The Need for Verbal Robot Explanations and How People Would Like a Robot to Explain Itself

ZHAO HAN, University of Massachusetts Lowell, USA

ELIZABETH PHILLIPS, George Mason University, USA

HOLLY A. YANCO, University of Massachusetts Lowell, USA

Although non-verbal cues such as arm movement and eye gaze can convey robot intention, they alone may not provide enough information for a human to fully understand a robot's behavior. To better understand how to convey robot intention, we conducted an experiment ($N = 366$) investigating the *need for robots to explain*, and the *content* and *properties* of a desired explanation such as *timing*, *engagement importance*, *similarity to human explanations*, and *summarization*. Participants watched a video where the robot was commanded to hand an almost-reachable cup and one of six reactions intended to show the unreachability: doing nothing (No Cue), turning its head to the cup (Look), or turning its head to the cup with the addition of repeated arm movement pointed towards the cup (Look & Point), and each of these with or without a Headshake. The results indicated that participants agreed robot behavior should be explained across all conditions, *in situ*, in a similar manner as what human explain, and provide concise summaries and respond to only a few follow-up questions by participants. Additionally, we replicated the study again with $N = 366$ participants after a 15-month span and all major conclusions still held.

CCS Concepts: • **Human-centered computing** → **Empirical studies in interaction design**;

Additional Key Words and Phrases: Robot explanation, behavior explanation, system transparency

ACM Reference format:

Zhao Han, Elizabeth Phillips, and Holly A. Yanco. 2021. The Need for Verbal Robot Explanations and How People Would Like a Robot to Explain Itself. *ACM Trans. Hum.-Robot Interact.* 10, 4, Article 36 (September 2021), 42 pages.

<https://doi.org/10.1145/3469652>

1 INTRODUCTION

Explaining our behavior is a natural part of daily life and the lack of explanations can be unsettling and disturbing [24]. It is important to provide explanations to improve understanding. This need is especially true for robots, often designed to be perceived as intelligent entities with physical appearances that resemble humans. Past work in **Human-Robot Interaction (HRI)** has illustrated that improving understanding of a robot makes it more trustworthy [16] and the resulting

This work has been supported in part by the Office of Naval Research (N00014-18-1-2503) and in part by the National Science Foundation (IIS-1909847).

Authors' addresses: Z. Han and H. A. Yanco, University of Massachusetts Lowell, 1 University Ave, Lowell, MA, USA, 01854; emails: zhao_han@student.uml.edu, holly@cs.uml.edu; E. Phillips, George Mason University, 4400 University Dr., Fairfax, VA, USA, 22030; email: ephill3@gmu.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.

2573-9522/2021/09-ART36 \$15.00

<https://doi.org/10.1145/3469652>

Fig. 1. Three of the six handover conditions after the almost-reachable cup is detected, without the head-shake included. *Left*: Robot does nothing (No Cue). *Middle*: Robot's head turns towards the cup (Look). *Right*: Robot's head turns towards the cup and its right arm is extended repeatedly (Look & Point).

interaction more efficient [3]. Human-agent interaction researchers in the artificial intelligence community have started to enable virtual agents to provide explanations [5]. Unlike virtual agents, however, robots have physical embodiment, which has been shown to have effects on metrics such as empathy [43] and cooperation [7].

Yet to our best knowledge, it remains unclear how robots should be enabled to *verbally* explain intention and behavior, or how to couple verbal explanations with other embodied explanatory cues, despite the growing collection of HRI research on non-verbal motion cues to indicate a robot's intent (e.g., References [19, 33, 41]). Before empowering robots to explain themselves, we need to answer *what and how would humans like a robot to explain?* In this study, we investigated participants' desired verbal explanations after being presented with robot non-verbal motion cues.

We drew inspiration from psychology: Specifically, research that examines the purposes of desired explanations from another. Psychologists have found that people hope to gain causal knowledge from explanations [8, 36] rather than from statistical evidence [25]. Koslowski [32] illustrated that when told that a car's color was a determining factor in the car's gas mileage, children and adults did not believe this to be true. Rather, they only believed that the size of the car impacted gas mileage. Their beliefs could be changed, however, when given the explanation that car color affects the mood of the driver, which can change whether or not the car is driven in a fuel-efficient manner. With causal knowledge, understanding is improved, and "people can simulate counterfactual as well as future events under a variety of possible circumstances" [37].

