Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Mirror Eyes: Explainable Human-Robot Interaction at a Glance

Matti Krüger, Daniel Tanneberg, Chao Wang, Stephan Hasler, Michael Gienger

2025

Preprint:

This is an accepted article published in 2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). The final authenticated version is available online at: https://doi.org/10.1109/RO-MAN63969.2025.11217810 Copyright 2025 IEEE

Mirror Eyes: Explainable Human-Robot Interaction at a Glance

Matti Krüger, Daniel Tanneberg, Chao Wang, Stephan Hasler, and Michael Gienger

***Abstract*— The gaze of a person tends to reflect their interest. This work explores what happens when this statement is taken literally and applied to robots. Here we present a robot system that employs a moving robot head with a screen-based eye model that can direct the robot's gaze to points in physical space and present a reflection-like mirror image of the attended region on top of each eye. We conducted a user study with 33 participants, who were asked to instruct the robot to perform pick-and-place tasks, monitor the robot's task execution, and interrupt it in case of erroneous actions. Despite a deliberate lack of instructions about the role of the eyes and a very brief system exposure, participants felt more aware about the robot's information processing, detected erroneous actions earlier, and rated the user experience higher when eye-based mirroring was enabled compared to non-reflective eyes. These results suggest a beneficial and intuitive utilization of the introduced method in cooperative human-robot interaction.**

I. INTRODUCTION

When multiple agents work together, they can accomplish goals that would be unattainable individually. However, such cooperation typically depends on coordinated actions and mutual understanding [1]–[5]. In the context of human-machine cooperation, natural language can serve as a powerful and intuitive medium for information exchange. Recent advancements in computational language processing [6], [7] have enabled machines to process and perform reasoning over natural language input, supporting sophisticated decision-making and action planning (e.g., [8], [9]). Yet, as a symbolic and context-dependent system, natural language is inherently ambiguous and susceptible to misinterpretation. For example, if a human requests *"pass me the can on the left"*, a robot might interpret *"left"* from the speaker's perspective, while the speaker actually intended the robot's left. Without further cues, the misunderstanding might not become evident until the action is carried out.

Multimodal communication that provides additional grounding can help to resolve some misunderstandings. For instance, in the above example, an additional glance or pointing gesture towards the object could have resolved the ambiguity. Such combinations of language and gaze or pointing gestures have been investigated previously (e.g., [10]–[14]) to enable faster and more natural communication from human to machine. Conversely, robot-generated gestures [15]–[18] and purposeful gazing [19]–[24] are being explored as communication media.

Focusing on the task of spatial referencing with the aid of robot gaze, a critical challenge consists of conveying the gaze target vector with sufficient precision. Physical models of eye balls require robust and accurate control to generate a consistent alignment of the model's pupil with the reference target in space (e.g., [25]). Screen-based virtual eye models (e.g., [21], [22], [26]–[28]) can reduce implementation complexity while adding expression flexibility (see, e.g., Figure 3) but their lack of physical depth, when shown on 2D displays, introduces other challenges for creating accurately interpretable references in 3D space [24].

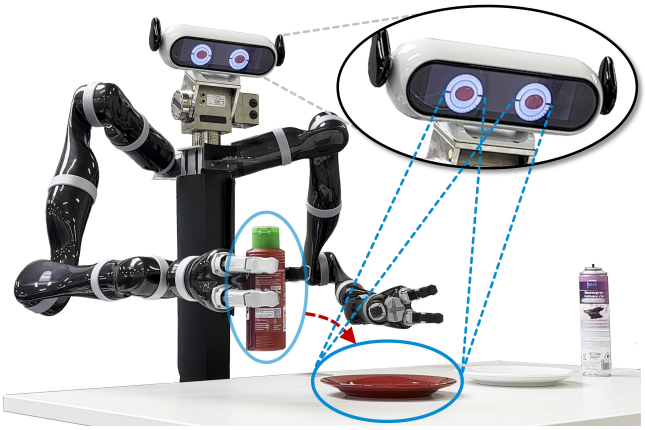

Fig. 1. Robot platform with the proposed Mirror Eye-enhanced head. Here the robot's task is to put the red bottle onto the red plate. In the displayed snapshot the robot has grasped the bottle and plans to put it onto the red plate next. The Mirror Eyes indicate this plan by showing the target location prior to plan execution.

To overcome spatial referencing limitations of screen-based eye models, Krüger et al. [24] have introduced a reflection-like overlay for virtual eye models that presents a live-image of the current focus of attention centered on each eye's pupil. The use of this method, termed Mirror Eyes, improved reference identification accuracy and user experience in a group interaction task, in which the eyes of the artificial agent were the only available reference cue. These results offer a first glance into the efficacy of Mirror Eyes and open new avenues for further investigation.

One such avenue is the consideration of ocular references across different object classes. In the study by Krüger et al. [24], the faces of the participants were the only attention targets that required action responses by users. However, faces may represent a special class of attention targets, as the human brain possesses highly specialized face-processing abilities [29], [30], and there is evidence for a link between gaze direction and face processing [31] (see [24]). It is unclear whether virtual mirror images of objects that lack such a dedicated neural processing architecture yield a comparable benefit. Secondly, the eyes were embedded in

Honda Research Institute Europe GmbH, Offenbach am Main, Germany. firstname.lastname@honda-ri.de Supplementary material: hri-eu.github.io/MirrorEyes

a virtual agent shown on a fixed screen that did not use any other means of spatial communication. The accuracy of screen-eye-based spatial communication could likely also be improved by embedding the eyes on mobile screens that can be physically oriented towards the attention target (e.g., [26]), thus partially elevating screen-based eyes to the third spatial dimension. This may lessen the need for disambiguation and call into question the necessity of additional stimuli.

