# Explainability Analysis for Skill Execution

Khatina Sari, Paul G. Plöger, and Alex Mitrevski
Department of Computer Science
Hochschule Bonn-Rhein-Sieg
Sankt Augustin, Germany
e-mail: khatina.sari@smail.inf.h-brs.de, {paul.ploeger; aleksandar.mitrevski}@h-brs.de

***Abstract*—Explainability holds significant importance for autonomous robots deployed in human-centered situations, particularly when errors occur during execution. In the context of robot action, it is important to consider various levels and types of explainability. The social dimension of Artificial Intelligence (AI) and robotic explanations, which highlights how they affect social interaction, values, and decision-making, has received little to no attention in prior research. With a particular emphasis on item handover, we hypothesize that users prefer systems with explanations and that explanations in natural language are more appealing than heatmaps. A user study, involving participants from diverse backgrounds and levels of expertise, is conducted to evaluate different levels and preferred types of explainability. The study results support our hypotheses and offer additional valuable information for future system development.**

***Keywords-Explainable Artificial Intelligence; Natural Language Processing; Heatmaps; Human-Robot Interaction.***

## I. INTRODUCTION

There have been notable developments in the disciplines of Artificial Intelligence (AI) and robotics in recent decades, which are both largely affiliated. Future robotics systems are anticipated to be far more advanced and adaptable as AI and robotics continue to grow. Rule-based systems, also referred to as white-box artificial intelligence, place an emphasis on transparency, making their logic processes clear and accessible to users. On the other hand, black-box AI, such as neural networks, often does not specify its decision-making process. Therefore, researchers are actively refining the interpretability of black-box AI, which can be used to improve transparency in robot actions, especially when failures occur [1-5].

Some challenges in Human-Robot Interaction (HRI) necessitate transparent communication. Varying user knowledge and expectations pose challenges in maintaining the right level of detail in the explanations. Another challenge is to determine the most effective explanation format for each user [6][7]. Explainability can be classified as local (usually focused on a single input dataset), global (describing how a model behaves generally), model-specific [8] (limited to particular model classes), model-agnostic [9] (may be local or global and independent of machine learning models), and counterfactual [10] (offering an alternate input scenario that would have produced a different model prediction). Meanwhile, there are three common levels of explainability [8]: low-level (which includes techniques like linear model coefficients or feature importance scores), medium-level (which delves deeper into how specific features impact the model's predictions), and high-level (which highlights intricate decision-making processes within the model).

This paper is focusing on robot object handover tasks, with the intention to enhance user understanding and trust in robot actions. A user study was conducted to evaluate the effectiveness of multiple levels of explainability in such tasks. This study aims to encourage innovation in autonomous robotics by providing access to more adaptable, flexible, and user-centered systems.

The remainder of this paper is organized as follows. Section II offers an overview of literature related to the challenging topic this study addresses. Section III describes the general approaches used in our methodology. Section IV outlines our experimental results, both qualitative and quantitative, as well as hypothesis testing. Section V summarizes our findings and includes possible future work.

## II. RELATED WORK

Transparent or white-box models refer to algorithms that provide users with both the end decision and a summary of the steps used to get there. One of the most common methods used for this is Bayesian network [11][12]. However, this method often requires substantial manual effort from users to explore the robot's behavior [13]. It lacks scalability and generalizability because it involves hand-annotating every domain-specific context up front, which hinders application to new circumstances.

On the other hand, opaque or black-box models are machine learning models that are difficult to explain and understand by experts in practical domains [14][15]. These models include random forest, support vector machine, multilayer neural network, etc. One of the ways to obtain information from such models are to use post-hoc interpretability. Although this approach provides useful information for end users, it often does not clarify precisely how a model works. Therefore, a more thorough analysis of a better strategy for building trust, reliance, and performance for human-AI teams needs to be conducted.