Knowing that the purpose of seeking explanations is often for causal knowledge, we designed and conducted an online experiment using Amazon's **Mechanical Turk (MTurk)** to investigate the perceived need for robot explanations, whether that perceived need is similar to what we would expect of a human, as well as how people want a robot to provide such explanations even when non-verbal causal motion cues are provided by the robot.

Participants first watched a scenario in which a Baxter robot was not able to hand over a cup. Participants then watched one of six reactions (Figure 1) from the robot containing motion cues intended. Information about why the robot was not able to complete the cup handover (causal information) was missing, i.e., the fact that the cup was slightly out of reach was not known to participants.

Participants then watched one of six reactions from the robot containing motion cues intended to convey why the robot was unable to complete the handover task: doing nothing (No Cue), turning its head to the cup (Look), or turning its head to the cup with the addition of repeated arm movement pointed towards the cup (Look & Point), and each of these with or without the addition of the robot providing a Headshake.

In a questionnaire, we asked participants if they perceived unexpected things throughout the task that they felt the robot should verbally explain. We then asked about some of the desired properties and the content of the robot explanations following a failed handover. Specifically, we asked about desired explanation timing (at the end, *in situ*, *a priori*, or other), engagement importance, whether and how robot explanations should be different from human explanations, explanation summarization, and explanation verbosity.

Results revealed that participants felt robot behavior should be explained in all conditions. The addition of a headshake and arm movement motion cues to indicate that the robot could not complete the task did not result in reduced need for explanations or less perceived unexpectedness, but rather confused participants, as the headshake was often interpreted as the robot having disobeyed the handover request, and the intention of the arm movement was unclear to participants.

Regarding the properties of robot explanations, participants reported that they wanted robots to explain *in situ*, not at the end of the task. They also thought that engagement was important. The robot should get the participants' attention by looking at them and possibly address them by name before explaining, which needs more investigation. People thought robots should explain the same content as humans explain, wanted concise summaries, and were willing to ask a few (1 to 3) follow-up questions after a summarized explanation was provided by the robot to gain more information. Participants thought the robot should explain why it failed to complete the task, why it disobeyed them during headshake executions without any arm motion, and why it kept moving its arm. When the robot did nothing, people wanted to know about the robot's previous behavior.

2 RELATED WORK

Psychologists have long studied human explanations. As categorized by Malle [37], humans either explain what was unexpected to have a complete and coherent understanding—meaning-finding explanations—or use explanations as communicative acts to create shared meaning and manage social interactions to influence the explainee's mind or behavior—interaction-managing explanations. From an early age, people seek explanations for clarification and further understanding when they are surprised or confused [31, 37, 47]. In contrast to the unsettling experience due to not having any explanation from other people [24], Moerman [40] found that patients feel better when they receive explanations about their illness.

Researchers have also started to enable virtual agents to provide explanations. Ofra, Finale, and David [5] used the Pac-Man platform to summarize and explain Pac-Man's turning behavior through videos. A human-subjects study [4] on this technique showed that participants preferred summaries from the model trained on participant-provided data rather than author-provided data. Indeed, artificial intelligence researchers (e.g., Reference [38]) and the human-computer interaction community (e.g., Reference [1]) are also contributing towards explainable or interpretable systems, including autonomous agents. Due to the embodiment of physical robots, some research in these fields (e.g., References [6, 46]) is less applicable to human-robot interaction. In a literature review about explainable agents and robots [6], approximately half (47%) of the explanation systems examined, used text-based communication methods, which may be less relevant for robots that are not usually equipped with display screens. For some other work in the AI community (e.g., References [13, 39]), however, the main audience has been machine learning experts interpreting trained models or the "black-box" systems [2], although recently there is a shift to non-expert end-users for human-in-loop AI systems [48] and some in automated planning specifically [11].

However, not much is known about what robots should explain and how robots should produce verbal explanations as humans do. De Graaf and Malle [15] concisely discussed the theory behind human behavior explanation developed by psychologists in the past and proposed to apply the

theory behind human behavior explanation to autonomous intelligent systems. But robot embodiment and the difference between human and robot explanations remain largely unclear in the literature. Additionally, Thellman et al. [45] compared people's interpretations of humanoid robot behavior to human behavior in static images with text descriptions and found that conscious goals are the perceived causes of robot behavior, while human behavior seems to be caused by dispositions. Hayes et al. [29] proposed explaining robot controller policy by manually annotating functions during programming, but this method is limited to programmers, and the explanations are constrained to programmers' logic, which may not be helpful for non-expert or non-roboticist explainees. Chakraborti et al. [12] treated explanation as a problem to the suggestion of humans' mental models of a robot [44] to align with the robot's model. The proposed algorithm was claimed to generate explanations with desirable requirements, such as completeness, conciseness, monotonicity, and computability, which were mathematically formulated but not verified with human participants.

