Generating Explanations for Autonomous Robots using Assumption-Alignment Tracking

Xuan Cao[1], Jacob W. Crandall[1], and Michael A. Goodrich[1]

***Abstract*— As the techniques of autonomous robots advance, there is an increasing demand for robots to provide explanations for their behavior. There are two commonly used explanation types. The first type emphasizes that a robot's policy is the best (or only) option that satisfies a specific property produced by its decision-making algorithms. The second explanation type is used when a robot fails and describes the cause of an error state that led to the failure. This paper proposes a new explanation type derived from a robot's proficiency self-assessment. The proposed explanation type not only supplements the first explanation type under typical operating conditions but also includes the second explanation type when the robot fails. The proposed explanation type is based on assumption-alignment tracking (AAT), a novel method for robot proficiency self-assessment. AAT provides three pieces of information for explanation generation: (1) assessment of assumptions veracity on which the robot's generators rely; (2) proficiency assessment measured by the probability that the robot will successfully accomplish its task; (3) counterfactual proficiency assessment computed by hypothetically varying assumptions. The information provided by AAT fits the situation awareness-based framework for explainable artificial intelligence. Examples of generated explanations are demonstrated using a simulated robot setting up a table with different blocks.**

I. Introduction

The rapid development of autonomous robots in recent years [1] has accelerated their deployment into the real world [2]–[5]. Consequently, there is an increasing demand for robots to provide *explanations* for their behavior, aiming to enhance their *transparency* [6] and *trustworthiness* [7].

There are two commonly used explanation types for robots. The first explanation type justifies that a robot's *policy* is the best (or only) outcome that its decision-making algorithms could produce that satisfies a specific property such as soundness or optimality [8]. The second explanation type describes the cause of an error state that led to a robot's *task failure* [9]. This paper adds to the two explanation types by proposing a new type of explanation from the perspective of robot *proficiency self-assessment* (PSA). The proposed explanation type both acts as a crucial supplement to the first explanation type and includes the second explanation type when the robot fails.

The proposed proficiency-oriented type of explanation is based on *assumption-alignment tracking* (AAT) [10], [11], a novel method of robot PSA. The basic idea of AAT is that a robot's proficiency is sensitive to how well its decision-making algorithms, which we call *generators,* align with the environment, its hardware, and the task. Assessing alignment is done by tracking the veracity of the assumptions upon which the robot's generators rely. Assumption veracity is performed by generator-specific functions called *assumption checkers*, which map the robot's observations (from sensory data) to boolean or real-valued alignment scores. A model encoding the correlation between the veracity of these assumptions and the robot's proficiency is then established following a data-driven approach.

AAT provides three useful pieces of information for generating explanations: (1) *veracity assessment* of the robot generator's assumptions; (2) *proficiency assessment* measured by the task success probability; and (3) *counterfactual proficiency assessment* computed by hypothetically varying the outputs of some assumption checkers. Thus, AAT provides information for each of the three situation awareness (SA) levels [12] used in Sanneman et al.'s explainable artificial intelligence (XAI) framework [13]: *perception*, *comprehension*, and *projection*.

This paper makes three contributions. First, a new PSA-based explanation type is proposed. Second, the relation between the proposed explanation type and the SA-based XAI framework is discussed. Third, examples of generating explanations using AAT are given. The examples use simple templates to convert the information provided by AAT to plain text. The examples are from a problem where a simulated robot sets up a table with different blocks.

II. Related Work

A. Impact of Explanations on Trust in Robots

Explanations for a robot's decision-making process and behavior enhance the robot's transparency [6] and increase humans' trust in the robot [14]. In a simulated human-robot team task formulated as a partially observable Markov decision problem, explanations for the robot's decision-making process enhanced the human's understanding of the robot and helped the human build proper trust in the robot, which improved the performance of the human-robot team [7], [15]–[17]. Similar effects of explanations on humans' trust were also observed in other task domains, such as a robot system opening bottles [18], a resource management task where humans were assisted by robots [19], and an interactive game-playing environment in which a human-robot team competed against a team of two humans [20], [21].

