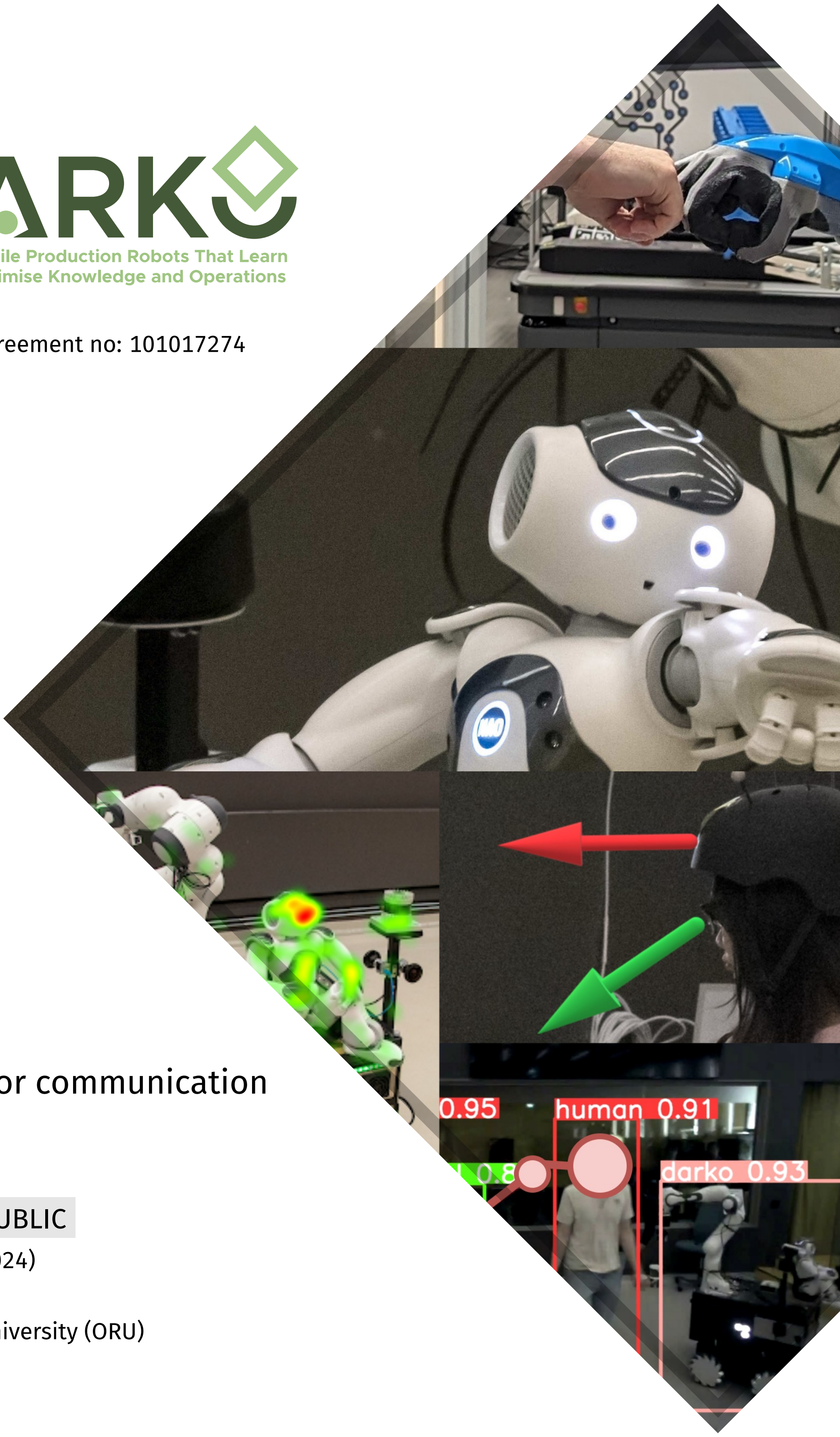
Dynamic Agile Production Robots That Learn and Optimise Knowledge and Operations

H2020-ICT-2020-2 Grant agreement no: 101017274



DELIVERABLE 5.2
Report on system for communication of robot intent

Dissemination Level: PUBLIC

Due date: month 43 (July 2024)

Deliverable type: Report

Lead beneficiary: Örebro University (ORU)

Contents

List of Abbreviations

Abbreviation	Meaning
6-DoF	6 degrees of freedom (3D position, 3D orientation)
AGV	Autonomous Guided Vehicle
AOI	Area of Interest
API	Application Programming Interface, the public interface provided by a library for use by software developers
ARENA 2036	A large research campus in the form of a modern factory hall in Stuttgart-Vaihingen, Germany. Provides an innovation platform for mobility & production of the future and hosts DARKO project demonstrations.
ARMoD	Anthropomorphic Robotic Mock Driver
DIN	Deutsches Institut für Normung
EN	European Norm (Standard)
FOV	Field of view of a sensor
sHRI	spatial Human-Robot Interaction
ILIAD	EU Horizon 2020 project (2016–2020) which deployed a heterogeneous fleet of mobile service robots in intralogistics environments.
ISO	International Organization for Standardization
LiDAR	Light Detection And Ranging, a time-of-flight-based sensor that produces point clouds. Also spelled “lidar”.
LLM	Large Language Model
NAO	A type of humanoid robot
ORU	Örebro University, member of the DARKO consortium
RGB-D	Red, Green, Blue - Depth
ROS	Robot Operating System, see www.ros.org
RRS2	(Refers to a package used in the statistical analysis, possibly a typographical variant of WRS2)
SDK	Software Development Kit
SPENCER	EU FP7 project (2013–2016) which deployed a mildly humanized service robot in a busy airport terminal at Amsterdam Schiphol Airport.
TobiiProLab	Software used for eye-tracking data analysis
TUM	Technische Universität München, member of the DARKO consortium
UNIPI	Università di Pisa, member of the DARKO consortium
WP	Work package
YOLO	A series of 2D object detectors developed by J. Redmon

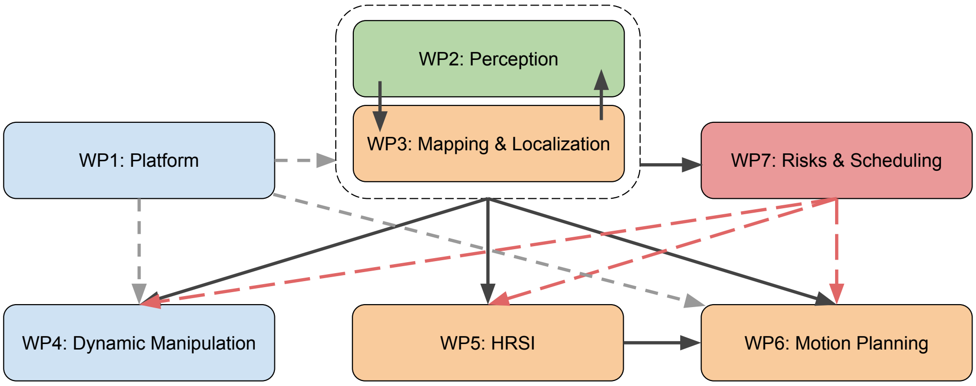


Figure 1: Relation of WP5, which this deliverable reports on, to other work packages in DARKO. Black arrows denote data flow during operation, dashed red arrows indicate constraints and orchestration, and dashed grey arrows indicate hardware dependencies.

1 Introduction

The deliverable reports on the final system for communication of robot intent developed in the EU H2020 task T5.2, including its scientific results and the final software prototype. Project partners contributing to this deliverable are Robert Bosch GmbH (BOSCH, lead responsible), Örebro University (ORU), and Technical University of Munich (TUM).

1.1 Relation to other work packages

Figure 1 illustrates the relation of work package WP5, which this deliverable reports on, to the other technical work packages in DARKO.

1.2 Key highlights and improvements over prior work

In the first three periods of DARKO, the consortium developed several novel methods, experiments, and datasets that directly addressed the key objectives of DARKO. Combined together, they have led to the DARKO intent communication system presented in this deliverable. Key highlights and improvements over prior work include:

- **Robotic Intent Communication with an Anthropomorphic Robotic Mock Driver (ARMoD):** Enhanced trust and interaction quality using human-like gestures, gazes, and speech.
- **Analysis of the THÖR-Magni Multimodal Dataset of Human Motion:** Provided insights for designing intuitive robotic systems by analyzing human visual attention and engagement.
- **LLM-powered HRI in Dynamic Settings:** Improved reasoning and communicative capabilities for more adaptive interactions.

1.3 Structure of this deliverable

The following sections detail the research and results on robotic intent communication obtained in Task T5.2 of Work Package WP5. The structure is as follows:

Robotic Intent Communication Section 2 explores the concept of robot intent communication we developed for the DARKO robot. It begins by identifying the native communication channels available to the DARKO robot, such as manipulators, LED strips, motion, and mounting spaces for additional solutions. We evaluate the potential of each channel to convey the robot's intentions in industrial contexts. The section also delves into anthropomorphic robot intent communication, discussing how human-like characteristics can enhance interaction and trust between robots and humans.

Designing the Anthropomorphic Robotic Mock Driver (ARMoD) Section 3 presents a detailed account of the Anthropomorphic Robotic Mock Driver (ARMoD) development. This section introduces the concept and motivation behind using anthropomorphic features to improve human-robot interaction. It provides a detailed account of the initial design and methodology, which were employed to assess the impact of the ARMoD on trust and interaction quality prior to initial interactions. We discuss the results of our initial investigation and their implications for enhanced robotic intent communication.

Experimental Evaluation of the ARMoD Section 4 presents a comprehensive experimental evaluation of the ARMoD concept. It outlines the methodology and design of experiments to validate the ARMoD's interactive capabilities. We tested two interaction styles: a verbal-only and a multimodal one, including robotic gaze and pointing gestures. The results indicated that the multimodal interaction style led to more natural fixation behavior and faster reaction times in collaborative tasks. This gaze behavior demonstrates the ARMoD's potential to enhance engagement and social interaction in workplace settings.

Insights from the THÖR-MAGNI Dataset Section 5 presents an analysis of the THÖR-MAGNI dataset, which offers valuable insights into spatial human-robot interactions involving the DARKO robot. The dataset includes scenarios that anticipate human movements and intentions, providing a comprehensive repository for developing intuitive robotic systems. The analysis focuses on aspects of human visual attention and engagement in shared environments, highlighting how these factors influence the perception and effectiveness of the DARKO robot conveying information. Through an examination of human gaze patterns and cognitive engagement, we highlight the importance of effective communication protocols for robotic intent communication.

Ongoing and Future Work The final Section 6 discusses the ongoing and future work within the work package. We describe the ongoing investigation of large language models (LLMs) for bidirectional human-robot interaction (HRI), examining how these models can augment the communication and reasoning capabilities of our developed system. Furthermore, future work will investigate the potential of using LEDs as a communication channel, focusing on their efficacy in various operational contexts. We highlight the potential of research and development to further enhance the efficacy of robotic intent communication.

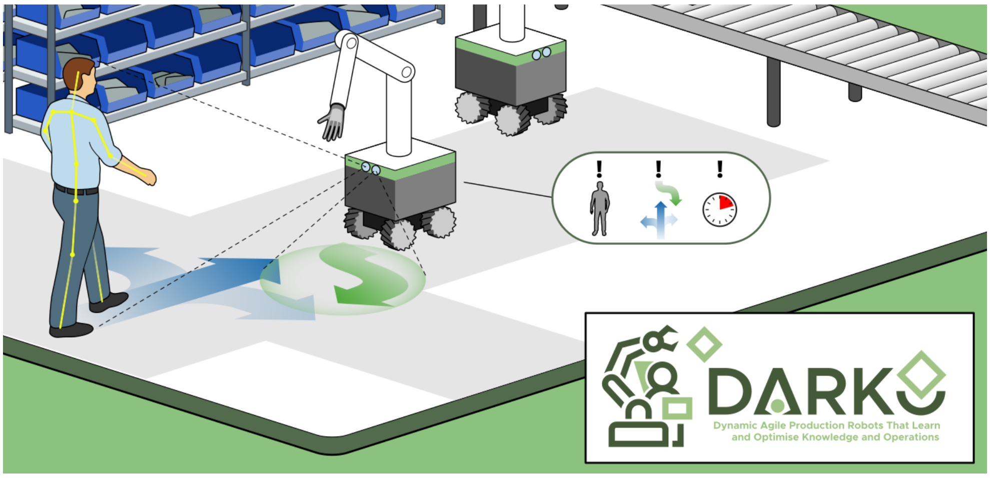


Figure 2: DARKO Concept image: A factory worker encounters the DARKO robot at an intersection. The robot must perceive the human and their intentions, determine an efficient route, and communicate this to the worker. Ideally, this interaction should not disrupt the worker's workflow (efficiency) and should be completed in the shortest time possible (timeliness).

2 (Robotic) Intent Communication in the context of DARKO

Effective communication of intent is a critical aspect of human-robot interaction, ensuring that robots can operate safely and efficiently in dynamic environments alongside humans. Intent communication refers to how a robot conveys its planned actions, goals, and status to human collaborators and other robots. This capability is vital for fostering trust, enhancing collaboration, and preventing accidents in shared spaces [1].

According to the survey by Pascher et al. [1], understanding and effectively communicating a robot's motion intent is crucial for avoiding task failures and collaboration in human-robot interactions. Multiple channels should be used to achieve cooperation between humans and mobile robots in a timely and efficient manner [2], highlighting the need for a unified language and systematized approach to communicating robot motion intent.

Figure 2 illustrates this challenge with a scenario where a factory worker encounters the DARKO robot at an intersection. In this situation, the robot must perceive the human and their intentions, determine an efficient route, and communicate this to the worker without disrupting their workflow and as quickly as possible. This scenario exemplifies the complexity of negotiating trajectories and communicating intent, a traditionally challenging task for humans, especially in narrow environments [3]. Designing and evaluating effective communication and universal channels for robots sharing environments with humans is the key to resolving these situations.

The first step in this process is to assess which communication channels are available to a robot by its native design and evaluate if these channels suit the potential tasks that must be resolved or situations that could occur in daily encounters. If the native channels are lacking, additional channels must be considered to ensure the robot can communicate effectively in various operational contexts.

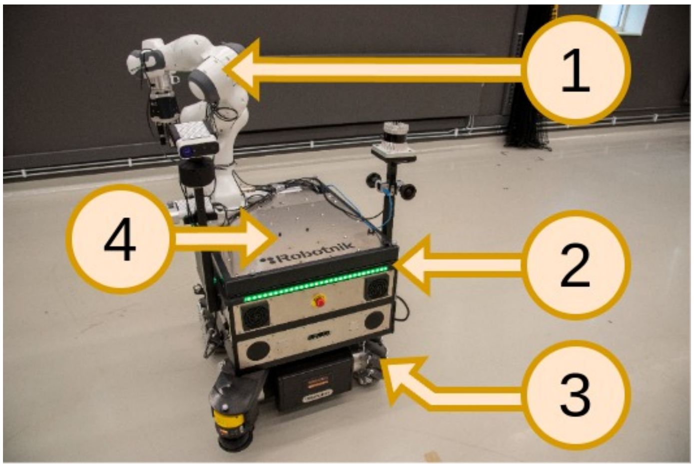


Figure 3: Intent Communication Channels on the DARKO Robot: The DARKO robot has several native design features to communicate intent. 1: Manipulator, 2: LED Stripes, 3: Mecanum wheels, 4: Mounting Space for On-Robot intent communication using Robot-Attached Solutions.

2.1 Native communication channels available to the DARKO Robot

The DARKO robot, our primary subject of study, is designed with several built-in features to communicate its intent displayed in Figure 3. These features include 1 a manipulator, 2 LED stripes, 3 mecanum wheels for driving, and 4 mounting space for additional on-robot solutions. Each component is a potential channel through which the robot can express its intentions to its surroundings, especially in industrial contexts where precise and clear communication is essential.

2.1.1 Robotic Manipulator

Intent communication using a robotic manipulator involves advanced methods such as recognizing human intent through force exchanges in collaborative manipulation [4], utilizing a gesture pseudo-language [5], and establishing intent communication models based on different intent dimensions [1]. While these traditional approaches are practical in scenarios with proximity or direct physical interactions between humans and manipulators [1], they are partially impractical for the DARKO project, where the robot navigates the factory floor and frequently has to interact with workers from a distance. Additionally, using the manipulator for intent communication could interfere with other critical work packages, such as WP1 (developing an elastic manipulator for throwing) and WP4 (efficient and safe dynamic manipulation).

2.1.2 LED Stripes

The use of LEDs for intent communication in robots has been widely discussed and explored due to their effectiveness in providing clear and intuitive signals to human collaborators. According to the survey by Pascher et al., [1], LEDs are often employed to communicate various types of robot states and intentions in a manner easily perceivable by humans. LEDs can change colors, blink, or form patterns to convey messages about the robot's current state and upcoming actions or provide warnings. For instance, colored LED stripes can be used to indicate the robot's operational state, such as active or inactive, which helps humans predict future motions and identify potential conflicts before they occur (e.g., the robot communicates its movement activity with the help of a colored LED stripe [6]). Additionally, LED signals can be employed to catch the user's attention before a robot's movement or activity [7], thereby preparing them for subsequent actions and improving overall safety and coordination (e.g., by moving its whole body or blinking LED lights to signal the intention to move). LEDs are a versatile, cheap, and effective tool for enhancing human-robot interaction by making the robot's intentions clear and easily understandable. This facilitates smoother and safer collaboration between humans and robots in various environments.

2.1.3 Intention communication through motion

The DARKO robot uses mecanum wheels to navigate around. Driving motion and direction inherently convey information about direction and speed. By refining how the motion signals the robot's intentions, we can enhance the clarity and predictability of its actions. This is particularly important in industrial settings where robots and humans share close quarters, and precise movements are necessary to avoid accidents [8]. The DARKO robot's wheels enable it to move directionally and omnidirectionally. Both modes of operation are valid for different situations. To allow a study of how intent can be communicated using these driving styles, it is necessary to investigate the human perception of these movements first, as humans potentially do not understand the displayed signals of an autonomous vehicle without prior knowledge [9].

2.1.4 Mounting Space for Robot-Attached Solutions

This feature provides a flexible platform for integrating additional intent communication tools like projectors, sound emitters, display screens, and other devices. The adaptability of the mounting space allows for the implementation of various systems tailored to specific operational contexts or user requirements. Using this additional space, the robot's communication capabilities can be enhanced with multi-modal and redundant signals, ensuring clear and compelling intent communication even in noisy, cluttered, or visually complex environments. This continuous evolution of the robot's communication methods enables real-time updates and modifications to meet emerging needs and technologies, significantly improving interaction with human co-workers in workplace settings.

2.2 Anthropomorphic Robotic Intent Communication

The native communication channels, such as the manipulator, LED stripes, and driving motion, offer a range of methods for conveying intent. However, they each have specific challenges and constraints in dynamic and complex industrial environments and may not be generalized between different mobile robots. Recognizing these limitations proves the need for effective communication methods universally applicable to mobile robots that need to communicate with humans in various work environments. Research has explored additional robot-attached channels such as floor projections [10, 11]. Despite these advancements, the need for approaches that can be validated and used across a range of mobile robots remains open [8, 12]. Building on this research, we designed the concept of the "Anthropomorphic Robotic Mock Driver" (ARMoD) to facilitate intuitive communication between non-humanoid robots and human co-workers in workplace settings [2]. The ARMoD concept involves integrating a humanoid robot onto the DARKO robot to leverage anthropomorphic features for more natural and intuitive communication. This approach was initially inspired by ongoing work at ORU on human-robot interaction and the promising results of established research such as [13, 14].

We pursued the development of the ARMoD as an innovative approach to augment any mobile robot's communication capabilities (see Figure 4). The ARMoD concept involves integrating a humanoid robot onto a mobile robot to leverage anthropomorphic features for more natural and intuitive communication. Mounting an ARMoD on the DARKO robot provides additional communication channels such as a Head for robotic gaze, a Text-To-Speech Engine for Dialogues, Additional LEDs and Arms for pointing, and other gestures.

Enhanced Social Cues: The ARMoD can use anthropomorphic features such as head movements, robotic gaze, and gestures to communicate intent, providing implicit cues similar to those used in human communication, enhancing the clarity and naturalness of a robot's communication in shared spaces. Especially naturalistic cues, like gaze direction, might be more easily interpreted than traditional signals, such as LED turn signals, providing a more intuitive and effective means of communication [15].

Versatility and Adaptability The ARMoD can be programmed to perform a wide range of communicative behaviors independently and in parallel to the DARKO robot, allowing it to adapt to various situational needs. Whether directing attention, providing instructions, or signaling warnings, the ARMoD's flexible design enables it to handle diverse communication requirements.