The need for user-centered design practices when creating explanations for AI systems was emphasized by [16]. They suggest involving users in the AI system design process

through user studies, interviews, and feedback sessions to understand their needs, mental models, and expectations. Even so, they primarily focused on design practices and guidelines for creating user experiences in explainable AI systems and did not delve deeply into technical solutions or algorithms to achieve explainability. As a result, the technical aspects of implementing the proposed guidelines may require further exploration.

In Human-Robot Collaboration (HRC), human workers should have the ability to naturally converse with robots, since they are the most crucial members of any HRC team. According to [17], while there are currently few means of communication between human workers and robots, gesture recognition has long been used as an efficient human-computer interaction. In conclusion, they believe that HRC will operate in a safer environment if a depth sensor and body-model technique are combined to track human movements.

As part of the machine learning adaptation in the robot's motion planning, our approach proposes the utilization of a neural network. This is an alternative approach to the genetic algorithm utilized by [14]. The adjustment in methodology highlights our dedication to investigating different and practical approaches that may result in improved responsiveness and flexibility of robotic systems in dynamic settings. In addition, inspired by [16] user-centric principles, we conducted a user study to uncover user preferences regarding different approaches in robot motion planning. Our questionnaire aims to uncover user preferences regarding the different approaches employed in robot motion planning, shedding light on which method resonates more effectively with particular users.

## III. APPROACH

The scope of our study concentrates on the usage of autonomous robots for object handover tasks from robot to human, an important use case that requires an effective explanation strategy. Giving our Toyota Human Support Robot (HSR) a skill set that corresponds to different levels of explainability—or, in some cases, no explainability at all—is the current challenge at hand. Our explainability analysis for skill execution takes into account a number of important factors, one of which is the recognition that explainability in our case is inherently local.

### A. Proposed Approach

The current approach used in our robot to determine the handover position is done by factoring in context-dependent (based on the posture of the detected person) and context-independent (static; based on the context-dependent outcome). However, the handover position in a context-independent approach does not consider any surrounding environment variables; thus, we propose to train a neural network to dynamically set the end-effector position based on the values obtained from the 3D bounding box. By allowing the neural network to generate random handover positions, we can collect input-output pairs dataset that can be used to fine-tune the model until it can automatically generate optimal handover positions based on the user's needs. This strategy would increase the effectiveness and usability of the robotic system. Regrettably, a prolonged mechanical issue in our Toyota HSR has forced us to delay the implementation of our neural network interpretation. Upon its resumption of operations, we shall resume our work and implement our planned approach.

### B. Explainability Setup

One of the primary concerns that drives our research is how to determine the robot's reasoning behind certain decisions, especially why it stops at a specific point in relation to the detected human position during object handover. To carry out this research, an advanced built-in program created by [18] is used, which generates a 3D bounding box to locate the detected person in front of the robot. It follows the right-handed coordinate system, which includes the depth ($x$-axis), horizontal ($y$-axis), and vertical ($z$-axis). Once the person is detected, their position will be determined; in our case, there are three possible positions: standing, sitting, and lying down.

Within our research framework, several notations play an important role in influencing how we perceive the spatial connection between humans and the robot during the handover task. Figure 1 illustrates the configuration in which $W_p$ represents the robot's end-effector location where the object is held, $W$ is the robot's base frame, $B$ denotes the bounding box, and $p$ is the relative position between the end effector and the center point of the bounding box.

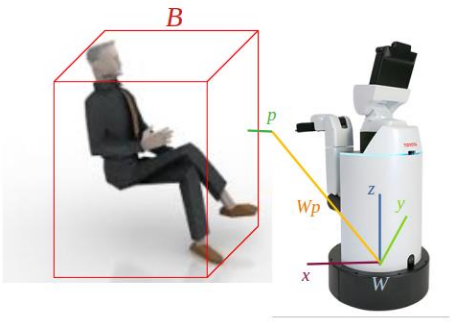


Figure 1. Illustration of the parameters on handover skill.

