Toward Socially-Aware Explainability for Autonomous Robots

Khatina Sari, Paul Gerhard Plöger, and Alex Mitrevski
Department of Computer Science
Hochschule Bonn-Rhein-Sieg
Sankt Augustin, Germany
e-mail: khatina.sari@gmail.com, paul.ploeger@h-brs.de, alemitr@chalmers.se

***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. However, the social dimension of Artificial Intelligence and robotic explanations, which highlights how they affect social interaction, values, and decision-making, has received little to no attention in prior research. Our research addresses the need for multiple types of explainability in relation to robot activity. The aim is to develop a framework that enables the robot to provide explanations at varying levels of detail and complexity, depending on the type of end-user involved. Extending upon previous research on a representation for skill parameterization, we replicate the desired functionality to obtain information about user preferences by introducing simulated explanations that are translated from logical expressions. We explore two potential modalities that could be effectively integrated into the robotic system in the future: heatmaps and natural language explanations. 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; Autonomous Systems; Human-Robot Interaction.

I. INTRODUCTION

This paper builds upon and extends our previously published work in [1]. In addition to the original contributions, we incorporate a more comprehensive review of related literature, provide a detailed description of the research methodology, and present an expanded user study. The user study now extends both quantitative and qualitative analyses, offering deeper insights and a more robust evaluation of the proposed approach.

In recent decades, the fields of robotics and artificial intelligence (AI), which are closely related, have seen significant advancements. 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 AI, 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 [2-6].

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 [7][8]. It is worth noting that the level of complexity in the explanations provided may need to vary based on the specific users and settings involved. The technical proficiency and domain knowledge of different users may differ. Tailoring the complexity of explanations to the intended users can enhance their comprehensibility and ensure that the information conveyed is meaningful and useful [9]. By providing explanations that reveal the robot's intentions and decision-making processes, users can gain insights into its actions, enabling a higher level of trust and facilitating smoother collaboration. Additionally, to enhance productivity and team effectiveness, it becomes essential for the robots to provide explanations for their actions and decisions. These explanations serve as a means to bridge the gap in understanding between humans and robots, enabling better coordination and shared knowledge.

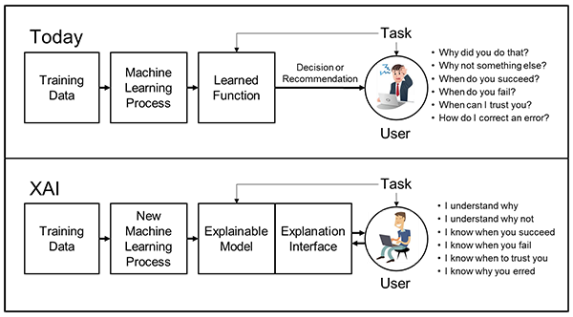

Figure 1. Concept of Explainable Artificial Intelligence (XAI) [6].

Explainability can be classified as local (usually focused on a single input dataset), global (describing how a model behaves generally), model-specific [10] (limited to particular model classes), model-agnostic [11] (may be local or global and independent of machine learning models), and counterfactual [12] (offering an alternate input scenario that

would have produced a different model prediction). Meanwhile, there are three common levels of explainability [10]: low-level (includes techniques like linear model coefficients or feature importance scores), medium-level (delves deeper into how specific features impact the model's predictions), and high-level (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. The aim is to develop a framework that enables the robot to provide explanations at varying levels of detail and complexity, depending on the type of end-user involved. Essentially, a key element of our work is our user study, which allows us to verify with a diverse and representative set of participants the benefits of offering multiple levels of explainability during item handover. 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 [12][13]. Lomas et al. [13] employed a Bayesian network model to evaluate how well the user understands the robot's actions. The algorithm calculates the amount of detail the robot should provide in its explanations based on a number of variables, including the user's prior knowledge, attention span, and verbal and non-verbal feedback. After determining the user's understanding level, the robot creates an explanation using a combination of pre-defined rules and machine learning methods (decision trees and classification algorithms), as well as graphical representations of its operations. While their approach does provide users with some ways of exploring what actions a robot will take in different scenarios, they do not attempt to carefully select which behaviors to demonstrate. Thus, they require substantial manual effort from users to explore the behavior of the robot and do not ensure that users understand the overall strategy of the robot [14]. Moreover, this method lacks scalability and generalizability because it involves hand-annotating every domain-specific context up front, which hinders application to new circumstances. In contrast, opaque or black-box models are machine learning models that are difficult to explain and understand by experts in practical domains [15][16]. These models include random forest, support vector machine, multilayer neural network, etc. One of the ways to obtain information from such models is to use post-hoc interpretability.