Potential of Anthropomorphic Features The anthropomorphic appearance of robots plays a crucial role in influencing users' emotional experiences and attitudes towards them. Research suggests that moderately anthropomorphic service robots evoke more positive emotions than highly or minimally anthropomorphic ones (Native DARKO Robot), as they induce higher pleasure, arousal, and physiological responses in users [16]. Furthermore, initiating interaction through social humanoid robots facilitates human-robot interactions, making them more natural and efficient, which is essential for seamless integration into daily human life or activities [17]. The anthropomorphic design of the ARMoD has been found to enhance appearance-based trust in the platform [18], which is especially important

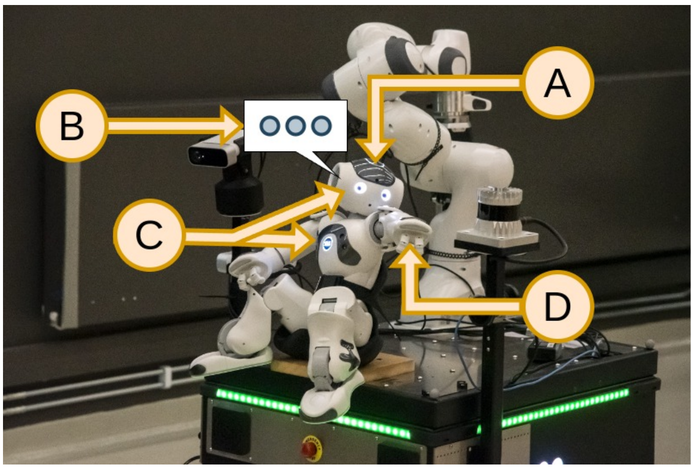


Figure 4: Additional Intent Communication Channels with an ARMoD: Equipping an ARMoD provides additional channels to leverage its anthropomorphic features to communicate intent. A: Head for robotic gaze, B: Text-To-Speech for Dialogues, C: Additional LEDs, D: Arms for pointing and other gestures.

for integrating robots into human-centric workspaces, where acceptance and cooperation from human co-workers are essential. Given these initial findings, we continued to develop the ARMoD concept, building on ongoing advancements in humanoid robotics.

2.2.1 Structure of the Development Process of ARMoD

In the following sections, we will outline the engineering and academic development process of the ARMoD concept and its potential to enhance communication in human-robot interaction and collaboration in industrial settings. Our evaluation shows that incorporating humanoid elements improves attention and engagement, leading to faster reaction times and efficient cooperation. The ARMoD also unifies cognitive activation and elevates the engagement among participants by using intuitive human-like gestures and expressions, benefiting fast-paced industrial environments. Lastly, the ARMoD facilitates future research into large language model (LLM)-empowered HRI and enables advanced, natural communication between humans and robots.

3 Designing the Anthropomorphic Robotic Mock Driver

Summary: Robots are increasingly deployed in spaces shared with humans, including home settings and industrial environments. In these environments, human-robot interaction (HRI) is crucial for safety, legibility, and efficiency. Trust is a critical factor in HRI, significantly influencing system acceptance. While anthropomorphism has enhanced trust development in robots, industrial robots are typically not anthropomorphic. To address this, we initiated a study to test the ARMoD concept by mounting it on an autonomous guided vehicle (AGV) in an industrial environment, as shown in Figure 5. This first proof of concept was part of ongoing work at the time of the DARKO project proposal. In the study, we designed a simple interaction where a human and the AGV had to negotiate trajectories in a narrow corridor with or without the ARMoD mounted on top. This required the human to attend to the robot's trajectory to avoid collisions. The results demonstrated a significant increase in reported trust scores when the ARMoD was present. This finding indicates that the presence of an anthropomorphic robot is sufficient to modulate trust, even in brief interactions. This initial study provided valuable insights and validated the potential of the ARMoD concept, laying the groundwork for further development and integration into the DARKO project.

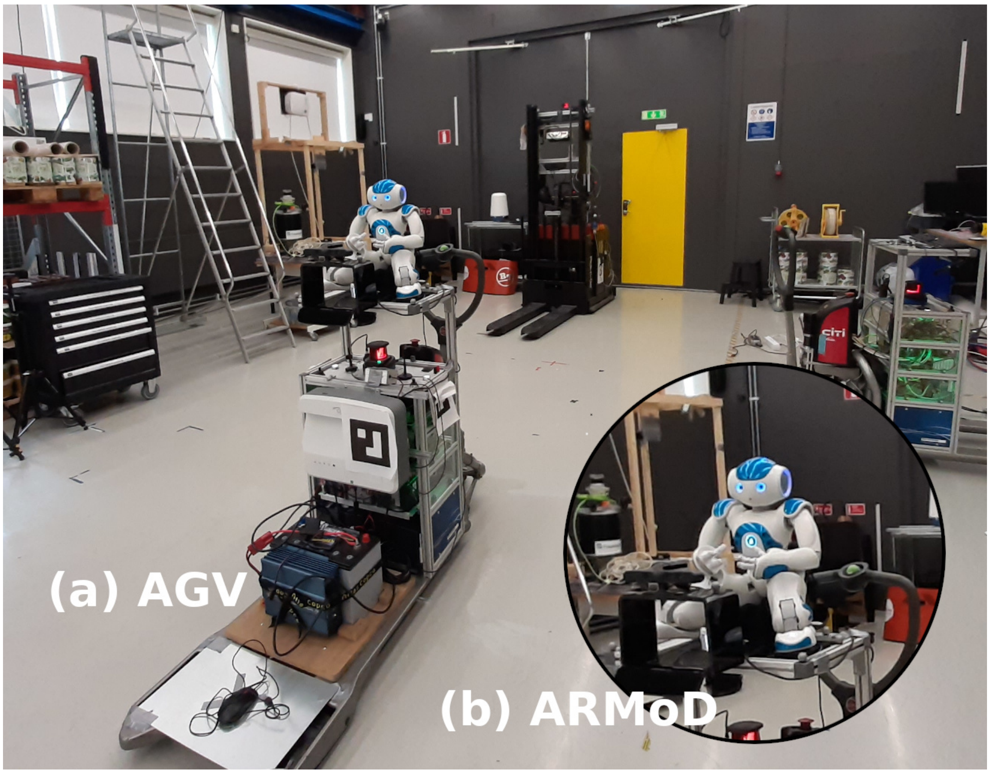


Figure 5: Setup of the (a) Autonomous Guided Vehicle (AGV) with the (b) Anthropomorphic Robotic Mock Driver (ARMoD) placed on top of the vehicle. The AGV can potentially communicate intent through the Driver's gestures, gazes, and speech.

3.1 Introduction

Non-humanoid robots usually have various non-verbal channels to communicate their intent to humans, such as light signals, floor projections, or auditory signals [8]. When navigating in a shared environment, such signals help coordinate their motion with people to avoid collision and increase legibility and task efficiency. As for the tasks that require cooperation and active coordination, like the handover task, more complex communication channels might be necessary, such as verbal communication with additional gestures and gazes. Humanoid robots with anthropomorphic features, e.g., arms, legs, and facial features, are often used in this context to improve interaction with human users [19], [20], and their anthropomorphism leads to an increase of users' trust in the interaction [21, 22].

As the usage and complexity of industrial robots increase, they take on unfamiliar shapes and, thus, complicate the interaction and establishment of trust in shared environments with humans. Robot-related factors were shown to be the most relevant for developing trust in these interactions [23]. Among these factors, the design of the robot is essential to get the human interaction partners to trust the robot appropriately [24]. Anthropomorphic features may aid the trust, but they are not often present in industrial robots. As the complexity of the new systems increases, the perception of these systems as collaborators rather than machines has been deemed positive [25]. For example, they are adding a pair of sunglasses to an industrial robotic hand and gripper, along with a set of breathing-like movements and gaze behavior, improved metrics from participants such as the perceived sociability and likeability of the system [26]. However, the authors did not find differences in trust in their study. One possible reason is that they included a scale not initially designed for industrial collaborations [27]. Another study showed that trust does not seem to be affected due to anthropomorphism in industrial settings [28]. In this case, the authors used a validated scale to trust in industrial collaboration, developed by Charalambous et al.[29]. Nevertheless, the study employed a limited form of anthropomorphism in which a face appeared on a screen attached to a robotic arm and gripper. In contrast to previous research, our study does not include tactile interaction. Our research explores a new perceived navigation modality for a non-humanoid robot that can potentially improve trust in the system by using a real humanoid robot, NAO, in industrial settings.

We propose a combined approach of a humanoid robot with an Autonomous Guided Vehicle (AGV), used, for instance, in the intralogistic settings[1]. We refer to this combination as a "robot-on-robot platform" (see Figure 5). By combining an AGV with sophisticated social robots with anthropomorphic features, successfully used in trust-related user studies [30], we expect to increase human users' trust. This is the first approach to study the interaction of a navigating AGV equipped with an Anthropomorphic Mock Driver (ARMoD) with participants in a shared environment. To this extent, we chose an NAO robot with a human-likeliness score of 46% [31], as it is the subject of recent user-based studies [32], it is small enough to be mounted on our AGV and posses a software development kits to develop custom modules. We designed an encounter in a narrow corridor to measure the impact of the ARMoD on trust reported by participants. In our scenario, as the Participant approached the robot, it looked at the Participant's head and traced it until the vehicle passed the Participant. We used the scale developed by Charalambous et al.[29] to measure trust in industrial collaborations. With the data we obtained from 33 participants, we found that the users reported higher trust in the interaction with the ARMoD than in the AGV alone.

[1] http://iliad-project.eu

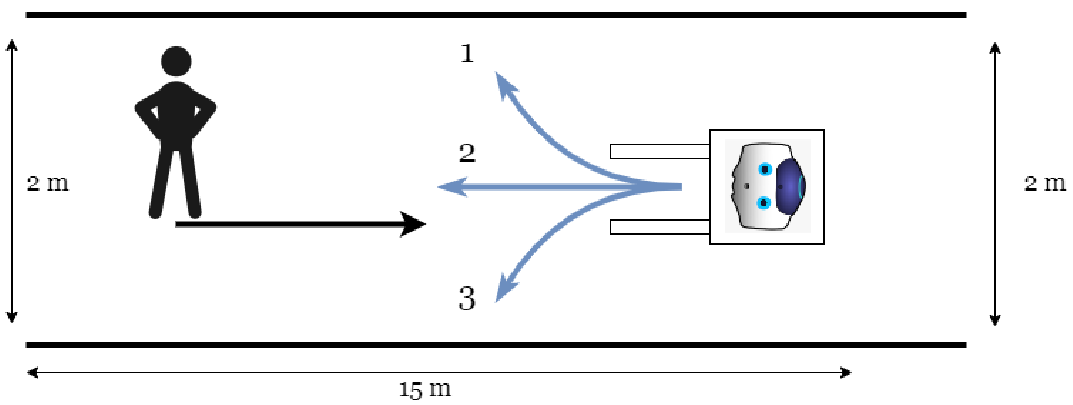


Figure 6: Participant encountering the robot-on-robot platform in a 2-meter wide and 15-meter long corridor. The Participant has to decide on a side to pass the platform. The robot takes one of three different trajectories **(1)** Curve to the right side, **(2)** Straight ahead, **(3)** Curve to the left side.

3.2 Experiment design

This study aimed to explore the impact that an Anthropomorphic Robotic Mock Driver (ARMoD) seated on top of an AGV has on users' trust during an essential human-robot interaction consisting of walking and avoiding the moving platform. The AGV we used fits the definition of "mobile platform" after ISO8373:2021. On top of the platform, a seat was mounted to hold the NAO robot in place, which enabled a fixed and repeatable placement (see Figure 5). We used the same AGV as in our previous studies of intent communication [10], a retrofitted Linde CitiTruck with a SICK S300 scanner at the back, to ensure the safety of humans approaching from behind. The AGV has sensor modules to localize itself in the laboratory and an onboard RGB-D camera for short-range person detection (≈2 meters).

We designed a scenario where participants encountered a moving robot platform in a hallway to study trust as a result of the appearance of the platform, either as it is (AGV) or with an anthropomorphic robot on the top (ARMoD). In the experiment, we chose a setup that reflected a potential encounter between humans and robotic workers in an industrial environment. The chosen width of the hallway was 2 meters, as it matches the regulations for corridors proposed in the DIN-18040-1 and the EN-ISO-24341 (former EN 426) standards for meeting areas. The participants and the platform started 14 meters apart, a feasible length for defining a corridor encounter. During this sequence, the participants saw the platform as it approached them, and they were instructed to walk by its side in the opposite direction. In the ARMoD condition, if the participants got close enough for the short-range person detection (≈2 meters), the robot used its head and simulated awareness to trace the participants' movement until they crossed paths. This encounter was repeated three times with the platform taking one of the different routes: (1) the platform moved in a curve to the right side of the hallway, (2) the platform moved to the left side of the hallway, and (3) the platform headed straight. The platform moved at a constant speed of 0.6 m/s.

After the task, participants completed an adapted scale to measure trust in industrial human-robot collaboration developed by Charalambous et al. [29]. In this version, the items referring to the robot's grip (C, E, G, J) were removed, as no gripper was used in our scenario. We also used a 7-point Likert scale ranging from 1 "strongly disagree" to 7 "strongly agree". Additionally, we obtained some demographic information (See table 1).

The age of participants ranged from 18 to 56 years (M=28.7, SD=7.88), and all of them were fluent in English. Participants were recruited at Örebro University, and participation was voluntary. All participants were informed about the task, consented to participate, and were aware of the possibility of leaving at any time. We analyzed the trust scores of 33 participants divided into two groups: one that walked by the ARMoD on top of the AGV (n=19) and one that walked just by the AGV (n=14).

3.3 Results

To ensure that the trust scale of Charalambous et al [29] was reliable despite removing certain items, we calculated the Cronbach's α. The scale yielded a score of 0.76, beyond the acceptable level of 0.7 [33].
The trusted scale is composed of three major components: the robot's motion and pick-up speed (1), safe cooperation (2), and robot and gripper reliability (3). Because the first and last components involve the gripper and the pickup action, which were not part of our experiment, we just used those items in these that applied to the robot but not to the gripper: one item for the first component (robot's motion and pickup speed, two items in the original scale), and one item for the third component (robot and gripper reliability, four items in the original scale). To calculate the final trust score, we multiplied each of these items' scores by the number of items belonging to that component in the original scale, two and four, respectively, and added these to the sum of the scores of the second component (safe co-operation).

Once each participant's trust score was obtained, we performed the analysis in R [34]. Because the trust scores in both groups were not normally distributed (see violin plots in Fig.7), we opted for a Robust variation of Welch's t-test [35] to compare the reported trust between the two groups. Based on bootstrapping, the *yuenbt* function from the WRS2 [36, 37] package was used for the analysis. We kept the default bootstrapping value of 599 samples of 20% trimmed means. The $\hat{\xi}$ measure was used as an explanatory measurement of robust effect size, as suggested by Wilcox and Tian [38]. This measure was calculated using the *yuen.effect.ci* function of the WRS2 package. Values of $\hat{\xi}$= 0.1, 0.3, and 0.5 correspond to small, medium, and large effect sizes, respectively.

On average, participants reported higher levels of trust for the ARMoD condition (M = 59.73, SE = 2.22) than for the AGV alone (M = 52.5, SE = 2.78). This difference was marginally significant t = -1.68, p = .051, 95% CI[-17, 0.09]; nevertheless, this difference did represent a medium-large effect, $\hat{\xi} = 0.41$.

3.4 Discussion

In this study, we explored how trust from users varied due to the anthropomorphic features of a robot in an industrial environment. Hancock et al. [23] outlined the importance of robot-related factors for developing trust in a human-robot interaction. In [24], the authors state that improving trust in a robot starts with appropriately designing it. Instead of developing a new robot from scratch to increase users' trust in industrial settings, we modified the design of an "Autonomous Guided Vehicle" (AGV) by adding

Table 1: Demographic information from the participants

Group	N	Age (SD)	Women	Other gender	Left-handed
ARMoD	19	29.7 (9.8)	13	1	1
AGV	14	27.3 (4)	5	0	1

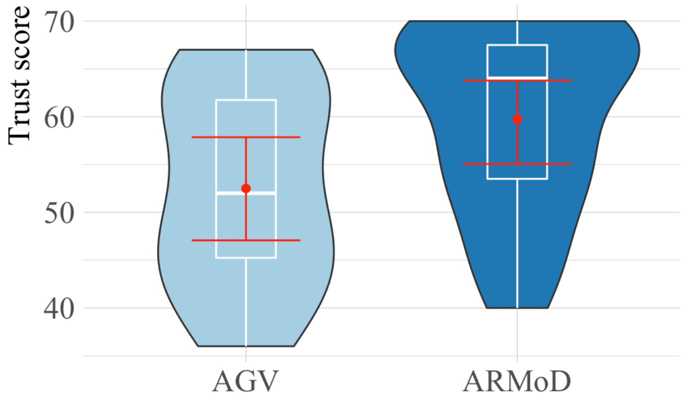


Figure 7: Violin and box plots of the trust scores for both conditions. Means and corresponding error bars are in red. Error bars show 95% bootstrapped confidence intervals.

an "Anthropomorphic Robotic Mock Driver" (ARMoD). AGVs, such as the forklift in our experiment, are frequently deployed in shared industrial environments alongside human co-workers. We used the popular anthropomorphic social robot "NAO" [32, 30] as the ARMoD.

The results of our study showed that the use of the ARMoD increased the reported trust in the interaction of participants with the platform. The simple addition of a robot on the top of the AGV, alongside basic gaze behavior, was enough to increase users' trust in the system in an industrial setting. Our results align with recent research emphasizing the role of anthropomorphism in trust [21, 22]. Contrary to other research set in industrial environments [26, 28], our results showed that perceived trust varies due to anthropomorphism in fundamental interactions such as the avoidance of a robot. Although not complex, this way of interaction will probably be expected in busy industrial settings. The difference in results is perhaps explained by the different nature of the interactions, as previously involved tasks such as handovers and precise object manipulations in which the success of the interaction may not have been taken for granted by participants.

Our suggested solution increased trust through anthropomorphic features and gaze behavior. However, other features can lead to setting an appropriate level of confidence during human interactions. For example, we previously explored a different method of communicating intent for the AGV using "Spatial Augmented Reality (SAR)" by projecting patterns on the floor in front of the robot [10]. However, this form of communication is limited by the environment's lighting conditions and can only be deployed to communicate navigational intent. Using the ARMoD, we can overcome the disadvantages of the previously studied SAR to design future experiments. e.g., independence from the lighting conditions or two-dimensional floor patterns. The ARMoD can interact with participants proactively through nonverbal communication, gazes, and gestures to communicate any intent. Future research should explore how these social features beyond plain anthropomorphism might impact users' trust in robots in industrial environments.

This research comes with two limitations. First, although high levels of trust are desirable, we just focused on the robot appearance component that modulates it. Appropriate functioning and the minimization of failures by the system can have a more significant impact on perceived trust. Moreover, manipulating trust purely by appearance while ignoring other aspects could lead to over-trust, which can be dangerous and undesirable in potentially threatening situations, such as the platform not breaking when headed toward a person. Second, we designed a basic encounter that did not involve tactile interaction or manipulation, contrary to previous research with industrial robots. Nevertheless, we believe that the scenario of a corridor encounter with a robot will likely become common every day. This situation may occur in various industries and with different types of workers, even those not directly involved in close collaborative work with the robot.

3.5 Conclusion and Continuation

This section presents the findings of a study examining a novel interaction method for an autonomous guided vehicle (AGV) using an anthropomorphic robotic mock driver (ARMoD). Involving 33 participants, the experiment simulated a hallway encounter where participants interacted with the AGV both with and without the ARMoD mounted on it. The results demonstrated that participants exhibited heightened levels of trust when the ARMoD was present, indicating that incorporating anthropomorphic characteristics enhances appearance-based trust in industrial contexts.

Building on these encouraging outcomes, the next step was to further examine the interactive capabilities of the ARMoD concept. Specifically, we aimed to compare different interaction styles: a machine-like style that communicates solely verbally and a multimodal style that combines verbal communication with human-like gestures and expressions. This comparison was motivated by the need to determine the most effective method for enhancing human-robot interaction, understanding that multimodal communication could potentially offer more intuitive and engaging interactions than verbal communication alone.

Before conducting large-scale evaluations, we sought to validate these interaction styles by surveying European citizens. This survey was designed to measure EU citizens' preferences for different interaction styles and to ensure our approach aligns with the public's expectations and needs. The survey results informed the design of subsequent experiments and facilitated the refinement and optimization of interactions in these experiments.

3.6 Robotics4EU DARKO Citizen Survey

The Robotics4EU project, funded by the EU's Horizon 2020 program under grant agreement No 101017283, aims to promote the widespread adoption of robotics in healthcare, infrastructure maintenance, agri-food, and agile production by advocating for responsible robotics principles. As part of this initiative, Robotics4EU conducted a European-wide citizen consultation in collaboration with the DARKO project to gather public input on 11 different robotic applications. Through an online, informed survey platform that included educational materials and guided questions, citizens provided valuable feedback. This approach underscores the importance of involving public perspectives in technological development to ensure that new robotic solutions meet societal expectations and needs, thereby facilitating their acceptance and successful integration. The survey received answers from at least 8 different countries, with Denmark coming in at the top with 44% of the total answers. Following this, Lithuania accounted for 12followed by both France and Norway with 10% and Estonia, Isle of Man, Latvia and Portugal each representing 1% each. 18% of respondents chose not to disclose from which country they came. Citizens from both Central and Eastern Europe, Northern Europe, Southern Europe, and Western Europe have answered the survey, indicating a diversity across Europe.