These arguments provide a basis for considering whether the previously observed benefits of Mirror Eyes can be expected to generalize to other gaze targets and more physically integrated eye models. Therefore, we investigate the Mirror Eyes under extended conditions, where the eyes are...

a. implemented on a physical robot that can orient its head towards directions of interest,
b. used to attend to a wider range of object classes, and
c. used in conjunction with natural language

We present an implementation of Mirror Eyes under these conditions and examine the following hypotheses:

For a robot agent that generates spatial references through coordinated physical head movements and screen-based eye gaze, an addition of virtual reflections of attended objects on the eye models...

H1: improves system explainability by facilitating reference disambiguation

H2: improves people's ability to detect situations that require constructive interference with a cooperating robot's actions

H3: improves the experience of object references by interaction partners

We further interpret the outcomes of the individual hypotheses as mutually dependent in the order presented: A more interpretable system (**H1**) has practical benefits for interaction (**H2**) that should raise the experience people have during interaction (**H3**). Due to the Mirror Eyes' eschewal of symbolic communication, we anticipate corresponding effects to be present after brief exposure and without a need for instructions or training.

II. METHODOLOGY

To investigate these hypotheses, we needed a robot system with spatial gazing capabilities, natural language interaction, basic manipulation capabilities, and the option to interfere with the system during action execution. This system is described in the following section.

A. System Description

We base the implementation of the robot system on our prior work [9], [32] on a Large Language Model (LLM)-based robot agent that can physically and verbally interact with one or several users. The robot is depicted in Figure 1. It is a humanoid-type torso robot with two 7 degrees of freedom (DoF) arms and a pan-tilt unit carrying the expressive head. The head's front surface contains a 24x7 cm LCD screen that is used to display its eyes. It also holds two articulated ears to display character traits such as curiosity [8]. A 3D camera is mounted above the robot with a range large enough to see all relevant scene items and the persons located around a table in front of the robot. We use the Azure Kinect Body Tracking SDK to track the postures of interaction partners, and a combination of Segment Anything 2 [33], depth information and OpenAI's multimodal gpt4o model to locate and identify objects in the scene. Spoken language and sound direction are detected with a microphone array and converted to text with Google Cloud Speech.

The software implementation contains a set of python functions (aka "tools") that allow to query information about the scene, to act and manipulate the objects within its reach, and to express its internal state using cues such as speaking, pointing, and head gestures. These tools, along with a natural language text prompt describing the agent's character, constitute a run-time system that is triggered by external events (in our case detected spoken text), and responds by LLM-based reasoning about what information to gather, which tools to use, and how to reply in the given context. We found that this concept based on "Character" (prompt) and "Capabilities" (tools) in combination with advanced LLM reasoning capabilities allows natural and intuitive interactions, also comprising physical interactions. It is also flexible, making it a natural choice for the user study we present in this paper.

1) Continuous-time Interruptibility: We consider system interruptibility to constitute one basic form and enabler of constructive interference. Accordingly, for the present work we decided to extend the described system with the capability of being interruptible through speech. The capability uses a keyword detector that catches a set of interrupt keywords (like "stop"). If such a keyword is detected, the system stops the interaction with the LLM, and also possibly ongoing physical interactions. On the motion planning level, we let the robot place down all objects that it currently holds in its hands. Subsequently, we stop the robot trajectories so that the robot's motion will come to a halt.

2) Head-Eye Coordination: To provide a consistent real-world robot gazing behavior that allows the user to interpret the robot's spatial referencing correctly, we devised a novel kinematic gaze model with two key features: Firstly, it consistently computes the planar on-screen motion of the eyes to realistically refer to a spatial 3D scene. Secondly, it distributes the eye gaze consistently to the motion of the eyes on the screen as well as on the motor joints driving the head's motion. The model is depicted in Figure 2: We model two eye balls that are virtually located behind the screen. Each eye ball has two DoF, allowing its gaze vector to point at any location in the scene. We also model an attention target as a point that the system should look at. Using inverse kinematics (IK), we compute the motion of all DoFs (including the physical joints) so that the left and right eye's gaze vectors intersect the attention target. The location of the eyes on the LCD screen is then the intersection point of the corresponding eye's gaze vector with the screen surface. We use a standard differential IK with a weighted pseudo-inverse (see [34]) to compute the velocities of the DoFs:

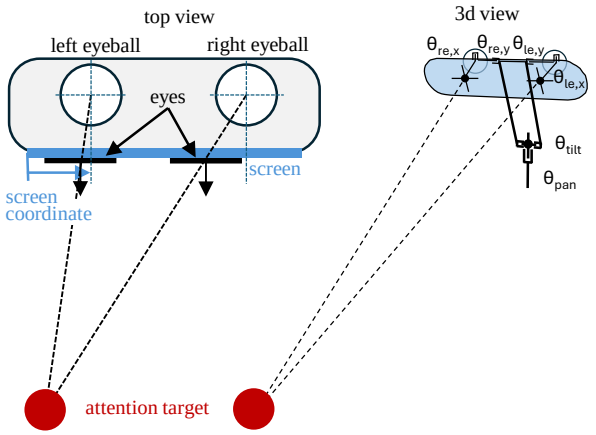

Fig. 2. Kinematics of the head-eye coordination model. Left: top view, right: 3D view including neck articulation with pan and tilt joints.