B. Desired Content and Properties of Explanations

In psychology, behavior can be classified as *unintentional* or *intentional* [22], [23]. Unintentional behavior is usually explained using the *cause* of the behavior, meaning the

[1]Computer Science Department, Brigham Young University, Provo, UT, USA. Contact: caoxuan8872@gmail.com

internal or external factors that made the behavior happen. By contrast, intentional behavior is usually explained using *enabling factors, causal history factors,* or *reasons* [22]–[27]. Keil [28] proposed that explanations can be categorized by the "explanatory stance" [29], or "mode of construal" [30] adopted for framing an explanation. Three explanatory stances were proposed in [29]: the mechanical, design, and intentional stances. Miller [31] argued that explainable AI can build on existing research in philosophy, psychology, and cognitive science about how people define, generate, select, evaluate, and present explanations. The psychological elements of belief, desire, and intention (BDI) are often used in agent design [32]. Harbers *et al.* [33] demonstrated that preferred explanations of virtual BDI agents focus on intention, i.e. beliefs and goals behind underlying behavior.

SA involves the *perception* of relevant elements in the current state of the system, the *comprehension* of what these elements mean, and a *projection* of future states of the system, and has shown to be crucial for good decision-making [12]. Chen *et al.* [34] proposed an SA-based model for increasing autonomous agent transparency. An SA-based framework for XAI was proposed in [13], suggesting that good explanations should provide the three SA information levels. In general, explanations for the perception level are about model input and output, explanations at the comprehension level are about the model itself, and explanations at the projection level are about model behavior in the future or how changes in model input would affect model output [13].

Sridharan and Meadows [35] advocated that explanations should (1) present context-specific information, (2) be able to provide online descriptions of decisions, rationale for decisions, knowledge, beliefs, and experiences, (3) not be task or domain specific, (4) consider human feedback and (5) be adjusted according to task execution. Hoffman *et al.* [36] proposed a few properties that make a good explanation, including accuracy, completeness, etc.

C. Methods for Explanation Generation

In the *Debrief* XAI system [37], an agent's internal state was continually stored during a mission and was reviewed after mission completion to generate explanations. Explanations for a specific decision were generated by recalling the agent's state in which the decision was made and systematically varying different aspects of the state to determine which aspects were critical to the decision. Similarly, the XAI system in [38] used logs of agents' activities to answer queries about the status of the agents and their tasks at any point in time during task execution.

For BDI agents, Harbers *et al.* [39] proposed that explanations can be generated by constructing a behavior log that stores the history of internal states of an agent and applying goal-based and belief-based explanation algorithms to the behavior log. A similar approach to explanation generation for BDI agents was proposed in [40].

Han et al. [41] proposed utilizing Behavior Trees (BTs), a tree structure that encapsulates behavior by control nodes that contain child execution nodes, to generate explanations. They adapted BTs by framing them as a set of semantic sets {goal, subgoals, steps, actions} and developed explanation generation algorithms that focus on causal information.

III. Explanation Types

This section formalizes a state-transition system, reviews the *assumption-alignment tracking* (AAT) method for PSA, defines three AAT-based explanation types, and discusses the relations among the three explanation types.

A. State-Transition System

Suppose a robot is running in an environment modeled as a discrete time state transition system, denoted by $E = S \times A \times S \times \mathbb{R}$, where S denotes a suitable state space, A denotes the robot's available actions, and $\mathbb{R}$ is a real number that denotes the reward for the transition. Elements of E are (present state, action, next state, reward) tuples.

The robot has a set of decision-making algorithms, or *generators*, denoted by $\mathbb{D} = \{\mathbb{D}_1, ..., \mathbb{D}_n\}$, which produces a *policy* $\pi = S \times A$ for the robot based on E subject to a desired *property* τ, i.e. $\mathbb{D} : E \times \tau \rightarrow \pi$. Examples of τ are soundness and optimality. Elements of π are (state, action) pairs, where each state is associated with at least one action.

