2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Planning of Explanations for Robot Navigation

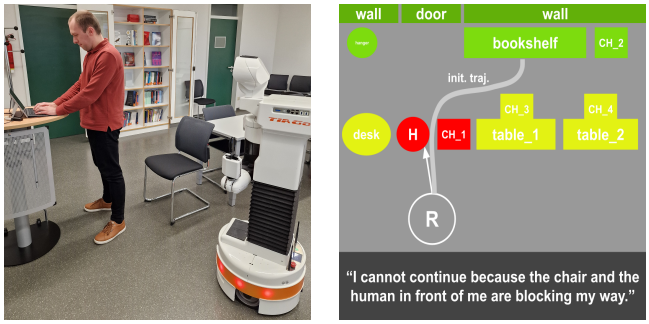
Amar Halilovic[1] and Senka Krivic[2]

***Abstract*— The choices made by autonomous robots in social settings bear consequences for humans and their presumptions of robot behavior. Explanations can serve to alleviate detrimental impacts on humans and amplify their comprehension of robot decisions. We model the process of explanation generation for robot navigation as an automated planning problem considering different possible explanation attributes. Our visual and textual explanations of a robot's navigation are influenced by the robot's personality. Moreover, they account for different contextual, environmental, and spatial characteristics. We present the results of a user study demonstrating that users are more satisfied with multimodal than unimodal explanations. Additionally, our findings reveal low user satisfaction with explanations of a robot with extreme personality traits. In conclusion, we deliberate on potential future research directions and the associated constraints. Our work advocates for fostering socially adept and safe autonomous robot navigation.**

I. Introduction

Offering explanations for robot actions has been shown to yield favorable effects on both human trust [1] and comprehension [2]. Moreover, it plays a pivotal role in nurturing effective human-robot interaction (HRI) [3]. An explainable robot is also perceived as more socially adept [4]. In alignment with this perspective, the IEEE Guideline for Ethically Aligned Design [5] advocates for the integration of *accountability*, *transparency*, and *justifications* in robots' design and decision-making. In the realm of various robot behaviors, our focus lies on motion planning in indoor environments and on explaining robot navigation to humans [6]–[9]: Consider an autonomous robot scenario where the task at hand involves retrieving a book from a library shelf and navigating through its surroundings while contending with potential obstacles. Unforeseen events, such as the abrupt presence of an obstruction in its path (e.g., an unexpected chair or human), can cause the robot's planning system to diverge from its original trajectory or encounter failures (as depicted in Fig. 1a). Such behavioral deviations have the potential to catch individuals interacting with the robot off guard, leading to a diminished level of confidence in the robot's intentions. To address this concern and cultivate a more secure environment for humans and robots, it becomes imperative for the robot to furnish explanations for its actions [10]. Figure 1b presents a multimodal explanation illustrating the robot's behavior in the context of the depicted failure scenario. The essence of multimodality resides in the simultaneous provision of various types of explanations, encompassing both visual and textual elements.

[1]Amar Halilovic is with the Institute of Artificial Intelligence, Ulm University, 89081 Ulm, Germany, amar.halilovic@uni-ulm.de
[2]Senka Krivic is with the Faculty of Electrical Engineering, University of Sarajevo, 71000 Sarajevo, BiH, senka.krivic@etf.unsa.ba



(a) Navigation scenario (b) Corresponding explanation

Fig. 1: The autonomous robot (TIAGo by PAL Robotics, marked by the white-bordered circle in visual explanation in Fig. 1b) encounters a stoppage trying to follow its original path (depicted in light grey in Fig. 1b). The motion planner cannot find an alternative path triggered by the presence of the chair in front of the robot and the nearby human. The primary causes (the chair and the human) of this planning failure are emphasized in the visual explanation (highlighted in red). Other objects are color-coded with a heat map, which reflects their distances from the robot. The textual explanation provides the key reason (in terms of objects) underlying the robot's stoppage.

We characterize a robot's personality by quantifying its levels of extroversion and introversion [8]. In the context of extroversion-introversion traits, studies [11], [12] have revealed that individuals prefer robots with personalities akin to their own. Robots must possess the capacity to comprehend both their intrinsic mental attributes and the surrounding environment. Given that extroversion and introversion represent two sides of the same personality trait spectrum, we employ the term "extroversion" predominantly throughout the remainder of this paper. We conceptualize the process of generating explanations as an AI planning problem, where explanations are modeled as one of the possible actions of a robot. We adopt the PDDL2.1 [13] to model the robot's decision-making as a deterministic process within this context. The main contributions of this paper are:

- We formulate the process of explanation generation as an automated planning problem and render the resulting multimodal explanations through a visual explanation layer coupled with a textual interface (Section III);
- We test the impact of visual (unimodal) and visual-textual (multimodal) explanations on user satisfaction (Section IV);
- We measure user satisfaction with the explanations of robots with extreme personality traits (Section IV).