In order to address issues with slow learning and poor interpretability of some learned controllers, [15] proposed a solution that utilizes a two-stage methodology. In the first stage, a genetic algorithm is used to generate a large number of candidate rules. In the second stage, a subset of the most effective and interpretable rules is selected using a set of novel interpretability metrics. Additionally, they applied ant colony optimization (ACO) search algorithm to cut down on learning time. Then they analyzed the interpretability of their proposed approach using several metrics, including the number of rules, the number of fuzzy sets used for each input variable, and the complexity of the rule base. Through a set of simulation experiments where they compared the performance of their proposed approach against other rule-based and model-based approaches, they demonstrated how their approach produced fuzzy controllers that are more interpretable than those generated by other methods, while maintaining comparable or better performance. Nonetheless, they only focused on fuzzy controller design and did not consider non-fuzzy approaches, which may provide effective and interpretable control strategies for mobile robotics applications. As an alternative, we propose the utilization of neural networks and reinforcement learning as part of the machine learning adaptation in robot motion planning. This adjustment highlights our dedication to investigating different and practical approaches that may result in improved responsiveness and flexibility of robotic systems in dynamic settings.

Alvanpour et al. [17] investigated how machine learning can be used to predict robot grasp failure and studied the trade-off between accuracy and interpretability by comparing interpretable ML models that are inherently explainable with black box ML models that tend to be more accurate but do not come with explanations. Their approach consists of two stages: feature extraction and failure mode prediction. Even though they provided explanations for the predictions made by their approach, they did not test their approach on a physical robot, but rather on a simulation environment. Therefore, it remains to be seen how well their approach would perform in a real-world scenario, where sensor noise and other factors could affect the reliability of the system [18]. The evaluation of our proposed approach, on the other hand, spans across both simulated and real-world scenarios, aiming to provide a comprehensive assessment of its applicability and effectiveness across diverse operational contexts. We applied this assessment strategy in order to validate the dependability and applicability of our technique in both the complex and unpredictable settings found in real-world applications, as well as under controlled, simulated situations.

The need for user-centered design practices when creating explanations for AI systems was emphasized by [19]. 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. Inspired from the user-centric principles they advocated, and given the importance of user trust in human-robot interaction, we conducted a user study to obtain

insightful feedback from users. This 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. Through incorporating user feedback into our assessment framework, we hope to ensure that our method not only satisfies technical requirements but also aligns with user expectations and promotes a constructive and reliable human-robot collaboration.

Based on [9], methods for enabling humans and robots to fully understand each other's internal states should be developed so that robots can communicate with humans without only understanding human instructions at a surface level. In their opinion, a robot needs to have two mechanisms: one that can predict users' thoughts and the other that enables it to convey its own ideas to people in a way that is simple to comprehend. Their survey reveals that the explainability of autonomous robots is a multidisciplinary field that lacks structure. Furthermore, there is currently little to no research on the social dimensions of explanations provided by AI and robots. While a mechanism for creating an interpretable decision-making space has been created, there is currently no way to create a decision-making space that can be used universally in any setting without the need for humans to be involved. Moreover, they make suggestions that it is crucial to model humans who obtain explanations because no approaches have been put out that successfully estimate a human internal model up to this point. But given the interconnection of explanations, it is still a challenging task for the robot to be able to adapt its own knowledge system while delivering explanations. One must take into account the other person, their own knowledge, and the reliability of the information.

Ideally, 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 [20], 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. They break down the gesture recognition process, which started with gesture raw data collection by sensors. Image-based methods (such as markers, depth sensors, stereo cameras, and single cameras) and non-image-based methods (such as bands, gloves, and non-wearable cameras) are the two main types of data collecting. Since the majority of human motion are dynamic, gesture categorization is the final and most crucial stage in gesture identification. Actuation classification must be completed either in conjunction with or after gesture tracking because a single dynamic gesture is always made up of many frames. 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.