Combining E and π produces a set of time-indexed state and action *trajectories* with elements, respectively,

$$\mathbf{s}_t = [s_0, s_1, \ldots, s_{t+1}] \text{ and } \mathbf{a}_t = [a_0, a_1, \ldots, a_t] \ .$$

where s_0 and a_0 are an initial state and action, respectively. Suppose the robot is tasked with reaching a specific *goal state*, denoted by $s^* \in S$. The robot successfully completes its task if the final state of a trajectory $s_{t+1} = s^*$, and otherwise fails. Let $o \in \{\text{success}, \text{failure}\}$ denote the *outcome* of a robot trial. Robot *proficiency* is measured by its success probability, denoted by $P(o = \text{success})$.

B. Assumption-Alignment Tracking

As asserted in many No-Free-Lunch Theorems (e.g., [42]), all generators for autonomous robots are based on assumptions or biases that dictate the generators' performance and affect the robots' proficiency [43]. In AAT [11], system designers identify the assumptions made by the robot's generators and create generator-specific functions, or *assumption checkers*, to track the veracity of these assumptions over time. Formally, suppose there are m assumptions on which the robot's generators rely and that each assumption has one checker that evaluates its veracity. Let $\mathbf{v}(t) = [v_1(t), \ldots, v_m(t)]$ denote the *veracity assessment vector* of the m checkers at time t. The time series of assessments $\mathbf{v}(t)$ generated in a single trial correlate to the robot's proficiency.

A proficiency assessment model that maps $\mathbf{v}$ to $P(o = \text{success})$ can be established using machine learning. Let $X = \{(\mathbf{v}_i, o_i)\}$ denote the training set collected beforehand where i represents a sample index, $\mathbf{v_i}$ represents the model input and o_i represents the training target. Data-driven algorithms such as k-Nearest Neighbor (kNN) and Random Forest [44] can be used to predict both o and $P(o = \text{success})$ given a novel input $\mathbf{v}$. Note the proficiency assessment model described above is a simplification of that proposed in [11].

TABLE I: Text templates for generating explanations using the information provided by AAT.

Information	Text Template(s)
Veracity Assessment	• *All my assumptions are satisfied.* • *The following assumptions are violated:* `[assumption 1], ..., [assumption k].`
Proficiency Assessment	• *My current success probability is about* `[proficiency assessment result]`, *therefore I am likely to succeed/fail in my task.*
Counterfactual Proficiency Assessment	• *My success probability would have been* `[proficiency assessment result]`, *if the status of the following assumptions are changed:* `[assumption 1], ..., [assumption k].`

C. Defining Three Explanation Types

This subsection defines three AAT-based explanation types. The first two types are from [8] and [9], respectively, and the third is contributed by this work. The notation is adapted from [9] and uses $\mathcal{E}$ to indicate an explanation plus a subscript to indicate the explanation type.

Explanation concerning policy property ($\mathcal{E}_\pi$). This type of explanation answers questions such as "Why π?" or "Why not π'?" The explanation is a justification that given the environment E and the property τ, π is either the only solution to $\mathbb{D}$ or better than any other solution π' with respect to some criteria such as cost and preferences. Unfortunately, $\mathcal{E}_\pi$ does not validate that $\mathbb{D}$ is suitable for E, which will be addressed by the third explanation type in Section III-D.

Explanation of failure cause ($\mathcal{E}_{\text{err}}$). This type of explanation is provided when the robot reaches a *failure state* that halts the execution of π. This explanation type answers questions such as "The robot is at the table, but why did it not pick up my beverage?" and provides the cause of the failure/error. As will be discussed in Section III-D, $\mathcal{E}_{\text{err}}$ is included in the third explanation type when the task fails.

Explanation by AAT ($\mathcal{E}_{\text{AAT}}$). This type of explanation indicates the alignment of $\mathbb{D}$ or π with E, and answers the question "Will executing π lead to success?" using the robot's generators $\mathbb{D}$ and proficiency estimate $P(o = \text{success})$. First, the veracity of the assumptions of $\mathbb{D}$ provides information about how suitable $\mathbb{D}$ is for E. Second, the estimation of $P(o = \text{success})$ provides information about how good π is with respect to task success.