Logical predicates describing the requirements for a successful handover interaction are adopted from [19] to define the success preconditions. The predicates include $in\_front\_of_{x,y}(p,B)$, $far\_in\_front\_of_{x,y}(p,B)$, $behind_{x,y}(p,B)$, $far\_behind_{x,y}(p,B)$, $above_{x,y}(p,B)$, $below_{x,y}(p,B)$, and $centered_{x,y}(p,B)$. Using the success preconditions, the natural language explanation for each position is generated manually, as shown in Tables I-III.

In addition to manual natural language translation, ChatGPT 3.5 [20] is employed to generate automated translation and evaluate the results using the Bilingual Evaluation Understudy (BLEU) score [21]. The first few initial tests did not produce close translations to the manual translation. Therefore, more detailed definitions of each logical expression were provided, as well as separating each predicate that consists of two or more coordinates; for

example, $centered_{x,y}(p,B)$ becomes $centered_x(p,B) \wedge centered_x(p,B)$. The outcome of the last iteration was then used for assessment.

TABLE I. PRECONDITIONS FOR STANDING POSITION

| Types | Success Preconditions |
|---|---|
| **Logical Predicates** | $centered_{y,z}(p,B) \wedge in\_front\_of_x(p,B) \wedge \neg centered_x(p,B) \wedge \neg below_{x,y}(p,B) \wedge \neg behind_{x,y}(p,B) \wedge \neg far\_behind_{x,y}(p,B) \wedge \neg above_{x,y}(p,B) \wedge \neg in\_front\_of_y(p,B) \wedge \neg far\_in\_front\_of_y(p,B)$ |
| **Natural Language** | The robot's arm should be in front of and centered around a person (corresponding to the person's height and width). It should not be behind, above, beneath, or to the right/left of a human. |

TABLE II. PRECONDITIONS FOR SITTING POSITION

| Types | Success Preconditions |
|---|---|
| **Logical Predicates** | $centered_{y,z}(p,B) \wedge in\_front\_of_x(p,B) \wedge \neg centered_x(p,B) \wedge \neg below_{x,y}(p,B) \wedge \neg behind_{x,y}(p,B) \wedge \neg far\_behind_{x,y}(p,B) \wedge \neg above_{x,y}(p,B) \wedge \neg in\_front\_of_y(p,B) \wedge \neg far\_in\_front\_of_{x,y}(p,B)$ |
| **Natural Language** | The robot's arm is positioned in front of and around the middle of a sitting person (according to the person's height and width). It is not behind, above, beneath, and to the right or left of the person. |

TABLE III. PRECONDITIONS FOR LYING DOWN POSITION

| Types | Success Preconditions |
|---|---|
| **Logical Predicates** | $above_{x,y}(p,B) \wedge centered_y(p,B) \wedge \neg centered_z(p,B) \wedge \neg below_{x,y}(p,B) \wedge \neg behind_{x,y}(p,B) \wedge \neg far\_behind_{x,y}(p,B) \wedge \neg in\_front\_of_y(p,B) \wedge \neg far\_in\_front\_of_{x,y}(p,B)$ |
| **Natural Language** | The robot's arm is positioned above and centered around the person's width. It is not below or around their head or feet. It should not extend all the way to the opposite side from where a robot is standing next to. |

BLEU provides a quantitative measure by comparing the output of machine translation systems (candidate translation) against reference translations, offering insights into the degree of overlap in n-gram or word sequences with human-generated counterparts [21]. The length of candidate sentences that are shorter than the reference phrases is penalized in the BLEU metric (Brevity Penalty), which is based on the modified $n$-gram precision measure. The following formula determines the BLEU score:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} \frac{1}{N} \cdot logP_n\right), \qquad (1)$$

where $BP$ = Brevity Penalty and $P_n$ = Precision for $n$-gram.