3.6.1 Demographics of the Survey and how it presented DARKO

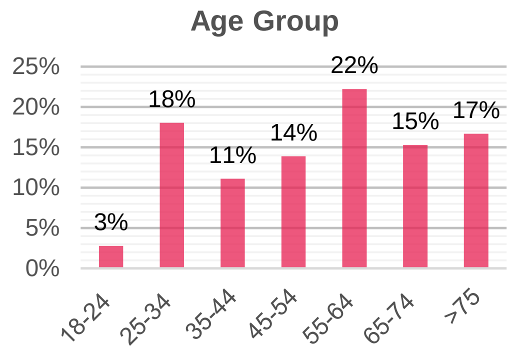
Demographics

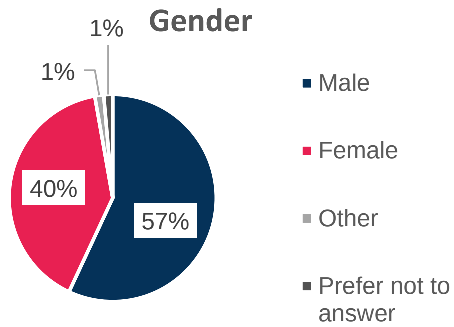
72 respondents answered the online consultation. The highest representation of citizens was the age group 55-64, accounting for 22%. While the distribution of the other age groups was divided closer to each other. The younger generations from 18-24 years were not as well represented.

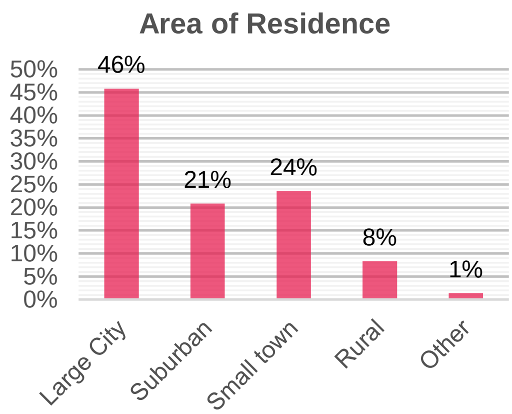
The gender distribution of citizens was leaning towards a little larger representation of male respondents, with male participants accounting for 57% and female participants accounting for 40%. The remaining either answered ‘other’ or did not specify their gender.

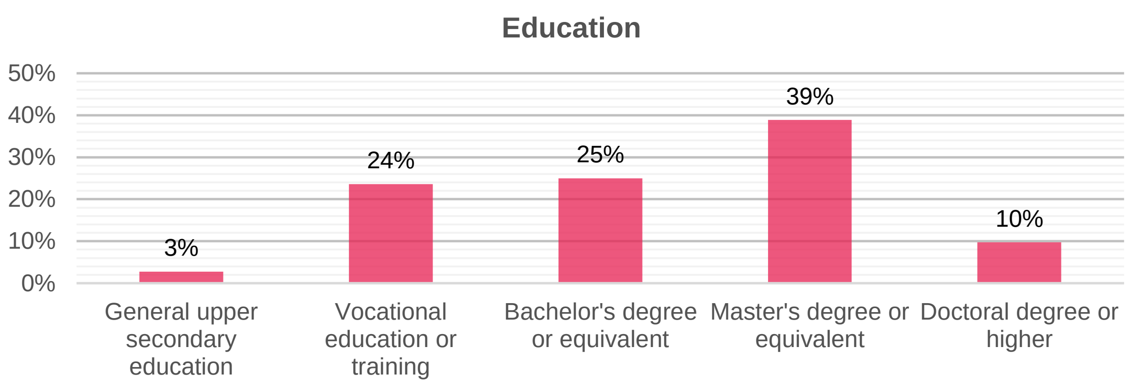
Looking at distribution of areas of residence, a total of 46% of the respondents answered that they lived in a large city. The second most chosen option was small town with a total of 24%, these were followed by suburban with 21% and rural with 8%. The remaining 1% entered ‘other’ as their area of residence. These results reflect the expectations when taking the distribution of the age groups into account.

The educational level of the respondents was high with 39% having a master’s degree, a quarter of the respondents having finished a bachelor’s degree, and 24% had a vocational education or training. Every tenth of the respondents had a doctoral degree. The last 3% had a general upper secondary degree.





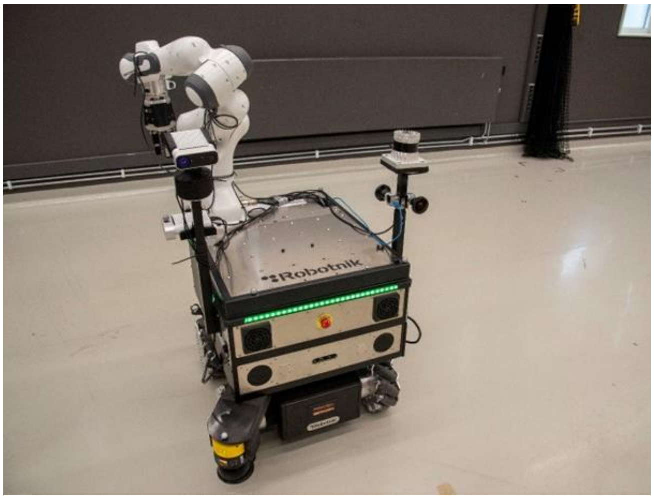




Presentation

DARKO is a European research project that develops new methods for robots that should work efficiently together with people, particularly in logistics and production.

The central theme for the DARKO robot is efficiency. The robot should navigate efficiently around people – comfortably driving among them in a way that doesn't disturb its co-workers, while still reaching its goals on time. This includes being able to efficiently communicate its intents to the people around it, as well as recognizing their intents.



The robot should be efficient at handling objects – which also includes throwing an object into the target tray, rather than driving there to drop the object. Throwing will save both time and energy. The robot should also be easy for anyone to install at a new site – increasing efficiency by reducing the work effort and modifications that might otherwise be needed to adapt the environment for the robot.

3.6.2 Questions regarding the DARKO Project

The survey aimed to gather respondents' preferences and feelings about working with robots in different scenarios. The first question focused on whether respondents preferred robots that move on predefined paths or navigate flexibly like humans. The second question asked if robots should adapt to their surroundings by learning human activity patterns. The third question explored whether respondents would feel safer working alongside robots that track human movements.

Question 1 Results: Respondents were asked if they preferred robots that move on predefined paths or navigate flexibly. Almost half (47%) preferred robots to follow clearly marked predefined paths for predictability and safety. About 35% preferred robots that could plan and navigate freely, valuing technological advancement and efficiency. The remaining 18% favored predefined paths without floor markings. Responses varied; some emphasized safety and predictability, while others wanted more advanced, flexible robots.

Question 2 Results: More than half of the respondents were positive about the idea when asked if robots should adapt to their surroundings by learning human activity patterns. Approximately 29% were neutral, possibly due to uncertainty or difficulty understanding the concept, while 14% expressed concerns about privacy and the robot's ability to adapt accurately. Only 4% thought it was a bad idea, reflecting a general openness to adaptive technology despite some reservations.

Question 3 Results: The third question explored safety perceptions regarding robots that track human movements. About 65% of respondents felt safer with tracking robots, citing increased reliability and better adaptation to human behavior. However, 14% felt safer without tracking due to privacy and safety concerns, such as the risk of data misuse and mistrust in the robot's technical capabilities. The remaining respondents either felt safe in either scenario (15%) or did not feel safe around robots at all (6%).

Testing different levels of communication To explore the use of different levels of communication, the respondents were presented with two different appearances of the DARKO robot and a reply to a questionnaire. The first was a picture of the DARKO robot. The second was the same picture, but now it has the humanoid robot NAO on top of the DARKO robot. This was done to test whether some of the functionalities of a humanoid-looking robot can have a positive impact on people's first impression of a robot or if it is indifferent to their feelings towards it.

Figure 8: Left: The DARKO robot in the factory stage appearance is called "Robot 1 in its current appearance". **Right:** The DARKO robot with the humanoid-looking robot NAO (from SoftBank Robotics) on top is called "Robot 2 with humanoid robot."

The respondents were also informed that they should be aware that the addition of NAO was only to test a concept and not necessarily how the developers envision the final product.

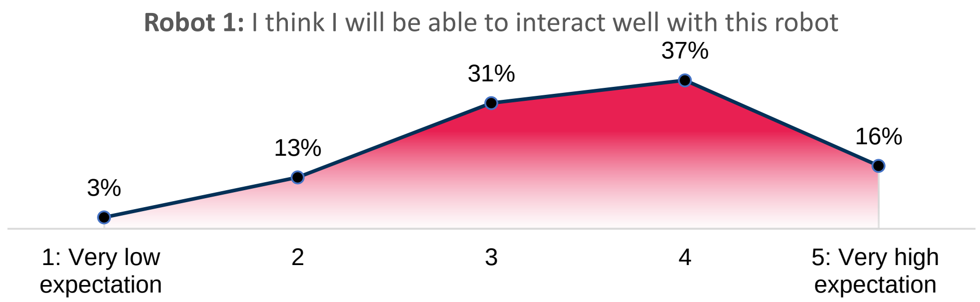
The respondents were asked to react to three statements indicating on a scale from 1-5 how very high or very low they expected to do the following:

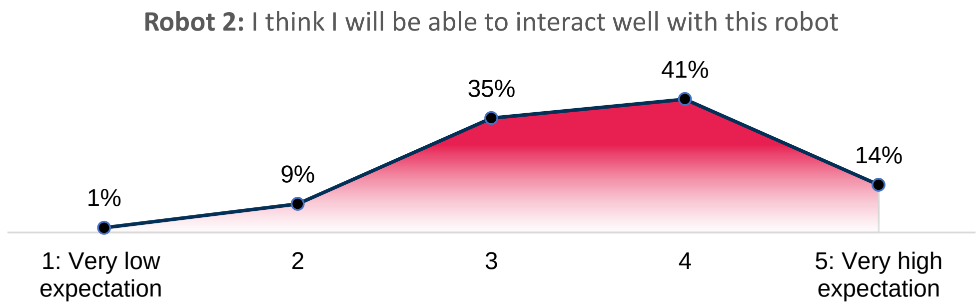
1) **I think I will be able to interact well with this robot**
2) **I would find this robot trustworthy**
3) **I would like to work alongside this robot**

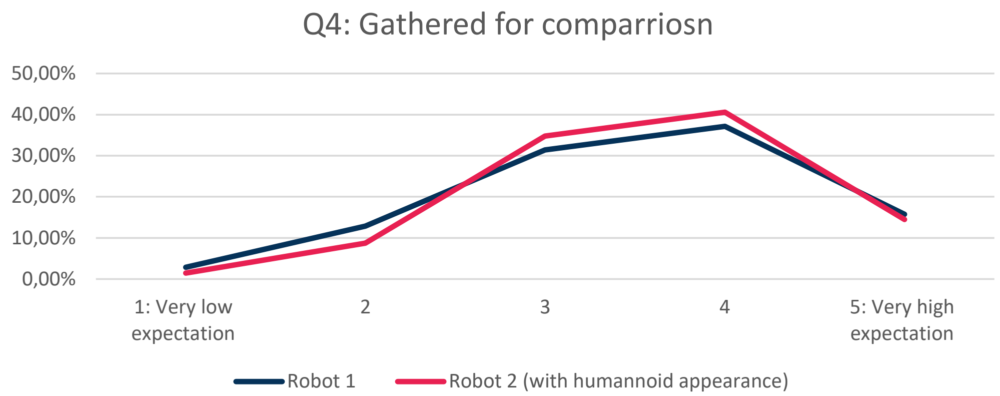
In the survey the respondents were first asked the above questions for the first robot and then they were presented with the second picture of the robot and asked the same questions again. In the report, we will however present one question at a time and then compare the responses of the two robots.

Question 4: I think I will be able to interact well with this robot

To explore the use of different levels of communication when interacting with the robot, the respondents were asked to indicate on a scale from 1-5 how very high or very low they expected to interact well with the robot. Below you can see the results from the two robots presented.



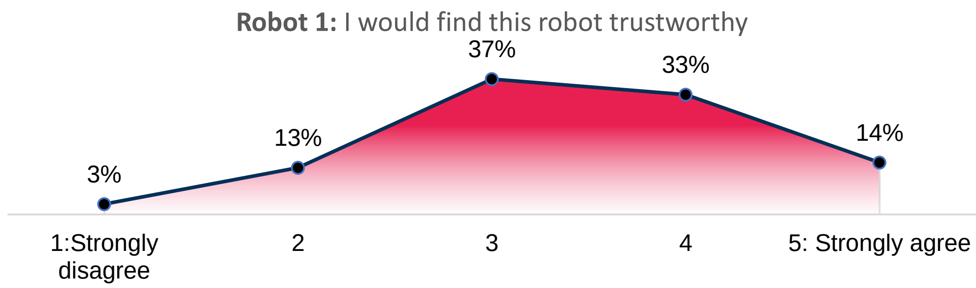


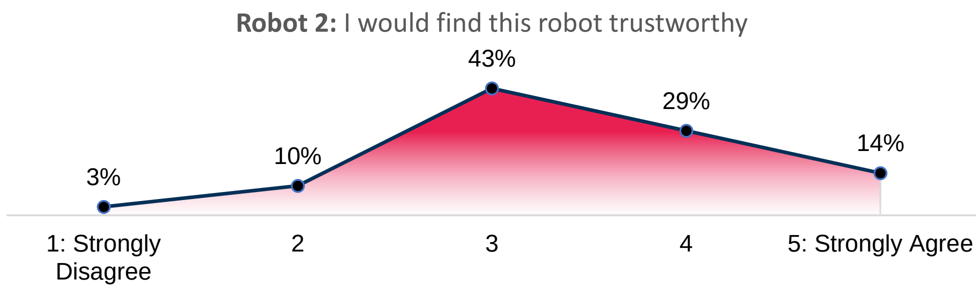


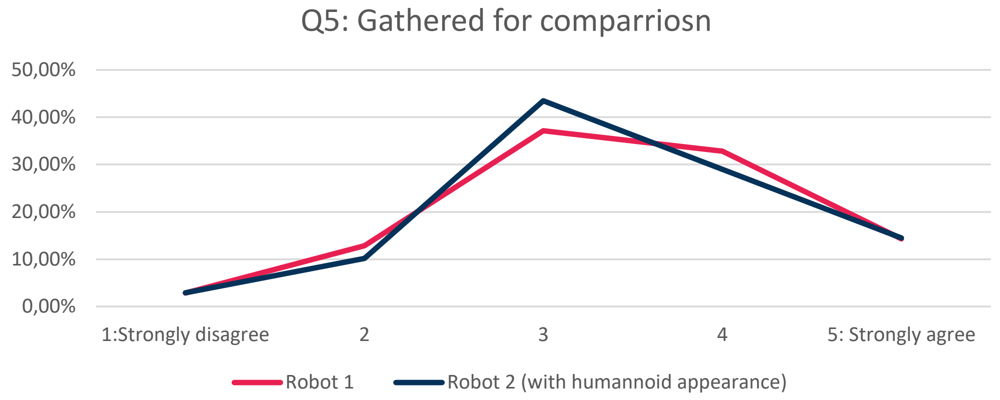
As can be seen from the comparison there's only a small difference in the respondents' answers. Robot 2 with the humanoid appearance scores marginally higher having fewer people answer that they had low or very low expectations towards interacting well with the robot but at the same time it also scores slightly lower in the other end of the scale with the very high expectations. Because the margin is so small and taking the number of respondents into consideration the result implies that the respondents generally are positive towards interacting with the robot regardless of the humanoid features brought by the NAO robot.

Question 5: I would find the robot trustworthy

The respondents were asked to enter on a scale from 1-5 how strongly they agreed or disagreed to the statement of finding the robot trustworthy.





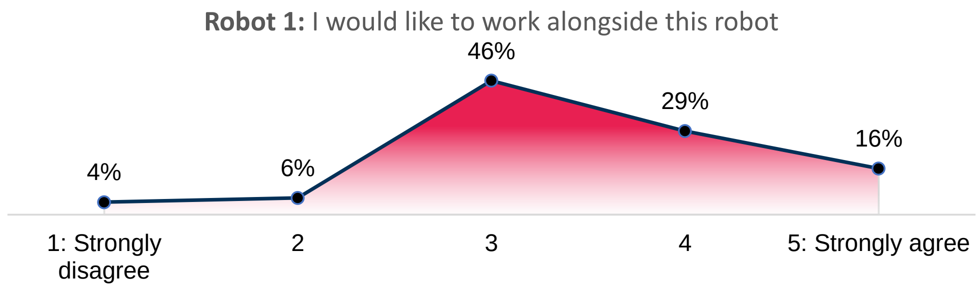


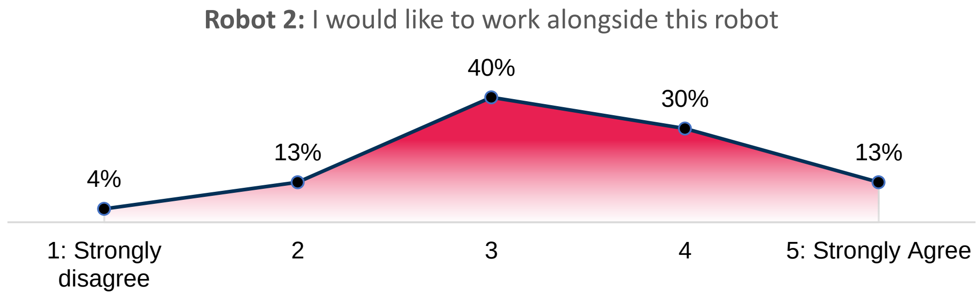
Once again, the results are very close to each other, the number of respondents strongly disagreeing to the statement are identical and the same goes for the number of respondents strongly agreeing to the robot being trustworthy. While respondents entering their score in the middle of the scale are a little higher towards the humanoid robot, this might be linked to the respondents being a bit more unsure how they should feel towards the humanoid robot not knowing what it can and can't do. It can therefore not be concluded whether giving the robot a face can generate a feeling of security and familiarity. However, this is something that could be further explored with real-life testing with regular citizens where they can get a better feel of the robot and explore it better than what can be done through a picture.

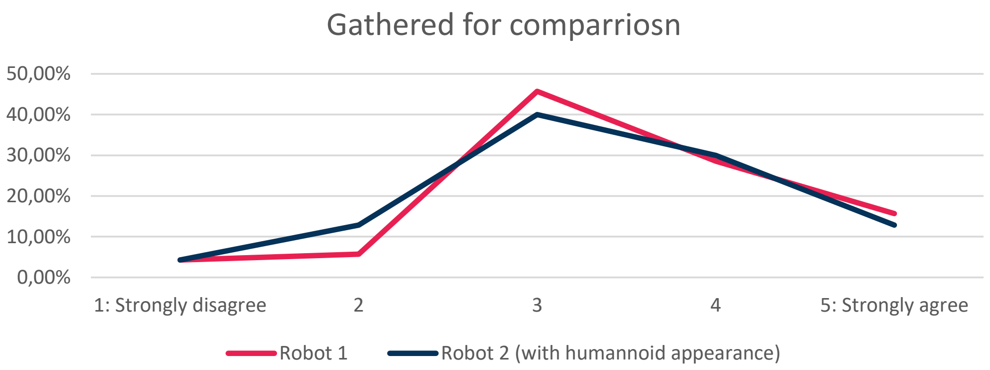
A focus group interview conducted among the participants of the Robotex International festival reveals that it is difficult to evaluate rudimentary robots and their functions when the first impression is of a machine in a very early stage of development. Trust is created by the need to see that the robot is mature. "The first thing that strikes me about him is that he is, as it were, at an early stage in its development. It is hard for me to understand what he is made for and what he does. Even if there is a description, the first feeling is that it is still too raw."

Question 6: I would like to work alongside this robot

On a scale from 1-5 respondents were asked to enter how strongly they would disagree or how strongly they would agree to like working alongside the robot.







Once again, the results are very similar but here the respondents were a little more positive about working alongside the robot without the small humanoid robot on top of it.