Instead of generating explanations directly and assuming certain qualities of explanations, we explore and attempt to better understand desired robot explanations by conducting a user study. Implicit nonverbal cues to express robot intent have also been investigated in the HRI community. Dragan et al. [19] proposed legible motion that adds extra robot arm motion to reveal a robot's reaching intent to a specific object. Kwon et al. [33] used the similar concept of adding legible motion to indicate whether a robot could lift a cup or push a bookshelf. For some conditions, the added legible motion may have been unexpected to participants rather than helpful in aiding their understanding. While our focus has been on motion cues, for a comprehensive review of all non-verbal cues, we refer readers to Reference [10].

Like humans, the robot should be capable of explaining unexpected things. But, as Malle [37] pointed out, knowing why and what generates the behavior of a robot will further improve one's prediction of the resulting behavior, particularly when the cause is not easily apparent, such as in the handover task we designed for an almost-reachable cup. In addition, interaction-managing explanations must be verbally expressed [37], further suggesting implicit nonverbal cues (like motion cues) need to be accompanied by more explicit verbal explanations.

3 HYPOTHESES

Driven by the prior related work, partly on the motion cues, we formalized the following hypotheses:

3.1 The Need for Robot Explanations

Hypothesis 1 — Robot behavior needs to be explained. In general, robot behavior will be considered unexpected to people, and there will be a desire for it to be explained.

Hypothesis 2 — As more causal information about robot behavior is provided, there will be less need for an explanation from the robot. There will be an association between the amount of causal information provided in robot behavior and participants' reported need for explanation. Specifically, as the robot's behavior provides participants with more causal information, participants will report less need for explanation.

Hypothesis 3 — Adding a headshake to the robot's explanation will result in less need for an explanation. Including a negative headshake, which implies that the robot cannot complete the handover task, will give participants more information about the robot's behavior than execution alone. When the robot couples its behavior with a headshake, participants will report less need for explanation than when it does not.

3.2 Expected Properties of Robot Explanations

Hypothesis 4 – Explanations offered at multiple points in time will be desirable. Having robots explain *a priori*, *in situ*, and at the end will be more desirable than at any one single point in time.

Hypothesis 5 – Engagement prior to providing an explanation will be important. People will prefer that the robot get their attention prior to explaining behavior as opposed to explaining behavior without getting their attention.

Hypothesis 6 – Similarity to human explanations will be expected. Participants will report that they expect that robot explanations should be similar to human explanations.

Hypothesis 7 – Summarization will be preferred. Participants will prefer that the robot provides an explanation that is presented as a summary as opposed to a detailed explanation.

Hypothesis 8 – Fewer number of follow-ups will be preferred. Participants will prefer to ask fewer clarifying follow-up questions as opposed to more follow-up questions after an explanation is given from the robot.

4 METHOD

4.1 Power Analysis, Participants, and Participant Recruitment

We used G*Power 3.1.9.4 [20] to perform two *a priori* power analyses, because we planned to run two types of hypothesis tests.

We first performed an *a priori* power analysis for “Goodness-of-fit tests: Contingency tables.” The parameters were: Effect size $w = 0.5$ for large effect size, α error probability = 0.05, Power ($1 - \beta$ error probability) = 0.95, Df = 2, which reflected the number of fixed choices in our measures described in Section 4.3. The output parameters in G*Power showed that the sample size to reach desired power $1 - \beta = 0.95$ was 62 for a single goodness-of-fit test. Thus, for the six conditions of our 3×2 experiment design, we needed at least $6 \times 62 = 372$ participants.

We also performed an *a priori* power analysis for “ANOVA: fixed effects, special, main effects and interactions” tests. The parameters were: Effect size $f = 0.4$ for large effect size, α error probability = 0.05, Power ($1 - \beta$ error probability) = 0.95, Numerator df = 2 (maximum for main effects and interactions), and Number of groups = 6, reflecting the number of independent conditions in our study. The output parameters showed that the total sample size needed was 100.

Thus, our study would need approximately $N = 372$ participants to be sufficiently powered for both types of statistical tests.