The structure of this paper is following: In Section II, we examine prior research. Section III introduces our explanation methodology, emphasizing the incorporation of environmental context and the robot's personality into the explanation generation planning process. Section IV presents experimental results. Finally, Section V discusses the limitations and broader implications of our research.

II. Related work

Our research lies at the crossroads of explainable navigation, the influence of personality on explanations in HRI, and automated planning. As such, we review relevant literature encompassing *Explainable Autonomous Navigation*, *The Influence of Personality in HRI*, and *Explainable AI Planning* to construct a holistic comprehension of the previous research.

a) Explainable Autonomous Navigation: Several approaches have integrated natural-language explanations. Noteworthy instances include works by Perera et al. [14] and Rosenthal et al. [15], which center on elucidating complete global trajectories through narrative summaries of robot paths. Gavriilidis et al. [16] employ surrogates to generate explanations agnostic to autonomous agents' behavior, deconstructing behavior into natural-language components. Furthermore, Das et al. [17] employ natural-language explanations to bolster human support in rectifying navigational errors. Stein's research [18] delves into model-informed natural language explanations, utilizing the algorithms underpinning navigation. Robb et al. [19] explain navigational planning failures by checking users' cognitive frameworks when confronted with explanations of failures of remote navigational robotic agents. In a related vein, Brandao et al. [20] contribute a taxonomy for explainable motion planning, formalizing explanations not only for instances of planning failures but also encompassing deviations and trajectory preferences. Halilovic and Lindner investigate local deviations from initial paths, employing visual [6] and visual-textual [7] explanations. Furthermore, a study of global path optimality is explored in [21], where authors present to users an alternative non-optimal global path while comparing it against the optimal counterpart. Building upon our prior research efforts [6]–[8], we extend our contributions by modeling the generation of real-time multimodal explanations of its navigational decisions as an automated planning formulation. The multimodal approach encompasses both visual and textual explanations, where timing, representation, and duration of explanations are dynamically influenced by the robot's personality and its perception of the social context in its immediate environment.

b) The Influence of Personality in HRI: Extroversion, a potent and finely delineated trait, significantly molds human peer relationships [22]. It holds substantial relevance in the HRI as well [23]. Extroverted individuals tend to exhibit heightened energy levels and an inclination for speaking more loudly, quickly, and at a higher pitch [23]. This disposition is often characterized by minimal pauses, employment of concise sentences, and utilization of simpler vocabulary [24]. Notably, extroverts are prone to initiating conversations more frequently [25], with an affinity for focusing on personal topics and self-discussion, as opposed to discussions about others [26]. Furthermore, their discourse leans toward more positive language usage [27] and an elevated propensity to accept intrusion into their personal space, a trait distinct from introverted individuals [28]. To our knowledge, there is no previous research on the influence of personality and specifically extroversion on explanations in HRI.

c) Explainable AI Planning (XAIP): In motion planning literature, explanations encompass failure instances and contrastive scenarios. Failure explanations explain the cause of failure, while contrastive explanations elucidate why a planner selects trajectory A over an expected trajectory B. User queries regarding plans typically take a contrasting form, such as "why A over B?" [29]. Although there is research on explaining plans, there is no research on modeling the explanation process as a planning problem to our knowledge. Brandao et al. [20], [30] present a preliminary taxonomy of explainable motion planning techniques, encompassing failure and trajectory-contrastive explanations as most researched explanation approaches. We adopt their taxonomy and propose modeling the process of explanation generation in terms of automated plans.

III. Explanations of Robot Navigation

Robots should be able to explain their actions even when they have not made any errors to reduce the probability of unwarranted criticism towards themselves [31]. Nevertheless, the research on explainable robotics remains rather limited [32]. Addressing this, in [8], we introduced "HiXRoN", a hierarchical framework (as depicted in Fig. 2) for generating contextually tailored explanations of robot navigation. The modularity inherent in the HiXRoN framework facilitates adaptable communication interfaces among recipients, robots, and their environment. Its explanation generation process is organized hierarchically and sequentially, with each step enhancing the explanation in distinct dimensions:

- **Step 1:** *HiXRoN*'s input encompasses a robot model, navigation framework, and the environment ontology.
- **Step 2:** Selecting an explanation target (*what to explain?*) involves scoping through context comprehension and inputs. It forms the explanation core, which is later refined using temporal and qualitative strategies.
- **Step 3:** Timing the explanation (*when to explain?*) is pivotal for human communication. The delivery moment depends on contextual factors.
- **Step 4:** After proper timing, the chosen explanation representation (*how to explain?*) significantly influences its perception and comprehension.
- **Step 5:** The fifth step entails the determination of the explanation's duration (*how long to explain?*) based on the context and the user's state.
- **Step 6:** The recipient receives socially and contextually attuned explanation via an explanation interface.
- **Step 7:** Users can revisit the whole process, incorporating additional queries and hyperparameters if desired.