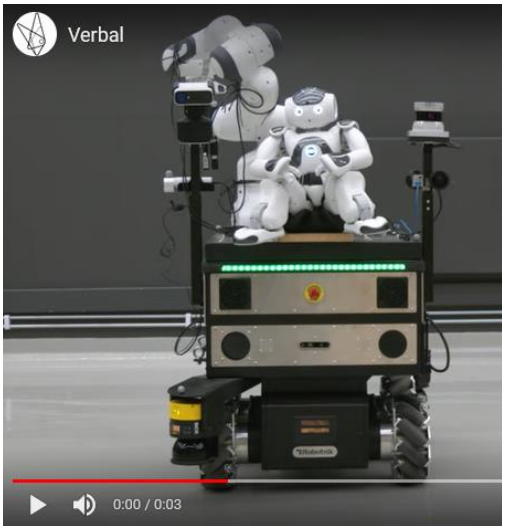
Looking at the results from the former 3 questions most of the answers were placed in the middle of the scale ranging from high expectations to low expectations and from strongly disagreeing to strongly agreeing. The distribution of the results is expected with conceptual questions where the respondents still haven't experienced in real life the situations they are being asked to respond to. The first 2 comparisons most of the respondents had a slightly lesser negative response to have the little humanoid robot on top of the robot, while with the last comparison a small margin of the respondents preferred working with the robot without the humanoid robot.

Looking back at the former questions in the survey, the respondents have expressed they preferred having the robot adapt to the human's contra having the humans doing the adaptation. So, when the robot is doing most of the adaptation this can potentially help with the trust building towards the robot. Given a situation where they must work alongside the robot it does not seem to be as important to have a humanoid robot sitting on top.

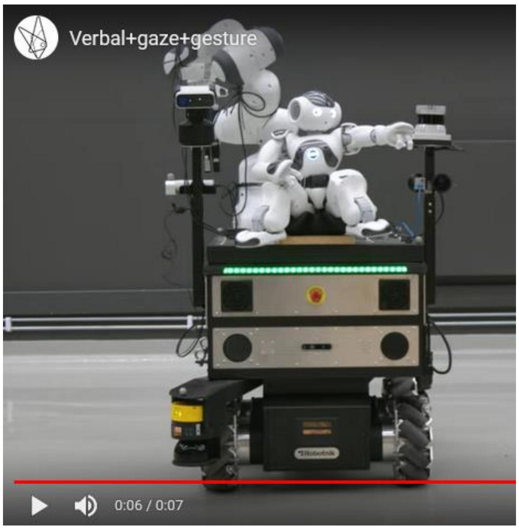
Testing of verbal and gestures as means of communication

To further test the functionalities a humanoid robot can provide, the respondents were introduced to two videos of the robot with NAO in function. The first video has NAO informing its intentions by using a voice saying, "let's go to goal number 5".

In the second video presented to the respondents, NAO informs its intention by using the same voice and a gesture by looking and pointing in a direction.



Video 1 (only verbal)

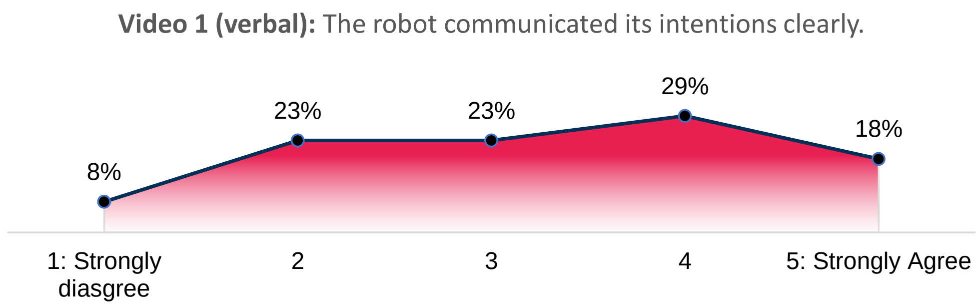


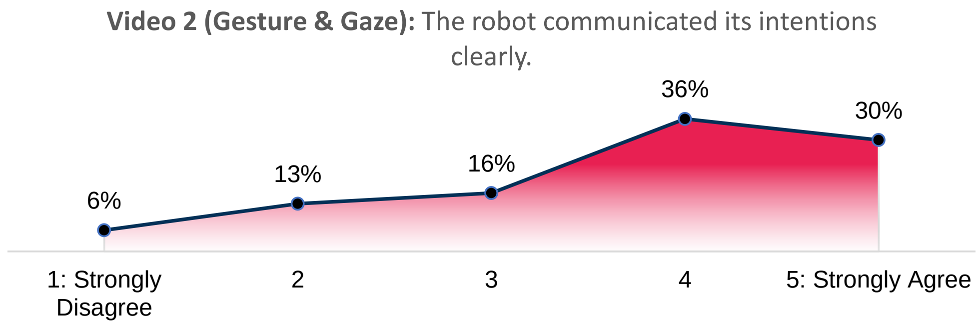
Video 2 (Verbal, Gesture & Gaze)

Respondents were asked both how strongly they agreed or disagreed with the robot clearly communicating its intentions and whether they thought this was appropriate way to communicate where the robot will go next on a scale from 1-5.

Question 7: Did the robot communicate its intentions clearly?

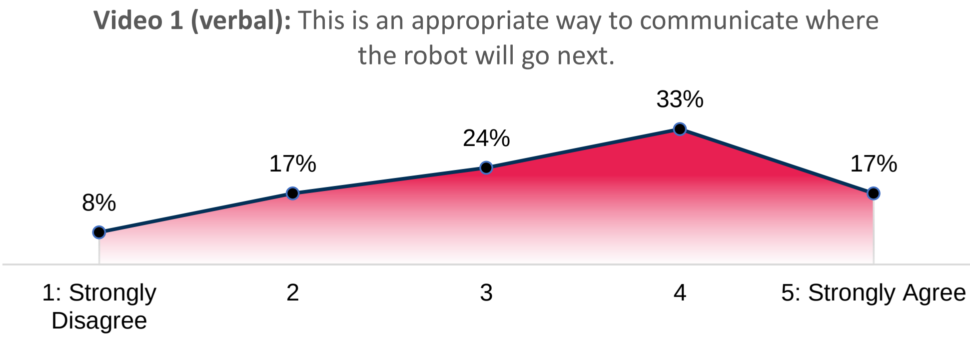
Looking at the results below the respondents were more prone to having the robot communicating its intentions by using both a voice and gesture. In the first video only 47% of the respondents agreed or strongly agreed that the robot communicated its intentions clearly, whereas in video 2 66% thought the communication was clear. Also, in the middle and the other end of the scale we see a clear difference between the two. However, we can also conclude that there still is a group of people who do not think the robot is clear in its communication one way or the other.

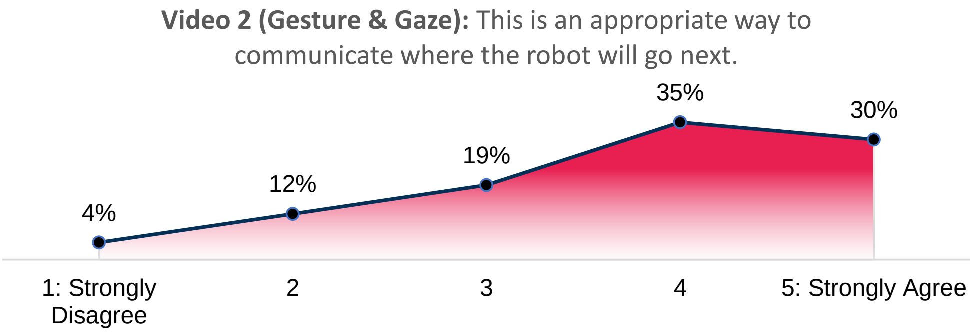




Question 8: Do you think this is an appropriate way to communicate where the robot will go next

The respondents are again introduced to two videos, one with the robot communicating by using a voice and one video with the robot communicating using a voice and gesture.

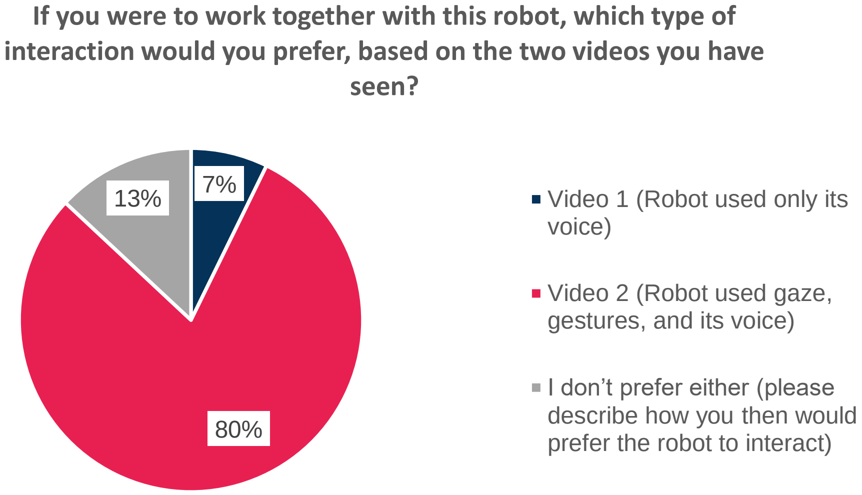




The results are similar to the former question. The respondents are more prone to having the robot communicating by using a voice and gesture. A higher number of the respondents disagree with communication being an appropriate way of communicating, when only using a voice to inform of where the robot will go next. Almost twice as many strongly agreed to having the robot use bot the voice command and gesture compared to only using the voice. Looking at the two previous questions we can conclude that there is an indication that the respondents would rather have communication in more than one way. In addition to this people with hearing or visual impairment should potentially also be considered.

Question 9: If you were to work together with this robot, which type of interaction would you prefer, based on the two videos you have seen?

Lastly the respondents were asked to evaluate which type of communication with the robot they preferred. They needed to answer if they wanted the robot to communicate using its voice or by using its voice along with gestures. If they didn't prefer either, respondents could describe how they wanted to interact with the robot.



80% of the respondents preferred the robot to communicate by using a voice and gesture. While 7% preferred the robot to use its voice to interact. 13% of the respondents didn't prefer either way of interacting with the robot.

Looking into the elaborative answers one respondent mentions how: *"we perceive differently, so therefore good with different actions"* supporting why using a gesture along with the voice is preferred. Some of the respondents do have reservations towards the voice being used. The voice needs to be clearer and asked for it to be in a more serious tone. Other respondents are asking for the use of led light for the robot to communicate where it is going. While another respondent mentions how humans do not communicate which directions on where they are moving, so this might not be needed with a robot. With this comment it should be considered that humans do use a lot of indirect body language and mimics that can show our intentions, which the robot does not have. To mimic this the robot can perhaps be accommodated by using lights, a display on the robot or by placing the humanoid robot on top of the robot by communicating its movements with gestures. One answer stands out from the others: *"it is easiest with only one indication",* the comment separates itself from the other comments and the 80% preferring the robot to communicate using different approaches. The Comment might be in relation to the former comment on humans not explicitly expressing their movements, and therefore the communication might not be necessary.

3.7 Summary and Conclusion of Survey Results

The EU4Robotics Survey provided valuable insights into the diverse opinions and expectations of technology among respondents, which is crucial for developing the ARMoD in the DARKO project. The results indicate that safety concerns and technological advancements are critical factors in determining preferences for robot movements, emphasizing the importance of developing a robot to address these concerns. Additionally, many respondents answered that they would feel safer working alongside a robot that tracks and records human movements, potentially indicating that it is aware of surrounding humans. Interestingly, the survey suggests that adding humanoid features does not significantly impact respondents' willingness to interact with the robot or their perception of its trustworthiness, potentially contradicting our initial findings [18]. However, the study was conducted online, and participants did not interact with an embodied robot, which is known to have a substantial effect, especially on subjective user ratings [39]. When asked about their preference regarding verbal and gestural communication, it was found that the latter improves the clarity and acceptability of the robot's intentions. Most respondents prefer robots to communicate using both voice and gesture, highlighting the need to consider people with hearing or visual impairments when designing robot communication systems. Overall, the survey highlights the importance of considering diverse perspectives and needs in technology development and suggests that further testing with real-life interactions is necessary to fully understand the impact of the ARMoD's features on human-robot interactions.

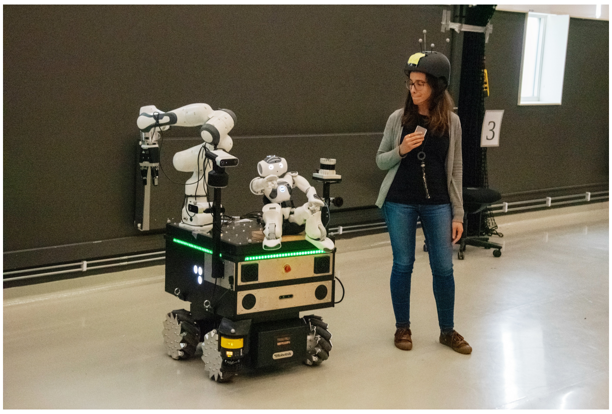


Figure 9: Participant encountering a mobile robot with an NAO robot mounted on top as the "Anthropomorphic Robotic Mock Driver" (ARMoD). The mobile robot communicates with participants through the ARMoD.

4 Experimental Evaluation of the ARMoD

In the previous sections, we outlined the communication channels available for the DARKO robot and the design process of the ARMoD. In the following section, we present our initial experimental evaluation of the ARMoD concept.

Summary: Robots are increasingly used in shared environments with humans, making effective communication necessary for successful human-robot interaction. Our work studies a crucial component: active communication of robot intent. We evaluated our anthropomorphic solution, which is when a humanoid robot communicates the purpose of its host robot, acting as the "Anthropomorphic Robotic Mock Driver" (ARMoD). We evaluate this approach in two experiments where participants work alongside a mobile robot on various tasks, with the ARMoD communicating a need for human attention or giving instructions to collaborate on a joint task. The experiments feature two interaction styles of the ARMoD: a verbal-only mode using only speech and a multimodal mode that includes robotic gaze and pointing gestures to support communication and register intent in space. Our results show that the multimodal interaction style, including head movements, eye gaze, and pointing gestures, leads to a more natural fixation behavior. Participants naturally identify and fixate longer on the areas relevant to intent communication and react faster to instructions in collaborative tasks. Our research further indicates that the ARMoD improves engagement and social interaction with mobile robots in workplace settings.

4.1 Introduction

Mobile robots are becoming increasingly common in today's workplaces, working alongside human colleagues. However, while humans use a complex set of social cues to interact with each other, mobile robots are often limited by their native design, making it difficult for them to produce legible social cues. Designing efficient communication methods is paramount to enable mobile robots to convey critical information about their environment and the task at hand to their human co-workers. Therefore, ensuring seamless and productive interactions between robots and humans requires the development of suitable methods to bridge the communication gap between them.

The need for effective communication between mobile robots and humans in different work environments has led to research into various approaches, including native communication channels such as LEDs [40, 41] and robot-attached channels such as floor projections [10, 11]. However, these cues may not be universally understood or applicable to all robots. The need for approaches that can be validated and used across a range of mobile robots remains open [8]. In this study, we evaluate the use of an "Anthropomorphic Robotic Mock Driver" (ARMoD) as seen in Figure 9 to facilitate intuitive communication between non-humanoid robots and human co-workers in workplace settings, building on the previous research in this area [13, 14, 10, 18].

This study evaluates the incorporation of communication channels for mobile robots using the ARMoD [18] without affecting their primary functionality. Prior research showed that adding anthropomorphic features can enhance communication with pedestrians [42]. Our initial validation of the ARMoD concept concluded an increase in appearance-based trust in the robot [18]. We investigated the interactive capabilities of the ARMoD and examined its effects on participants' attention by measuring their eye gaze during the interaction in a collaborative task. To frame our experiments, we draw on the terminology of the intent communication model introduced by Pascher et al. [1] to categorize the robot's conveyed intents.

In human-robot interaction (HRI), eye tracking is a powerful tool for analyzing visual attention and perception. Researchers can gain valuable insights into how people perceive and interact with their environment by recording and analyzing fixations, brief periods when the eye remains relatively stable, and visual information is acquired [43]. Previous research used eye tracking to investigate how a robot's intent communication affected human bystanders' gaze [10] and participants' engagement [44]. In our study, we use eye tracking to analyze how participants' fixations are distributed between the ARMoD and the mobile robot and how the interaction style of the ARMoD affects participants' reaction times to cues relevant to collaborative tasks.

To validate the interactive capabilities of the ARMoD as an intention communication entity for mobile robots, we design two different styles of interaction: a purely verbal one, where the intention is communicated using only the speech of the humanoid robot, and a multimodal one, where the intention is supported by the robotic gaze and pointing gestures of the robot. These align with recent literature on humanoid robots [14, 44]. We aim to investigate their effect on communicating different types of intentions to human users. The ARMoD is mounted on two different mobile robots and interacts with human participants in either verbal-only or multimodal communication styles, depending on the experimental condition.

To investigate the impact of these two different interaction styles on the quality of human-robot interaction aided by ARMoD, we conducted two experiments in which participants worked alongside the robot on various tasks that required collaboration with the robot. In these experiments, we address the following three research questions:

1. How do different interaction styles influence participants' fixation duration on the ARMoD during an attention-grabbing greeting behavior?

2. Does an interaction style that registers communicated intent in space lead to faster reaction times than a style that does not?

3. To which extent do participants fixate on the ARMoD vs. the mobile robot during HRI, and how are two different interaction styles affecting this behavior?

Our study validates the observation of prior research by Salem et al. [14] that a multimodal interaction style of a humanoid robot leads to participants interacting in a "fairly natural way". Furthermore, we find additional evidence that eye contact established by a humanoid robot leads participants to longer fixate on its face, which Kompatsiari et al. [44] correlated

Figure 10: In Experiment A, participants interact with a robotic forklift. The ARMoD instructs the participants to place an object on the forks of the mobile robot.

with increased engagement. Equipping mobile robots with an ARMoD that utilizes a multimodal interaction style to communicate with users results in faster reaction times in collaborative tasks, where the robotic gaze registration of communicated instructions enables quicker localization of goal points and objects of interest.

4.2 Evaluation Methodology and Design

This study evaluates the ability of the “Anthropomorphic Robotic Mock Driver" (ARMoD) to communicate intentions for mobile robots in a workplace setting. We examine the impact of two interaction styles – verbal-only and multimodal – on conveying various intentions, including attention, motion, and instruction. In this context, attention refers to when a robot aims to catch the user’s attention for a subsequent movement activity. The ARMoD is mounted on a different mobile robot in each Experiment and interacts with human participants using one of the two interaction styles based on the experimental condition. This section provides a detailed description of our experimental design and methodology.

This paper presents the results of two experiments to evaluate the ARMoD concept and answer the research questions. The initial Experiment A investigates the interaction styles of the ARMoD in one-on-one interactions in a narrow corridor. Intriguing fixation patterns are observed, such as more prolonged fixation on the face of a humanoid robot when eye contact was established and faster reaction times in collaborative tasks when using a multimodal interaction style with an ARMoD. Experiment B is designed to confirm these findings using a different mobile robot in repeated interactions in a more open workplace setting.

In both experiments, participants work alongside the mobile robot as coworkers and work on various tasks. When the robot encounters a situation requiring assistance to complete its task, the ARMoD communicates the need for the human’s attention to initiate an interaction. Once the interaction starts, the human collaborates in a joint task with the robot. Participants are instructed to cooperate with the robot if the ARMoD requests it. In both experiments, the ARMoD communicates instructional and motion intent to coordinate the fulfillment of the collaborative task with the human.

The experiments occur under two different conditions, each modulating the interaction style of the ARMoD. In the verbal-only condition, the ARMoD communicates solely verbally with the participants. In the multimodal condition, we combine verbal communication with gaze cues and pointing gestures from the NAO robot to register communicated intent in space if necessary. This multimodal interaction style builds on the one proposed by Salem et al. [14].

Experiments A and B differ primarily in the mobile robots used, the nature of the collaborative task, and the workspace design. In Experiment A, participants transport an aluminum tin can (diameter 160 mm, filled with 750 ml canned vegetables) to a table and then collaborate with a robotic forklift, which must transport a box (see Figure 10 and Figure 11) to the other side of a corridor. The ARMoD instructs the humans to place the box on the forklift's forks and, once the box is loaded, guides the human's path to avoid a collision by using its voice to say "Pass on my left" and pointing to its left in the multimodal interaction style. In contrast, in Experiment B, participants interact with a smaller, more agile mobile robot with different physical appearance and driving characteristics, see Figure 14. This mobile robot, equipped with a robotic arm in its resting position, navigates in a 10 × 9 meter open workplace setting and requires the assistance of a human at a specific goal point. The ARMoD communicates the robot's next goal point and instructs the human to accompany it.

In our experiments, we use Tobii eye-tracking glasses (versions 2 and 3) to capture the participants' gaze behavior while interacting with the robots. The data obtained from the Tobii glasses requires post-processing for suitable data analysis, as described in Section 4.2.3. Otherwise, the results are susceptible to misinterpretation. We deploy the standard Tobii IVT attention gaze filter with a classification threshold of 100°/s. For the evaluation, we use the software "TobiiProLab"[2]. We describe the preparation of the eye tracking data in Section 4.2.3 and its analysis in Sections 4.3 and 4.4.