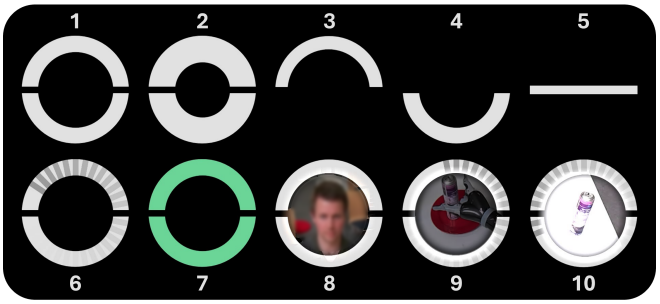

Fig. 3. Selected eye-based expressions. *(Top row)* Human-inspired expression primitives: 1. stylized iris and pupil, 2. reduced pupil size, 3. positive state, 4. negative state, 5. eyes closed, *(Bottom row)* Superhuman augmentations: 6. processing animation, 7. color coding of robot state, 8. Mirror Eyes focused on a person (blurred), 9. Mirror Eyes focused on objects at lower reflection opacity + loading animation indicating processing, 10. Mirror Eyes focus on an object with a brief overexposure (flash) indicating first registration.

$$\dot{q} = W_q \, J^T \left(J W_q J^T \right)^{-1} e \; + \; N \, W_J \dot{q}_0 \tag{1}$$

Vector $q = \{\theta_{pan} \; \theta_{tilt} \; \theta_{re,x} \; \theta_{re,y} \; \theta_{le,x} \; \theta_{le,y}\}^T$ comprises the physical and virtual DoFs of the model. The Jacobian J contains the kinematic constraint equations that force the gaze vector to intersect the attention target, and vector e is the constraint error. The diagonal weighting matrix W_q is parameterized to distribute the velocities on the individual DoFs. We use low weights for the motorized neck joints, and high weights for the virtual eye-ball joints. This leads to a gaze behavior with rapidly moving eyes and an accordingly slower moving head, resulting in an impression of purposeful gazing. The nullspace velocity vector q_0 models a resting pose in the sense to bring the gaze vector into a forward-pointing direction (orthogonal to the screen). The diagonal matrix W_J removes nullspace motion for the neck joints. This makes the head eventually converge to frontally face the attention target. This control concept also allows to seamlessly combine gaze with head gestures such as nodding or head shaking. To accomplish this, we simply add constraint equations for the DoFs related to the gestures in the kinematic model (typically pan and tilt DoFs). By forcing the gesture trajectory into these constraint variables, IK will adjust the gaze vector so that the gaze attention target is tracked throughout the head gesture, effectively mimicking the behavior of the vestibulo-ocular reflex (VOR) [35].

3) Eye Reflections: To realize the Mirror Eye effect, we added a mirror module that complements the rendering of the eye projections by creating an overlay of a horizontally flipped segment of a camera feed at a chosen transparency level. This module is based on the virtual reflection module introduced by Krüger et al. [24] but requires a different method for pupil-image alignment due to the underlying eye movement control described in Section II-A.2. Additionally, we extended the mirror module to be able to dynamically scale camera images of attended regions to the pupil size of the eyes in the case of attending objects of varying size or distance to the camera. These changes resulted in the following method for pupil-image alignment:

Given a point of interest (p_x, p_y) on a camera image with size c_w x c_h, the segment of its (optionally rescaled) mirror image with the top left corner (tl_x, tl_y) and target size t_w x t_h for rendering on the pupils, is given by

$$tl_{x|y} = o_{x|y} + \left(s \cdot t_{w|h} \cdot p_{x|y} \right) - t_{w|h} \tag{2}$$

In case of rescaling, the scale s and offset o are given by

$$s = min \left(\frac{t_{w|h}}{p_{w|h}}, 1 \right) \tag{3}$$

$$o_{x|y} = min \left(c_{w|h}, max \left(0, p_{x|y} - s \cdot p_{x|y} \right) \right) \tag{4}$$

Just as the original virtual reflection module introduced by Krüger et al. [24], the described technique has the added benefit of mapping potentially large camera images to relatively small eye models without loss of resolution. The added dynamic rescaling enables recognizable reflections of objects of arbitrary size or proximity. We implemented the mirror module using the *Python* programming language [36] with the *OpenCV* library [37] for image processing.

4) Eye Expressions: Virtual eye models enable a wide range of means for information encoding. Figure 3 shows a selection of features we implemented for our robot system. While some of these features, such as changes in pupil size or a simplified depiction of a closed eye, have counterparts in human eyes, others aim to explore expressions beyond the human repertoire. For instance, color changes may encode changes in specific system states such as turning green (Figure 3.7) when listening to the person towards whom the eyes are directed. A subsequent appearance and disappearance of stylized iris-bands (Figure 3.6) mimics aspects of a loading animation that can be used to indicate periods of information processing. In the Mirror Eyes condition (Figure 3.8-10), various filters, e.g., for blur, opacity, and brightness, may be used to contextually adapt saliency or visibility of mirrored content as needed. Within the system used in the presented study some of these features were enabled. The robot used the "smile"-expression of Figure 3.3 after successful task

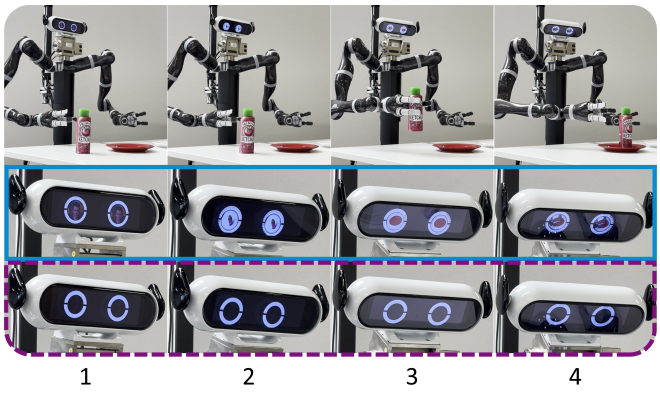

Fig. 4. Exemplary action sequence and close up images of the robot head in the Mirror Eyes (solid blue frame) and Eyes-Only (dashed purple frame) conditions. Sequence description: 1. The robot looks at a person standing in front of it. The person gives the instruction "put the red bottle onto the red plate" 2. The robot looks at the bottle it intends to pick up. 3. The robot looks at the plate it intends to place the bottle on while moving the bottle. 4. The robot looks at the joint location of the bottle and the plate after completing the pick-and-place action.

completion. The loading animation was used during periods of information processing such as after the user request and upon system interruption. In the Mirror Eyes condition the brief overexposure effect (Figure 3.10) was used when the robot first scanned objects in the scene.