D. Relationships Between the Three Explanation Types

Relationship of $\mathcal{E}_\pi$ and $\mathcal{E}_{\text{AAT}}$. $\mathcal{E}_\pi$ and $\mathcal{E}_{\text{AAT}}$ evaluate π using complementary perspectives. $\mathcal{E}_\pi$ focuses on how π was derived and emphasizes that π is the best outcome of $\mathbb{D}$ that satisfies τ given E. By contrast, $\mathcal{E}_{\text{AAT}}$ focuses on the outcome of executing π by providing the alignment between $\mathbb{D}$ and E and the probability of success for the robot. Combing $\mathcal{E}_\pi$ and $\mathcal{E}_{\text{AAT}}$ should produce more comprehensive information about π.

Relationship of $\mathcal{E}_{\text{err}}$ and $\mathcal{E}_{\text{AAT}}$. $\mathcal{E}_{\text{err}}$ can be elicited only after task failure while $\mathcal{E}_{\text{AAT}}$ can be elicited at any time during task execution. When a task failure occurs, the veracity assessment (see Sec. IV-A) part of $\mathcal{E}_{\text{AAT}}$ from a narrow time window before the task failure can be retrieved to produce $\mathcal{E}_{\text{err}}$. Thus, $\mathcal{E}_{\text{AAT}}$ includes $\mathcal{E}_{\text{err}}$.

IV. Explanation Generation

This section describes the information provided by AAT, describes the text templates for generating $\mathcal{E}_{\text{AAT}}$, and discusses the relationship between $\mathcal{E}_{\text{AAT}}$ and the SA framework.

A. Information Provided by AAT

AAT provides three pieces of information that could be used to generate $\mathcal{E}_{\text{AAT}}$. First, AAT *assesses the veracity* of each generator assumption. Second, AAT provides the robot's *proficiency assessment* by predicting the probability of success by combining the veracity assessment and the proficiency assessment model. Third, AAT allows *counterfactual proficiency assessment* for queries such as "what would the robot's proficiency have been if some assumptions are violated/not violated" by changing the veracity assessment and re-assessing the robot's proficiency.

B. Templates

As in [35], [41], simple first-person narrative templates are used to convert the information provided by AAT to plain English text, as shown in Table I.

C. $\mathcal{E}_{\text{AAT}}$ and the SA-Based XAI Framework

We claim that $\mathcal{E}_{\text{AAT}}$ addresses each of the three explanation levels in the SA-based XAI framework. Fig. 1 depicts the relation between $\mathcal{E}_{\text{AAT}}$ and the three explanation levels. Future work should evaluate the claims via user studies.

Perception Level. Veracity assessment is based on mapping robot sensors to assumption evaluations. Reporting which assumptions are satisfied or violated provides information about how the robot perceives the environment and itself. This type of information theoretically enables a user to perceive why the robot is behaving the way it is.

Comprehension Level. Proficiency assessment provides a success probability at any point in time during execution. Thus, proficiency assessment is a further interpretation of how well the situation in the world aligns with the robot's assumptions. This type of information theoretically enables a user to comprehend how the robot's history has shaped the robot's current attempts to succeed at a task.

Counterfactual-based proficiency assessment provides even further interpretation of how well the situation in the world aligns with the robot's assumptions. This type of information theoretically provides a contrast that can be used by a user to comprehend why the current situation is either compatible with success or makes success unlikely.

Projection Level. The success probability provided by proficiency assessment is an explicit projection of the future

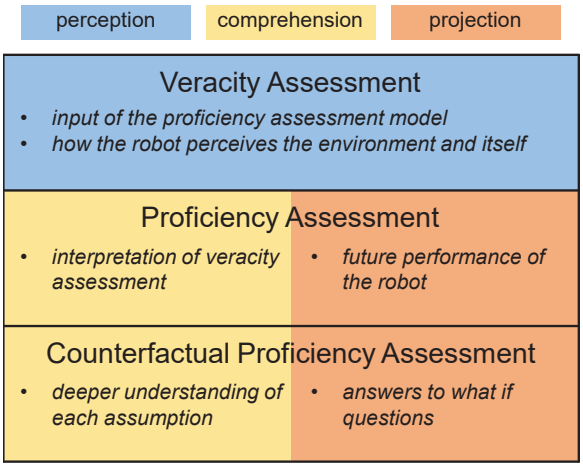

Fig. 1: The relation between the three pieces of information provided by AAT (represented by different blocks) and the three levels of explanations in the SA-based XAI framework (indicated by different colors as shown in the legend).

outcomes that the robot is able to produce. It provides a user with a prediction of how well the robot expects to perform, which theoretically allows a user to understand how the current situation in the world shapes the likely outcome of the robot's behaviors. Interestingly, counterfactual proficiency assessment theoretically allows a user to predict how changes in the world would affect likely outcomes.