In addition to the eye gaze trackers, in both experiments, we measured subjective ratings and perceptions of the robot using questionnaires. For Experiment A, we deploy the same trust scale for "Trust in Industrial Human-robot Collaboration" by Charalambous et al. [29] as for our prior work [18] to assess how an interaction is affecting the subjective user ratings. In Experiment B, we add Bartneck's "Godspeed questionnaire" [45] to better understand participants' perception of the robot system and to check for potential differences in interaction styles.

4.2.1 Experiment A: Request for human assistance

In Experiment A, we explore the two interaction styles of the ARMoD, giving simple instructions. To counterbalance learning effects, each participant participates randomly in both conditions. One interaction style is verbal-only, while the other is multimodal and includes pointing gestures, robotic gaze, and eye contact with participants. The interactions occur in a 15 m long and 2 m wide corridor. Participants approach a table to pick up a box and correctly place it on a marked area on the robot's forks. The ARMoD then instructs participants to disengage. The interaction is initialized when the distance between the participant and the forklift is less than or equal to five meters, based on the social distance model by Hall [46]. Before the experiment, a human instructor explains how to place objects on the robotic forklift's forks, as participants are not expected to have prior experience with forklifts. Figure 11 shows the experimental setup.

The ARMoD deploys various gazes and gestures during the interaction with participants. When the ARMoD's distance to the participant is less than or equal to five meters, the ARMoD starts giving instructions. In the multimodal interaction style, the ARMoD performs referential gestures and gazes while speaking, making eye contact with the participants, and tracing them using head movements. The spoken instruction "Pass on my left" accompanies an optional referential gesture. In the verbal-only interaction style, the ARMoD only gives

[2] https://www.tobiipro.com/product-listing/tobii-pro-lab/

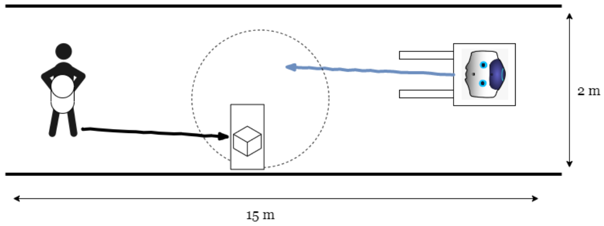


Figure 11: Experimental setup for **Experiment A**, in which a human participant interacts with an Anthropomorphic Robotic Mock Driver (ARMoD) seated on a mobile robotic forklift. The participant begins at one end of a corridor, and the forklift and ARMoD are at the opposite end. The experiment involves transporting a tin can and later collaborating with the robot to place a box according to instructions on the forklift.

spoken instructions while looking in the driving direction. The program sequence plan, shown in Figure 12, details the sequence of actions and behaviors of the ARMoD during the interaction with participants in Experiment A. Interactions ranged from 74 s to 104 s with a median duration of 89 s in with the verbal-only and 96 s with the multimodal interaction style.

4.2.2 Experiment B: Mediating joint navigation

Experiment B verifies Experiment A's findings by testing the interaction styles with a different mobile robot and repeated interactions. It evaluates the difference between multimodal and verbal-only styles for collaborative tasks and compares user ratings and perceptions. Participants navigated freely with the robot in an open room with seven-goal points (see Figure 13). Participants drew cards from decks at designated goal points, which indicated their next navigation goal. Each deck had a varying number of cards, with goal points (1) and (7) having 15 cards each, goal point (3) having 12 cards, and goal points (4), (5), and (6) having nine cards each. Two special cards instructed participants to look for the robot in the room and interact with it.

Upon encounter, the ARMoD initiated the interaction in either a multimodal or verbal-only style. An experimenter monitored the scene and adjusted the ARMoD's behavior by entering the next goal point for the mobile robot. This was communicated to participants through the ARMoD. If too many participants were at a goal point, the experimenter interrupted the mobile robot's autonomous navigation shortly before reaching it. If interrupted prematurely, the mobile robot would tell the participant to abort the interaction and continue drawing cards. The mobile robot would navigate alone to the goal point once less crowded.

In Experiment B, we explore the two interaction styles (multimodal and verbal-only) of the ARMoD, giving simple instructions. In the multimodal style, the ARMoD greeted the participant while establishing eye contact, communicated attention intently, and used head and pointing gestures to instruct the participant to go to the next goal point and draw a card. At the goal point, the ARMoD again used head and pointing gestures to instruct the participant to go to the goal on the card. In the verbal-only style, the ARMoD only greeted the participant and provided final instructions at the goal point without eye contact, robotic gaze, or pointing gestures. Depending on the interaction style and the distance between goals, interactions lasted around 30-40 seconds, with a median duration of 37 seconds for the multimodal style and 32 seconds for the verbal-only style.

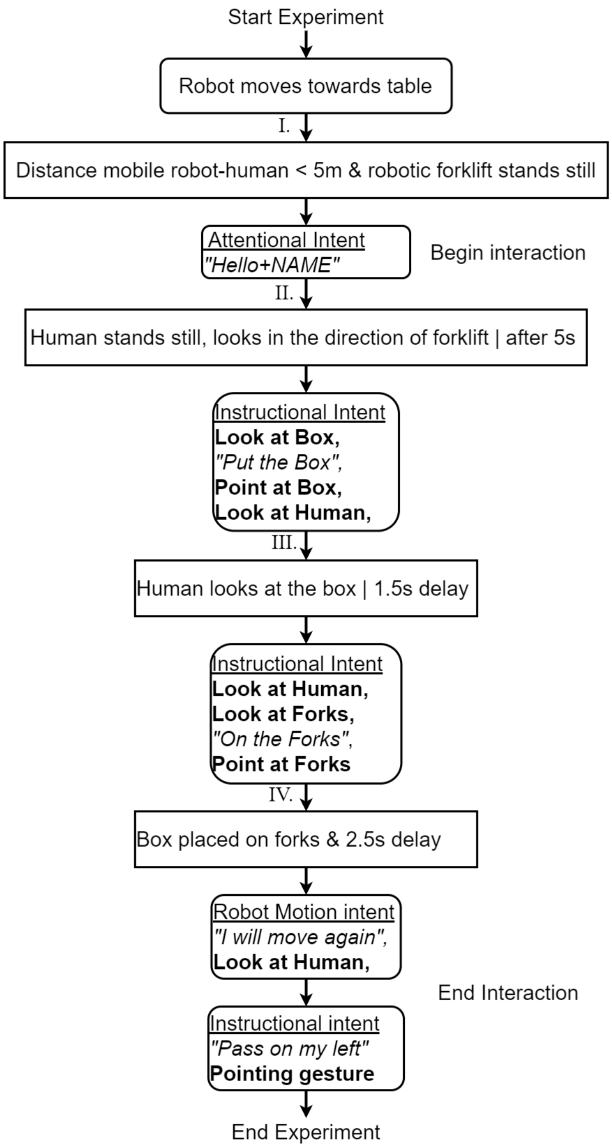


Figure 12: Flow chart illustrating the programmed behavior of the ARMoD during **Experiment A** in a hallway encounter. The sequence of events during each step of the interaction is shown from top to bottom. Dialogue spoken by the ARMoD is indicated by *italicized text in quotes*, while **bold text** indicates movements only present in the multimodal interaction style condition.

4.2.3 Eye Tracker recordings

We generate heatmaps from the eye-tracking data to analyze how the different interaction styles influence participants' attention patterns and reaction times to ARMoD's instructions. To obtain these heatmaps, we label essential events in the recordings captured by the Tobii Pro Glasses camera. We use Tobii Pro Lab's assisted mapping tool to map the user's gazes from the eye-tracker global camera onto 2D images. We also use the software's AOI (area of interest) annotation tool to define regions of interest in the snapshots. This allows us to analyze fixation count and duration on certain robot parts. Finally, we generate heatmaps for the count of participants' fixations on the snapshots (see Figure 14). This

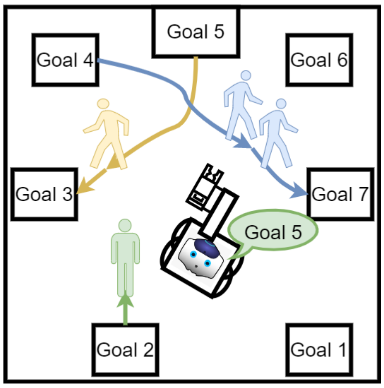
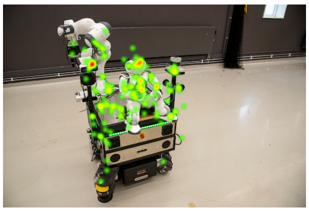
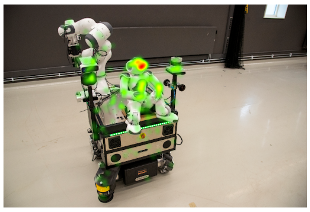


Figure 13: Experimental setup for **Experiment B**, which investigates the interaction between multiple participants and robots in a shared workplace setting. Participants navigate between designated goal points by drawing cards, as described in [47, 48]. Two special cards instruct participants using the phrase "Go to the robot" to look for the robot, approach and interact with it. The study aims to examine participants' behavior and perceptions during these interactions in a dynamic, realistic environment.



(a) Heatmap condition verbal-only interaction style



(b) Heatmap condition multimodal interaction style

Figure 14: Heatmaps showing participant gaze distribution on the robot platform in two conditions. In the **verbal-only condition (left)**, fixations are spread widely across the robot and its sensory equipment, with multiple red blobs on the ARMoD's body and one on the RGBD camera. In the **multimodal condition (right)**, participants focus more strongly on the ARMoD, as indicated by the red blob on the robot's face. Red blobs indicate centers of high fixation counts in both heatmaps.

process enables us to analyze the influence of interaction styles on participants' eye gaze and reaction times.

4.2.4 Participants

We recruited 25 participants for Experiment A and 9 for Experiment B. Participants' ages ranged from 18 to 56 years (M = 28.7, SD = 7.88) in Experiment A and 23 to 38 years (M = 30.2, SD = 4.73) in Experiment B. All participants are fluent in English and identify as female (14/25; 4/9), male (10/25; 5/9), or non-binary (1/25; 0/9). In Experiment B, participants interact twice with each interaction style in four four-minute long sessions in randomized order. In Experiment A, participants interact once with each interaction style in two-minute long sessions.

4.3 Results

We present the results of qualitative questionnaires and quantitative eye-tracking measurements. The questionnaires provide limited insights due to the small and heterogeneous sample size. Therefore, we primarily rely on eye-tracking measurements to address our research questions.

Questionnaires We gathered subjective user ratings in the system using Charalambous' questionnaire [29] for "Trust in Industrial Human-robot Collaboration" in both experiments. In Experiment A, we tested for significant differences in subjective user ratings between the two proactive interaction styles with different modalities and the data for interaction with no modalities from our prior work [18] using a one-way ANOVA. The median scores were 42 for the interaction with no modalities and 43 for verbal-only and multimodal interactions. This may indicate a slight improvement in subjective trust using either interaction style. However, no significant difference between the groups was found in the statistical test (F-statistic = 0.22, p = 0.80).

In Experiment B, we added Bartneck's Godspeed questionnaire to evaluate participants' subjective perceptions of the ARMoD's interaction styles. We used a Mann-Whitney U tests to compare sub-scales between verbal-only and multimodal interaction styles. The analysis shows slight, non-significant differences for some constructs in the questionnaire. We use Shapiro-Wilk tests to confirm that all data was not normally distributed before performing the tests. Results show no significant difference between the groups in any subscales (all p-values > 0.05). The median scores are 10 for both conditions in the Anthropomorphism subscale, 13 for verbal-only and 16 for multimodal in the Animacy subscale, 18 for both conditions in the Likeability subscale, 14 for verbal-only, and 15 for multimodal in the Intelligence subscale. Each subscale is rated on a scale from 1 to 25, with higher scores indicating more favorable levels of the measured attribute. For the Safety subscale (1 to 15), there are 10 for the verbal-only and 11 for the multimodal interaction style.

Gaze Behavior of Participants during the interactions We found that participants fixated on the robots differently between the verbal-only and multimodal interaction styles in both experiments. Figure 14 shows sample heatmaps generated from the gaze data of participants in experiment B. The heatmap on the left (Figure 14a) for the verbal-only interaction style shows scattered fixation counts across the robots. In contrast, the heatmap on the right (Figure 14b) for the multimodal interaction style shows a prominent center of high fixation counts around the head of the robot. Similarly, in experiment A, the heat maps clearly focus on the head of the ARMoD for the multimodal interaction style. With the absolute fixation count per heatmap, we calculate how much percent of these fixations land in some areas of interest. With the respective median durations of interactions, we calculate the fixation frequencies as 1.62 Hz and 1.67 Hz for the multimodal and 2.83 Hz and 2.5 Hz for the verbal-only interaction style in Experiments A and B.

Figure 15 shows the percentage of the total fixation count for each region of the analyzed heatmaps in the experiments. Verbal-only interaction saw a higher fixation count on the platform and sensors, while multimodal interaction saw a higher percentage of fixations on the ARMoD. T-tests found significant differences between verbal-only and multimodal interaction styles for both ARMoD ($p = 0.01$) and Mobile Robot AOI counts ($p = 0.02$), with small and medium effect sizes (Cohen's d: 0.29 and 0.49). These results suggest that visual and gestural cues in the multimodal interaction style shift participants' fixations towards the ARMoD as the entity communicating intent. This finding is consistent with the heatmap analysis and further supports the effectiveness of the multimodal interaction style in directing participants' attention toward the communication interface.

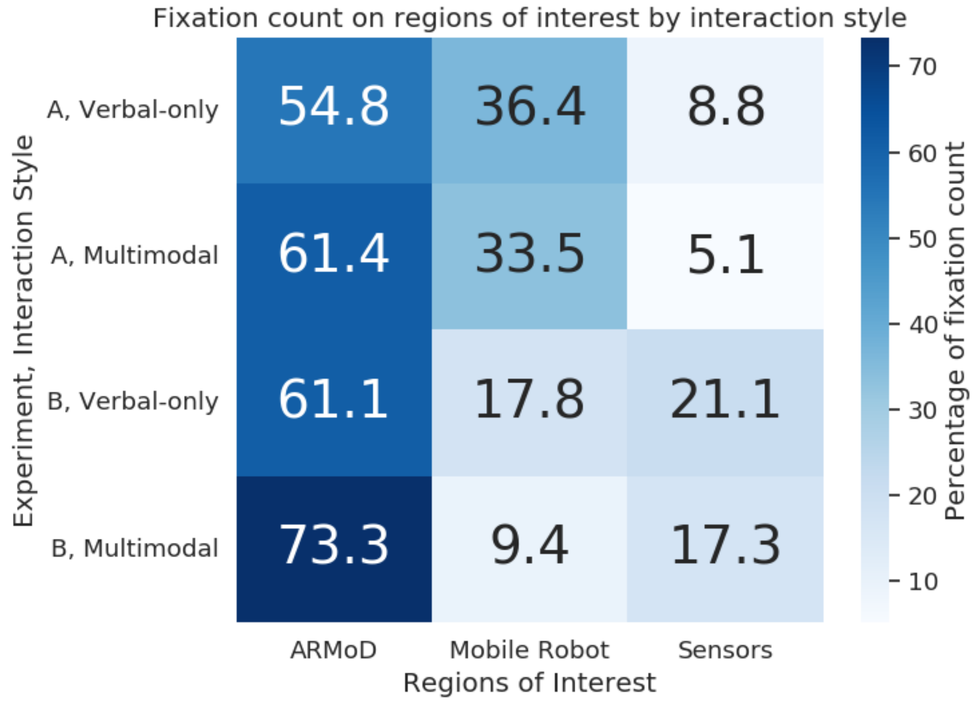


Figure 15: Matrix comparing the percentages of fixation counts on regions of interest for verbal-only and multimodal interaction styles. Fixations on the background or other parts of the scene that receive minimal fixations are excluded from the analysis to focus on how participants fixate on the robots during the interaction. In the multimodal interaction style, the ARMoD receives more fixations, suggesting that participants interact with it in a "fairly natural way" (as per Salem et al. [14]).

We also analyzed the duration of fixations on the ARMoD based on its interaction style. The duration of all fixations during the interactions with the ARMoD was extracted for each condition in the two experiments. Independent t-tests were then performed for each condition to test for statistical significance. Participants underwent the conditions in a randomized order to counterbalance learning effects. During Experiment A, participants fixated slightly longer on the ARMoD (M = 232 ms, SD = 159 ms) in the multimodal interaction style than in the verbal-only interaction style (M = 226 ms, SD = 153). However, this difference was insignificant ($t = 0.77$, $p = 0.44$). However, in Experiment B, we found a statistically significant difference ($t = -3.38$, $p = 0.01$, Cohen's $d = 0.34$) between the mean fixation duration of verbal-only and multimodal interaction styles of the ARMoD. Participants fixated on the robot significantly longer during the multimodal interaction style (M = 278 ms, SD = 192) than during the verbal-only interaction style (M = 212 ms, SD = 136).

We analyzed the participants' time to first fixation on a point or object of interest after a world-centered instructional intent communicated by the ARMoD in both experiments. We measured the time between the instruction and the first fixation on the target point or object using two event markers. We tested the data for normality using the Shapiro-Wilk test and deployed a Mann-Whitney U test, as the Shapiro-Wilk tests did not indicate normality. Our analysis found significantly shorter times to first fixation (reaction times) for the multimodal interaction style than verbal-only in both experiments. Experiment A showed a decrease from M = 2317 ms, SD = 853 (verbal-only) to M = 1237 ms, SD = 524

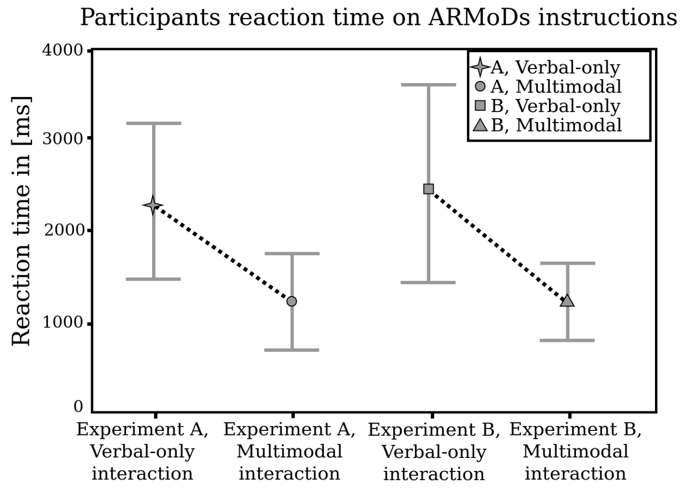


Figure 16: Lineplot to compare participants' reaction time between the ARMoD instruction and the first fixation on the target. Error bars show standard deviation. **Left:** Experiment A, ARMoD gave instructions to place a box. **Right:** Experiment B, ARMoD gave instructions regarding the next common goal point.

(multimodal), a difference of 1080 ms ($U = 6$, $p = 0.03$, Cohen's $d = 0.39$). Experiment B showed a decrease from M = 2505, SD = 1095 (verbal-only) to M = 1232, SD = 436 (multimodal), a difference of 1273 ms ($U = 9$, $p = 0.03$, Cohen's $d = 0.41$). Figure 16 illustrates the decrease in reaction time. These results suggest that multimodal interaction styles facilitate faster and more efficient communication of intent than verbal-only.

4.4 Discussion

4.4.1 Subjective user ratings

In our previous work, we studied appearance-based trust in a mobile robot equipped with an ARMoD, which revealed higher levels of trust from participants [18]. This study examined subjective user ratings for verbal-only and multimodal interaction styles during collaborative tasks. We found no statistically significant difference in subjective user ratings between the two interaction styles. However, the multimodal style showed slightly higher median ratings regarding animacy, likability, intelligence, and safety. These findings support previous research by Salem et al. [14], who observed that participants had more positive perceptions and evaluations of robots with multimodal interaction styles. Previous user studies evaluating the text-to-speech, appearance, and performance of the NAO robot have shown that users desire more natural speech and gesture capabilities [32]. Therefore, future research could investigate how users' subjective evaluations would vary with a more sophisticated robot, such as the iCub robot used by Kompatsiari et al. [44].

4.4.2 Gaze Behavior of Participants during Interactions

The results of both experiments support the idea that eye contact can "freeze attentional focus on the robot's face" [44], suggesting that incorporating head movements and robotic gaze cues into the ARMoD's interaction style could enhance its ability to engage users. This finding addresses our first research question: "How do different interaction styles influence participants' fixation duration on the ARMoD during an attention-grabbing greeting behavior?". A multimodal interaction style, which includes gaze cues and eye contact for the ARMoD, may be more effective in capturing and holding participants' attention than verbal-only interactions. This is supported by previous research [44] that eye contact is crucial in facilitating engagement and social interaction with robots. Therefore, combining a verbal greeting and establishing eye contact via head movements might be sufficient for the attention-grabbing behaviors described by Pascher et al. [1] to precede the delivery of motion and instructional intents.