Using MTurk, we recruited a total of 460 MTurk workers to participate in the study. We purposefully recruited extra participants to account for potential data loss due to some MTurk workers failing data quality assurance checks and/or not completing the entire study. All of the 460 participants completed the study. However, we had two participants complete the study twice, providing us with 458 unique cases. Seventy-eight participants did not pass the data quality assurance checks in the form of “attention check” questions (described below), which resulted in 380 valid cases used in data analyses. To ensure that we had an equal number of participants in each of the 6 conditions, we trimmed the data from the last participants who entered into the study. This procedure resulted in a sample size of $N = 372$ with 62 participants in each of the 6 between-subjects conditions.

However, after completing a strict replication study (see Section 6) and re-inspecting our data, we found four participants who responded to a question by providing a partial or full copy and paste of the common definition of a robot giving when searching the web for the word “robot” (e.g., entry for robot on Wikipedia). We thus removed these four participants and then tried to

replace their data with data from four participants who had previously been trimmed from the original dataset, as described above. However, we were unable to replace data equally across conditions. Therefore, we trimmed the data again to balance the number of participants in each of the experimental conditions. This process resulted in 61 participants in each condition, $N = 366$. The analyses reported in subsequent sections were conducted on this final dataset.

The final sample included 209 males, 153 females, 3 participants who preferred not to say, and 1 transgender person; with ages ranging from 18–74, $M = 37$, $median = 34$, $skewness = 1.07$. Seventy-three participants (20%) agreed with the statement, “I have experience with robots,” 204 disagreed (56%), and 89 (24%) responded that they neither agreed nor disagreed.

Specified qualifications for participation on MTurk included being over 18 years old, living in the United States, which provided a reasonable assumption of some English language comprehension, having performed at least 1,000 **Human Intelligence Tasks (HITs)**, and a 95% approval rating. Each MTurk worker, whether or not they passed data quality assurance checks, was paid U.S. \$1 for their participation.

4.2 Robot Platform

A Rethink Robotics Baxter humanoid robot (humanlikeness score = 27.30 on a scale of 0 “Not human-like at all” to 100 “Just like a human” [22, 42]) depicting a digital smiling face from References [21, 28] was used in the experiment. Baxter has a large appearance: 1.8 m tall with two 128 cm arms, measured from its shoulder joint to gripper tip. It is taller than an average adult male over age 20 in the U.S. (175 cm) [23], and Baxter’s arm is around two times longer than an average human arm [35].

4.3 Measures

To test our hypotheses, we designed a measure consisting of items about participant perceptions of robot explanations. To create the measure, we searched the existing HRI and robotics literature for scales in the context of robot explanation, but few have been developed and validated by the community. However, we did find a subjective scale of predictability in Reference [18] by Dragan and Srinivasa; because of its high internal reliability (Chronbach’s $\alpha = 0.91$), we adapted two of its questions (Table 1 in Reference [18]) for our Unexpectedness measure (the last two questions in the first row in Table 1):

- (1) “The robot’s behavior matched what I expected” is adapted from “Trajectory ‘x’ matched what I expected.”
- (2) “The robot’s behavior *surprised* me” is adapted from “I would be *surprised* if the robot executed Trajectory ‘x’ in this situation.”

Our measure consisted of four subscales, each with items tailored to gather information about participant perceptions of the unexpectedness of the robot’s behavior, the need for explanations, desired similarities and differences between robot and human explanations, and the desired level of detail of robot explanations. For each subscale, multiple contradicting or very similar questions were designed to help establish internal consistency. All items are listed in Table 1. Participants responded to each item using a 7-point Likert-type scale that ranged from –3 “Strongly Disagree” to 3 “Strongly Agree,” with 0 representing neither agreement nor disagreement at the mid-point of the scale. The order of these items was presented to participants at random.

Additionally, we asked participants to respond to several items about desired properties of robot explanations. These included questions about explanation timing, engagement, and summarization, as follows:

Table 1. Explanation Measure Items

Unexpectedness (Cronbach's $\alpha = 0.80$) 1. <i>I found the robot's behavior confusing.</i> 2. <i>The robot's behavior matched what I expected.</i> (Reversed) 3. <i>The robot's behavior surprised me.</i>
Need (Cronbach's $\alpha = 0.74$) 1. <i>I want the robot to explain its behavior.</i> 2. <i>The robot should not explain anything about its behavior.</i> (Reversed)
Human-robot Difference (Cronbach's $\alpha = 0.49$) 1. <i>There should be no difference between what a robot says to explain its behavior and what a person would say to explain the same behavior.</i> 2. <i>If a person did what the robot did, they should both explain the same behavior in the same way.</i>
Summarization (Cronbach's $\alpha = -0.57$; 0.65 if Q1 is dropped) 1. <i>The robot should give a very detailed explanation.</i> (Reversed) 2. <i>The robot should concisely explain its behavior.</i> 3. <i>The robot should give a summary about its behavior before giving more detail.</i>

Likert items are coded as -3 (Strongly Disagree), -2 (Disagree), -1 (Moderately Disagree), 0 (Neutral), 1 (Moderately Agree), 2 (Agree), and 3 (Strongly Agree).