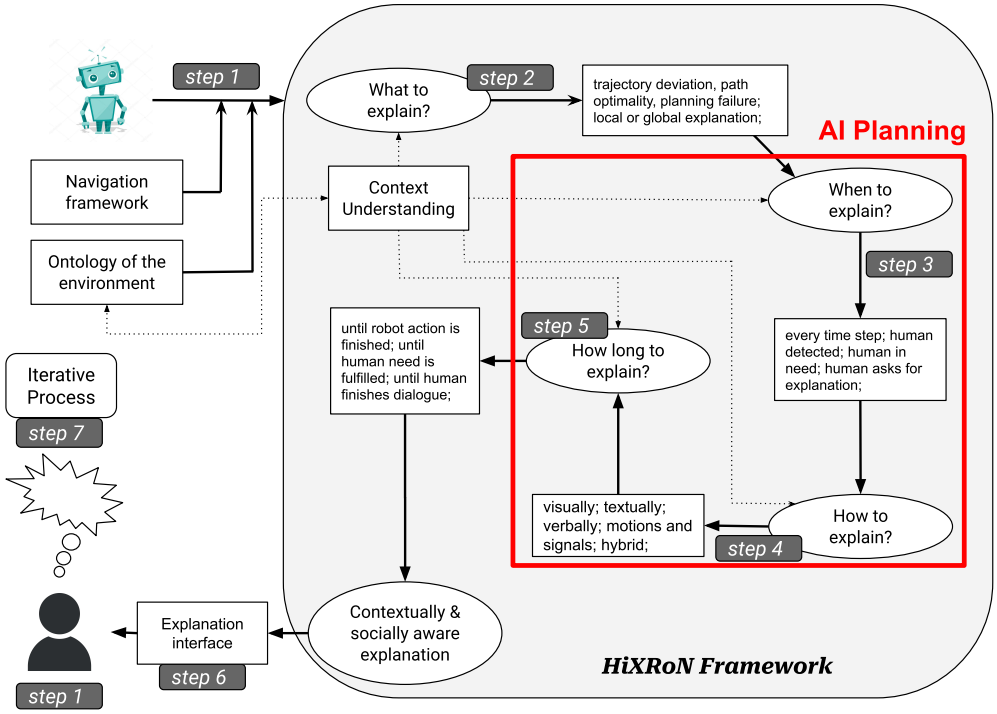


Fig. 2: **HiXRoN** - **Hi**erarchical e**X**plainable **Ro**bot **N**avigation Framework. We are modeling three hierarchical levels of explanation generation in HiXRoN (timing, representation, and duration) with automated planning.

In this paper, our primary focus lies in delivering real-time multimodal explanations while modeling the impact of the robot's personality on these explanations through automated plans. The concept of multimodality is realized through the utilization of both visual and textual explanations, wherein their combined characteristics collectively shape the representation of the explanation. We represent a robot's personality through its extroversion level. Within our framework, we model how this personality trait influences the key explanation aspects: timing (Step 3), representation (Step 4), and duration (Step 5) (see Fig. 2).

A. Visual explanations

We represent visual explanations through a bird-eye-view local visual *explanation layer* that envelops the robot at each moment during its navigation. Within this layer, objects, that are potential obstacles in the robot's vicinity, are visually annotated based on their proximity to the robot, effectively forming a localized heat map. We employ a color scheme inspired by the yellow-green heat map [35] (refer to Fig. 1b). Objects situated closer to the robot, indicating a higher likelihood of becoming obstacles, are shaded closer to yellow, whereas objects farther away lean toward a green hue. The color red is reserved for instances when an object transforms into an obstacle, inducing a substantial alteration in the robot's navigational behavior, i.e., planning failure or trajectory deviation. This color scheme aligns with basic principles of color psychology, where red serves as a potent indicator of a critical event (failure or deviation in robot motion planning), while yellow and green, sequentially increasing in effect, signify a more tranquil and stable state.

B. Textual explanations

In addition to furnishing visual explanations, the robot also generates corresponding textual explanations. As the robot navigates, our system continuously tracks the positions of both the robot and nearby objects, keeping the foundational ontology updated (see [7], [8]). It facilitates the creation of spatial and semantic links between objects and the robot. The textual explanations we provided were made up of brief, clear statements designed to effectively communicate information to those receiving the explanations. These statements aimed to shed light on the robot's current activities and the nearby objects, offering clarity on why the robot might have strayed from its planned path or encountered planning issues (refer to Fig. 3a). While we developed these textual explanations, our main focus was on following established guidelines from existing literature, rather than inventing new textual explanations approaches.

