Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

IARIA

b-it Bonn-Aachen International Center for Information Technology

Explainability Analysis for Skill Execution

Khatina Sari, Prof. Dr. Paul G. Plöger, Dr. Alex Mitrevski

Barcelona, Spain, October 30, 2025

Presenter
Khatina Sari
Department of Computer Science
Email: khatina.sari@smail.inf.h-brs.de

Khatina Sari

Master of Autonomous Systems (M.Sc.), H-BRS
Fullstack Developer at GiPsy Software Solutions

Research Interests: *Explainable AI, Human-Robot Interaction, Deep Learning.*
Publications & Projects

- *Assessing the Robustness of Learning-Based Robot Skill Models*, 2025
- *Developing Facial Recognition with CNN and Liveness Detector for Attendance App*, 2022
- *Developing Chatbot for PT. Everbright Jambi*, 2021
- *Analyzing Local E-government Success Using DeLone and McLean Information System Success Model*, 2019

Introduction

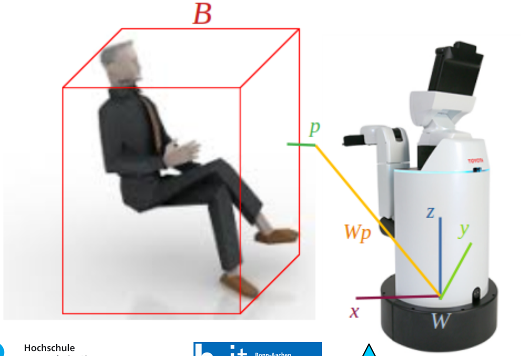
Explainability in Object Handover Tasks aims to enhance user understanding and trust in robot actions.

- Autonomous robot behavior is often difficult to explain due to dynamic and uncertain environments.
- Diverse user knowledge and expectations make it challenging to maintain the appropriate level of explanation detail.
- Emphasizes the importance of bridging the knowledge gap between humans and robots to enable effective collaboration in HRI.

User Study: Evaluating the effectiveness of multiple levels of explainability in item handover tasks.

Basic Setup

- A 3D bounding box is used to **detect people within the robot's environment**.
- Distinguishes three **human positions:** standing, sitting, and lying down.
- **Notations:** B = bounding box, W = robot base frame, W_p = robot end-effector, p = relative point between end-effector and bounding box.



Spatial Predicates
$in\_front\_of_{x,y}$ (p,B)
$far\_in\_front\_of_{x,y}$ (p,B)
$behind_x$ (p,B)
far_behind_x (p,B)
$above_{x,y}$ (p,B)
$below_{x,y}$ (p,B)
$centered_{x,y}$ (p,B)

Natural Language Explanation

Success precondition & its translation for each position

Position	Logical Predicates	Natural Language Translation
Standing	$centered_{y,z}$ (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)	"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."
Sitting	$centered_{y,z}$ (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)	"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."
Lying Down	$above_{x,y}$ (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)	"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 Score Evaluation

We **automatically generated translations using ChatGPT 3.5** and evaluated their quality using the **Bilingual Evaluation Understudy (BLEU)** metric, where scores range from 0 to 1.

BLEU measures the similarity between a machine-generated translation (candidate) and a human reference translation. Shorter candidate sentences are penalized through the **Brevity Penalty** component of the metric.

Table 1: BLEU scores for ChatGPT 3.5-generated translations by position

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

Neural Network Interpretation

Proposed Approach: *Neural Network & Grad-CAM*

- **Automatic generation of handover positions** using a neural network model.
- **Post-hoc visual explanation** through Grad-CAM heatmaps to highlight decision regions.

Implementation Note: *A simulated heatmap was used due to a temporary technical issue with the robot.*

Figure 1: Left: Original handover scenario. Right: Simulated Grad-CAM heatmap visualization.

Experimental Design

- **Three explanation methods evaluated:** no explanation, partial explanation, and detailed explanation.
- **Explanations integrated** into a previously collected dataset[1].
- **Two hypotheses** formulated to guide the study.

Hypothesis 1

Users prefer to use robotic systems that have explanations over those without.

Hypothesis 2

Natural language explanations are preferred by users over alternate visualization methods like heatmap.