Beyond the emphasis on trust, our study considers broader implications for user acceptability and cooperation in interaction between humans and robots; incorporating innovative approaches aims to meet users' expectations, preferences, and comfort levels in addition to raising technological criteria. By actively seeking user feedback and insights, our research aspires not only to advance the technical limitations of robotic systems but also to foster an inclusive and user-friendly approach for the comfortable integration of humans and robots in various domains. Additionally, we intend to measure how much these methods add to a generalized increased degree of confidence in the human-robot interaction paradigm.

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. Moreover, we investigated the domain of behavior explainability, focusing on cases of execution failures in tasks involving the handover of objects between humans and robots. Although it is obvious how important it is to explain robot behavior, it can be difficult to decide how much information is suitable to provide in these explanations. Our research intends to explore the complicated field of model-agnostic explainability since we acknowledge that various users may have varied expectations regarding the quantity of explanation needed and varying degrees of expertise with autonomous systems.

Within our explainability analysis for skill execution, model agnosticism means taking an approach that goes beyond the complexities of the machine learning models that are used as the foundation for the autonomous robot's decision-making. Model-agnostic approaches offer a general framework that allows us to examine the autonomous robot's decision-making process and behaviors without being limited by the internal model architecture's specific details. These are the methods that we would like to use to our Toyota HSR, which is the physical representation of our research into the problem of robot-human object handover. As we explore the advantages of explainability at several levels, our research methodology highlights the necessity of adaptability and flexibility in giving explanations that are specific to the needs of various users and circumstances.

Our goal in implementing model-agnostic explainability is to guarantee that our understanding of skill execution failures is relevant to all situations, regardless of how complicated the machine learning models are at their core. This methodology is in harmony with the diverse skill sets and demands of users interacting with autonomous robots in human-centered settings, and it eventually leads to a more open, understandable, and intuitive paradigm of interaction.

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 conducting numerous training cycles, the neural network discovers the underlying patterns and connections between various contextual parameters and related handover positions. Through exposure to a broad range of random positions, the network improves its capacity to generalize and make appropriate choices in new scenarios encountered in real-world user interactions. Consequently, the trained neural network gains competency in automatically producing handover positions that properly correspond with user requirements, improving the efficiency and user-friendliness 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 [21] 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, a number of notations play an important role in influencing how we perceive the spatial connections and decision-making processes associated with handing over objects between humans and robots. Figure 2 shows an illustration of the configuration where W_p represents the location on the robot's effector where the object to be given to the human is held, and W is the robot's base frame. Denoted as B, the bounding box is a three-dimensional structure containing the detected individual. $\bar{B}$ corresponds to the center point of this bounding box, providing a simple reference for the person's overall physical position. The minimum and maximum points of the bounding box, represented as $min(B)$ and $max(B)$ respectively, define the spatial extent of the detected individual within the bounding box. These points serve as key parameters in defining the boundaries of the person's posture and guide the robot's subsequent decision-making.

The relative position between the end effector and the center point of the bounding box is denoted as p. During the handover task, this relative position plays a key role in figuring out the spatial relationship between the identified person and the robot. Additionally, c_p is a threshold parameter that specifies an extra distance allowance in relation to p, enabling the robot's judgment criteria to be flexible and fine-tuned. To account for potential noise and variations in the center position estimate, a small threshold that controls for fluctuations and uncertainties in the center point's determination is introduced as ϵ_p. It serves as a mechanism to mitigate the impact of noise, ensuring a more robust and reliable analysis of the spatial dynamics between the robot and the detected person. Together, these notations improve the robot's accuracy and flexibility in the complex task of object handovers between humans and robots.

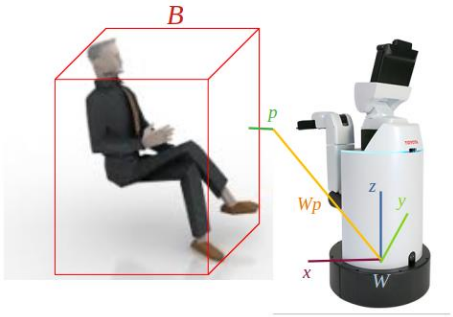

Figure 2. Illustration of the parameters on handover skill.