The Natural Language Toolkit (NLTK) [22] and spaCy [23] are used in our BLEU score computation to provide an unbiased evaluation of machine-generated translations. The translation produced by ChatGPT 3.5 (as a candidate translation) is compared with our original translation (as a reference). The results of the BLEU score for each translation performed by ChatGPT in comparison to the manual translation are presented in Table IV.

TABLE IV. BLEU SCORE OF CHATGPT 3.5 TRANSLATION

| No. | Position | BLEU Score |
|---|---|---|
| 1 | Standing | 0.85 |
| 2 | Sitting | 0.81 |
| 3 | Lying Down | 0.88 |

The final translation output from ChatGPT 3.5 provides a good starting point for future developments. Despite the fact that the translations produced by the first few iterations were not satisfactory, adding further specific information made it generate a translation that was similar to the one that was done manually. The key realization is that it is possible to train models, like ChatGPT, to translate technical terminology into natural languages effectively.

When it comes to interpreting the neural network's decisions about handover position, Grad-weighted Class Activation Mapping (Grad-CAM) [24] integration shows itself to be an effective tool for insight. It offers a transparent and insightful lens into the decision-making processes of complex models. Grad-CAM fills this gap by giving an illustration of the areas in the input data that have a major impact on a certain outcome. Unfortunately, the problem with our Toyota HSR prevented us from implementing this method. Despite this obstacle, a previously collected dataset from our research team [25] was leveraged, and the video content was edited to achieve the same heatmap effect (as seen in Figure 2). This decision allowed us to simulate and observe the intended outcomes, ensuring the continuity of the research despite the technical constraints.

Figure 2. Additional heatmaps on one of the handover scenarios.

The dataset, which includes relevant information but lacks explanations, was then extended by adding explanations in both heatmap and natural language formats. This improvised solution allows us to proceed with our user study within the designated timeframe, preserve the research objectives, and ensure the timely execution of the study.

### C. Experimental Design

In our comprehensive user study aimed at investigating user preferences in interacting with AI-based or robotic systems, two distinct hypotheses were formulated to guide our research. The first hypothesis is that users have a preference for systems that offer explanations while they are using them. The second hypothesis is about the preferred explanation format among users; in particular, we hypothesize that people

prefer explanations in natural language over alternative visualization techniques like heatmaps.

Our user study adopts a mixed-methods strategy to gather quantitative data and qualitative insights through surveys in order to experimentally validate our hypotheses. After being presented with simulated robotic interfaces that include heatmaps and natural language explanations, participants' preferences, satisfaction, and understanding were carefully examined.

Through selectively crafted survey questions, user experiences, preferences, and challenges are explored, allowing us to obtain insights into the factors that contribute to a positive or negative interaction. Additionally, scenarios that are meant to replicate real-world interactions were chosen by giving users experiences that were contextually appropriate and reflected the difficulties and complexities of real-world circumstances. Ten videos and three different explanation varieties were presented to help construct a more comprehensive understanding of user preferences: no explanation, partial explanation using heatmaps, and detailed explanation using natural language. In order to prevent any potential biases, 8 out of 10 videos were purposefully presented in a random order. Following every video, participants were asked to rate how confident they were in their understanding of the robot decision-making process.

## IV. EXPERIMENTAL RESULTS

Our user study involved a total of 33 participants, ages ranging from 18 to 40 years old, education ranging from high school to Ph.D., and different academic and professional backgrounds. Our participants' demographic profiles show a variety of age groups, gender identities, levels of education, and fields of study. This diversity attempts to determine whether there is any relationship between the preferred explanation technique and the educational background.

### A. Quantitative Analysis

In terms of the participants' experiences and expectations in the realms of robotic systems and Artificial Intelligence (AI), 75.8% of them have prior hands-on experience with robotic systems, while an overwhelming 84.8% are familiar with AI or machine learning in their practical lives. In a survey on comfort levels, 72.7% of the respondents said they felt uneasy when AI systems made decisions without providing an explanation, highlighting the significance of transparency.