V. Demonstration

This section presents a demonstration of how proficiency assessment can be used to generate explanations, $\mathcal{E}_{\text{AAT}}$, for a problem where a simulated robot system is tasked with setting up a table with different blocks.

A. Robot System

Fig 2 illustrates a task where a simulated robot must manipulate nine unique blocks with three shapes (square, circle, and triangle) and three colors (red, blue, and black). The robot must put those blocks in desired positions in the center area (dotted-line-square) of a table. The robot uses the AlegAATr algorithm [45] to choose one from the following actions at each step: (1) moving blocks that are outside of the center area into the center area, (2) flipping overturned blocks that are in the center area, (3) separating blocks that are next to each other, and (4) putting blocks that are in the center area into correct positions.

B. Data Collection

AAT data for the simulated robot are collected from 66 robot trials with different initial configurations of blocks on the table, and then randomly split into training and test datasets with a ratio of 7:3. The training set includes 31 success trials and 16 failure trials, while the test set includes 12 success trials and 7 failure trials.

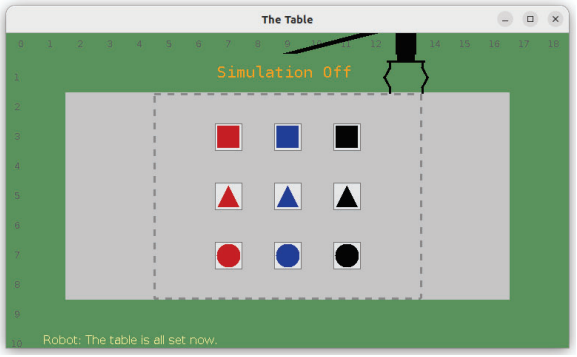

Fig. 2: A simulated robot setting up a table with various blocks.

C. Proficiency Assessment Model

The proficiency assessment model is implemented using the Random Forest algorithm. Replicating the methods from [46], two metrics, AUC-ROC (area under the ROC curve) and ECE (expected calibration error), are used to evaluate the OSA model. Higher AUC-ROC and lower ECE indicate better model performance. A perfect model is with an AUC-ROC of 1.0 and an ECE of 0%. The AUC-ROC and ECE of the proficiency assessment model are 0.916 and 8.6%, respectively.

D. Generated Explanations

We sampled explanations from the 12 successful and 7 unsuccessful trials in the test set. We looked at inflection points in the success probability, like those that happen around time-step 6 and time-step 11 in Figure 3, because inflection points occur when veracity assessments change. We subjectively evaluated the explanations before and after the inflection points to see whether the change in explanations allowed us to perceive and comprehend what was happening to the robot and what was likely to happen in the future. Fig. 3 illustrates typical results. Around time-step 6, the explanation (in the orange box) indicates a success probability of about 0.05, which would have been 0.75 if the "setup_grippable" assumption was not violated. Then around time-step 11 the explanation (in the green box) demonstrates that the success probability has increased to about 0.75 since the "setup_grippable" assumption has become satisfied, and would have improved to about 0.90 if the "scatter_reachable" assumption was not violated. These subjective evaluations support the claims made in Section IV-C, which hypothesizes that a user study should reveal a high correlation between successful responses to SA probes and the actual events in the world.

VI. Limitations and Future Work

A limitation with the subjective assessment is that we designed the generators, identified assumptions, and veracity assessment algorithms, which means that we understand what the assumption variables like "hasPossession" mean.