Logical predicates describing the requirements for a successful handover interaction are adopted from [22] 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 $centered_{x,y}(p,B)$. These predicates are made up of a collection of restrictions and conditions that capture the main elements affecting the robot's decision-making. A variety of elements, including spatial connections, postural considerations, and dynamic adaptability, are included in the qualitative representation of success, which is defined by the logical predicates. The following predicates are applied:

$$in\_front\_of_{x,y}(p,B) \coloneqq p_{x,y} < \min(B)_{x,y} \wedge |p_{x,y} - \min(B)_{x,y}| < C_p$$
$$far\_in\_front\_of_{x,y}(p,B) \coloneqq p_{x,y} < \min(B)_{x,y} \wedge |p_{x,y} - \min(B)_{x,y}| > C_p$$
$$behind_{x,y}(p,B) \coloneqq p_{x,y} > \max(B)_{x,y} \wedge |p_{x,y} - \max(B)_{x,y}| < c_p$$
$$far\_behind_{x,y}(p,B) \coloneqq p_{x,y} > \max(B)_{x,y} \wedge |p_{x,y} - \max(B)_{x,y}| > c_p$$
$$above_{x,y}(p,B) \coloneqq p_z > \max(B)_z$$
$$below_{x,y}(p,B) \coloneqq p_z < \min(B)_z$$
$$centered_{x,y,z}(p,B) \coloneqq |p_{x,y,z} - \max(\bar{B})_z| \leq \epsilon_p$$

Overall, these logical predicates include conditions like "the robot maintains a safe distance from the human recipient," or "the handover is executed in accordance with the detected posture". Such conditions provide a qualitative basis for success, allowing us to comprehensively assess the robot's behavior in real-world scenarios. Because these preconditions are explicit, it is easier to analyze the robot's behavior and provides a transparent and interpretable lens through which to assess the handover task's success. Consequently, we present logical predicates that correspond to every posture in order to describe the factors driving the robot's actions. And based on this, we transformed it into an explanation in natural language to give the robot's decision criteria a more understandable and natural flow, as shown in Tables I-III.

In addition to manual natural language translation, ChatGPT 3.5 [23] was employed to generate automated translation and evaluate the results using the Bilingual Evaluation Understudy (BLEU) score [24]. 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	$\boldsymbol{centered_{y,z}(p,B)} \wedge \boldsymbol{in\_front\_of_x(p,B)} \wedge \neg\boldsymbol{centered_x(p,B)} \wedge \neg\boldsymbol{below_{x,y}(p,B)} \wedge \neg\boldsymbol{behind_{x,y}(p,B)} \wedge \neg\boldsymbol{far\_behind_{x,y}(p,B)} \wedge \neg\boldsymbol{above_{x,y}(p,B)} \wedge \neg\boldsymbol{in\_front\_of_y(p,B)} \wedge \neg\boldsymbol{far\_in\_front\_of_y(p,B)}$
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	$\boldsymbol{centered_{y,z}(p,B)} \wedge \boldsymbol{in\_front\_of_x(p,B)} \wedge \neg\boldsymbol{centered_x(p,B)} \wedge \neg\boldsymbol{below_{x,y}(p,B)} \wedge \neg\boldsymbol{behind_{x,y}(p,B)} \wedge \neg\boldsymbol{far\_behind_{x,y}(p,B)} \wedge \neg\boldsymbol{above_{x,y}(p,B)} \wedge \neg\boldsymbol{in\_front\_of_y(p,B)} \wedge \neg\boldsymbol{far\_in\_front\_of_{x,y}(p,B)}$
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	$\boldsymbol{above_{x,y}(p,B)} \wedge \boldsymbol{centered_y(p,B)} \wedge \neg\boldsymbol{centered_z(p,B)} \wedge \neg\boldsymbol{below_{x,y}(p,B)} \wedge \neg\boldsymbol{behind_{x,y}(p,B)} \wedge \neg\boldsymbol{far\_behind_{x,y}(p,B)} \wedge \neg\boldsymbol{in\_front\_of_y(p,B)} \wedge \neg\boldsymbol{far\_in\_front\_of_{x,y}(p,B)}$
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 [24]. 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) [25] and spaCy [26] 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) was 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. 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.