C. The influence of robot personality on explanations

We translate extroversion properties, outlined in Section II, into a mental model for a robot's explanations, one that is influenced by both its extroversion level and the contextual presence of people around it. Our representation of the robot's extroversion level is graded on a scale ranging from 0 to 1, with increments of 0.1, effectively defining a discretized, stepwise "extroversion probability". Additionally, the robot maintains an understanding of social space, adhering to Hall's theory of social spaces [34] (as depicted in Fig. 3b). Within this framework, the robot's social spaces are categorized into four zones, including the personal and social zones. Typically, the social zone extends from distances

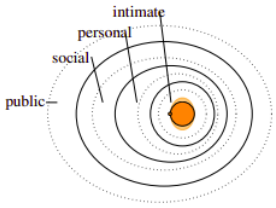
of 1.2 to 3.6 meters from the robot. In our model, we ascertain the presence of humans based on the breach of the robot's "explanation representation threshold", which is determined as a linear function of the extroversion level: $explanation\_representation\_threshold = 3.6 - 2.4 \times extroversion\_probability$. For a robot classified as entirely extroverted, this threshold amounts to 1.2 meters, whereas for a completely introverted robot, it stands at 3.6 meters. Consequently, a totally extroverted robot adjusts its explanation representation when its personal zone is breached, while a totally introverted robot does so upon breaching its social zone. This aligns with the observation that extroverts tend to be more accommodating of intrusions into their personal space compared to introverts.

Both a robot's personality and the presence of humans influence the explanation generation process. Drawing from the observation that extroverted individuals tend to minimize pauses, the timing, denoting the delay in articulating an explanation after its formation, is modeled as a linear function of the robot's extroversion level. Consequently, a more extroverted robot exhibits shorter delays or pauses in its explanations. Given that extroverted individuals speak more rapidly, this leads to shorter explanations by extroverted robots, contrasting with the longer explanations typically provided by introverted counterparts. Both the timing and duration are computed as descending linear functions of extroversion probability: $timing(duration) = N - N \times extroversion\_probability + 1$, where N is the maximum delay whose value is implementation-dependent.

The size of the visual explanation layer corresponds directly to the dimensions of the explanation representation zone. It is in line with the notion that more extroverted robots formulate less detailed explanations, maintaining a narrower visual explanation layer (narrower explanation scope). Furthermore, the presence of humans impacts the explanation representation. Robots tend to adopt a more reserved approach when their explanation representation zone

- **Heading straight.**
- **The chair_8 has been moved.**
- **Obstruction: chair_8.**
- **Rerouting.**
- **New route found.**
- **Right turn between bookshelf_5 and bookshelf_6.**

(a) Textual explanations



(b) Zones of Hall's social space

Fig. 3: **a)** Textual explanations consist of carefully crafted, straightforward statements that provide participants with information about the robot's current action, objects in the vicinity, and the rationale behind any unexpected behavior it exhibits; **b)** A robot's social space is divided into four zones [33], [34]: intimate, personal, social and public, respectively increasing in the distance from the robot. Each zone can be divided into two regions: near and far.

Algorithm 1 The influence of robot extroversion and human presence on explanation representation

if distance_to_human $<$ exp_representation_threshold **then**
    representation: visual
else
    representation: visual + textual
end if

is breached, opting to rely solely on visual representation (see Algorithm 1). A breach is detected if the distance between the robot and the nearest human falls below the explanation representation threshold, which is determined by the extroversion probability.

D. Planning of Explanations

We pose the problem of deciding what, when, how, and how long to generate an explanation as part of the problem of a robot's high-level decision-making system. We use AI planning formalisms to model the robot's world and its capabilities, including explaining. The planning model describes a transition system consisting of *(a)* a set of possible states of the world, S; *(b)* a set of possible (labeled) transitions between these states, $T \subseteq S \times S$; (*c*) an initial state, $s_0 \in S$; and *(d)* a set of goal states, $G \subseteq S$. A plan is a sequence of transitions (each corresponding to an action) that leads from the initial state s_0 to some goal state $s_f \in G$ [36].

We use the Planning Domain Definition Language (PDDL) [37] to represent the planning problem of our librarian robot (refer to Fig. 1). PDDL splits the definition of a planning problem into two parts: the domain file D, and the problem file P. We model explanation generation actions within a domain file, among other actions within this scenario. In the problem file, specific initial and goal values are initialized to steer the planning process from the current or given initial state toward the goal or wanted state.