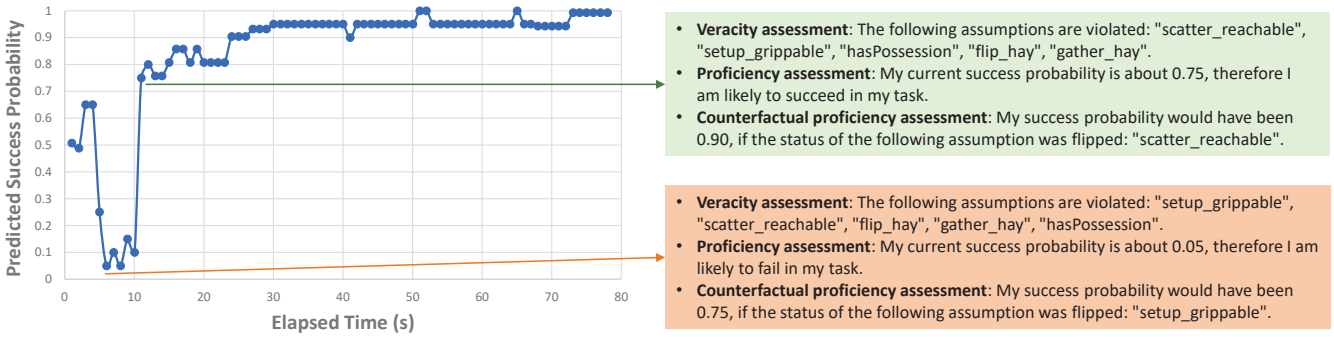

Fig. 3: $\mathcal{E}_{\mathrm{AAT}}$ generated during a robot trial.

A user study would need to map these abstract assumption variables to narrative sentences about what the variables mean in terms of what the robot is sensing, what is happening in the world, or what the robot believes it is doing. Future work should evaluate whether using large language models to replace the text templates and abstract assumption variables leads to better SA when compared to training users to understand the robot's abstract variables.

The information provided by AAT discussed in Sec. IV-A is independent of the specific algorithm used to establish the proficiency assessment model, and could potentially be enriched by the proficiency assessment model's own explainability. For example, if the proficiency assessment model is based on kNN, then the chosen nearest neighbor data points and their distances to the test data point could also contribute to the information for explanation generation. Such additional information belongs to the comprehension level of the SA-based XAI framework.

Techniques for explaining black-box models' outputs could also be applied to the proficiency assessment model to enhance the information provided by AAT for explanation generation. For example, SHAP (SHapley Additive exPlanations) [47] could be exploited to demonstrate how each assumption satisfaction/violation contributes to the success probability predicted by the proficiency assessment model. However, to apply SHAP to AAT for explanation generation, system designers must make sure that checkers are independent of one another, which is a precondition for efficiently approximating SHAP values [47].

The method of computing counterfactual proficiency assessment described in Sec. IV-A serves as an alternative way to measure how each assumption satisfaction/violation affects the proficiency assessment result that does not require checker independence. Computing counterfactual proficiency assessment is similar to the method for explanation generation proposed in [37], and future work should evaluate how much explanations based on counterfactual proficiency assessments contribute to or detract SA probes.

Malle's framework [22] distinguishes intentional and unintentional behavior. This paper simply considers a robot's behavior as unintentional and leaves the relation between $\mathcal{E}_{\mathrm{AAT}}$ and Malle's framework for future work. Future work should look deeper into the relation between $\mathcal{E}_{\mathrm{AAT}}$ and Malle's framework.

VII. Summary

This paper proposes a new type of explanation that is complementary to existing explanation types from the perspective of a robot's proficiency. The proposed type of explanation is based on assumption-alignment tracking (AAT), which provides three pieces of proficiency-related information for explanation generation: (1) veracity assessment of the assumptions on which the robot's generators rely; (2) proficiency assessment measured by the probability that the robot will accomplish its task; (3) counterfactual proficiency assessment computed with the veracity of some assumptions varied hypothetically. The information provided by AAT covers the three levels of the situation awareness-based framework for XAI. Examples of generated explanations are demonstrated using a simulated robot setting up a table with different blocks.

Future work should also look deeper into the relations between the proposed explanation type and existing explanation types. Another interesting direction of future work is using large language models to replace the text templates for text generation.

Acknowledgment

This work was supported by the U.S. Office of Naval Research (grants N00014-18-1-2503 and N00014-16-1-3025).