In our scenario-based questions, two identical videos served as starting points. The first was without explanation, whereas the second included a natural language explanation. The majority indicated that they were unclear about the robot's action in the first video, though it was a successful object handover scenario. However, the participant's confidence level improved after watching the second video, which revealed a positive beginning. Table V summarizes participants' confidence levels after eight more videos were shown in a random order. It reveals that individuals feel more confident when they are given an explanation of how the robot makes decisions. Less than 40% of the participants felt confident about their understanding of the robot decision-making process in the three videos without an explanation, in both successful and unsuccessful handover scenarios. More than 50% of the participants in the two videos where heatmaps were used as an explanation type expressed confidence in the successful handover scenario. However, in the case of an unsuccessful handover, only 34.6% of participants reported feeling confident. With natural language explanations, on the other hand, 48.4% of those surveyed expressed confidence in the unsuccessful scenarios. In the successful scenario, over 80% of the participants expressed confidence and none of them indicated lack of confidence.

TABLE V. AN OVERVIEW OF PARTICIPANTS' CONFIDENCE LEVEL

| Video | Outcome | Explanation Type | Confidence Level (%) | | | | |
|---|---|---|---|---|---|---|---|
| | | | 5 | 4 | 3 | 2 | 1 |
| 3 | Succeed | None | 9.1 | 24.2 | 48.5 | 18.2 | 0.0 |
| 4 | Succeed | Heatmap | 24.2 | 27.3 | 36.4 | 12.1 | 0.0 |
| 5 | Failed | None | 12.1 | 15.2 | 30.3 | 24.2 | 18.2 |
| 6 | Failed | Natural Language | 24.2 | 24.2 | 15.2 | 15.2 | 21.2 |
| 7 | Succeed | None | 3.0 | 18.2 | 15.2 | 36.4 | 27.3 |
| 8 | Succeed | Natural Language | 27.3 | 57.6 | 15.2 | 0.0 | 0.0 |
| 9 | Failed | Natural Language | 24.2 | 24.2 | 18.2 | 27.3 | 6.1 |
| 10 | Failed | Heatmap | 18.2 | 21.2 | 30.3 | 27.3 | 3.0 |

To conclude, compared to visual explanation (using a heatmap), natural language explanation improves their confidence by over 30% (shown in Figure 3).

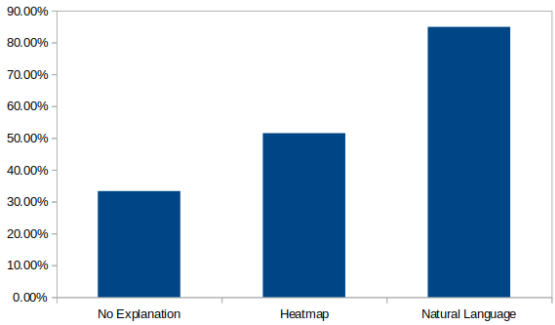


Figure 3. Participants' overall confidence in understanding the robot decision-making process.

### B. Hypotheses Testing

We conducted the first hypothesis test to investigate users' preferences regarding the type of videos when seeking information. The hypothesis aimed to determine whether users prefer videos with explanations over videos without explanations. The participants were presented with the question "Which type of video do you prefer when seeking information?" and the response options: videos with explanation, without explanation, and depending on the context.

A chi-square test [26] for independence is employed to analyze the association between the type of video and user preference, where $H_0$ = no preference difference and $H_1$ = there is a preference for videos with an explanation. If the $p$-

value of a given dataset is less than 5%, the null hypothesis is rejected because it is assumed that there is a preference difference among the options. To calculate the $p$-value using chi-square formula (2), the observed value ($O$) needs to be identified first, which represents the actual counts derived from the sample, and the expected value ($E$), which represents the values of each category in the event that there was no preference difference between all categories. $E$ is obtained by dividing the total number of observed values by the number of categories. The following calculation can then be used to get its chi-square statistic ($\chi^2$) based on the observed and expected values:

$$X^2 = \sum \frac{(O-E)^2}{E}. \quad (2)$$

The result, along with the degrees of freedom ($df$), which is a number representing how much variation is involved in the research ($n$) minus 1,

$$df = n - 1, \quad (3)$$

is used to calculate the $p$-value from the chi table.