- *Timing* (forced-choice). When would be the best time for the robot to explain its behavior? Participants responded with "At the end"; "Whenever something unexpected happens"; "Before something unexpected happens"; or "Other (Please elaborate)."
- *Engagement Importance* (true/false). Do you think it is important for the robot to get your attention before starting to explain its behavior?
- *Engagement Approach* (forced-choice). How should the robot get your attention before starting to explain its behavior? Possible responses included, "Look at me"; "Raise volume"; or "Other (Please elaborate)."
- *Summarization* (forced-choice). After the robot gives a summary about its behavior, how many questions would you be willing to ask to get more details from the robot? Response choices included, "None"; "A few (1 to 3)"; "As many as needed (4 or more)."
- *Content* (open-ended). What would you like the robot to explain specifically?
- *Content Reasoning* (open-ended). Please explain why. (Why would you like the robot to explain the things you mentioned above?)

Finally, we asked several attention-check questions to help us ensure participant attention to the experimental stimuli, similar to those used in Brooks et al. [9]. Participants were asked to indicate the color of the robot depicted in the experimental stimuli (Red and black; Blue and yellow; Green and white), the identity of certain objects in the scene: if there were any robots shown in the scene (Yes; No), and if the robot moved (Yes; No). Failure to answer any of those questions correctly resulted in the removal of the participant's data from the analysis. These attention-check questions were randomly intermixed with the items asking about robot explanations and the properties of robot explanations.

4.4 Experimental Design

This study followed a 3×2 between-subjects design. Participants were asked to imagine that they were interacting with the Baxter robot while viewing a video of Baxter executing a handover

Fig. 2. To establish the handover task context, the image above was first presented to participants with the three yellow pop-ups slowly fading in one after another clockwise from the top left.

task. In the handover task, the robot attempted to grasp and hand over a cup under one of the study conditions. The handover task was chosen because robots will often be expected to complete handover tasks, especially as they enter factories and homes. The handover task is one of the most common interactions among humans. Additionally, the handover task involves manipulation, which is characteristic of typical tasks that many robots currently do. We manipulated the robot's handover task *execution type*. The execution type conditions were designed to provide participants with a variety of robot motion cues giving causal information for why the robot was unable to complete the handover task in each video. We also manipulated whether or not the robot shook its head negatively (*Headshake*) while completing each execution type. This resulted in six between-subjects experimental conditions. The code for this study is available on GitHub at <https://github.com/uml-robotics/takeit>.

4.4.1 Study Conditions. For each condition, participants were provided with a video (see Figure 2), where they were informed that the Baxter robot was organizing a desk by placing several small toy robots into a transparent bin. Participants were then told to imagine that they were with Baxter in the scene by imagining that they were busy watching Netflix near Baxter. They were then asked to imagine that they were thirsty and have asked Baxter to pass them a cup.

Also seen in Figure 2, in the video, the three pieces of information described above were presented by pop-up text box annotations: From the top left, each pop-up text box slowly faded in, in clockwise order, and displayed for five seconds. The duration was selected to make sure participants had ample time to finish reading a pop-up before the next was shown. Participants could also pause the video to review the information.

Note that for the pop-up content, only the left-bottom pop-up (ask the robot to pass the cup) was important to establish the handover task. The other two were intended to make the imagined interaction in the scenario more complete and used to remove potential bots or careless responders via attention-check questions.

After viewing these annotations, the participants viewed the following white text on a black background for six seconds: "The robot understands your request. Now please watch how the

Table 2. Study Conditions across the Two Factors

Execution Type: <i>Motion Cues</i>	Headshake
Look & Point: <i>Head turning & arm movement</i>	<i>With Headshake</i>
Look: <i>Head turning</i>	<i>Without Headshake</i>
No Cue: <i>None of above</i>	

robot responds.” Note that we carefully selected the neutral word of “understand” rather than “acknowledge” or “accept” to avoid the impression that the robot would unquestionably finish the handover task. After presenting this information, the video depicted the Baxter dropping the small toy robot it was previously holding and about to place in the transparent bin. Then Baxter attempted to pass the cup under one of the study conditions described below.