The effect of ARMoD, registering instructional intent in space, on participants' reaction times was examined according to our second research question: "Does an interaction style that registers communicated intent in space lead to faster reaction times compared to a style that does not?". Two styles of interaction used by the ARMoD were compared: a verbal-only style and a multimodal style in which the robot used head movements and pointing gestures to register intent. Pascher et al. [1] argue that unregistered intent requires additional mental steps to establish a spatial link, potentially slowing reaction times. Specifically, using head movements and robotic gaze in the multimodal interaction style is vital in this effect. Participants took 0.8 – 1 s less to fixate on an ARMoD-referenced target when these cues were used. However, the relative contributions of head movements and pointing gestures to this effect cannot be determined from this study and require further investigation. This finding is particularly relevant to industrial HRI contexts, where fast and effective communication is critical for productivity and safety.

The heatmap analysis from participants' gaze data revealed that the multimodal interaction style was more effective at capturing and directing participants' attention than the verbal-only interaction style. This finding aligns with our third research question: "To what extent do participants fixate on the ARMoD and the mobile robot during HRI, and how are two different interaction styles affecting this behavior?". The heatmaps of Figure 14 show that most fixations were on the ARMoD's face, particularly in the multimodal interaction style. These results align with research by Gullberg and Holmqvist [49], suggesting that participants tend to fixate on a speaker's face rather than their gestures during interactions. Our findings indicate that by using head movements and pointing gestures, the multimodal interaction style can effectively direct participants' attention to critical spatial cues while maintaining a natural interaction style. Overall, our results highlight the potential of an ARMoD deploying a multimodal interaction style in enhancing human-robot interactions by improving attentional focus and facilitating natural communication.

4.5 Conclusion: Developing the ARMoD

Our research on robotic intent communication has made significant strides in understanding and enhancing Human-Robot Interaction (HRI) through integrating anthropomorphic features. The initial study demonstrated that an Anthropomorphic Robotic Mock Driver (ARMoD) mounted on an Autonomous Guided Vehicle (AGV) increases appearance-based trust among participants in industrial settings.

Further insights from the EU-citizen survey revealed a preference for robots that adapt to their surroundings and learn human activity patterns for increased safety. The humanoid features alone were found to not significantly impact trust without an interaction with the embodied robot. The combination of verbal and gestural communication enhanced the rating of clarity and acceptability of the robot's intentions as accepted. This highlights the importance of multimodal communication in designing effective HRI systems. Our final study further investigated the effectiveness of the ARMoD in providing additional communication channels for mobile robots. The results indicated that multimodal interaction styles using the ARMoD lead to more focused attention on the robot's face and quicker response times to instructions, demonstrating its efficacy in directing attention and enhancing communication in industrial settings. Due to the positive results achieved, the concept was integrated into the THÖR-MAGNI dataset collections.

Applying the ARMoD to real-world applications necessitates understanding the long-term effects of repeated interactions with the concept. Future work should include prolonged studies to examine how different ARMoD designs impact user perception over time. Additionally, addressing the dynamic nature of everyday interactions requires advanced reasoning capabilities, such as those provided by human gaze tracking or large language models. There is an ongoing investigation into integrating LLM-informed HRI with ARMoD to further enhance the robot's ability to understand and respond to human behavior in highly dynamic environments.

5 Insights from the THÖR-MAGNI regarding Robotic Intent Communication

The THÖR-MAGNI dataset is a comprehensive resource designed to facilitate advancements in research about social human navigation, human-robot interaction, and robotic intent communication [50]. The dataset addresses a significant gap in existing datasets by including various contextual features and scenario variations essential for modeling and predicting human motion, analyzing goal-oriented interactions between humans and robots, and studying visual attention in social contexts. The dataset's comprehensive contextual annotations and multi-modal data, including walking trajectories, gaze tracking, LiDAR, and camera streams, present a distinctive opportunity to develop robust models that explain the relationship between contextual cues and human behavior in diverse scenarios. Using the data from THÖR-MAGNI, we assess visual attention and engagement in shared environments by exploring human activity and engagement via pupil dilation and then examining visual attention through computer vision in these environments.

The THÖR-MAGNI dataset consists of 52 four-minute recordings of various activities, totaling over 3.5 hours of motion data for 40 participants and 8.3 hours of eye-tracking data for 16 participants. The data was acquired in a laboratory setting with two room configurations and varied obstacle layouts, facilitating the study of human navigation and interaction patterns. The dataset provides detailed information on the context of human motion, f.e. in the form of eye tracking and trajectory data recorded for various activities. Table 2 shows the full extent of recorded data, 548 eye tracking, and 1416 minutes of trajectory data.

5.1 Scenario Design in the THÖR-MAGNI Dataset

The dataset comprises five designed scenarios exploring various human and robot interactions in shared environments. These scenarios aim to capture the dynamics of motion, role-specific behavior, and the impact of different robot motion styles on human activities. The following analysis will provide a summary of the scenarios. For a comprehensive overview of the scenarios, please refer to Figure 17

5.1.1 Scenario 1: Capturing Motion Dynamics in the Environment

Scenario 1 establishes a baseline for goal-directed social navigation by examining how the semantic attributes of the environment influence human movement. It includes two

Table 2: Amount of eye tracking- and trajectory data recorded for various activities with all three devices: Tobii 2, Tobii 3, and Pupil Invisible glasses

Activity	Eye tracking (min.)	Trajectory data (min.)
Visitors–Alone	108	392
Visitors–Group 2	124	344
Visitors–Group 3	52	168
Visitors–Alone HRI	64	112
Carrier–Bucket	32	96
Carrier–Box	60	96
Carrier–Large Object	92	192
Carrier–Storage Bin HRI	16	16
Total	548	1416

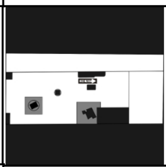
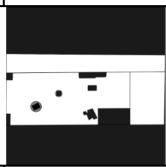
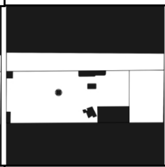
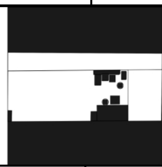
Information	**Scenario 1:** Capturing Motion Dynamics in the Environment	**Scenario 2:** Role-Specific Motion Patterns in Industrial Environments	**Scenario 3:** Impact of Mobile Robot Motion on Human Behavior	**Scenario 4:** Spatial HRI and Navigation in a Shared Environment	**Scenario 5:** Spatial HRI, Proactive Robotic Assistance
Roles	*Visitors-Alone* *Visitors-Group 2* *Visitors-Group 3*	*Visitors-Alone* *Visitors-Group 2* *Visitors-Group 3* *Carrier-Box* *Carrier-Bucket* *Carrier-Large Object*	*Visitors-Alone* *Visitors-Group 2* *Visitors-Group 3* *Carrier-Box* *Carrier-Bucket* *Carrier-Large Object*	*Visitors-Alone* *Visitors-Alone HRI* *Visitors-Group 2*	*Visitors-Alone* *Visitors-Group 2* *Carrier-Storage Bin HRI*
Robot-Motion	Stationary (Obstacle)	Stationary (Obstacle)	Condition based (Teleoperated)	Directional (Semi-Autonomous)	Directional (Semi-Autonomous)
Environment-Layout					
Conditions	Condition A Layout without- Condition B with ***semantics***	No conditions	Condition A Differential- Condition B Omnidirectional- Driving	Condition A Verbal-Only HRI Condition B Mutlimodal HRI	No conditions
Duration and Recording Day	64 min. on Day 1-4	32 min. on Day 1-4	64 min. on Day 1-4	32 min. on Day 5	16 min. on Day 5

Figure 17: The THÖR-MAGNI dataset consists of various scenario definitions detailing the roles of participants, the robot motion status (autonomous or teleoperated), environment configuration (obstacle maps), specific scenario conditions, recording duration, and the days on which the recordings took place. Each recording day involves a unique set of participants, with nine participants on Day 1 and seven on Days 2 to 4. Three mobile eye-tracking devices were used for three participants each day, except on Day 5, when two devices were used for two sets of participants. For more detailed information on the recorded trajectory and eye-tracking data duration, please refer to Table 2

conditions:

- **Condition A**: A static environment with obstacles like tables, stationary robots, and goal points.
- **Condition B**: Adds floor markings and stop signs in a one-way corridor to the elements of Condition A. This setup allows studying participants' natural reactions to environmental cues and motion patterns.

5.1.2 Scenario 2: Role-Specific Motion Patterns in Industrial Environments

Building on Scenario 1A's layout, Scenario 2 introduces role-specific tasks for participants, emulating industrial activities:

- Participants perform tasks such as carrying small objects (buckets), medium objects (boxes), and large objects (poster stands) between goal points.
- This setup enables studying how human occupation and specific roles influence motion profiles and interactions in a shared environment.

5.1.3 Scenario 3: Impact of Mobile Robot Motion on Human Behavior

Scenario 3 explores the effect of different robot motion styles on human behavior by making the previously stationary DARKO robot mobile:

- **Condition A**: The robot uses directional differential-drive kinematics.
- **Condition B**: The robot employs omnidirectional movement, allowing it to drive in any direction.
- Participants perform the same roles as in Scenario 2, and a human operator controls the mobile robot to ensure safety.

5.1.4 Scenario 4: Spatial HRI in a Shared Environment

This scenario examines human-robot interactions in a shared space with a semi-autonomous mobile robot:

- Participants move individually or in pairs between designated goal points.
- The ARMoD interacts with participants using two styles
- **Condition A**: Verbal-Only Interaction Style (see [2])
- **Condition B**: Multi-Modal Interaction Style (see [2])
- The interactions aim to understand how different communication styles influence participant behavior and navigation tasks.

5.1.5 Scenario 5: Spatial Human-Robot Interaction, Proactive Robotic Assistance

In this scenario, participants and the robot engage in collaborative tasks:

- Roles include navigating between goal points and carrying storage bins in a simulated factory environment.
- The ARMoD proactively offers assistance to participants, such as transporting storage bins on the mobile robot.
- This scenario investigates the effectiveness of proactive robotic assistance in enhancing human-robot collaboration.

These scenarios provide a comprehensive framework to study human-robot interactions, capturing subtle behaviors and responses in various settings. The systematic design ensures a broad exploration of interactions, contributing valuable insights into developing intuitive and socially acceptable robots.

5.2 Visual Attention and Engagement in Shared Environments

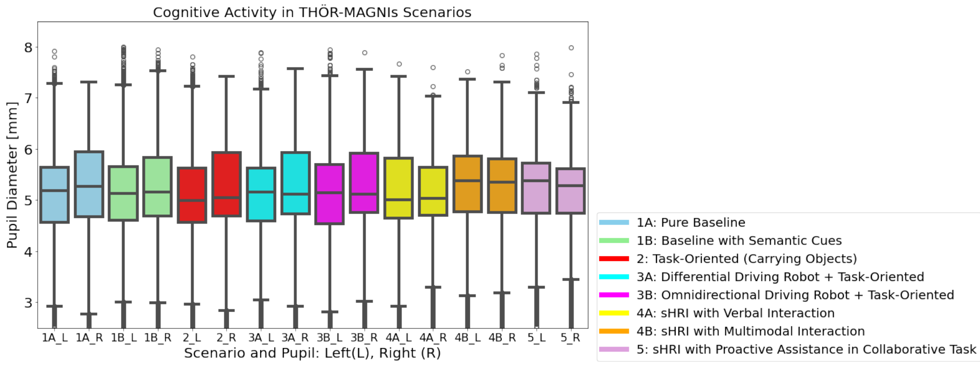
This section explores the complex dynamics of visual attention and user engagement in shared environments, as documented in the THÖR-MAGNI dataset. The analysis concentrates on how participants interact with their surroundings and the robots, thereby offering insights into cognitive activities and the distribution of visual attention based on the respective roles. Examining pupil dilation allows for inferring users' cognitive load and engagement levels across different scenarios. Furthermore, we will employ gaze-overlay videos and a contemporary computer vision methodology to investigate how visual attention fluctuates depending on participants' roles in human-robot interaction (HRI) scenarios. These analyses are crucial for comprehending the nuances of human-robot interactions and optimizing the design of intuitive, communicative robotic systems.

5.2.1 Eye-Tracking Data Collection and Specifications in the THÖR-MAGNI Dataset

Eye-tracking data of the dataset was collected using three distinct models of eye-tracking devices: Tobii Pro Glasses 2 and 3 and Pupil Invisible. The Tobii Glasses models record raw gaze data at a frequency of 50 Hz and camera footage at 25 Hz, while the Pupil Glasses record gaze data at 100 Hz and camera footage at 30 Hz. We used the I-VT Attention filter to export Tobii Glasses data, optimized for dynamic situations, to classify gaze points into fixations and saccades based on a velocity threshold of $100°/s$. All eye trackers have an IMU comprising an accelerometer and a gyroscope operating at 100 Hz. In addition, the Tobii Glasses 3 has a magnetometer that operates at 10 Hz. The infrared cameras in these devices capture the human gaze, which is then superimposed onto a 2D video by the scene cameras. The Pupil Invisible Glasses' scene camera has a resolution of 1088 × 1080 pixels, with both horizontal and vertical field of view (FOV) angles measuring 80°. In contrast, the Tobii Glasses offer a resolution of 1920 × 1080 pixels. The Tobii 3 Glasses feature FOV angles of 95° horizontally and 63° vertically, while the FOV of the Tobii 2 Glasses is 82° horizontally and 52° vertically. For the following analyses, only the data of the Tobii glasses is considered.

5.2.2 Exploring Human Activity and Engagement via Pupil Dilation

This section aims to gain insight into the cognitive processes engaged by distinct interaction styles and tasks involving the full communication system of the DARKO robot. By examining pupil dilation, a well-established physiological marker of cognitive load, insights into the mental effort required during these tasks can be obtained. Our analysis encompasses a range of scenarios, from baseline conditions to task-oriented activities and direct human-robot interactions (HRI). Each scenario captures specific aspects of human movement and engagement, providing a comprehensive understanding of the impact of different environments and robot behaviors on human cognitive activity. This investigation is vital for advancing the field of robotic intent communication and for the development of more effective and intuitive human-robot interaction systems. Pupil dilation has been the subject of extensive study to indicate cognitive load and mental effort. The existing research demonstrates that pupil size increases in response to elevated cognitive demands, reflecting the activity of the autonomic nervous system (Beatty and Lucero-Wagoner, 2000; van der Wel and van Steenbergen, 2018). This physiological response is advantageous for real-time assessments of cognitive load, as it provides continuous and non-invasive measurements of mental effort (Granholm and Steinhauer, 2004). In the context of human-robot interactions, an understanding of pupil dilation patterns can elucidate the impact of different interaction styles, such as verbal and multimodal communication, on user engagement and cognitive activity (Wang and Tsiotras, 2020).



(a) Pupil Diameter Comparison Plot **(b)** Legend

Figure 18: Comparison of Pupil Dilation Measurements Across the THÖR-MAGNI Scenarios. The plot shows the distribution of pupil diameters for each scenario, while the legend on the right gives a short description of each scenario: **1A** - Pure Baseline, **1B** - Baseline with Semantic Cues, **2** - Task-Oriented (Carrying Objects), **3A** - Differential Driving Robot + Task-Oriented, **3B** - Omnidirectional Driving Robot + Task-Oriented, **4A** - sHRI with Verbal Interaction, **4B** - sHRI with Multimodal Interaction, **5** - sHRI with Proactive Assistance in Collaborative Task.

Methods After extracting the pupil diameters from the raw eye-tracking data (Not part of the actual THÖR-MAGNI dataset), we employed pairwise t-tests using a function from a Python library to analyze the differences in pupil dilation measurements in the different scenarios. Table 18 lists the extracted mean pupil diameters plus spread size and is visualized in Figure 3. This statistical test compares the means of two independent samples to determine if there is a statistically significant difference between them. We conducted separate t-tests for each pairwise comparison for the left and right pupil diameters. Specifically, the left and right pupil diameter data were extracted for each pair of groups, and any missing values were removed. The resulting t-tests provided p-values, which indicate the probability that the observed differences between groups occurred by chance. A p-value of less than 0.05 was considered statistically significant.

In addition to the t-tests, we calculated Cohen's d to measure the effect size of the differences between groups. Cohen's d is a standardized measure that expresses the difference's magnitude relative to the data's variability. To interpret the effect sizes, we used the following thresholds for Cohen's d: a small effect ($|d| \geq 0.2$), a medium effect ($|d| \geq 0.5$), and a large effect ($|d| \geq 0.8$). This dual approach of evaluating both statistical significance and effect size provides a comprehensive understanding of the impact of different experimental conditions on cognitive load, as measured by pupil dilation.

Furthermore, the experimental design ensured all scenarios were conducted under controlled laboratory conditions, minimizing external variables that could influence pupil dilation. The environments included adjustable lighting and window blinds to standardize ambient light levels. This controlled setup allowed for more reliable comparisons across different scenarios and ensured that observed differences in pupil dilation were attributable to the experimental conditions rather than extraneous factors.

Table 3: Summary of Pupil Dilation Measurements Across the THÖR-MAGNI Scenarios. This table presents each scenario's mean and standard deviation of pupil dilation. The findings indicate cognitive activity and engagement levels associated with the different scenarios.

Scenario	Description	Mean Pupil Dilation	SDEV
1A	Pure Baseline	5.15	0.71
1B	Baseline with Semantic Cues	5.18	0.77
2	Task-Oriented (Carrying Objects)	5.16	0.76
3A	Differential Driving Robot +Task-Oriented	5.15	0.70
3B	Omnidirectional Driving Robot +Task-Oriented	5.16	0.75
4A	sHRI with Verbal Interaction	5.14	0.70
4B	sHRI with Multimodal Interaction	5.26	0.74
5	sHRI with Proactive Assistance in Collaborative Task	5.19	0.68

Results

Baseline and Task-Oriented Comparisons: The baseline scenarios (1A and 1B) and the task-oriented scenario (2) are the fundamental measures for pupil dilation. The mean pupil dilation values for these scenarios were approximately 5.16, with slight variations in standard deviations. The introduction of semantic cues in Scenario 1B (conducted first on each recording day to prevent any potential biasing of participants) and the task of carrying objects in Scenario 2 did not result in a notable change in the overall mean pupil dilation compared to the pure baseline (1A). These findings indicate that the baseline cognitive activity associated with navigating a room with obstacles remains relatively consistent when introducing semantic cues or simple tasks. The observed differences were statistically significant but minor, indicating that while semantics and tasks influenced cognitive activity, the impact was relatively limited.

Impact of Robot Motion: Introducing a mobile robot in Scenario 3 (3A and 3B) presented a scenario where participants must navigate alongside an interactive element. The mean pupil dilation for this scenario was slightly higher than that observed in the baseline and task-oriented groups, indicating a modest increase in cognitive activity due to the robot's presence. Furthermore, there was an increase in the variability of pupil dilation, particularly in Scenario 3B, where the robot moved omnidirectionally. An indication that the unanticipated movements of the robot necessitated additional cognitive processing on the part of the participants.

Human-Robot Interaction (HRI): There are notable differences in pupil dilation between Scenarios 4 and 5, the latter involving more direct interactions with the robot. In Scenario 4, participants interacted with the DARKO robot through the ARMoD and proceeded to goal points in conjunction with the robot. This scenario demonstrated higher mean pupil dilation with lower variability, particularly in the multimodal interaction style (4B), indicating increased cognitive activity and engagement compared to the verbal-only interaction style (4A). Scenario 5 has the highest mean pupil dilation, where participants were required to collaborate with the robot to transport storage bins. This finding highlights the elevated cognitive activity associated with collaborative tasks that necessitate continuous engagement and robot coordination. The consistently higher mean pupil dilation suggests that active collaboration induces sustained attention and cognitive processing, thereby underscoring the importance of effective communication in such scenarios.

Discussion

Comparative Insights: Cognitive Activity Across Scenarios: A comparison of the various scenarios indicates that human-robot interaction (Scenarios 4 and 5) typically results in higher mean pupil dilation and lower variability than baseline and task-oriented scenarios (Scenarios 1 and 2). An indication that the nature of the task and the complexity of the interaction modes exert an influence on cognitive activity. Specifically, verbal and multimodal interactions and active collaboration have increased participants' cognitive activation and engagement, as evidenced by their pupil dilation responses.

Furthermore, Scenario 3, which considers the influence of mobile robot motion on human behavior in addition to the solely task-oriented Scenario 2, also demonstrates an increase in mean pupil dilation, particularly in Condition B (omnidirectional movement). The elevated cognitive activity observed in Scenario 3 indicates that the mere presence and motion of a mobile robot, mainly when exhibiting intricate movement patterns, necessitates a greater degree of mental exertion on the part of the participants to interpret and comprehend its movements. This scenario introduces dynamic elements that require participants to continuously adapt and process new information, thereby contributing to overall cognitive activity.

The lower variability observed in HRI scenarios (Scenarios 4 and 5) indicates a greater consistency in cognitive activity among participants. This consistency is likely due to effective communication protocols, which suggests that the tasks are equally demanding for most individuals. This uniformity is advantageous for designing intuitive and user-friendly human-robot interactions, ensuring predictable user responses, and enhancing training and evaluation processes. A deeper comprehension and exploitation of this consistency can enhance the design and implementation of robotic intent communication systems, rendering them more accessible and effective for a broader spectrum of users.