B. User Study

To investigate our hypotheses using the described robot system, we designed a human-robot interaction experiment that consisted of two blocks, which differed with respect to the configuration of the robot head. In one block, the head of the robot described in Section II-A was configured to look at regions of interest through coordinated head- and eye movements (Eyes-Only condition). In the other block, the eyes were additionally augmented with reflection-like overlays showing mirror images of the attended regions (Mirror Eyes condition). The difference in head appearance for both conditions can be seen in Figure 4. In both conditions, the radius of the iris was set to 1.4 cm and the radius of the pupil to 1.23 cm.

Each experiment block consisted of three trials, in which participants were asked to stand in front of a table facing the robot, verbally make specific action requests to the robot

TABLE I

ACTION INSTRUCTIONS THAT COULD BE GIVEN TO THE ROBOT

Instruction - *Put the ...*	Expectation	Outcome	Error
red bottle onto the red plate	bottle → red	bottle → red	-
purple can onto the red plate	can → red	bottle → red	1
spray can onto the white plate	can → white	can → red	2
spray can onto the red plate	can → red	can → red	-
ketchup bottle onto the red plate	bottle → red	can → red	1
red bottle onto the white plate	bottle → white	bottle → red	2

Instructed actions (Expectation) and planned robot actions for the respective instruction (Outcome) are listed in a shorthand notation of the form "object to move → color of destination plate". Action errors are underlined. The error column shows the classification of the action error as no error (-), a step 1 (1) or a step 2 error (2).

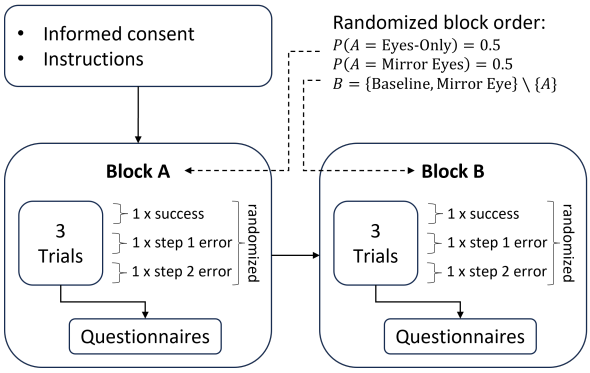

Fig. 5. Experiment procedure for individual participants.

when asked to do so by the experimenter, and monitor the robot's execution of each request. Each request involved two out of four objects available on the table such that one object (either a purple spray can or a red ketchup bottle) should be placed on another (either a red or a white plate). Table I lists the subset of requests that were utilized. With the listed requests, the robot could either succeed in carrying out the request or it could fail intentionally in two specific ways: 1. It could pick up the wrong object and place it on the correct target plate (Step 1 error - early). 2. It could pick up the correct object but place it on the wrong target plate (Step 2 error - late). Figure 6 illustrates these three potential outcome categories for a specific request.

By prompting the LLM to misinterpret specific labels on purpose, namely *purple can*, *ketchup bottle* and *white plate*, we enforced occurrences of the two action error types in two trials per block. For each trial, participants were asked to monitor whether the robot would carry out the given instructions correctly and interrupt the robot action in case they should notice an action error by saying *stop*. Correct

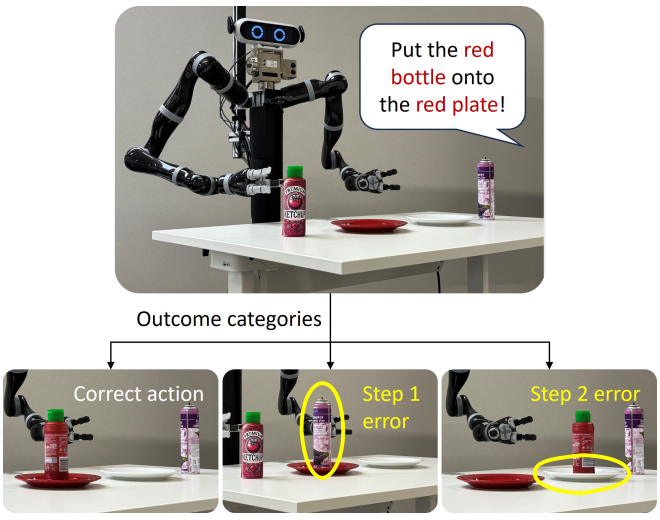

Fig. 6. Potential outcomes for a specific user request without action interruption. The robot was configured to either carry out a task (speech bubble) correctly as requested (left), make an early mistake by misinterpreting the object it should pick up (middle) or make a late mistake by misinterpreting the target location of the picked object (right).

robot actions should not be interrupted. The order of blocks and the order of trials within each block were randomized. The durations of individual robot actions were normalized so that key events occurred at the same times relative to function onset in all trials (see Figure 8). To prevent bias, participants were not instructed about the existence or purpose of the eyes in either condition. Exposure to each condition was intentionally kept short to minimize learning effects and assess whether understanding and utilization of the information given by the eyes would arise intuitively. After each block, participants were given two questionnaires to sample system explainability and user experience with respect to the robot head and their task of monitoring the robot actions for errors. Figure 5 illustrates the study procedure for each participant. The procedure took about 15 minutes in total. Each block, as well as the instructions and consent took about 5 minutes. The experiment involved 33 participants (8 women) aged between 22 and 65 years, mainly consisting of students and researchers from engineering and natural sciences disciplines.