References

[1] N. Maslej, L. Fattorini, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, H. Ngo, J. C. Niebles, V. Parli, Y. Shoham, R. Wald, J. Clark, and R. Perrault, "The AI index 2023 annual report," tech. rep., AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2023.

[2] G. P. Moustris, S. C. Hiridis, K. M. Deliparaschos, and K. M. Konstantinidis, "Evolution of autonomous and semi-autonomous robotic surgical systems: a review of the literature," *The international journal of medical robotics and computer assisted surgery*, vol. 7, no. 4, pp. 375–392, 2011.

[3] C. D. Bellicoso, M. Bjelonic, L. Wellhausen, K. Holtmann, F. Günther, M. Tranzatto, P. Fankhauser, and M. Hutter, "Advances in real-world applications for legged robots," *Journal of Field Robotics*, vol. 35, no. 8, pp. 1311–1326, 2018.

[4] S. Srinivas, S. Ramachandiran, and S. Rajendran, "Autonomous robot-driven deliveries: A review of recent developments and future directions," *Transportation research part E: logistics and transportation review*, vol. 165, p. 102834, 2022.

[5] S. Zhao, Q. Wang, X. Fang, W. Liang, Y. Cao, C. Zhao, L. Li, C. Liu, and K. Wang, "Application and development of autonomous robots in concrete construction: Challenges and opportunities," *Drones*, vol. 6, no. 12, p. 424, 2022.

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

[7] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 109–116, IEEE, 2016.

[8] T. Chakraborti, S. Sreedharan, and S. Kambhampati, "The emerging landscape of explainable AI planning and decision making," *arXiv preprint arXiv:2002.11697*, 2020.

[9] D. Das, S. Banerjee, and S. Chernova, "Explainable AI for robot failures: Generating explanations that improve user assistance in fault recovery," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 351–360, 2021.

[10] A. Gautam, T. Whiting, X. Cao, M. A. Goodrich, and J. W. Crandall, "A method for designing autonomous robots that know their limits," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 121–127, IEEE, 2022.

[11] X. Cao, A. Gautam, T. Whiting, S. Smith, M. A. Goodrich, and J. W. Crandall, "Robot proficiency self-assessment using assumption-alignment tracking," *IEEE Transactions on Robotics*, 2023.

[12] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human factors*, vol. 37, no. 1, pp. 32–64, 1995.

[13] L. Sanneman and J. A. Shah, "A situation awareness-based framework for design and evaluation of explainable AI," in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9–13, 2020, Revised Selected Papers 2*, pp. 94–110, Springer, 2020.

[14] A. Ezenyilimba, M. Wong, A. Hehr, M. Demir, A. Wolff, E. Chiou, and N. Cooke, "Impact of transparency and explanations on trust and situation awareness in human–robot teams," *Journal of cognitive engineering and decision making*, vol. 17, no. 1, pp. 75–93, 2023.

[15] N. Wang, D. V. Pynadath, S. G. Hill, and A. P. Ground, "Building trust in a human-robot team with automatically generated explanations," in *Proceedings of the interservice/industry training, simulation and education conference (I/ITSEC)*, vol. 15315, pp. 1–12, 2015.

[16] N. Wang, D. V. Pynadath, and S. G. Hill, "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*, pp. 997–1005, 2016.

[17] N. Wang, D. V. Pynadath, E. Rovira, M. J. Barnes, and S. G. Hill, "Is it my looks? or something I said? the impact of explanations, embodiment, and expectations on trust and performance in human-robot teams," in *Persuasive Technology: 13th International Conference, PERSUASIVE 2018, Waterloo, ON, Canada, April 18-19, 2018, Proceedings 13*, pp. 56–69, Springer, 2018.

[18] M. Edmonds, F. Gao, H. Liu, X. Xie, S. Qi, B. Rothrock, Y. Zhu, Y. N. Wu, H. Lu, and S.-C. Zhu, "A tale of two explanations: Enhancing human trust by explaining robot behavior," *Science Robotics*, vol. 4, no. 37, p. eaay4663, 2019.

[19] L. Zhu and T. Williams, "Effects of proactive explanations by robots on human-robot trust," in *Social Robotics: 12th International Conference, ICSR 2020, Golden, CO, USA, November 14–18, 2020, Proceedings 12*, pp. 85–95, Springer, 2020.