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

When it comes to interpreting the neural network's decisions about handover position, Grad-weighted Class Activation Mapping (Grad-CAM) [27] 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. By emphasizing the areas of the input data that have a substantial impact on a certain output, it makes the focus and decision-making processes of the neural network easier to visualize. Grad-CAM allows us to produce heatmaps that show the regions of the 3D bounding box that affected the network's choice of a particular handover position. It primarily serves as a diagnostic tool, revealing the internal mechanisms of the network and helping identify potential areas for improvement or refinement.

Figure 3. Simulated heatmaps on one of the handover scenarios.

Grad-CAM performs best in models with a Global Average Pooling (GAP) layer following the last convolutional layer, which provides a spatial localization of the model's

focus. Grad-CAM would be implemented by first loading a neural network model that has already been trained, together with its weights, using TensorFlow, a deep learning framework that is commonly used for this kind of operation. The last classification layer is then eliminated from the model since Grad-CAM requires the output of the last convolutional layer. The next step involves constructing a new model that includes the layers of the pre-trained model up to the final convolutional layer and the GAP layer. This model is intended to produce the model's prediction as well as the output of the final convolutional layer. Unfortunately, the problem with our Toyota HSR prevented us from implementing this method. Despite this obstacle, a previously collected dataset from our research team [28] was leveraged, and the video content was edited to achieve the same heatmap effect (as seen in Figure 3).

This decision allowed us to simulate and observe the intended outcomes, ensuring the continuity of the research despite the technical constraints. 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 allowed us to proceed with our user study within the designated timeframe, preserved the research objectives, and ensured 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. The feedback collected was then analyzed to confirm or deny our hypotheses on users' inclination for technical explanations and their preferred format. Beyond hypothesis testing, the survey plays a crucial role in unraveling the intricate entanglement of users' needs and expectations. For technology to be designed that fundamentally appeals to users, it is extremely important that these demands be understood.

We adopted 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' experiences, preferences, and challenges were carefully examined, 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. The videos intentionally featured a mix of both successful and unsuccessful handovers to capture a broad spectrum of user experience. In order to prevent any potential biases, eight out of ten 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.

The demographic data we were able to obtain from participants via our questionnaire presents a varied view and offers important new perspectives on the traits of our study participants (shown in Figure 4). In terms of age distribution, the vast majority of participants (72.7%) are between the ages of 26 and 40, which indicates that this group is well-represented. In addition, 21.2% of participants are between the ages of 18 and 25, indicating a representation of younger people, while 6.1% are over 40, adding to the age diversity of the study.

Meanwhile, gender representation showcases a predominant presence of males, constituting 66.7% of the participant pool. On the other hand, 27.3% of respondents identify as female and the remaining participants identified as non-binary or opted not to disclose their gender. In terms of education, 45.5% hold a Master's degree, 39.4% have a Bachelor's, 12.1% completed high school, and 3% hold a PhD.

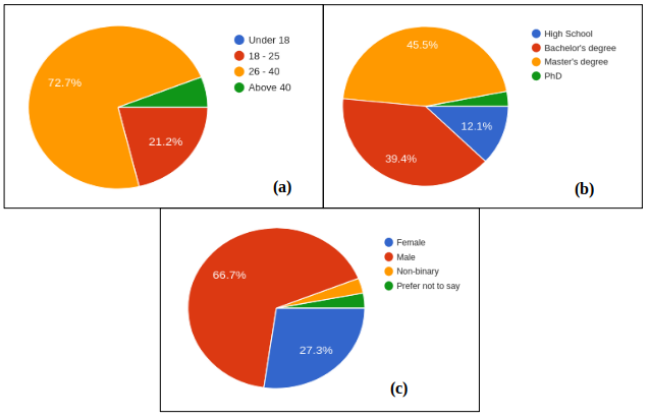

Figure 4. (a) Age range of our user study participants (in years); (b) Gender representation of our user study participants; (c) Education background of our user study participants.

In conclusion, our participants' demographic profiles show a variety of age groups, gender identities, levels of education, and fields of study. This diversity ensures that the responses gathered are representative of a wide cross-section of people, which strengthens the reliability and generality of our user study findings. Based on this diversity, we attempt to determine whether there is any relationship between the preferred explanation technique and the educational background.