III. RESULTS

A. Information Processing Awareness

We were interested in the explainability that each version of the head may provide about the system's information processing and behavior. For this purpose, we applied the subjective information processing awareness (SIPA) scale [38], which was developed to assess how well participants feel enabled by a system to perceive, understand, and predict its information processing. The SIPA score is acquired from a 6-item questionnaire, which was given to participants after completion of each experimental block.

Figure 7 presents the results of the SIPA ratings of 33 participants in the two display conditions of the eyes. SIPA scores were significantly higher ($\Uparrow$) for the Mirror Eyes condition ($M = 4.9, SD = 1.01$) than for the Eyes-Only condition ($M = 3.42, SD = 1.16$), $t(32) = 8.17, p < 0.001$. In support of our first hypothesis **H1**, these results indicate that participants perceived the Mirror Eyes condition to provide additional system explainability.

B. Interruption Speed

To assess people's ability to intuitively utilize the information provided by the robot eyes for constructive interference, we measured the time it took participants to interrupt the robot action when they observed the robot to make mistakes. The times of uttering the stop command were recorded

TABLE II

MEAN STOP TIMES WITH STANDARD DEVIATIONS AND LIMITS (N = 33)

Error step	Condition	Mean ± SD	Minimum	Maximum
1	Eyes-Only	5.52 ± 2.77	2.45	12.73
1	Mirror Eyes	4.66 ± 1.85	2.38	9.44
2	Eyes-Only	15.76 ± 1.31	14.09	19.61
2	Mirror Eyes	14.58 ± 2.9	4.29	17.71

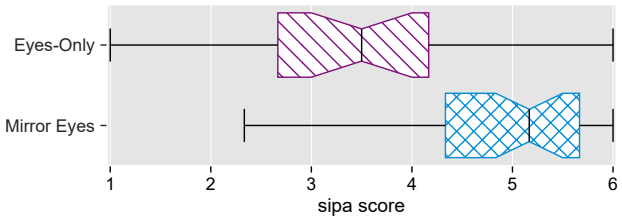

Fig. 7. Subjective information processing awareness (SIPA) scores ($\Uparrow$).

relative to the onset of the function that would execute the respective action. Figure 8 shows the distributions of error interruption times of 33 participants across conditions from function onset until the potential action completion. Table II shows the mean interruption times, standard deviations, minima, and maxima for each error step and condition.

A two-way ANOVA revealed significant main effects for display condition ($F(1, 124) = 6.1253, p < 0.05$) and error step ($F(1, 124) = 609.6492, p < 0.001$) but no significant interaction ($F(1, 124) = 0.1518, p = 0.6975$). Thus, interruptions (naturally) occurred earlier for step 1 errors ($M = 5.09, SD = 2.38$) than for step 2 errors ($M = 15.17, SD = 2.31$) and participants interrupted a mistaken robot earlier in the Mirror Eyes condition ($M = 9.47, SD = 5.54$) than in the Eyes-Only condition ($M = 10.32, SD = 5.7$). These results support the hypotheses **H1** and **H2**. First, the detection of erroneous actions prior to their full execution relies on the ability to correctly interpret the robot's actions or intentions. Secondly, a mismatch between such actions or intentions and an intended goal must be recognized to warrant interference in the form of a stop command.

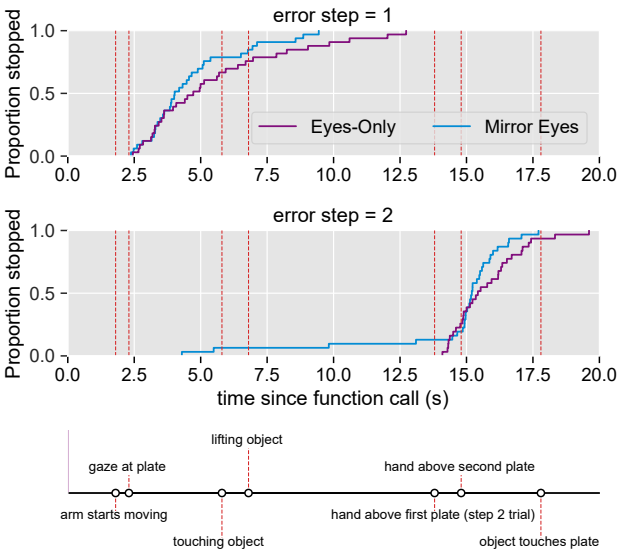

Fig. 8. Cumulative distributions of error interruption times (s) between function onset and potential action completion. Dashed vertical lines indicate start times of robot actions and scene events following function onset.

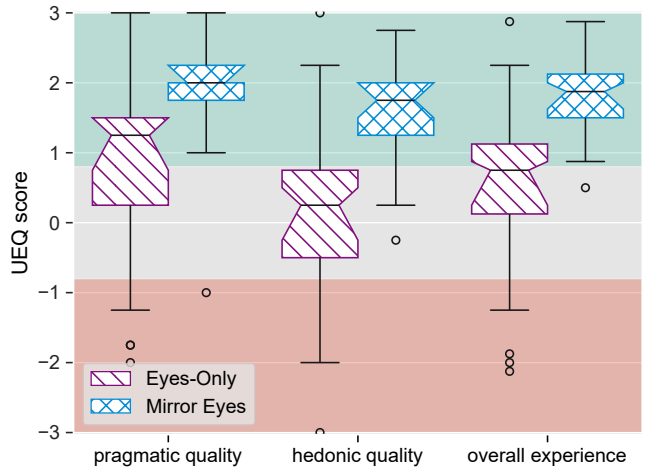

Fig. 9. Box plots for UEQ-S rating of pragmatic and hedonic quality, and overall user experience across the two different conditions of the eye models. Different colors represent evaluation ranges: values between −0.8 and 0.8 indicate neutral ratings (grey), values above 0.8 signify positive ratings (green), and values below −0.8 indicate negative ratings (red).