To solve the planning problem, we use the forward search temporal planner POPF [38]. The current state is taken as the initial state. While the plan is executing the state changes are observed. In case of deviations from the plan such as action failures or external events, the replanning is performed by taking the current state as the initial state. The planning approach has been implemented in a system that integrates HiXRoN with a task planning framework in Robot Operating System (ROS), ROSPlan [39], used for generating and dispatching plans on a robot as well as replanning.

Figure 4 shows two explanation plans of the planning failures of an introverted and extroverted robot. For both introverted and extroverted robots the domain file is the same, while the problem (instance) file determines the robot's personality traits. The main personality variable is *extroversion probability*, which is defined as a real-valued variable in the domain and initialized by a concrete value in the problem file. This means that for every robot with different extroversion levels, a new problem file should be defined. There is a possibility for a fluctuating value of extroversion probability, but in this case, the fluctuation strategy should

```
0.000: (detect_failure tiago)  [1.000]
1.000: (detect_human tiago explainee)  [1.000]
2.000: (explain_failure_start tiago explainee) [1.000]
3.000: (calculate_explanation_timing tiago) [1.000]
4.000: (pick_visual_textual_representation tiago explainee) [1.000]
5.000: (calculate_explanation_duration tiago) [1.000]
6.000: (wait tiago)  [1.000]
7.000: (explain_failure tiago explainee) [1.000]
8.000: (explain_failure_ended tiago explainee)  [1.000]
```

(a) Explanation plan of a totally extroverted robot

```
0.000: (detect_failure tiago)  [1.000]
1.000: (detect_human tiago explainee)  [1.000]
2.000: (explain_failure_start tiago explainee) [1.000]
3.000: (calculate_explanation_timing tiago) [1.000]
4.000: (pick_visual_representation tiago explainee) [1.000]
5.000: (calculate_explanation_duration tiago) [1.000]
6.000: (wait tiago)  [11.000]
17.000: (explain_failure tiago explainee) [11.000]
28.000: (explain_failure_ended tiago explainee)  [1.000]
```

(b) Explanation plan of a totally introverted robot

Fig. 4: The figures illustrate possible plans for the example failure situation illustrated in Fig. 1. The robot first detects the failure and a nearby human (explainee) and then starts the explanation generation process. First, it calculates the explanation timing value, proceeds with choosing an explanation representation, and finishes with calculating the explanation duration value. Explanation timing is expressed as the duration of the durative action *wait*, while explanation duration is expressed as the duration of the durative action *explain*. Both these values are calculated from a robot extroversion probability value. The durations of other functions are set to 1.0 for the sake of simplicity. The chosen explanation representation is realized by low-level actionable code; **(a)** The robot has an extroversion probability of 1.0. Its explanation representation threshold is small and it usually produces visual-textual explanations. In this example, its distance to the explainee was bigger than the threshold. The robot's timing and duration are both 1 second (N=10; see red-bordered rectangular areas); **(b)** The robot has an extroversion probability of 0.0. Its explanation representation threshold is higher and it is expected to more often produce visual-textual explanations than the extroverted counterpart. In this example, its distance to the explainee was lower than the threshold. The robot's timing and duration are both 11 seconds (N=10; see red-bordered rectangular areas).

be defined by an external code. The ROSPlan sensing interface offers the possibility of continuous sensing and an update of extroversion probability. However, personality fluctuation control is a topic for future work. Extroversion probability directly influences the explanation timing and duration values, which we express as durations of actions (see Fig. 4). However, in PDDL, the durations of actions must be initialized in the domain file, as PDDL does not offer the action duration calculation during planning time. This means, that for every problem file, timing and duration values (and actions' durations) have to be calculated and manually coded in problem and domain files. This does not pose a problem in our setting but could present a challenge for planning applications with dynamic personality, where extroversion probability is sensed. Explanation representation threshold is also calculated using the extroversion probability and it stays the same for one instance, i.e., planning problem. Distance between the robot and the person interacting with the robot during explaining (explainee) is calculated from the robot sensor measurements and sensed by ROSPlan. The if-then logic for determining explanation representation (see Algorithm 1) can be expressed as a precondition in actions for choosing explanation representation. An explanation instantiation is achieved by the low-level action interfaces connected to the high-level planner interface.