Our observed and expected values based on the survey results are displayed in Table VI. The total observed values—33 in this case—and the number of categories—3 in this case—are then used to compute the expected values, yielding the value $E$ = 11.

TABLE VI. THE OBSERVED AND EXPECTED VALUES

| User Preference | $O$ | $E$ | $O-E$ | $(O-E)^2$ |
|---|---|---|---|---|
| With Explanation | 22 | 11 | 11 | 121 |
| Without Explanation | 2 | 11 | -9 | 81 |
| Depend on the Context | 9 | 11 | -2 | 4 |

These observed and expected values were used to calculate the chi-square statistic, which was then used to test the hypothesis. The result yielded $\chi^2$= 28.1; with $df$ = 2, the resulting $p$-value was 0.0000008. Since the $p$-value is less than $\alpha$ = 5% or 0.05, it is determined that the null hypothesis is rejected.

The second hypothesis is tested based on two identical videos with two distinct explanations—one using a heatmap (video 4) and the other using natural language (video 8). Participants were asked to choose which of the two videos gave them a better understanding of the robot decision-making process. Participants who selected video 8 are considered to prefer the natural language explanation. A one-sample proportion test ($Z$) [27] is employed to analyze whether the proportion of users who prefer video 8 differs significantly from 50% (no preference). The null hypothesis ($H_0$) assumed no preference difference, while the alternative hypothesis ($H_1$) assumed a preference for videos with natural language explanation.

To conduct the test, we need to estimate the proportion $\hat{p}$ as:

$$\hat{p} = \frac{x}{n}, \quad (4)$$

where $x$ is the number of participants who have chosen video 8 and $n$ is the total number of participants. After that, the test statistic can be calculated with the following formula:

$$Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \quad (5)$$

where $p_0$ is the pre-specified value; in this case, it is 50% to indicate that if half of the total participants chose video 8, there is no significant preference for that particular video. From there, the calculated $Z$-value is compared with critical values, which can be obtained from the $Z$ table, from the standard normal distribution. Given that the sampling distribution of our data is a normal distribution with a significant value of 0.05, the critical values are in a range of -1.96 to 1.96. Based on the result of our survey, a one-sample proportion test was calculated with $x$ = 23 and $n$ = 33, which yielded a $Z$-value of 2.46. Because the $Z$-value is larger than the maximum critical value, the null hypothesis is rejected.

A post hoc sensitivity analysis [28] was conducted to evaluate the statistical power of our study. Cohen's $w$,

$$w = \sqrt{\sum \frac{(p_i-p_{0i})^2}{p_{0i}}}, \quad (6)$$

where $p_i$ is the observed value in category $i$ and $p_{0i}$ is the expected value under the null hypothesis in category $i$, is used to measure the effect size for the chi-square test of the first hypothesis. The thresholds are 0.10 for a small effect, 0.30 for a medium effect, and 0.50 for a large effect. The result yielded $w$ = 0.75, which represents a large effect.

Furthermore, we assess the effect size for the one-sample proportion test of the second hypothesis with Cohen's $h$,

$$h = 2\,x\left(\arcsin\left(\sqrt{p_1}\right) - \arcsin\left(\sqrt{p_2}\right)\right), \quad (7)$$

where $p_1$ and $p_2$ are the two proportions being compared. The thresholds are 0.20 for a small effect, 0.50 for a medium effect, and 0.80 for a large effect. From our user study result, 23 out of 30 participants preferred video with natural language explanation; thus, $p_1$ = 69.7%. Then we compare it with $p_2$ = 50% for the proportion that shows no preference difference. The result yielded $h$ = 0.40, which indicates a moderate effect size.