C. User Experience

Having found Mirror Eye-linked improvements in system explainability as well as practical advantages in terms of error detection performance, we would expect to also see a gain in the user experience. To evaluate the user experience with the two variants of the robot head, we used the 8-item version of the User Experience Questionnaire (UEQ-S [39]). After each experimental block, participants filled out the UEQ-S, consisting of eight items, four representing a pragmatic scale and the other four representing a hedonic scale. Each item is rated on a 7-point semantic differential scale, ranging from -3 (most negative) to +3 (most positive).

Figure 9 presents the results of the UEQ-S ratings of 33 participants who evaluated the user experience in the two display conditions of the eyes in terms of pragmatic quality, hedonic quality, and overall experience. Table III shows the scale means and standard deviations for the UEQ-S ratings in the two display conditions of the eyes. Pragmatic quality scores were significantly higher for the Mirror Eyes condition ($M = 2.01$, $SD = 0.72$) than for the Eyes-Only condition ($M = 0.88$, $SD = 1.32$), $t(32) = 5.32$, $p < 0.001$. Also, hedonic quality scores were significantly higher for the Mirror Eyes condition ($M = 1.67$, $SD = 0.66$) than for the Eyes-Only condition ($M = 0.08, SD = 1.32$), $t(32) = 8.11$, $p < 0.001$. Consequently, the overall user experience was significantly higher for the Mirror Eyes condition ($M = 1.84$, $SD = 0.52$) than for the Eyes-Only condition ($M = 0.48$, $SD = 0.52$), $t(32) = 8.22$, $p < 0.001$. In relation to a benchmark dataset of 468 studies [40], these ratings put the user experience in the Mirror Eyes condition into the "Excellent" category and within the top 10% range of all studies in the user experience benchmark dataset. These results support hypothesis **H3** about the positive impact of Mirror Eyes on user experience in human-robot interaction.

TABLE III
SCORE MEANS AND STANDARD DEVIATIONS OF THE UEQ-S (N = 33)

Condition	Pragmatic quality	Hedonic quality	Overall
Eyes-Only	0.88 ± 1.32	0.08 ± 1.32	0.48 ± 1.17
Mirror Eyes	2.01 ± 0.72	1.67 ± 0.66	1.84 ± 0.52

IV. DISCUSSION

Here we introduced a system that embeds the Mirror Eyes concept [24] in a robot equipped with a moving head, speech processing, vision processing, multimodal reasoning, and physical manipulation capabilities to investigate potential effects of Mirror Eyes on human-robot interaction settings. We conducted a human-robot interaction study with 33 participants in which participants were asked to verbally instruct the robot to carry out specific pick-and-place tasks, monitor the robot's execution of the request for correctness, and interrupt it verbally in case of erroneous actions. The task was carried out in two experimental conditions, one with an enabled Mirror Eye feature and one that included the eyes but lacked the mirroring effect. In one third of the trials the robot was programmed to make an error early during the pick-up phase, in one third it made a late error during the object placement phase, and in the last third it carried out the request correctly. We found that participants felt more aware about the robot's information processing, detected erroneous actions earlier, and rated the user experience higher when mirroring was enabled. These results are consistent with our hypothesis about a transfer of Mirror Eye benefits to physically embedded robot systems and confirm that the benefits are not exclusive to faces as reference targets. In fact, compared to findings on screen-based Mirror Eye use in a group interaction setting [24], the user experience with the Mirror Eyes embedded in a physically moving robot head was even higher in the present study.

One possible explanation for the transfer and even amplification of Mirror Eye benefits lies in the nature of human multimodal perception. Human sensory experience is partly attributable to an integration of information from multiple senses, including percepts that might feel strictly unimodal [41]–[43] (see [44]). Even in the absence of ambiguity, congruent [45]–[48] stimuli from multiple senses have been found to be advantageous, e.g., by increasing reaction speed [49]–[54] and reducing the risk of sensory overload in demanding scenarios [55], [56]. A virtual reflection and an eye model are both perceived visually. Nevertheless, physiological models for visual processing suggest that they might not be processed in the same way: The two-stream model for visual processing distinguishes between two main neural pathways that process different aspects of visual information. The ventral stream processes foveal retinal input with a high spatial resolution for the recognition of objects and faces [57] whereas the dorsal stream is responsible for processing motion and spatial relationships, including observed eye movements [58] at high temporal resolution using input from the entire retina. These complementary and parallel roles might be exploited by stimuli that combine

location and identity information, such as those created by the Mirror Eyes. But even if such a within-sense facilitation should not be present or relevant, Mirror Eyes could still be supportive for a disambiguation of information from other senses such as speech and other sounds.

One key factor to take into account in the interpretation of the results is that they reflect interactions with very brief exposure to the system and each experimental condition. This restriction, coupled with the intentional omission of any explanation about the functionality or role of the eyes, was implemented to capture the intuition the Mirror Eyes might provide in human-robot interaction. The results suggest, that an intuitive use of the Mirror Eyes has indeed taken place and was met with high appreciation by the participants.

Nevertheless, this avoidance of learning effects also means that effects might change once participants get used to the presence of the eyes in either condition and learn what features to pay attention to at what points during an interaction. A first glance into the potential of more optimized use by human interaction partners might be present in the interruption times of a subset of participants in step 2 (late) error cases. Figure 8 shows that some participants were able to identify and interrupt step 2 errors around 10 seconds earlier in the Mirror Eyes condition than in the Eyes-Only condition. This was theoretically possible because the gaze on the placement destination preceded the corresponding movement. We observed that most participants in the Mirror Eyes condition appeared to have noticed the gaze towards the wrong placement target in step 2 error cases early, but hesitated with the stop command until receiving confirmation from a corresponding movement of the arms. Similarly, for pick-up errors (step 1), participants who showed signs of early error detection through the Mirror Eyes typically appeared to wait for arm movement confirmation before interruption. We therefore consider it likely that the proportion of early interruptions will increase once participants become more familiar with the scenario, e.g., by increasing the number of trials per participant. In contrast to the present study, which focused on early intuition, future research should attend to the development of Mirror Eye effects after prolonged or repeated exposure, as well as the role of robot kinematics in disambiguating eye behavior.

