guided interface for robot search. In *Proc. of the International Conference on Intelligent Robots and Systems (IROS).* 3305–3310.

[26] Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).

[27] Meghann Lomas, Robert Chevalier, Ernest Vincent Cross II, Robert Christopher Garrett, John Hoare, and Michael Kopack. 2012. Explaining robot actions. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction.* ACM, 187–188.

[28] Moshe Mash, Raz Lin, and David Sarne. 2014. Peer-design agents for reliably evaluating distribution of outcomes in environments involving people. In *Proc. of the International conference on Autonomous Agents and Multi-Agent Systems (AAMAS).* 949–956.

[29] Deborah L McGuinness, Alyssa Glass, Michael Wolverton, and Paulo Pinheiro Da Silva. 2007. A Categorization of Explanation Questions for Task Processing Systems.. In *ExaCt.* 42–48.

[30] Deborah L McGuinness, Alyssa Glass, Michael Wolverton, and Paulo Pinheiro Da Silva. 2007. Explaining Task Processing in Cognitive Assistants That Learn.. In *AAAI Spring Symposium: Interaction Challenges for Intelligent Assistants.* 80–87.

[31] Karen L Myers. 2006. Metatheoretic Plan Summarization and Comparison.. In *ICAPS.* 182–192.

[32] Stefanos Nikolaidis and Julie Shah. 2013. Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction.* IEEE Press, 33–40.

[33] Donald A Norman. 1983. Some observations on mental models. *Mental models* 7, 112 (1983), 7–14.

[34] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386* (2016).

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proc. of ACM International Conference on Knowledge Discovery and Data Mining.* ACM, 1135–1144.

[36] Bastian Seegebarth, Felix Müller, Bernd Schattenberg, and Susanne Biundo. 2012. Making hybrid plans more clear to human users-a formal approach for generating sound explanations. In *Twenty-Second International Conference on Automated Planning and Scheduling.*

[37] Shirin Sohrabi, Jorge A Baier, and Sheila A McIlraith. 2011. Preferred Explanations: Theory and Generation via Planning.. In *AAAI.*

[38] Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, Press William, Saxenian AnnaLee, Shah Julie, Tambe Milind, and Teller Astro. 2016. Artificial intelligence and life in 2030. *One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel* (2016).

[39] Kristen Stubbs, Pamela J Hinds, and David Wettergreen. 2007. Autonomy and common ground in human-robot interaction: A field study. *IEEE Intelligent Systems* 22, 2 (2007).

[40] Leila Takayama, Doug Dooley, and Wendy Ju. 2011. Expressing thought: improving robot readability with animation principles. In *Proceedings of the 6th international conference on Human-robot interaction.* ACM, 69–76.

[41] Lisa Torrey and Matthew Taylor. 2013. Teaching on a budget: Agents advising agents in reinforcement learning. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems.* 1053–1060.

[42] Daniel Urieli and P. Stone. 2014. TacTex'13: A Champion Adaptive Power Trading Agent. In *Proceedings of the Twenty-Eighth Conference on Artificial Intelligence (AAAI'14).* 465–471.

[43] Prasanna Velagapudi, Jijun Wang, Huadong Wang, Paul Scerri, Michael Lewis, and Katia Sycara. 2008. Synchronous vs. Asynchronous Video in Multi-robot Search. In *ACHI'08.* 224–229.

[44] Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. 2012. Making machine learning models interpretable.. In *ESANN*, Vol. 12. 163–172.

[45] Huadong Wang, Andreas Kolling, Nathan Brooks, Sean Owens, Shafiq Abedin, Paul Scerri, Pei-ju Lee, Shih-Yi Chien, Michael Lewis, and Katia Sycara. 2011. Scalable target detection for large robot teams. In *HRI'11.* 363–370. https://doi.org/10.1145/1957656.1957792

[46] Huadong Wang, Prasanna Velagapudi, Paul Scerri, Katia Sycara, and Michael Lewis. 2009. Using Humans as Sensors in Robotic Search. *FUSION'09* (2009), 1249 – 1256.

[47] Ning Wang, David V Pynadath, and Susan G Hill. 2016. The impact of pomdp-generated explanations on trust and performance in human-robot teams. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems.* 997–1005.

[48] Michael Wellman, Amy Greenwald, and Peter Stone. 2007. *Autonomous bidding agents - strategies and lessons from the trading agent competition.* MIT Press.

[49] Holly A. Yanco and Jill L. Drury. 2006. Rescuing interfaces: A multi-year study of human-robot interaction at the AAAI Robot Rescue Competition. *Autonomous Robots* 22, 4 (Dec. 2006), 333–352. https://doi.org/10.1007/s10514-006-9016-5

[50] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180, 3 (2017), 689–722.

[51] Yu Zhang, Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. 2017. Plan explicability and predictability for robot task planning. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on.* IEEE, 1313–1320.