Implications for Robotic Intent Communication: These findings have significant implications for designing and implementing robotic intent communication systems. An understanding of the cognitive processes associated with different interaction styles can inform the development of communication protocols that are more intuitive and effective. For example, while multimodal interactions can enhance communication richness, they also increase cognitive activity. It is essential to ensure that these interactions effectively manage the cognitive demands placed upon users to prevent overwhelming them and maintain high engagement levels.

Moreover, the insights derived from Scenario 3 underscore the necessity of incorporating robot motion patterns into the design of human-robot interactions. The additional cognitive demands imposed by complex robot movements indicate that simplifying or more effectively communicating these movements could assist in managing cognitive activity levels and enhancing the user experience. Incorporating these findings into the design of HRI systems enables developers to create more effective and user-friendly interfaces that support seamless and intuitive interactions, thereby enhancing the overall efficacy of robotic intent communication.

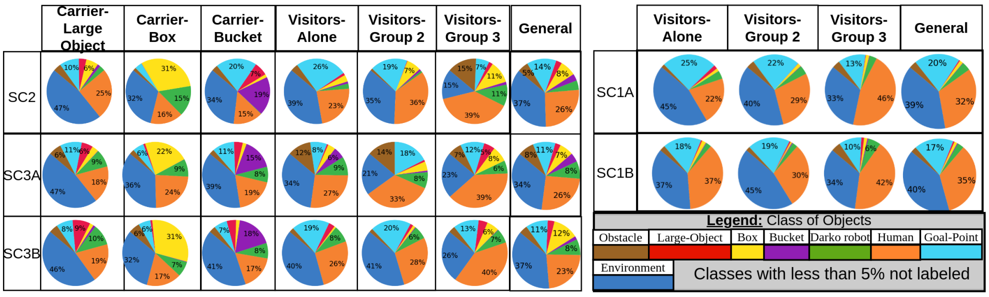


Figure 19: Distribution of fixations on objects by participants in Scenarios 2, 3A, and 3B (left) and 1A and 1B (right).

5.2.3 Exploring Visual Attention Through Computer Vision Techniques

In human-robot interaction (HRI), a fundamental understanding of how humans allocate their attention is essential for developing intuitive and effective robotic systems. The YOLO object detection model was employed to achieve a more granular decomposition of attention into classes of semantic objects. By identifying and categorizing objects or areas that attract significant visual focus, insights can be gained into the semantics of targets of participants' gaze, thereby enriching our understanding of attention allocation in dynamic settings, especially during locomotion. The application of contemporary computer vision methodologies to eye-tracking data represents a promising avenue for the contextual interpretation of human attention in human-robot interaction (HRI).

Methods We utilized the YOLOv8 object detection model [51], pre-trained on the COCO dataset [52], and refined with a custom dataset containing labeled objects from the THÖR-MAGNI project. Our custom dataset comprised 355 images annotated with seven classes: role-dependent objects (e.g., boxes and buckets), other walking people, and the mobile robot DARKO. The YOLOv8 model was applied to video frames captured by eye-tracking glasses to classify and quantify the visual focus on different objects, providing a detailed analysis of gaze distribution.

Results Applying the YOLOv8 model enabled the observation of significant alterations in the distribution of attention across diverse scenarios and activities. Figure 19 illustrates these shifts in attention distribution. It shows how attention allocation changes from focusing on the environment and other participants in Scenario 1 to more diversified attention toward the DARKO robot in subsequent scenarios.

In Scenario 2, participants observed a notable increase in attention directed toward large objects, including boxes and buckets, indicated by the substantial portions of the pie charts dedicated to these categories. The presence of other participants (visitors) also attracted considerable attention, indicating the dynamic nature of the interaction environment.

In Scenarios 3A and 3B, the introduction of the mobile robot DARKO resulted in a marked increase in attention directed towards the robot. Furthermore, attention distribution differed significantly based on the robot's driving style. Although overall attention towards the robot increased, there was no significant difference between the directional (3A) and omnidirectional (3B) driving styles, suggesting that the robot's mere presence was a significant factor in capturing attention rather than its specific movement patterns.

Scenarios 1A and 1B demonstrated that the initial exposure to the environment with added semantic cues (1B) did not significantly alter attention allocation compared to the pure baseline (1A). The pie charts for these scenarios demonstrate a balanced distribution of attention across various elements of the environment, providing further evidence to support the stability of the baseline cognitive activity.

The statistical analysis, which employed t-tests and calculated Cohen's d-effect sizes, corroborates these observations with significant findings. The transition from Scenarios 1 and 2 to Scenarios 3A and 3B revealed a marked increase in attention towards DARKO, with effect sizes of $[-1.6, -0.8]$, all with $p < 0.1$, respectively. These findings suggest that the robot's presence influences the allocation of participants' attention. Furthermore, there are no statistically significant differences in attention between the various driving styles of DARKO in Scenarios 3A and 3B, indicating that the robot's presence, whether static or dynamic and its motion pattern, rather than these factors, primarily capture human attention.

Discussion The presence of static and dynamic objects in an environment significantly impacts how humans orient their gaze. The results of our study indicate that the presence of the DARKO robot has a marked effect on attention allocation, evidenced by the increased focus on the robot in Scenarios 3A and 3B compared to the baseline and task-oriented scenarios. The minor yet statistically significant differences in attention directed towards various objects underscore the subtle ways in which robot presence and interaction modes influence cognitive processes.

The comparable distribution of gaze between directionally and omnidirectionally driving robots indicates that the perception of these two mobility styles is similar. This distribution suggests the possibility of versatility in human acceptance of different robotic mobility styles, thereby opening avenues for innovative robot designs without compromising the user experience. Confirming technological advancements in robot locomotion fosters optimism regarding their integration in human-centric environments.

In conclusion, applying YOLO object detection for gaze analysis has yielded significant insights into the allocation of attention during human-robot interaction (HRI). This approach enhances our comprehension of human cognitive processes in dynamic contexts, thereby facilitating the advancement of more intuitive and efficacious human-robot interaction systems. We recommend that researchers further explore integrating advanced computer vision techniques with eye-tracking data to clarify the complexities of human attention and engagement in HRI contexts.

5.3 Conclusion: Insights and Implications for Robotic Intent Communication

Our comprehensive analysis of the THÖR-MAGNI dataset offers valuable insights into visual attention and engagement in shared environments, particularly in human-robot interaction (HRI) and robotic intent communication. By examining pupil dilation and employing computer vision techniques to analyze gaze patterns, we have enhanced our understanding of how different interaction styles and tasks impact cognitive activity.

The study on human activity and engagement via pupil dilation demonstrated that cognitive activity remains relatively stable in baseline and task-oriented scenarios but increases with interactive elements, such as mobile robots. Specifically, scenarios encompassing human-robot interaction (Scenarios 4 and 5) exhibited heightened mean pupil dilation but with low variability, indicating consistent and elevated cognitive engagement. These findings highlight the significance of effective communication protocols and task design in maintaining user engagement and optimizing cognitive processing during HRI.

Applying the YOLO object detection model for gaze analysis provided further insight into the allocation of attention. This approach demonstrated that the presence and movement of the DARKO robot significantly influenced participants' gaze patterns, highlighting the robot's role in dynamic environments. The consistency in gaze distribution between different robot driving styles indicates that human acceptance of robotic mobility is versatile, which has significant implications for the design of future HRI systems.

Insights from previous sections emphasized that integrating anthropomorphic characteristics and multimodal communication strategies in the robotic intent communication system can facilitate trust and user acceptance in industrial contexts. The combination of verbal and gestural communication proved particularly effective in elucidating the robot's intentions and directing human attention. This effect is further supported by greater focused attention on the robot's face and reduced response times to instructions. These findings reinforce the importance of multimodal communication in effective human-robot interaction.

In conclusion, analyzing human gaze and cognitive engagement in the THÖR-MAGNI dataset provides evidence supporting the integration of advanced technologies and anthropomorphic features in robotic systems. These elements are crucial for enhancing human-robot interaction, ensuring effective communication, and promoting user trust and acceptance in shared environments.

6 Ongoing and Future Work

6.1 Large Language Model informed bidirectional HRI

Summary: Integrating multimodal foundation models has significantly enhanced autonomous agents' language comprehension, perception, and planning capabilities. However, while existing works adopt a *task-centric* approach with minimal human interaction, applying these models to developing assistive *user-centric* robots that can interact and cooperate with humans remains underexplored. We introduce "Bident", a framework designed to integrate robots seamlessly into shared spaces with humans. Bident enhances the interactive experience by incorporating multimodal inputs like speech and user gaze dynamics. Furthermore, Bident supports verbal utterances and physical actions like gestures, making it versatile for bidirectional human-robot interactions. Potential applications include personalized education, where robots can adapt to individual learning styles and paces, and healthcare, where robots can offer personalized support, companionship, and everyday assistance in the home and workplace environments.

6.1.1 Introduction

Designing and integrating assistive robots into daily life requires focusing on practical human-robot interaction (HRI) methods. This involves: (i) developing communication strategies that can handle complex interactions and interpret human intentions through ambiguous verbal and non-verbal cues (Theory of Mind), and (ii) designing robots with the flexibility to operate across various platforms, enabling adaptation to different environments and tasks. Traditional methods in HRI often rely on rigid, predefined schedules and struggle with novel scenarios [53], highlighting the need for more adaptive approaches. Despite advancements, significant challenges persist, often leading researchers to employ "Wizard of Oz" experiments, which covertly control robots to simulate advanced capabilities and gather data to refine these systems [54, 55]. While these experiments provide valuable insights, they highlight the limitations of current robotic systems and emphasize the need for more sophisticated solutions that can navigate the complexities of real-world human-robot interactions with little to no human intervention.

Recent advancements in Large Language Models (LLMs) have shown promise in addressing these challenges, particularly in Natural Language Processing [56] and reasoning [57]. LLMs have found applications in diverse robotics contexts, including task planning [58], manipulation [59], and improved perception [60, 61, 62]. In the domain of human-robot interaction (HRI), LLMs present opportunities for enhancing collaboration through multimodal inputs, potentially improving both communication and adaptability. However, most current approaches focus on task execution with minimal human interaction (i.e. task-centric). There is a significant need for an effective user-centric framework that integrates multimodal user input with task planning and action generation.

To this end, we focus on developing a user-centric framework for HRI, prioritizing the user's needs and intentions. Specifically, we introduce Bident, a framework that integrates multimodal user input – including verbal utterances and gaze dynamics – into its processing to fully capture the user's context. By analyzing both what the user says and where they look, Bident effectively tailors robot responses and actions that are contextually appropriate (see Figure 20). Particularly, eye tracking can be directly linked to shifts in human attention [63] and proven valuable in autonomous driving [64] and human motion analysis [65]. The "bi"directional aspect of Bident reflects the robot's ability to interpret the user's intent from inputs and support its own communication through verbal and gesture actions. This augments the user's abilities and improves the overall user experience, thereby making the user an integral part of the task execution process.

6.1.2 Methods and Ongoing Work

Our human-robot interface framework comprises different modules, which we will explain in the following section. It uses multimodal inputs to create a more immersive and collaborative HRI experience with interconnected modules communicating via a ROS2 network. Programmed in Python and tested with a simulated NAO robot, each module processes multimodal data for seamless interaction.

A) User Input: Vision and Audio The vision and audio modules of the framework work in tandem to capture and process multimodal inputs, enhancing the robot's understanding of its environment and interactions. The vision module utilizes visual inputs from mobile eye-tracking glasses and RGB-D cameras, integrating data from the user's perspective and the robot's contextual viewpoint. This module employs custom-trained models for object detection [66], segmentation [67], and tracking [68] to maintain accurate, real-time knowledge of both user focus and the broader scene. Meanwhile, the audio module transcribes verbal inputs using a local implementation of the Whisper module [69], enabling the robot to process and understand spoken language.

B) Reasoning Module The reasoning module is powered by advanced LLMs like GPT-3.5 [70] or Llama [71]. It receives inputs such as transcribed speech (from the audio module) and object and scene descriptions in natural language (from the vision module). Leveraging its reasoning capabilities and extensive world knowledge, the module then generates discrete physical actions (e.g., pointing to an object) and verbal actions (e.g., describing an object or posing a query to the user) to assist the user. We employ prompting approaches that guide the model through step-by-step processing [72, 73, 74, 75]. Additionally, it integrates user and environmental feedback to refine its actions [76]. We plan to explore both zero-shot [77] and in-context learning [78] capabilities to enhance performance. We measure performance by evaluating the robot's ability to accurately interpret inputs and generate appropriate, contextually relevant actions.

C) Action Module The action module enables the NAO robot to execute responses from the reasoning module, moving beyond the constraints of a pre-programmed action schedule [2]. Verbal responses are transformed from natural language into speech through the NAO robot's text-to-speech module [79]. The action module invokes the NAO robot's predefined functionalities, enabling it to point to and look at objects in the environment, state their categories, and provide further information upon request.

The action module incorporates a loopback prevention mechanism using a ROS message to prevent the loopback from detecting its vocalizations as speech input. Each round of communication is integrated into a feedback loop, informing and refining subsequent reasoning and execution processes. Feedback loops and active reasoning allow the robot to *actively perceive* and seek information in cases of ambiguous inputs (e.g., occlusions or ambiguous queries).

D) Future Evaluation: The evaluation of the framework will be conducted in two stages, providing a comprehensive assessment of its performance. The first stage will involve using a simulated NAO to test and fine-tune the modules for their intended purposes. The second stage will include a user study with an embodied NAO robot functioning as ARMoD for the DARKO robot. This stage will focus on measuring the accuracy, responsiveness, and contextual appropriateness of the robot's interactions, providing insights into its overall effectiveness and identifying potential areas for improvement.

Continuation and Future Work Our future work will refine our framework to enhance verbal communication and human gaze integration, realizing the concept of ARMoD 2.0. This will develop a robotic system capable of handling dynamic situations in industrial settings and potentially in healthcare, offering personalized support and companionship. Further experiments will validate the framework's effectiveness in bidirectional communication, safety, and comprehensibility of the NAO robot, identifying areas for improvement. We will conduct multiple user studies to compare the effectiveness of our approach over systems that rely on pre-programmed knowledge in interactions. While ensuring that ethical considerations like privacy and dependency are addressed, we will focus on creating a versatile and dependable robotic system for seamless integration into everyday industrial environments.

6.2 Investigating LEDs as a Communication Channel

This section provides a preliminary study design for a planned experiment towards the end of 2024. This study aims to provide a comprehensive understanding of the effectiveness of LED-based communication channels in human-robot interaction, with the goal of informing the design of future communication systems for mobile robots, enhancing their ability to operate seamlessly in shared environments with humans. We aim to investigate the effectiveness of different LED-based communication channels for the DARKO robot and the ARMoD. This study will involve participants recruited from the average visitors of the "German Museum" in Munich, Germany, and online participants via Prolific. We will recruit around 30 participants for the in-person study, and the online study will aim for the same sample size. See Figure 21 for an experimental design for the in-person study. The comparison between in-person and online participants will help us understand the effects of embodiment on the ratings of the robot and reaction times from users.

Participants will engage in a task where they observe the DARKO robot approaching an intersection. Their task will be to press a button when they believe they can correctly identify the robot's task or intention. Each encounter with the robot will take approximately 30 seconds. The environment will be set up with participants seated in one alley of an intersection, surrounded by boxes and industrial-like equipment to create a realistic setting. The environment will remain constant, with only the communication methods and robot tasks being manipulated. Online participants will watch videos of the encounters in randomized order.

The study will explore four different LED configurations: DARKO with Native LEDs (LED stripes on the DARKO robot), ARMoD using its LEDs (LEDs inside ARMoD's chest button and eyes), DARKO without LEDs, and a Combined Configuration (DARKO and ARMoD using both their LEDs). The LEDs will indicate status, motion intent, and instructional intent. Data collection methods will include eye-tracking using Tobii 3 Eye-Tracking Glasses to measure visual attention and gaze patterns, reaction times recorded via button presses, textual responses from participants on what they believe the robot's task is, and pre-and post-interaction questionnaires focusing on expected usability, perceived safety, trust, and overall user experience.

The experimental procedure will begin with an introduction to the study and the robot's communication methods, followed by a short training session to familiarize participants with the environment and the task. Participants will then observe the robot in each of the four conditions in a randomized order, with data being recorded for eye-tracking, reaction times, and textual responses for each condition. Participants will also complete pre- and post-interaction questionnaires. The data will be analyzed using Tobii Pro Lab Software to determine visual attention and gaze patterns, statistical analysis to compare reaction times across different conditions, and questionnaires to evaluate changes in perceived usability, trust, safety, and overall user experience before and after interaction.

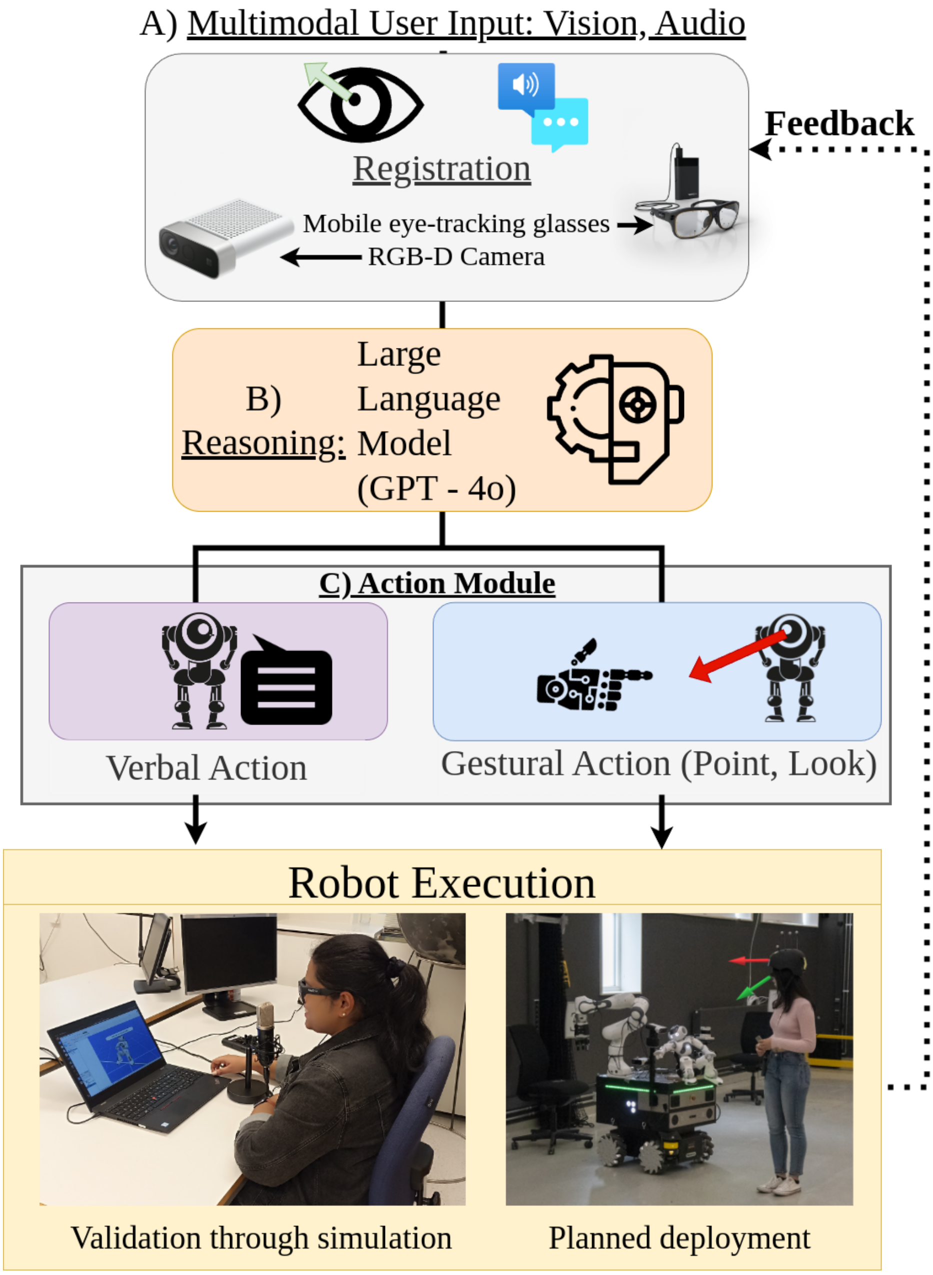


Figure 20: Bident framework for LLM informed dynamic interactions: Integrating verbal utterances and gaze (including head orientation **(red)** and eye-gaze direction **(green)**) allows an LLM to understand the situation through reasoning and generate action plans to appropriately respond to the user's input. Bident enables bidirectional communication by generating and refining plans through multimodal feedback **(dotted arrow)**, supporting closed-loop planning in dynamic environments. Participants interact with a simulated NAO robot to test the module. Final deployment will be withthe ARMoD [2]

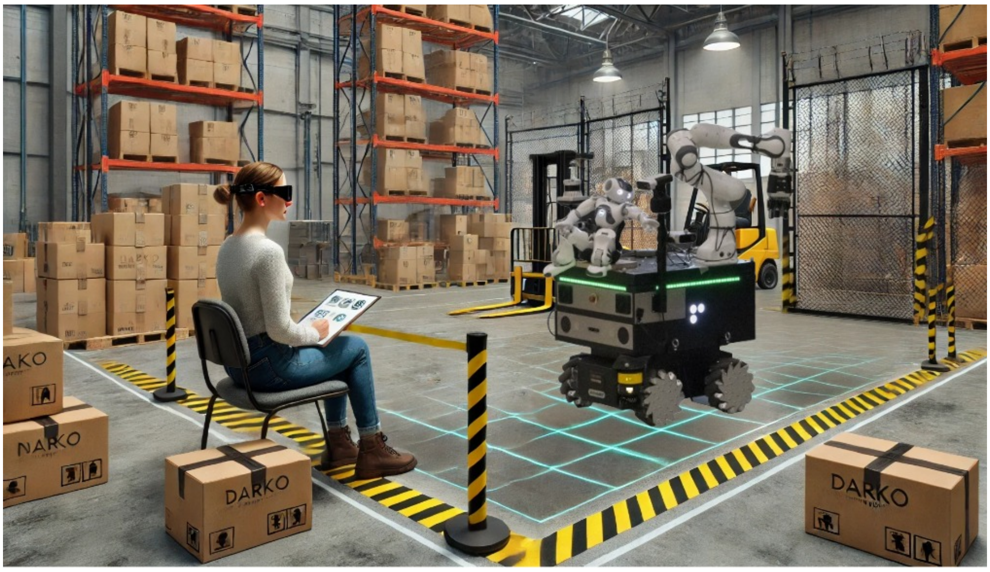


Figure 21: Concept Image for Study Investigating LED Communication Channels: This concept image depicts a study setup in an industrial-like warehouse environment. A human participant, wearing mobile eye-tracking glasses, is seated at an intersection, observing a mobile robot (DARKO) approaching. The mobile robot has LED stripes, and the NAO robot (ARMoD) mounted on top has LEDs in its chest and eyes. The LEDs can indicate status-, motion-, and instructional intent. The participant holds a button to press when identifying the robot's intent, capturing data on visual attention, reaction times, and task interpretation in a realistic setting. - Image partially generated by ChatGPT4o

7 WP5.2 Summary and Future Directions

This report outlines the research advances in human-robot interaction (HRI) and robot intent communication of WP5 T5.2. Each contributes to a comprehensive understanding of how robots can effectively communicate their intentions to humans. The work presented in the first five sections of this deliverable includes several key findings and implications for the HRI community.