IV. Evaluation

A. User Study on Explanation Representation

We conducted a comparative user study to assess the impact of explanation representation on user explanation satisfaction. Our study's design is grounded in the recommendation of Hoffman et al. [40] regarding metrics for Explanation Satisfaction, which, in this context, quantifies users' perceived satisfaction with the explanations of robot navigation. The study was conducted with 84 volunteers divided into two groups with 42 persons each. Participants' ages ranged from 19 to 40 years. The mean age of the participants was 22.75 ($SD = 4.23$) years, with 60.71% identifying as female and 38.10% identifying as male, while one person identified as other. Participants were randomly allocated to two distinct groups: the control group (G1), which was presented with visual explanations, and the experimental group (G2), which engaged with visual-textual explanations. A description of the environment and the robot's task was provided to all participants. Both groups were tasked with viewing a video depicting a robot's navigation journey from its initial position to a coffee machine in an office setting. Along this route, the robot encountered obstacles, prompting deviations from its intended path and eventual halts due to path planning system failures. The robot was explaining failure and deviations without the influence of its personality. We employed the Explanation Satisfaction Scale to assess participants' satisfaction with the explanations generated by the framework. This scale, developed by Hoffman, Klein, and Mueller [40] and tailored for Explainable AI systems, follows a 5-point Likert scale format, ranging from "strongly disagree" to "strongly agree," correlating linearly to values between 1.0 and 5.0. Following the Explanation Satisfaction scale, we used seven Likert scale questions given in Table I.

Quantitative Results and Discussion: The evaluation results with the Explanation Satisfaction Scale are shown in Figure 5. The mean of all responses for Group 1 is 4.24 ($SD = 0.99$), corresponding to the overall attitude of "somewhat agree" that visual explanations are satisfying. The mean of all responses for Group 2 is 4.62 ($SD = 0.58$), corresponding to the overall attitude of "strongly agree" that visual-textual explanations are satisfying. We also performed a t-test ($t = -5.6$, $p = 3.28e-08$). We

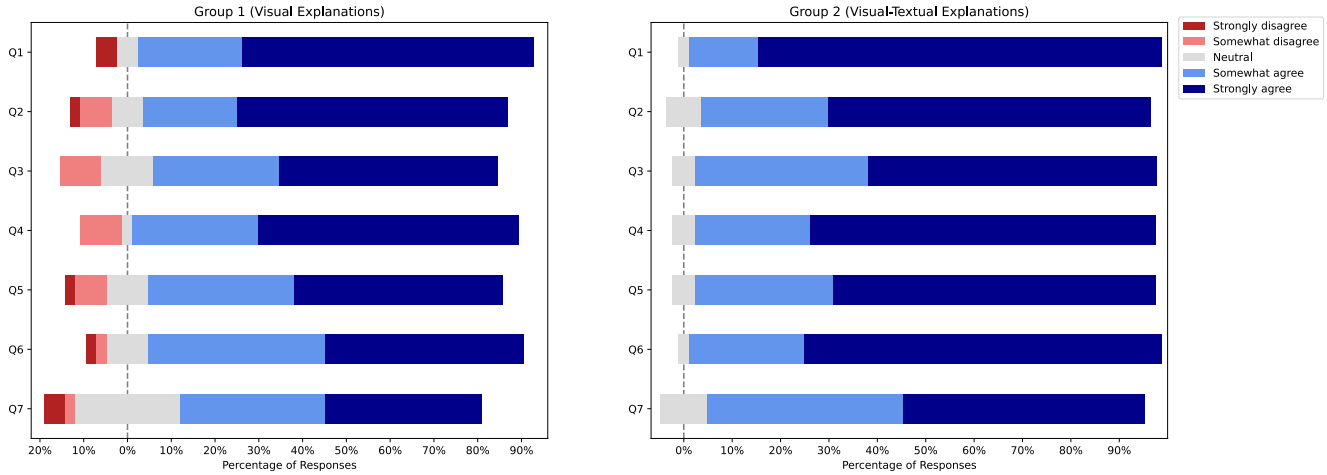


Fig. 5: We used the Explanation Satisfaction Scale in our study to gather data from both control (Group 1) and experimental (Group 2) groups. The Likert bar plot illustrates the distribution of users' attitudes towards received explanations.

Q1	I am able to understand the actions/behavior of the robot with the given explanation.
Q2	I am satisfied with the explanation provided.
Q3	The explanation provides sufficient details of the robot's actions and behaviors.
Q4	The explanation accurately describes the movement and actions of the robot.
Q5	The explanation provides reliable information about the robot's actions.
Q6	The explanation describes the robot's actions efficiently.
Q7	The explanation describes the robot's actions and situation completely.

TABLE I: Explanation Satisfaction Questionnaire

found a statistically significant difference in the explanation satisfaction between participant groups alluding to the higher satisfaction of visual-textual (multimodal) compared to only visual (unimodal) explanations. Although users are generally somewhat satisfied with visual explanations, the provision of textual together with visual explanations, adds another level of detail and natural language expression which increase user satisfaction. Based on these results, we proceeded to use visual-textual (multimodal) explanations.