One reason for hesitation in acting earlier on information provided by the Mirror Eyes could be the multitude of roles eyes play as communication interfaces: Eyes point towards regions that are observed, but the reasons for observation can be diverse, such as scanning the environment or preparing an action. While the Mirror Eyes may leave little doubt about *what* is observed, there can still be ambiguity about *why* it is observed. To an extent, we tried to counter this ambiguity by introducing brief moments of overexposure during initial object registration, as depicted in Figure 3.10. However, since no element of the eyes was explained to the participants, it is unclear whether the participants noticed or interpreted this detail in fulfillment of their monitoring task. Future work could focus on extensions that aim to circumvent remaining ambiguities while preserving the observed intuition.

V. CONCLUSION

The addition of a reflection-like overlay on the display of virtual eye models (Mirror Eyes) on a mobile robot head significantly improved people's subjective understanding of a robot's information processing, their ability to detect and interrupt erroneous robot actions, and their user experience with the robot head as a communication channel in interaction with the robot. These results suggest that even mobile robot heads, which can be physically oriented towards regions of interest, benefit from Mirror Eye utilization in human-robot interaction scenarios. Because these effects occurred after very short exposure to the system and without any explanations about the role of the eyes in the interaction, we believe that they are indicative of the presence of quick intuition. Future research should investigate the effects of prolonged exposure to Mirror Eyes as well as complementary means for reducing any remaining ambiguities the eyes may have as explanatory media.

REFERENCES

[1] M. E. Bratman, "Shared cooperative activity," *The philosophical review*, 1992.

[2] J.-M. Hoc, "Towards a cognitive approach to human–machine cooperation in dynamic situations," *International journal of human-computer studies*, 2001.

[3] M. Krüger, C. B. Wiebel, and H. Wersing, "From tools towards cooperative assistants," in *International Conference on Human Agent Interaction*, 2017.

[4] B. Sendhoff and H. Wersing, "Cooperative intelligence-a humane perspective," in *IEEE international conference on human-machine systems (ICHMS)*, 2020.

[5] P. Wollstadt and M. Krüger, "Quantifying cooperation between artificial agents using synergistic information," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2022.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Neural information processing systems*, 2017.

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah *et al.*, "Language models are few-shot learners," *Neural information processing systems*, 2020.

[8] C. Wang, S. Hasler, D. Tanneberg, F. Ocker, F. Joublin, A. Ceravola, J. Deigmoeller, and M. Gienger, "Lami: Large language models for multi-modal human-robot interaction," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024.

[9] D. Tanneberg, F. Ocker, S. Hasler, J. Deigmoeller, A. Belardinelli, C. Wang, H. Wersing, B. Sendhoff, and M. Gienger, "To help or not to help: Llm-based attentive support for human-robot group interactions," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[10] R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel, "Natural human-robot interaction using speech, head pose and gestures," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2004.

[11] I. Maurtua, I. Fernandez, A. Tellaeche, J. Kildal, L. Susperregi, A. Ibarguren, and B. Sierra, "Natural multimodal communication for human–robot collaboration," *International Journal of Advanced Robotic Systems*, 2017.

[12] C. Wang, M. Krüger, and C. B. Wiebel-Herboth, ""watch out!": prediction-level intervention for automated driving," in *International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2020.

[13] E. Menendez, M. Gienger, S. Martínez, C. Balaguer, and A. Belardinelli, "Semanticscanpath: Combining gaze and speech for situated human-robot interaction using llms," 2025.

[14] A. Belardinelli, C. Wang, D. Tanneberg, M. Krüger, S. Hasler, and M. Gienger, "Explaining to and being explained by a service robot: Four hri studies revisited under a framework for explainability," in *Proceedings of the 3rd TRR Conference Contextualizing Explanations (ContEx)*, 2025.

[15] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin, "Generation and evaluation of communicative robot gesture," *International Journal of Social Robotics*, 2012.

[16] C. P. Quintero, R. Tatsambon, M. Gridseth, and M. Jägersand, "Visual pointing gestures for bi-directional human robot interaction in a pick-and-place task," in *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2015.

[17] G. Hoffman, O. Zuckerman, G. Hirschberger, M. Luria, and T. Shani Sherman, "Design and evaluation of a peripheral robotic conversation companion," in *ACM/IEEE international conference on human-robot interaction*, 2015.

[18] C. T. Ishi, D. Machiyashiki, R. Mikata, and H. Ishiguro, "A speech-driven hand gesture generation method and evaluation in android robots," *IEEE Robotics and Automation Letters*, 2018.

[19] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The icub humanoid robot: An open platform for research in embodied cognition," in *8th Workshop on Performance Metrics for Intelligent Systems*, 2008.

[20] A. Zaraki, D. Mazzei, M. Giuliani, and D. De Rossi, "Designing and evaluating a social gaze-control system for a humanoid robot," *IEEE Transactions on Human-Machine Systems*, 2014.

[21] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: a review," *Journal of Human-Robot Interaction*, 2017.

[22] N. Yoshida, S. Yonemura, M. Emoto, K. Kawai, N. Numaguchi, H. Nakazato, S. Otsubo, M. Takada, and K. Hayashi, "Production of character animation in a home robot: A case study of lovot," *International Journal of Social Robotics*, 2022.