A. Quantitative Analysis

Participant profiles in the study reveal that 60.6% of participants have a background in computer science, which is a significant majority. With 18.2% of the total, engineering is the second most common field. Other fields that represent a wide range of academic and professional backgrounds among our participants are finance (9%), natural science (6.1%), social science (3%), and administration (3%). In terms of the participants' experiences and expectations in the realms of robotic systems and 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.

Table V indicates that the majority of individuals who have experience with robotic systems or AI, including computer science and engineering, prefer natural language over heatmaps. This suggests that, even if they could be considered experts in the topic, they would still rather receive a full explanation than a simplified one. In our qualitative analysis (Subsection C), we will discover more about their reasoning for this preference.

TABLE V. AN OVERVIEW OF PARTICIPANTS' EDUCATIONAL BACKGROUNDS AND THEIR PREFERRED TYPES OF EXPLANATIONS

No.	Educational Background	Preferred Type of Explanation	
		Natural Language	Heatmap
1	Computer Science	8	6
2	Engineering	5	1
3	Natural Science	1	1
4	Administration	1	0
5	Finance	1	2

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, despite it being a successful object handover scenario. However, the participant's confidence level improved after watching the second video, which revealed a positive beginning. Table VI summarizes participants' confidence levels after eight more videos were shown in a random order.

TABLE VI. 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

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. To conclude, compared to visual explanation (using a heatmap), natural language explanation improves their confidence by over 30% (shown in Figure 5).

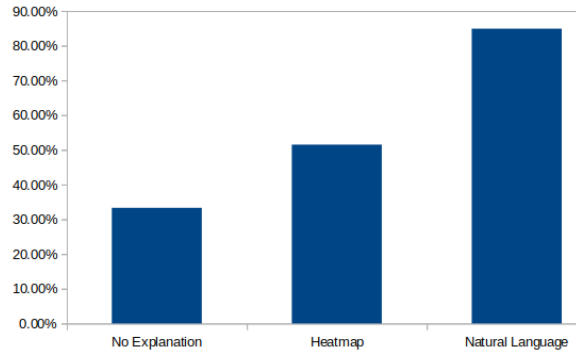

Figure 5. Participants' overall confidence in understanding the robot decision-making process.

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 [29] 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 (χ^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 VII. 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 VII. 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 χ^2= 28.1; with df = 2, the resulting p-value was 0.0000008. Since the p-value is less than α = 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) [30] 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 [31] 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. Using this result combined with significant value (α= 0.05) and number of participants (N = 33), we achieved statistical power of 0.978 or 97.8%, indicating that the test was adequately powered to detect strong differences in preference.

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. This effect size corresponds to statistical power of 0.369 or 36.9%, which suggests that our study may have been underpowered to reliably detect effects of this magnitude. Therefore, we suggest future studies to be conducted with a larger sample size to confirm our finding.

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. In addition to asking the respondents to rank their degree of confidence in each scenario, our scenario-based questions also inquired about any confusion or difficulty the respondents may have had comprehending the robot's decision-making process. This is an outline based on various levels of explanation given:

1) Without Explanation: many wondered why the robot did not come closer to the person and were curious as to how the robot decides when to release the object. In other instances, they questioned what the robot was trying to accomplish and what factors might have played a role when the robot was unable to correctly pass on the item.

2) With Heatmap: even if the heatmap makes the target area easier for them to understand, they believe that more thorough explanation is necessary in order to understand it completely. In their opinion, the heatmap did not help them understand why the object fell or why the robot was unable to detect the user's hand.

3) With Natural Language: Many of them highlighted that the explanation was too lengthy and that they were unable to follow it quickly. Moreover, the goal of the robot was not stated in the provided explanation, and it excluded the reason for the robot's inability to recognize a person's hand.

Given these uncertainties, we have also asked the participants to suggest ways we could improve the overall system. Of the 33 respondents, 8 people said that they have no suggestions regarding how the robot's performance or the explanations it provides could be made better. Most, nonetheless, recommended shorter or simpler sentences for an explanation in natural language or integrating a voice or audio option to improve the explanation process. In order to visualize information more clearly, participants suggested utilizing arrows or targets instead of heatmaps. For a more interesting and educational experience, some participants suggested an interactive graphical explanation that incorporates both written and graphic elements. Other recommendations centered on instructing the robot to recognize what type of item it is handling. To protect delicate objects like glasses or plates, for instance, the robot should handle them with caution and hold them upright.