B. User Study on Influence of Extroversion

We also conducted a user study to evaluate the influence of robot extroversion extremes on user explanation satisfaction. The study was conducted with 20 volunteers divided into two groups of 10 persons each. Participants' ages ranged from 21 to 40 years. The mean age of the participants was 27.45 ($SD = 5.47$) years, with 60% identifying as male and 40% identifying as female. Participants were randomly allocated to two distinct groups: the first group was provided with visual-textual explanations of an introverted robot, while the second was presented with visual-textual explanations of an extroverted robot. All participants received a detailed description of the environment and the robot's mission. They were provided with a video illustrating a robot's journey within a library, commencing from its starting point and culminating at a remote bookshelf. Its task was fetching a book for an interested reader. Throughout this journey, the robot encountered obstacles, leading to unplanned route changes and eventual stops caused by failures in the path planning system. To assess participants' satisfaction with the explanations generated by the framework, we employed the same explanation satisfaction Likert-based scale we used in the user study on explanation representation (see Table I).

Quantitative Results and Discussion: We did not find a significant difference in explanation satisfaction between the two groups. The mean of all responses for Group 1 is 3.47 ($SD = 1.21$), while the mean of all responses for Group 2 is 3.77 ($SD = 0.77$). Participants were mostly neutral to somewhat satisfied with the explanations of totally introverted and extroverted robots. These results correlate with the fact that the extroversion levels of most people do not hit extremes, thus they expect similar levels from robots.

V. Conclusion and Future Work

We have introduced an approach to modeling the explanation generation process through automated plans incorporating the influence of a robot's extroversion level and human presence on explanation timing, representation, and duration. Our explanations are multimodal providing both visual and textual insight into the robot's navigational decision-making. We have discovered that the use of multimodal (visual-textual) explanations increases user satisfaction compared to unimodal (visual) explanations. Furthermore, explanation plans allow for sequential calculation of the robot explanation strategy influenced by an extroversion level and human presence. The lower user satisfaction with the explanations of totally extroverted and introverted robots creates the need for future work, which will include testing user satisfaction with explanations of robots with different extroversion levels and the inclusion of different personality traits into the robot's planning of explanations. Future work also includes research on learning human explanation preferences and personalization of explanations based on these preferences.

REFERENCES

[1] B. Leichtmann, C. Humer, A. Hinterreiter, M. Streit, and M. Mara, "Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task," *Computers in Human Behavior*, vol. 139, p. 107539, 2023.

[2] W. Van Camp, "Explaining understanding (or understanding explanation)," *European Journal for Philosophy of Science*, vol. 4, pp. 95–114, 2014.

[3] R. Setchi, M. B. Dehkordi, and J. S. Khan, "Explainable robotics in human-robot interactions," *Procedia Computer Science*, vol. 176, pp. 3057–3066, 2020.

[4] J. Ambsdorf, A. Munir, Y. Wei, K. Degkwitz, H. M. Harms, S. Stannek, K. Ahrens, D. Becker, E. Strahl, T. Weber *et al.*, "Explain yourself! effects of explanations in human-robot interaction," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 393–400.

[5] K. Shahriari and M. Shahriari, "Ieee standard review—ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems," in *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*. IEEE, 2017, pp. 197–201.

[6] A. Halilovic and F. Lindner, "Explaining local path plans using lime," in *Advances in Service and Industrial Robotics: RAAD 2022*. Springer, 2022, pp. 106–113.

[7] ——, "Visuo-textual explanations of a robot's navigational choices," in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 531–535.

[8] A. Halilovic and S. Krivic, "The influence of a robot's personality on real-time explanations of its navigation," in *International Conference on Social Robotics*. Springer, 2023, pp. 133–147.

[9] J. Karalus, A. Halilovic, and F. Lindner, "Explanations in, explanations out: Human-in-the-loop social navigation learning," in *ICDL Workshop on Human aligned Reinforcement Learning for Autonomous Agents and Robots*, 2021.

[10] S. Tolmeijer, A. Weiss, M. Hanheide, F. Lindner, T. Powers, C. Dixon, and M. Tielman, "Taxonomy of trust-relevant failures and mitigation strategies," in *Proceedings of HRI 2020*, 2020.

[11] A. Tapus and M. J. Mataric, "Socially assistive robots: The link between personality, empathy, physiological signals, and task performance." in *AAAI spring symposium: emotion, personality, and social behavior*, 2008, pp. 133–140.

[12] A. Tapus, C. Ţăpuş, and M. J. Matarić, "User—robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy," *Intelligent Service Robotics*, vol. 1, pp. 169–183, 2008.

[13] M. Fox and D. Long, "Pddl2. 1: An extension to pddl for expressing temporal planning domains," *Journal of artificial intelligence research*, vol. 20, pp. 61–124, 2003.

[14] V. Perera, S. P. Selveraj, S. Rosenthal, and M. Veloso, "Dynamic generation and refinement of robot verbalization," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 212–218.