### C. Qualitative Analysis

As proven in our hypothesis 2, natural language explanations are preferable to heatmaps. In order to evaluate it on a qualitative level, the participants were asked why they

preferred one type of explanation over the other, and the majority of them responded that they preferred natural language because it is easier to understand and more elaborate. In addition, they believe that natural language explanations can be enhanced by an audio or speech component.

They were then asked to imagine a situation in which they would favor a different kind of explanation than the one they had previously selected. Those who have chosen natural language say that they prefer heatmaps when a robot performs a simple task, interacts with static objects, or is in a simulation. On the other hand, those who have chosen heatmaps say that they prefer natural language when failure occurs, when the robot is in a dynamic environment, or when the user has no background knowledge about the system.

When asked to imagine a situation in which they would prefer to have no explanation at all, the majority of respondents believe that in a straightforward or routine task that is repeated, there is no need for an explanation because the rationale is obvious. While some claim that they cannot think of any situation in which it is preferable not to have an explanation, others highlight this point by stating that, even in tasks that appear straightforward, having an explanation is desirable since it provides a clear reasoning behind the robot's chosen action.

## V. CONCLUSION AND FUTURE WORK

Our user study results supported our hypotheses, offering statistical evidence that users do, in fact, prefer explanations when interacting with robotic systems. These findings highlight that providing explanations improves users' trust and understanding of robot systems. Although the study demonstrates a clear preference for explanations in natural language as opposed to heatmap visualizations, respondents express a preference for heatmaps or no explanations at all when the robot is performing regular or routine tasks. This tendency implies that, in situations they are familiar with, participants think that the visual representations of the heatmaps are sufficient or that perhaps they prefer them more when the tasks are simple and require no extra information. Due to the wide range of participant preferences, flexible communication strategies that take into account varying user expectations and levels of experience with certain robotic tasks are necessary.

Even though the results suggest that users prefer systems that provide explanations over those that do not, it is important to acknowledge a potential bias in how this hypothesis was tested. The question itself highlights the presence or absence of an explanation, which might have led participants to gravitate toward the condition with explanations, independent of their actual utility in decision making. Future studies should aim to mitigate this bias by embedding explanations in more naturalistic tasks where the usefulness of the explanation emerges organically rather than being made explicit to participants.

While our findings indicate that participants preferred natural language explanations, it is important to recognize that this result may partly reflect differences in interpretability between formats. Natural language requires little effort to process, whereas heatmaps demand additional interpretation and prior familiarity. This asymmetry may have disadvantaged the heatmap condition. To address this imbalance, future studies should explore providing training or familiarization with visual explanations, refining visualization design to reduce cognitive effort, or presenting hybrid formats that combine textual and visual elements for complementary strengths.

Further studies could explore automating the translation of scientific terms into natural language to provide explanations for nonexpert users. To implement audio explanations effectively, future work may explore the integration of speech synthesis technologies or Natural Language Processing (NLP) models specialized in generating spoken content. Additionally, exploring the potential of machine learning techniques, such as reinforcement learning, could contribute to optimizing explanation selection. This way, the system could learn over time which combination of explanation modalities yields the most positive user responses or facilitates optimal task performance.

## ACKNOWLEDGMENT

The authors would like to express gratitude to all of the participants who willingly participated in the research, contributing their time and insights. Without their cooperation, this study would not have been possible.

## REFERENCES

[1] O. Loyola-Gonzalez, "Black-box vs. white-box: understanding their advantages and weaknesses from a practical point of view," IEEE Access, vol. 7, pp. 154096–154113, 2019.

[2] W. Samek, T. Wiegand, and K.-R. M¨uller, "Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models," arXiv preprint arXiv:1708.08296, 2017.

[3] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," IEEE access, vol. 6, pp. 52 138–52 160, 2018.

[4] R. Guidotti, et al., "A survey of methods for explaining black box models," ACM computing surveys (CSUR), vol. 51, no. 5, pp. 1–42, 2018.