Table 2 briefly lists the experimental conditions across the two factors *Execution type* and *Headshake*. The videos shown to participants are available on YouTube at <https://bit.ly/2U6VR0L>. A brief summary of each of the study conditions is provided below.

- *Look & Point without Headshake.* Inspired by the legible arm motion research in References [19, 33], the Look & Point execution conditions were designed to provide the most causal information about why the robot was not able to complete the handover task. The robot provided motion cues to help participants infer causal information for why the robot could not execute the cup handover task. Specifically, in this condition, during the handover task, the robot would *stop organizing, move its arm to reach towards the cup, simultaneously move its head to look at the cup, and keep extending its arm fully toward the cup*. However, it could not reach the cup on the table. The robot repeated this pattern of motion cues towards the cup three times, which was intended to show participants that the robot was trying to complete the task but was unable to do so because the cup was out of reach. Note that this type of motion is not what a robot would commonly do when it cannot complete a grasping or handover task; commonly, the robot would simply stop organizing and do nothing. Typically, its motion planner, e.g., MoveIt [14], would return a plan or execution failure status when the object is not reachable, rather than physically illustrating that an object is out of reach.
- *Look & Point with Headshake.* Under this condition, we added a Headshake to the Look & Point motion execution. In addition to arm and head motion cues described above, after reaching for the cup, the robot shook its head from left to right (i.e., a “No” pattern) to further communicate that it could not reach the cup. The robot repeated the reaching motion while shaking its head two more times before the video ended.
- *Look without Headshake.* In the Look execution condition, during the handover task, the robot would *stop organizing and turn its head towards the cup*. The robot did not reach its arm toward the cup. Roboticists may immediately understand why the robot was unable to complete the handover task: The robot turns its camera mounted on its head, probably an RGBD camera, to detect the cup with depth information, plans to maneuver its arm to the pose of the cup, and eventually failed to do so because the cup was out of reach. However, because roboticists represent a small group of specialists, this execution type is likely opaque to most people, who will be wondering why the robot turned its head toward the cup but did not pick it up.
- *Look with Headshake.* Similar to the Look & Point with Headshake condition, in the Look with Headshake condition, we added a Headshake to the robot’s looking motion cue. Specifically, the robot would *stop organizing, turn its head towards the cup, and shake its head negatively*. The robot did not reach its arm toward the cup.

- *No Cue without Headshake.* Under this condition, during the handover task, the robot would *stop organizing and do nothing*. Unlike the previous two execution conditions, the robot did not provide any motion cues, e.g., turn its head or reach with its arm.
- *No Cue with Headshake.* Again, we added a Headshake to the robot's execution. In the handover task, rather than doing nothing, the robot would *stop organizing and shake its head from left to right*. Note that, unlike the Look with Headshake condition, the robot did not turn its head toward the cup before shaking its head in a "No" pattern.

4.5 Procedure

The study was conducted on MTurk, where participants entered the study via an anonymous link to a Qualtrics survey. Once started, participants were presented with informed consent information. After reviewing this information and agreeing to participate, participants were randomly assigned to one of the experimental conditions. On Qualtrics, participants were provided with the following instruction: "Before answering questions, please watch the following video in full-screen mode." Participants then watched one video depicting the robot executing the handover task under one of the experimental conditions. After viewing the entire video, participants were then asked to complete the measure containing the explanation, properties of explanations, and attention-check items. On the survey page, participants were encouraged to review the video as many times as they needed. Specifically, they were told: "You may go back to the video by clicking the back button at the bottom of this page. If you need to do this, your answers on this page will be saved." Participants were then provided with a code to receive their payment. The entire study took approximately 10 minutes to complete for most participants, who were compensated U.S. \$1 in return for participating in the study. This study was approved by the Institutional Review Board at the University of Massachusetts Lowell.

5 RESULTS

We used R to analyze the data. Table 1 lists all of the items from the robot explanations measure, the anchors for the Likert-type scales, and Cronbach's alpha values of internal consistency for each of the items. M used without standard deviation values denotes median values throughout this section.

5.1 H1: Robot behavior will be considered unexpected and needs to be explained

Cronbach's alpha on the unexpectedness subscale was 0.80, a good level of internal consistency reliability [17]. We thus calculated an unweighted average score of responses across the items to achieve a composite score for the unexpectedness scale on our questionnaire, plotted in Figures 3 and 4.