Our investigation into the effects of anthropomorphism and multimodal communication on trust in industrial HRI has shown that adding anthropomorphic features to robots improves the quality and naturalness of interactions. This finding challenges traditional designs of industrial robots and suggests that incorporating human-like characteristics can lead to better acceptance and cooperation in shared workspaces. Furthermore, our developed concept of the ARMoD has highlighted the importance of multimodal communication strategies. By integrating speech, gaze, and gestures, robots can communicate their intentions more clearly and effectively, improving interaction quality. This research provides a framework for future robotic systems to adopt multimodal communication channels to ensure more natural and intuitive interactions, setting a new standard for how robots engage with humans, especially in dynamic and complex environments. These advances have practical implications for designing future industrial robots that are efficient, more relatable, and trustworthy to human coworkers.

The THÖR-MAGNI dataset has been an important asset of our research, providing data on spatial human-robot interaction. The detailed analysis of visual attention and engagement in shared environments has provided new perspectives on robot design to facilitate interactions with humans. The dataset is a critical resource for the community, fostering further research and development in spatial HRI and robotic intent communication.

Although the integration of multimodal Large Language Models (LLMs) is still in its early stages, our preliminary results indicate that these models hold great promise for improving HRI. The ability of LLMs to understand and generate human-like responses can largely improve the robot's interaction capabilities. LLMs' anticipated integration within the DARKO project will mark a leap forward in providing a robust system capable of reasoning and interaction in dynamic daily life-like HRI scenarios.

Impact of WP5.2 on Robotic Intent Communication

The research outlined in this report has far-reaching implications for the HRI and Robotic Intent Communication communities. First, it highlights the critical role of anthropomorphism and multimodal communication in building trust and facilitating effective human-robot interactions. Second, our findings from analyzing visual attention and engagement provide a new lens through which researchers can explore and understand human-robot interactions. Promoting the development of more adaptive and responsive robotic systems. Finally, integrating LLMs shows potential for more sophisticated and intelligent HRI applications capable of nuanced understanding and interaction.

In conclusion, this report presents significant advances in the field and sets the stage for future innovations. Our research contributes to creating more effective, engaging, and human-centered robotic systems by addressing the core challenges of trust, communication, and interaction quality. As the community continues to build on these findings, our findings will contribute to a future where robots, with their transformative potential, effortlessly integrate into human environments, enhancing productivity and collaboration.

8 References

Other references

[1] Max Pascher, Uwe Gruenefeld, Stefan Schneegass, and Jens Gerken. “How to Communicate Robot Motion Intent: A Scoping Review”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–17.

[2] Tim Schreiter, Lucas Morillo-Mendez, Ravi T Chadalavada, Andrey Rudenko, Erik Billing, Martin Magnusson, Kai O Arras, and Achim J Lilienthal. “Advantages of Multimodal versus Verbal-Only Robot-to-Human Communication with an Anthropomorphic Robotic Mock Driver”. In: *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2023, pp. 293–300.

[3] Emmanuel Senft, Satoru Satake, and Takayuki Kanda. “Would you mind me if i pass by you? socially-appropriate behaviour for an omni-based social robot in narrow environment”. In: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 2020, pp. 539–547.

[4] Zhanibek Rysbek, Ki-Hwan Oh, and Milos Zefran. “Recognizing intent in collaborative manipulation”. In: *Proceedings of the 25th International Conference on Multimodal Interaction*. 2023, pp. 498–506.

[5] Petr Vanc, Jan Kristof Behrens, Karla Stepanova, and Vaclav Hlavac. “Communicating human intent to a robotic companion by multi-type gesture sentences”. In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2023, pp. 9839–9845.

[6] Sichao Song and Seiji Yamada. “Effect of expressive lights on human perception and interpretation of functional robot”. In: *Extended abstracts of the 2018 CHI conference on human factors in computing systems*. 2018, pp. 1–6.

[7] K. L. Koay, G. Lakatos, D. S. Syrdal, M. Gácsi, B. Bereczky, K. Dautenhahn, A. Miklósi, and M. L. Walters. “Hey! There is someone at your door. A hearing robot using visual communication signals of hearing dogs to communicate intent”. In: *2013 IEEE Symposium on Artificial Life (ALife)*. 2013, pp. 90–97.

[8] Elizabeth Cha, Yunkyung Kim, Terrence Fong, Maja J Mataric, et al. “A survey of nonverbal signaling methods for non-humanoid robots”. In: *Foundations and Trends® in Robotics* 6.4 (2018), pp. 211–323.

[9] Maarika Oidekivi, Alexander Nolte, Alvo Aabloo, and Karl Kruusamäe. “Interpreting externally expressed intentions of an autonomous vehicle”. In: *2021 14th International Conference on Human System Interaction (HSI)*. IEEE. 2021, pp. 1–6.

[10] Ravi Teja Chadalavada, Henrik Andreasson, Maike Schindler, Rainer Palm, and Achim J Lilienthal. “Bi-directional navigation intent communication using spatial augmented reality and eye-tracking glasses for improved safety in human–robot interaction”. In: *Robotics and Computer-Integrated Manufacturing* 61 (2020), p. 101830.

[11] *Linde Safety Solutions: Linde BlueSpot*. [Online; accessed 10. Mar. 2023]. Mar. 2023.

[12] Georgios Angelopoulos, Francesco Vigni, Alessandra Rossi, Giuseppina Russo, Mario Turco, and Silvia Rossi. “Familiar acoustic cues for legible service robots”. In: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2022, pp. 1187–1192.

[13] Kerstin Severinson-Eklundh, Anders Green, and Helge Hüttenrauch. "Social and collaborative aspects of interaction with a service robot". In: *Robotics and Autonomous systems* 42.3-4 (2003), pp. 223–234.

[14] Maha Salem, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. "A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction". In: *2011 ro-man*. IEEE. 2011, pp. 247–252.

[15] Justin Hart, Reuth Mirsky, Stone Tejeda, Bonny Mahajan, Jamin Goo, Kathryn Baldauf, Sydney Owen, and Peter Stone. "Unclogging our arteries: using human-inspired signals to disambiguate navigational intentions". In: *arXiv preprint arXiv:1909.06560* (2019).

[16] Yun Zhang, Yaqin Cao, Robert W Proctor, and Yu Liu. "Emotional experiences of service robots' anthropomorphic appearance: a multimodal measurement method". In: *Ergonomics* 66.12 (2023), pp. 2039–2057.

[17] Fatma Bal, Mehmet Tekerek, Magdalena Palacz, Mehmet Gök, and Ramazan Şimşir. "Human Robot Interaction with Social Humanoid Robots". In: *El-Cezeri* 11.1 (2024), pp. 94–102.

[18] Tim Schreiter, Lucas Morillo-Mendez, Ravi T Chadalavada, Andrey Rudenko, Erik Alexander Billing, and Achim J Lilienthal. "The Effect of Anthropomorphism on Trust in an Industrial Human-Robot Interaction". In: *arXiv preprint arXiv:2208.14637* (2022).

[19] Martina Szabóová, Martin Sarnovskỳ, Viera Maslej Krešňáková, and Kristıéna Machová. "Emotion Analysis in Human–Robot Interaction". In: *Electronics* 9.11 (2020), p. 1761.

[20] Ruth Stock-Homburg. "Survey of Emotions in Human–Robot Interactions: Perspectives from Robotic Psychology on 20 Years of Research". In: *International Journal of Social Robotics* (2021), pp. 1–23.

[21] Manisha Natarajan and Matthew Gombolay. "Effects of anthropomorphism and accountability on trust in human robot interaction". In: *Proc. of the ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI)*. 2020, pp. 33–42.

[22] Lara Christoforakos, Alessio Gallucci, Tinatini Surmava-Große, Daniel Ullrich, and Sarah Diefenbach. "Can Robots Earn Our Trust the Same Way Humans Do? A Systematic Exploration of Competence, Warmth, and Anthropomorphism as Determinants of Trust Development in HRI". In: *Frontiers in Robotics and AI* 8 (Apr. 2021), p. 79.

[23] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. "A meta-analysis of factors affecting trust in human-robot interaction". In: *Human factors* 53.5 (2011), pp. 517–527.

[24] Bing Cai Kok and Harold Soh. "Trust in robots: Challenges and opportunities". In: *Current Robotics Reports* 1.4 (2020), pp. 297–309.

[25] Susanne Stadler, Astrid Weiss, Nicole Mirnig, and Manfred Tscheligi. "Anthropomorphism in the factory - A paradigm change?" In: *ACM/IEEE International Conference on Human-Robot Interaction*. 2013, pp. 231–232.

[26] Yunus Terzioğlu, Bilge Mutlu, and Erol Şahin. "Designing Social Cues for Collaborative Robots: The Role of Gaze and Breathing in Human-Robot Collaboration". In: *ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2020.

[27] Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. "Measuring acceptance of an assistive social robot: A suggested toolkit". In: *IEEE International Workshop on Robot and Human Interactive Communication*. 2009, pp. 528–533.

[28] Linda Onnasch and Clara Laudine Hildebrandt. “Impact of Anthropomorphic Robot Design on Trust and Attention in Industrial Human-Robot Interaction”. In: *ACM Transactions on Human-Robot Interaction (THRI)* 11 (1 Oct. 2021).

[29] George Charalambous, Sarah Fletcher, and Philip Webb. “The development of a scale to evaluate trust in industrial human-robot collaboration”. In: *International Journal of Social Robotics* 8.2 (2016), pp. 193–209.

[30] Takane Ueno, Yuto Sawa, Yeongdae Kim, Jacqueline Urakami, Hiroki Oura, and Katie Seaborn. “Trust in Human-AI Interaction: Scoping Out Models, Measures, and Methods”. In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 2022, pp. 1–7.

[31] Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F Malle. “What is human-like?: Decomposing robots’ human-like appearance using the anthropomorphic robot (abot) database”. In: *Proc. of the ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI)*. IEEE. 2018, pp. 105–113.

[32] Aida Amirova, Nazerke Rakhymbayeva, Elmira Yadollahi, Anara Sandygulova, and Wafa Johal. “10 Years of Human-NAO Interaction Research: A Scoping Review”. In: *Frontiers in Robotics and AI* 8 (2021).

[33] Jum. C Nunally. *Psychometric theory*. 3rd ed. McGraw-Hill, 1994.

[34] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021.

[35] B.L. Welch. “The generalization of ’student’s’ problem when several different population variances are involved”. In: *Biometrika* 34 (1-2 Jan. 1947), pp. 28–35.

[36] Rand Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, 2016.

[37] Patrick Mair and Rand Wilcox. “Robust statistical methods in R using the WRS2 package”. In: *Behav. Res. Methods* 52 (2 2020), pp. 464–488.

[38] R. Wilcox Rand and Tian S. Tian. “Measuring effect size: a robust heteroscedastic approach for two or more groups”. In: *Journal of applied statistic* 38 (7 July 2011), pp. 1359–1368.

[39] Eileen Roesler, Dietrich Manzey, and Linda Onnasch. “Embodiment matters in social hri research: Effectiveness of anthropomorphism on subjective and objective outcomes”. In: *ACM Transactions on Human-Robot Interaction* 12.1 (2023), pp. 1–9.

[40] Sichao Song and Seiji Yamada. “Designing led lights for a robot to communicate gaze”. In: *Advanced Robotics* 33.7-8 (2019), pp. 360–368.

[41] Elaheh Sanoubari, Byron David, Chase Kew, Corbin Cunningham, and Ken Caluwaerts. “From Message to Expression: Exploring Non-Verbal Communication for Appearance-Constrained Robots”. In: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2022, pp. 1193–1200.

[42] Chia-Ming Chang, Koki Toda, Xinyue Gui, Stela H Seo, and Takeo Igarashi. “Can Eyes on a Car Reduce Traffic Accidents?” In: *Proceedings of the 14th international conference on automotive user interfaces and interactive vehicular applications*. 2022, pp. 349–359.

[43] Julia Trabulsi, Kian Norouzi, Seidi Suurmets, Mike Storm, and Thomas Zoëga Ramsøy. “Optimizing fixation filters for eye-tracking on small screens”. In: *Frontiers in Neuroscience* 15 (2021), p. 578439.

[44] Kyveli Kompatsiari, Francesca Ciardo, Davide De Tommaso, and Agnieszka Wykowska. "Measuring engagement elicited by eye contact in Human-Robot Interaction". In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, pp. 6979–6985.

[45] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots". In: *International journal of social robotics* 1.1 (2009), pp. 71–81.

[46] Saul Greenberg, Nicolai Marquardt, Till Ballendat, Rob Diaz-Marino, and Miaosen Wang. "Proxemic interactions: the new ubicomp?" In: *interactions* 18.1 (2011), pp. 42–50.

[47] Andrey Rudenko, Tomasz P Kucner, Chittaranjan S Swaminathan, Ravi T Chadalavada, Kai O Arras, and Achim J Lilienthal. "Thör: Human-robot navigation data collection and accurate motion trajectories dataset". In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 676–682.

[48] Tim Schreiter, Tiago Rodrigues de Almeida, Yufei Zhu, Eduardo Gutierrez Maestro, Lucas Morillo-Mendez, Andrey Rudenko, Tomasz P Kucner, Oscar Martinez Mozos, Martin Magnusson, Luigi Palmieri, et al. "The Magni Human Motion Dataset: Accurate, Complex, Multi-Modal, Natural, Semantically-Rich and Contextualized". In: *arXiv preprint arXiv:2208.14925* (2022).

[49] Marianne Gullberg and Kenneth Holmqvist. "Keeping an eye on gestures: Visual perception of gestures in face-to-face communication". In: *Pragmatics & Cognition* 7.1 (1999), pp. 35–63.

[50] Tim Schreiter, Tiago Rodrigues de Almeida, Yufei Zhu, Eduardo Gutierrez Maestro, Lucas Morillo-Mendez, Andrey Rudenko, Luigi Palmieri, Tomasz P Kucner, Martin Magnusson, and Achim J Lilienthal. "TH\" OR-MAGNI: A Large-scale Indoor Motion Capture Recording of Human Movement and Robot Interaction". In: *arXiv preprint arXiv:2403.09285* (2024).

[51] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. *Ultralytics YOLOv8*. https://github.com/ultralytics/ultralytics. 2023.

[52] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft coco: Common objects in context". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755.

[53] Elena Corina Grigore, Kerstin Eder, Anthony G Pipe, Chris Melhuish, and Ute Leonards. "Joint action understanding improves robot-to-human object handover". In: *2013 IEEE/RSJ international conference on intelligent robots and systems*. IEEE. 2013, pp. 4622–4629.

[54] Laurel D Riek. "Wizard of oz studies in hri: a systematic review and new reporting guidelines". In: *Journal of Human-Robot Interaction* 1.1 (2012), pp. 119–136.

[55] Jauwairia Nasir, Pierre Oppliger, Barbara Bruno, and Pierre Dillenbourg. "Questioning Wizard of Oz: Effects of Revealing the Wizard behind the Robot". In: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2022, pp. 1385–1392.

[56] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. "Recent advances in natural language processing via large pre-trained language models: A survey". In: *ACM Computing Surveys* 56.2 (2023), pp. 1–40.

[57] Jie Huang and Kevin Chen-Chuan Chang. "Towards reasoning in large language models: A survey". In: *arXiv preprint arXiv:2212.10403* (2022).

[58] Rishi Hazra, Pedro Zuidberg Dos Martires, and Luc De Raedt. "SayCanPay: Heuristic Planning with Large Language Models using Learnable Domain Knowledge". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 2024, pp. 20123–20133.

[59] Peiyuan Zhi, Zhiyuan Zhang, Muzhi Han, Zeyu Zhang, Zhitian Li, Ziyuan Jiao, Baoxiong Jia, and Siyuan Huang. "Closed-loop open-vocabulary mobile manipulation with gpt-4v". In: *arXiv preprint arXiv:2404.10220* (2024).

[60] Xufeng Zhao, Mengdi Li, Cornelius Weber, Muhammad Burhan Hafez, and Stefan Wermter. "Chat with the environment: Interactive multimodal perception using large language models". In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2023, pp. 3590–3596.

[61] Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. "Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 14183–14193.

[62] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. "Language is not all you need: Aligning perception with language models". In: *Advances in Neural Information Processing Systems* 36 (2024).

[63] Michael I Posner. "Orienting of attention". In: *Quarterly journal of experimental psychology* 32.1 (1980), pp. 3–25.

[64] Feng Zhou, X Jessie Yang, and Joost CF De Winter. "Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving". In: *IEEE Transactions on Intelligent Transportation Systems* 23.3 (2021), pp. 2284–2295.

[65] Tim Schreiter, Andrey Rudenko, Martin Magnusson, and Achim J. Lilienthal. "Human Gaze and Head Rotation during Navigation, Exploration and Object Manipulation in Shared Environments with Robots". In: *arXiv preprint arXiv:2406.06300* (2024). This is the final version of the accepted manuscript for the 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN).

[66] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection". In: *arXiv preprint arXiv:2303.05499* (2023).

[67] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. "Segment anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

[68] Rishi Hazra, Brian Chen, Akshara Rai, Nitin Kamra, and Ruta Desai. "EgoTV: Egocentric Task Verification from Natural Language Task Descriptions". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 15417–15429.

[69] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. "Robust speech recognition via large-scale weak supervision". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 28492–28518.

[70] OpenAI. *GPT-3.5*. 2022.

[71] Hugo Touvron et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models". In: *arXiv* 2307.09288 (2023).

[72] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 24824–24837.

[73] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. "Least-to-Most Prompting Enables Complex Reasoning in Large Language Models". In: *The Eleventh International Conference on Learning Representations*. 2023.

[74] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. "ReAct: Synergizing Reasoning and Acting in Language Models". In: *The Eleventh International Conference on Learning Representations*. 2023.

[75] Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. *ADaPT: As-Needed Decomposition and Planning with Language Models*. 2023. arXiv: 2311.05772 [cs.AI].

[76] Aman Madaan et al. "Self-Refine: Iterative Refinement with Self-Feedback". In: *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 2023, pp. 46534–46594.

[77] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. "Large Language Models are Zero-Shot Reasoners". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 22199–22213.

[78] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[79] R. Gelin. "NAO". In: *Humanoid Robotics: A Reference*. Ed. by A. Goswami and P. Vadakkepat. Berlin: Springer-Verlag, 2017.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017274