[5] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," AI magazine, vol. 40, no. 2, pp. 44–58, 2019.

[6] A. Holzinger, "From machine learning to explainable ai," in 2018 world symposium on digital intelligence for systems and machines (DISA). IEEE, pp. 55–66, 2018.

[7] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," Journal of Artificial Intelligence Research, vol. 70, pp. 245–317, 2021.

[8] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," arXiv preprint arXiv:2006.11371, 2020.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," arXiv preprint arXiv:1606.05386, 2016.

[10] R. M. Byrne, "Counterfactuals in explainable artificial intelligence (xai): evidence from human reasoning," in IJCAI, pp. 6276–6282, 2019.

[11] M. Lomas, et al., “Explaining robot actions,” in Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot, pp. 187-188, 2012.

[12] 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.

[13] O. Amir, F. Doshi-Velez, and D. Sarne, “Summarizing agent strategies,” Autonomous Agents and Multi-Agent Systems, vol. 33, pp. 628–644, 2019.

[14] M. Mucientes and J. Casillas, “Quick design of fuzzy controllers with good interpretability in mobile robotics,” IEEE Transactions on Fuzzy Systems, vol. 15, no. 4, pp. 636–651, 2007.

[15] A. Alvanpour, S. K. Das, C. K. Robinson, O. Nasraoui, and D. Popa, “Robot failure mode prediction with explainable machine learning,” in 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE). IEEE, pp. 61–66, 2020.

[16] Q. V. Liao, D. Gruen, and S. Miller, “Questioning the ai: informing design practices for explainable ai user experiences,” in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–15, 2020.

[17] H. Liu and L. Wang, “Gesture recognition for human-robot collaboration: a review,” International Journal of Industrial Ergonomics, vol. 68, pp. 355–367, 2018.

[18] A. F. Abdelrahman, “Incorporating contextual knowledge into human-robot collaborative task execution,” Technical Report, H-BRS Sankt Augustin, 2020.

[19] A. Mitrevski, “Skill generalisation and experience acquisition for predicting and avoiding execution failures,” Ph.D. Dissertation, RWTH Aachen, 2023.

[20] OpenAI. (2023). ChatGPT (Oct 16 version) [Large language model]. [Online]. Available from:
https://chat.openai.com/chat. Retrieved: June 2024.

[21] M. Evtikhiev, E. Bogomolov, Y. Sokolov, and T. Bryksin, “Out of the bleu: how should we assess quality of the code generation models?” Journal of Systems and Software, vol. 203, p. 111741, 2023.

[22] S. Bird, “NLTK: the natural language toolkit,” in Proceedings of the COLING/ACL 2006 interactive presentation sessions, pp. 69-72. 2006.

[23] Y. Vasiliev, “Natural language processing with Python and spaCy: A practical introduction,” No Starch Press, 2020.

[24] R. R. Selvaraju, et al., “Grad-cam: visual explanations from deep networks via gradient-based localization,” in Proceedings of the IEEE international conference on computer vision, pp. 618–626, 2017.

[25] IROS 2022 HEART-MET Handover Failure Detection Challenge. Available from:
https://codalab.lisn.upsaclay.fr/competitions/6757#learn_the_details-evaluation. Retrieved: September 2025.

[26] Pandis, N., “The chi-square test,” American Journal of Orthodontics and Dentofacial Orthopedics, vol. 150, no. 5, pp. 898–899, 2016.

[27] David I, Adubisi O, Farouk B, and Adehi M., “Assessing MSMEs growth through rosca involvement using paired t-test and one sample proportion test,” J Soc Econ Stat, vol. 9, no.2, pp. 30–42, 2020.

[28] Cohen, D. “Culture, social organization, and patterns of violence,” Journal of Personality and Social Psychology, vol. 75, no. 2, pp. 408–419, 1998, doi:10.1037/0022-3514.75.2.408.