To analyze the data, we used a two-way between-subjects factorial ANOVA with the unexpectedness score as a dependent variable.

We did not find a statistically significant interaction between *Headshake* and *Execution Type*, but found statistically significant main effects for *Headshake* ($F(1, 360) = 15.91, p < 0.0001$) and *Execution Type* ($F(2, 360) = 11.90, p < 0.0001$) on unexpectedness scores.

Before we conducted pairwise comparisons across conditions, we used the ANOVA model to calculate estimated marginal means of all conditions and performed multiple comparisons with Holm-Bonferroni correction [30] to test whether these means significantly deviate from 0 ($H_0 : \mu = 0$). Without Headshake, we only found a statistically significant difference in No Cue execution type ($1.00 \pm 0.17, p < 0.0001$), but not in Look ($0.20 \pm 0.17, n.s.$) and Look & Point ($0.20 \pm 0.17, n.s.$) conditions. With Headshake, statistically significant differences were found in

Fig. 3. The distribution of Unexpectedness responses with median lines and estimated marginal means. Except for Look without Headshake and Look & Point without Headshake, all conditions were rated unexpected (first two lines of the annotation in each of the boxes, comparison against 0). Results of post hoc pairwise comparisons are also shown (second and third lines of the annotation in each of the boxes).

Fig. 4. Boxplot of Unexpectedness responses. All robot behaviors were rated unexpected except for Look without Headshake (the middle red box) and Look & Point without Headshake (the right red box).

all execution types—No Cue ($1.38 \pm 0.17, p < 0.0001$), Look ($1.16 \pm 0.17, p < 0.0001$) and Look & Point ($0.52 \pm 0.17, p < 0.01$).

The results above suggest that, when the robot performed the Look & Point and Look executions without Headshake, participants felt neutral about the unexpectedness (neither unexpected nor expected), but participants reported that the robot's behavior was significantly more unexpected for the No Cue execution than the other two execution types. Surprisingly, the addition of the Headshake behavior did not decrease the unexpectedness but rather made the robot's behavior more unexpected with strong evidence in the No Cue and Look with Headshake conditions and weak evidence¹ in the Look with Headshake condition (median scores in the top row in Figure 3).

¹This effect becomes non-significant in our replicated study.

Fig. 5. The distribution of Need responses with median lines and estimated marginal means, showing that robot behavior should be explained in all conditions. No statistical significance was found in post hoc pairwise comparisons.

We also performed post hoc pairwise comparisons using Tukey's test with Holm-Bonferroni correction ($H_0 : \mu_i = \mu_j$). Without Headshake, statistically significant differences were found between Look & Point and No Cue ($0.80 \pm 0.24, p < 0.01$) and Look and No Cue ($0.80 \pm 0.24, p < 0.01$). With Headshake, we only found the statistically significant difference between Look & Point and No Cue ($0.85 \pm 0.24, p < 0.01$) conditions.

The pairwise comparisons suggest the No Cue execution was more unexpected than the Look & Point and Look conditions. However, no statistically significant differences between Look & Point and Look execution conditions were found in either Headshake condition.

In summary, **H1**, which states that all robot behavior would be considered unexpected, was partially supported. For the without Headshake conditions, participants reported that the robot's behavior in only the No Cue condition was unexpected but rated it more neutral in the Look and Look & Point conditions. The robot's behavior was rated as unexpected across Headshake conditions, however.

5.2 H1, H2, & H3: Causal information, Headshake, and need for explanations

Similar to the unexpectedness responses, we calculated an average score for responses to the need for explanation items. Cronbach's alpha for this subscale was 0.74. Responses are plotted in Figures 5 and 6. We also used the between-subjects factorial ANOVA to analyze the data. Again, we did not find a statistically significant interaction between *Headshake* and *Execution Type*, but found statistically significant main effects for *Headshake* ($F(1, 360) = 14.99, p < 0.0001$) and *Execution Type* ($F(2, 360) = 3.31, p < 0.05$) for explanation scores.

Again, before we conducted pairwise comparisons across conditions, we used the ANOVA model to calculate estimated marginal means of all conditions and performed multiple comparisons with

Fig. 6. Boxplot of the Need for explanation responses. Participants in all conditions agreed that robot behavior should be explained. No significant differences were found pairwise between conditions.