[23] A. Belardinelli, A. R. Kondapally, D. Ruiken, D. Tanneberg, and T. Watabe, "Intention estimation from gaze and motion features for human-robot shared-control object manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[24] M. Krüger, Y. Oshima, and Y. Fang, "Virtual reflections on a dynamic 2d eye model improve spatial reference identification," *arXiv*, 2025.

[25] M. Teyssier, M. Koelle, P. Strohmeier, B. Fruchard, and J. Steimle, "Eyecam: Revealing relations between humans and sensing devices through an anthropomorphic webcam," in *CHI Conference on Human Factors in Computing Systems*, 2021.

[26] R. Gomez, D. Szapiro, K. Galindo, and K. Nakamura, "Haru: Hardware design of an experimental tabletop robot assistant," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2018.

[27] Y. Fang, L. Merino, S. Thill, and R. Gomez, "Designing visual and auditory attention-driven movements of a tabletop robot," in *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2023.

[28] Y. Fang, J. M. Pérez-Molerón, L. Merino, and R. Gomez, "Enhancing human perception of direct gaze from a social robot through eye-head coordination," in *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2024.

[29] G. McCarthy, A. Puce, J. C. Gore, and T. Allison, "Face-specific processing in the human fusiform gyrus," *Journal of cognitive neuroscience*, 1997.

[30] G. Besson, G. Barragan-Jason, S. J. Thorpe, M. Fabre-Thorpe, S. Puma, M. Ceccaldi, and E. J. Barbeau, "From face processing to face recognition: Comparing three different processing levels," *Cognition*, 2017.

[31] N. George, J. Driver, and R. J. Dolan, "Seen gaze-direction modulates fusiform activity and its coupling with other brain areas during face processing," *Neuroimage*, 2001.

[32] F. Joublin, A. Ceravola, P. Smirnov, F. Ocker, J. Deigmoeller, A. Belardinelli, C. Wang, S. Hasler, D. Tanneberg, and M. Gienger, "CoPAL: corrective planning of robot actions with large language models," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[33] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," *arXiv*, 2024.

[34] Y. Nakamura, *Advanced Robotics: Redundancy and Optimization*. Addison-Wesley, 1991.

[35] T. Shibata, S. Vijayakumar, J. Conradt, and S. Schaal, "Biomimetic oculomotor control," *Adaptive Behavior*, 2001.

[36] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[37] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[38] T. Schrills, M. Sieger, M. Gruner, and T. Franke, "Evaluation of a scale to assess subjective information processing awareness of humans in interaction with automation & artificial intelligence," *Artificial Intelligence and Social Computing*, 2024.

[39] M. Schrepp, J. Thomaschewski, and A. Hinderks, "Design and evaluation of a short version of the user experience questionnaire (ueq-s)," *International Journal of Interactive Multimedia and Artificial Intelligence*, 2017.

[40] A. Hinderks, M. Schrepp, and J. Thomaschewski, "A benchmark for the short version of the user experience questionnaire," in *International Conference on Web Information Systems and Technologies-APMDWE*, 2018.

[41] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, 1976.

[42] M. Botvinick, J. Cohen *et al.*, "Rubber hands ' feel' touch that eyes see," *Nature*, 1998.

[43] J. V. Verhagen and L. Engelen, "The neurocognitive bases of human multimodal food perception: sensory integration," *Neuroscience & biobehavioral reviews*, 2006.

[44] M. Krüger, "An enactive approach to perceptual augmentation in mobility," Ph.D. dissertation, Ludwig-Maximilians-Universität München, 2022.

[45] M. Meredith, J. Nemitz, and B. Stein, "Determinants of multisensory integration in superior colliculus neurons. i. temporal factors," *Journal of Neuroscience*, 1987.

[46] B. E. Stein and M. A. Meredith, *The merging of the senses.* The MIT Press, 1993.

[47] M. A. Meredith and B. E. Stein, "Spatial determinants of multisensory integration in cat superior colliculus neurons," *Journal of Neurophysiology*, 1996.

[48] D. Alais, F. Newell, and P. Mamassian, "Multisensory processing in review: from physiology to behaviour," *Seeing and Perceiving*, 2010.

[49] J. W. Todd, *Reaction to multiple stimuli*. Science Press, 1912.

[50] M. Hershenson, "Reaction time as a measure of intersensory facilitation." *Journal of experimental psychology*, 1962.

[51] R. S. Nickerson, "Intersensory facilitation of reaction time: energy summation or preparation enhancement?" *Psychological review*, 1973.

[52] I. H. Bernstein, M. H. Clark, and B. A. Edelstein, "Effects of an auditory signal on visual reaction time." *Journal of experimental psychology*, 1969.

[53] A. Diederich and H. Colonius, "Bimodal and trimodal multisensory enhancement: Effects of stimulus onset and intensity on reaction time," *Perception & Psychophysics*, 2004.

[54] J. L. Mozolic, C. E. Hugenschmidt, A. M. Peiffer, and P. J. Laurienti, "Multisensory integration and aging," *In: The Neural Bases of Multisensory Processes*, 2012.

[55] P. A. Hancock, J. E. Mercado, J. Merlo, and J. B. V. Erp, "Improving target detection in visual search through the augmenting multi-sensory cues," *Ergonomics*, 2013.

[56] B. Wahn, J. Schwandt, M. Krüger, D. Crafa, V. Nunnendorf, and P. König, "Multisensory teamwork: using a tactile or an auditory display to exchange gaze information improves performance in joint visual search," *Ergonomics*, 2016.

[57] N. Kanwisher, J. McDermott, and M. M. Chun, "The fusiform face area: a module in human extrastriate cortex specialized for face perception," *Journal of neuroscience*, 1997.

[58] O. Braddick, J. Atkinson, and J. Wattam-Bell, "Normal and anomalous development of visual motion processing: motion coherence and 'dorsal-stream vulnerability'," *Neuropsychologia*, 2003.