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.

Last but not least, the following are some further comments, suggestions, or feedback about the application of AI systems with various levels of explainability: in order to give users the flexibility to select the level of explanation they are looking for and when it should be provided, the majority of them recommended making explanations available upon request; a few more said that the whole experience would be enhanced by increasing the level of engagement: to make that happen, they recommended implementing a voice-activated technology; one other noteworthy feedback is that they recommend providing additional context or reference data, such as the user's height or the robot's expected reach, to help non-expert users who are less familiar with the system understand it better.

V. CONCLUSION AND FUTURE WORK

Effective communication of intent is crucial for successful interaction between humans and robots in applications involving HRI. Humans need to understand the robot's internal state, including its purposes and intentions, to engage in collaborative tasks effectively. This requirement has fueled the development of explanation techniques that make the robot's internal workings transparent to people. Given that explanations are essential for building user confidence and comprehension of robotic systems, our research emphasizes how important it is to include them in HRI. The capacity to explain robot decision-making becomes crucial for users to interact with these intelligent machines in a meaningful way as they become more and more integrated into diverse sectors. In order to fulfill this demand, our method entails offering two different kinds of explanations inside the robot handover framework. First, we added natural language explanation, translating complex logical expressions into easily understandable statements. This improves the interpretability of the robotic system's activities and makes them more consistent with human cognitive processes, which makes it easier for people to understand the robot's behavior more naturally. Second, we introduced a heatmap-like explanation in the style of Grad-CAM that could be applied to a trained neural network. This visual representation provides a supplementary and visual explanation by enabling users to identify the regions that are relevant within the input data that impacted the robot's decision-making.

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.

Our research lays the groundwork for compelling avenues of future work, particularly in the domain of automating the translation of scientific terms into natural language. Building on the success of employing ChatGPT 3.5 to translate logical expressions, the next logical step involves expanding this method to encompass a broader spectrum of scientific terminology. By refining and extending the capabilities of such model, we envision a system that can seamlessly convert intricate scientific terms into clear and accessible natural language, catering to both expert and non-expert users. One prospective approach for this development involves fine-tuning language models on specialized scientific domains. The model could be adapted to recognize and translate specific scientific terms, ensuring accurate and contextually relevant conversions. Additionally, incorporating domain-specific ontologies or knowledge graphs could enhance the model's understanding of relationships between scientific terms, contributing to more coherent translations.

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] K. Sari, P. G. Plöger, and A.Mitrevski, “Explainability Analysis for Skill Execution,” in EXPLAINABILITY 2025 The Second International Conference on Systems Explainability, pp. 14-20, 2025.

[2] 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.

[3] W. Samek, T. Wiegand, and K.-R. Mueller, “Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models,” arXiv preprint arXiv:1708.08296, 2017.

[4] 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.

[5] 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.

[6] D. Gunning and D. Aha, “Darpa’s explainable artificial intelligence (xai) program,” AI Magazine, vol. 40, no. 2, pp. 44–58, 2019.

[7] 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.

[8] 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.

[9] T. Sakai and T. Nagai, “Explainable autonomous robots: A survey and perspective,” Advanced Robotics, vol. 36, no. 5-6, pp. 219-238, 2022.

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

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

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

[13] M. Lomas et al., “Explaining robot actions,” in Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot, pp. 187-188, 2012.

[14] 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.

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

[16] 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.

[17] 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.

[18] S. B. Remman and A. M. Lekkas, "Robotic lever manipulation using hindsight experience replay and shapley additive explanations," in 2021 European Control Conference (ECC), pp. 586-593, 2021.

[19] 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.

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

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

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

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

[24] 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.

[25] S. Bird, "NLTK: the natural language toolkit," in Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, pp. 69-72. 2006.

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

[27] 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.

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

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

[30] 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.

[31] Cohen, D. "Culture, social organization, and patterns of violence," Journal of Personality and Social Psychology, vol. 75, no. 2, pp. 408–419, 1998.