[1] https://codalab.lisn.upsaclay.fr/competitions/6757#learn_the_details-dataset

Experimental Setting

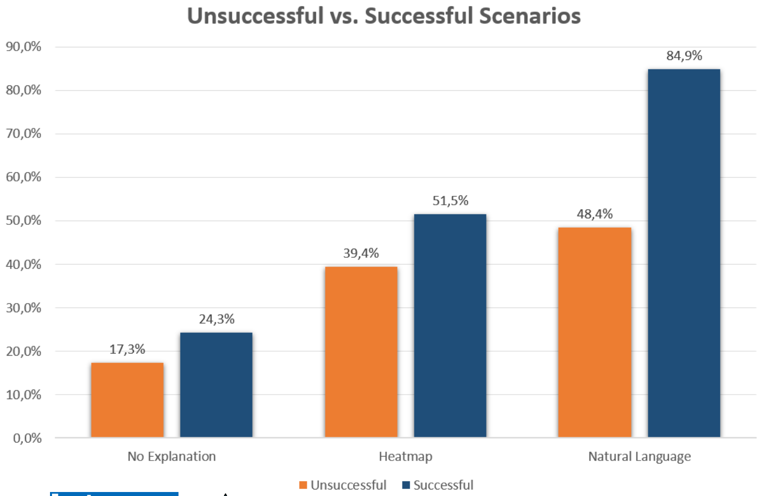
- **Online questionnaire** conducted for broad participant reach and real-time data collection.
- **Video 1 (no explanation)** and **Video 2 (with explanation)** were identical in content.
- Participants were shown **8 out of 10 videos** in random order to minimize potential bias.
- Participants **rated their confidence** in understanding the robot's decision-making process (5 = very confident, 1 = not confident at all).
- **Preference question:** "Which type of video do you prefer when seeking information?" Response options: video with explanation, without explanation, or depends on the context. (Used for Hypothesis 1)
- **Video 4 (heatmap)** and **Video 8 (natural language)** were identical and used for Hypothesis 2 testing.

Results

- **33 participants** aged between 18 and over 40 years.
- Educational background ranged from **high school to PhD** levels, representing diverse academic and professional fields.
- Most participants had **prior hands-on experience** with robotic systems (75.8%) and were **familiar with AI/ML concepts** (84.8%).
- Over two-thirds reported feeling **uneasy with AI systems lacking explanations**.
- After viewing **Video 1 (no explanation)**, most participants expressed uncertainty, whereas **Video 2 (with explanation)** led to a noticeable increase in confidence levels.

Overview of Participants' Confidence Levels

in relation to scenario-based questions (Videos 3–10)



Hypotheses Testing

Hypothesis 1 – Chi-square Test

- H_0: No preference difference among explanation types.
- H_1: Preference exists for videos **with explanations**.

Test result: $\chi^2 = 28.1$, $df = 2$, p-value = 0.0000008 $< \alpha = 0.05$
$\Rightarrow$ Reject H_0 (significant difference found).

Hypothesis 2 – One-sample Proportion Test

- H_0: No preference difference.
- H_1: Preference exists for **natural language explanations**.

Test result: $Z = 2.46$; with $\alpha = 0.05$, critical range = [-1.96, 1.96].
Since $Z > 1.96$, Reject H_0 (significant preference found).

Post-hoc Power Analysis

H1: Large effect ($w = 0.70$), power = 0.98 (robust)
H2: Medium effect ($h = 0.40$), power = 0.37 (limited sensitivity)

Conclusions

- Providing **explanations enhances** users' **trust and understanding** of robot actions.
- **Heatmaps** provide a **medium level of explainability**, suitable for simple or straightforward robot tasks.
- **Natural language explanations** offer **high-level interpretability** for more complex robot behaviors.
- Findings highlight the importance of **adaptive communication strategies** for effective Human–Robot Interaction (HRI).

Lessons Learned

- **Potential response bias** in Hypothesis 1 — the question wording may have influenced participants toward explanations.
- The outcome of Hypothesis 2 may reflect **asymmetry in explanation format** — heatmaps require greater cognitive interpretation than text.

Future Work

- Explore **automated generation of natural language explanations** for dynamic, context-aware communication.
- Incorporate **interactive explainability features** that enable user engagement, such as:
  - Combination of **natural language and heatmaps** for multi-modal feedback.
  - **Audio or speech-based** explanation delivery for accessibility and realism.
- Apply **machine learning techniques** to **personalize explanation types** based on individual user profiles.