[20] M. Javid, V. Estivill-Castro, and R. Hexel, "Enhancing humans trust and perception of robots through explanations," *Proceedings of the ACHI*, 2020.

[21] M. Javaid and V. Estivill-Castro, "Explanations from a robotic partner build trust on the robot's decisions for collaborative human-humanoid interaction," *Robotics*, vol. 10, no. 1, p. 51, 2021.

[22] B. F. Malle, "How people explain behavior: A new theoretical framework," *Personality and social psychology review*, vol. 3, no. 1, pp. 23–48, 1999.

[23] M. J. O'Laughlin and B. F. Malle, "How people explain actions performed by groups and individuals.," *Journal of personality and social psychology*, vol. 82, no. 1, p. 33, 2002.

[24] B. F. Malle, "Folk explanations of intentional action," *Intentions and intentionality: Foundations of social cognition*, pp. 265–286, 2001.

[25] B. F. Malle, *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press, 2006.

[26] B. F. Malle, "Time to give up the dogmas of attribution: An alternative theory of behavior explanation," in *Advances in experimental social psychology*, vol. 44, pp. 297–352, Elsevier, 2011.

[27] M. M. De Graaf and B. F. Malle, "How people explain action (and autonomous intelligent systems should too)," in *2017 AAAI Fall Symposium Series*, 2017.

[28] F. C. Keil, "Explanation and understanding," *Annual review of psychology*, vol. 57, p. 227, 2006.

[29] D. C. Dennett, *The intentional stance*. MIT press, 1987.

[30] F. C. Keil, "The growth of causal understandings of natural kinds.," 1995.

[31] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[32] A. S. Rao, M. P. Georgeff, *et al.*, "BDI agents: from theory to practice.," in *Icmas*, vol. 95, pp. 312–319, 1995.

[33] M. Harbers, K. v. d. Bosch, and J.-J. C. Meyer, "A study into preferred explanations of virtual agent behavior," in *International Workshop on Intelligent Virtual Agents*, pp. 132–145, Springer, 2009.

[34] J. Y. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia, and M. Barnes, "Situation awareness-based agent transparency," tech. rep., U.S. Army Research Laboratory, 2014.

[35] M. Sridharan and B. Meadows, "Towards a theory of explanations for human–robot collaboration," *KI-Künstliche Intelligenz*, vol. 33, no. 4, pp. 331–342, 2019.

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

[37] W. L. Johnson, "Agents that learn to explain themselves," in *AAAI*, pp. 1257–1263, Palo Alto, CA, 1994.

[38] M. Van Lent, W. Fisher, and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in *Proceedings of the national conference on artificial intelligence*, pp. 900–907, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

[39] M. Harbers, K. van den Bosch, and J.-J. Meyer, "Design and evaluation of explainable BDI agents," in *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 2, pp. 125–132, IEEE, 2010.

[40] J. Broekens, M. Harbers, K. Hindriks, K. v. d. Bosch, C. Jonker, and J.-J. Meyer, "Do you get it? User-evaluated explainable BDI agents," in *German Conference on Multiagent System Technologies*, pp. 28–39, Springer, 2010.

[41] Z. Han, D. Giger, J. Allspaw, M. S. Lee, H. Admoni, and H. A. Yanco, "Building the foundation of robot explanation generation using behavior trees," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 3, pp. 1–31, 2021.

[42] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.

[43] L. S. Fletcher, S. Teller, E. B. Olson, D. C. Moore, Y. Kuwata, J. P. How, J. J. Leonard, I. Miller, M. Campbell, D. Huttenlocher, A. Nathan, and F.-R. Kline, *The MIT – Cornell Collision and Why It Happened*, pp. 509–548. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[44] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[45] E. Pedersen, "Alegaatr the bandit," Master's thesis, Brigham Young University, 2023.

[46] X. Cao, J. W. Crandall, and M. A. Goodrich, "Improving robot proficiency self-assessment via meta-assessment," *IEEE Robotics and Automation Letters*, vol. 8, no. 11, pp. 7297–7303, 2023.

[47] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 4765–4774, Curran Associates, Inc., 2017.