[15] S. Rosenthal, S. P. Selvaraj, and M. M. Veloso, "Verbalization: Narration of autonomous robot experience." in *IJCAI*, vol. 16, 2016, pp. 862–868.

[16] K. Gavriilidis, A. Munafo, W. Pang, and H. F. Hastie, "A surrogate model framework for explainable autonomous behaviour," in *ICRA2023 Workshop on Explainable Robotics*, 2023.

[17] 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*, ser. HRI '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 351–360. [Online]. Available: https://doi.org/10.1145/3434073.3444657

[18] G. Stein, "Generating high-quality explanations for navigation in partially-revealed environments," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[19] D. A. Robb, X. Liu, and H. Hastie, "Explanation styles for trustworthy autonomous systems," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2023, pp. 2298–2300.

[20] M. Brandao, G. Canal, S. Krivić, and D. Magazzeni, "Towards providing explanations for robot motion planning," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 3927–3933.

[21] M. Brandao, A. Coles, and D. Magazzeni, "Explaining path plan optimality: Fast explanation methods for navigation meshes using full and incremental inverse optimization," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 31, 2021, pp. 56–64.

[22] L. A. Jensen-Campbell, K. A. Gleason, R. Adams, and K. T. Malcolm, "Interpersonal conflict, agreeableness, and personality development," *Journal of personality*, vol. 71, no. 6, pp. 1059–1086, 2003.

[23] K. M. Lee, W. Peng, S.-A. Jin, and C. Yan, "Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction," *Journal of communication*, vol. 56, no. 4, pp. 754–772, 2006.

[24] J.-M. Dewaele and A. Furnham, "Extraversion: The unloved variable in applied linguistic research," *Language Learning*, vol. 49, no. 3, pp. 509–544, 1999.

[25] B. Tay, Y. Jung, and T. Park, "When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction," *Computers in Human Behavior*, vol. 38, pp. 75–84, 2014.

[26] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, vol. 30, pp. 457–500, 2007.

[27] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference." *Journal of personality and social psychology*, vol. 77, no. 6, p. 1296, 1999.

[28] L. Takayama and C. Pantofaru, "Influences on proxemic behaviors in human-robot interaction," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 5495–5502.

[29] B. Krarup, S. Krivic, D. Magazzeni, D. Long, M. Cashmore, and D. E. Smith, "Contrastive explanations of plans through model restrictions," *Journal of Artificial Intelligence Research*, vol. 72, pp. 533–612, 2021.

[30] M. Brandao, G. Canal, S. Krivić, P. Luff, and A. Coles, "How experts explain motion planner output: a preliminary user-study to inform the design of explainable planners," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 299–306.

[31] T. Kim and P. Hinds, "Who should i blame? effects of autonomy and transparency on attributions in human-robot interaction," in *ROMAN 2006-The 15th IEEE international symposium on robot and human interactive communication*. IEEE, 2006, pp. 80–85.

[32] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1078–1088.

[33] F. Lindner, "A conceptual model of personal space for human-aware robot activity placement," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 5770–5775.

[34] E. T. Hall, *The hidden dimension*. Anchor, 1966, vol. 609.

[35] R. Netek, T. Pour, and R. Slezakova, "Implementation of heat maps in geographical information system–exploratory study on traffic accident data," *Open Geosciences*, vol. 10, no. 1, pp. 367–384, 2018.

[36] E. Karpas and D. Magazzeni, "Automated planning for robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, no. 1, pp. 417–439, 2020. [Online]. Available: https://doi.org/10.1146/annurev-control-082619-100135

[37] D. McDermott, M. Ghallab, A. E. Howe, C. A. Knoblock, A. Ram, M. M. Veloso, D. S. Weld, and D. E. Wilkins, "Pddl-the planning domain definition language," 1998. [Online]. Available: https://api.semanticscholar.org/CorpusID:59656859

[38] A. Coles, A. Coles, M. Fox, and D. Long, "Forward-chaining partial-order planning," in *Proceedings of the Twentieth International Conference on International Conference on Automated Planning and Scheduling*, ser. ICAPS'10. AAAI Press, 2010, p. 42–49.

[39] G. Canal, M. Cashmore, S. Krivić, G. Alenyà, D. Magazzeni, and C. Torras, "Probabilistic planning for robotics with rosplan," in *Towards Autonomous Robotic Systems: 20th Annual Conference, TAROS 2019, London, UK, July 3–5, 2019, Proceedings, Part I 20*. Springer, 2019, pp. 236–250.

[40] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: challenges and prospects," *CoRR*, vol. abs/1812.04608, 2018. [Online]. Available: http://arxiv.org/abs/1812.04608