Holm-Bonferroni correction [30] to test whether these means significantly deviate from 0 ($H_0 : \mu = 0$). Results show statistical significance across all conditions ($p < 0.001$):

- Without headshake:
 - Look & Point: $0.75 \pm 0.18, p < 0.0001$
 - Look: $0.62 \pm 0.18, p < 0.001$
 - No Cue: $1.18 \pm 0.18, p < 0.0001$
- With headshake:
 - Look & Point: $1.43 \pm 0.18, p < 0.0001$
 - Look: $1.25 \pm 0.18, p < 0.0001$
 - No Cue: $1.61 \pm 0.18, p < 0.0001$

This result suggests that people want the robot to explain and that the robot should explain its behavior across all conditions, even when additional arm movement (Look & Point), head turning (Look), and Headshake cues are included.

Pairwise comparisons were also conducted with Tukey's test ($H_0 : \mu_i = \mu_j$), not revealing any statistically significant differences between the execution type conditions. This suggests that people want the robot to explain in all conditions equally.

The analysis supports part of **H1** that robot behavior should be explained, even when robot behavior includes non-verbal motion cues. Together with responses to the Unexpected items, **H1 was partially supported**: When the robot's behavior is deemed neutrally unexpected, the robot should explain its behavior. However, **H2** was not supported. As more causal information was added across the conditions from No Cue to Look & Point, the perceived need for explanation did not drop accordingly. **H3** was also not supported: The pairwise comparisons showed no statistically significant differences between with Headshake and without Headshake groups, suggesting that the addition of the Headshake cue did not result in a less perceived need for an explanation from the robot.

5.3 H4: Explanation timing

We performed a chi-square goodness-of-fit test on the responses to the multiple-choice timing question, which revealed statistical significance ($p < 0.0001$). Post hoc binomial tests with Holm-Bonferroni correction for pairwise comparisons were performed and show significant differences between *in situ* ($p < 0.0001$), *a priori* ($p < 0.001$), and the other response options ($p < 0.0001$), except for explaining at the end ($p = 0.76$). Among the 13 participants who elaborated their choice for "other," four participants expressed that they wanted a *running commentary* from the robot from the beginning of its action, which is different from asking *in situ* questions. All other comments were either isolated or very similar to the other three choices.

Fig. 7. Timing and verbosity preferences. Around half the participants prefer the robot to explain *in situ* and most (75%) participants are willing to ask only a few clarifying follow-up questions.

Fig. 8. Engagement (Look: Look at me, Raise: Raise volume, Other: Other (Please elaborate)). Most participants (69%) agree it is important for the robot to engage with them before explaining. Slightly over one-third of all participants prefer looking at them to get their attention.

Thus, **H4** was not fully supported. As shown in Figure 7 left, approximately half (193/366, 53%) of the participants wanted the robot to explain *in situ* as unexpected things happened, and only 66 (18%) participants wanted explanations from the robot before something unexpected happens. However, the binomial test shows that the choice of subsequent explanations may have happened by chance: Whether more or fewer people wanted explanations after the robot's behavior remains unknown.

5.4 H5: Engagement importance/preference

A chi-square test was run on the engagement importance true/false responses and indicated that the proportion of "false" responses (115, 31%) to the item asking about whether it was important for the robot to get the participant's attention before giving an explanation was significantly lower than "true" responses (251, 69%; $p < 0.0001$), as shown in Figure 8 (left). Thus, **H5** was supported: It is important for the robot to get one's attention before explaining.

Regarding the preference for *how* the robot should get the attention of humans (Figure 8, right), a multinomial goodness-of-fit test was performed and shows statistical significance ($p < 0.0001$). Post hoc binomial tests with Holm-Bonferroni correction for pairwise comparisons were performed and revealed significant differences between all choices ($p < 0.0001$ for all) but not for Raising the volume ($p = 0.62$), which suggests people may have selected this choice at random. In summary, more than one-third of participants (135, 37%) preferred the robot to look at them to get their attention, and only 57 participants (16%) preferred the robot to raise its volume. Interestingly, no participants chose the combined option to "Look, Raise & Other."

Among the 43 participants who elaborated their answers about the "Other" choice, 30 participants wanted the robot to address them by name or title or with words such as "hey" or "excuse me"; 7 participants wanted a beep sound while the other responses were isolated.

5.5 H6: Similarity to human explanation

For the two questions in this scale, Cronbach's alpha reports 0.49, suggesting those two questions represent two independent subscales. Upon further reflection, we realized that the first question

Fig. 9. Robot vs. human explanations in what and how (green line indicates median). Half participants agreed on no differences in both. Please see Section 5.5 for more details.

was asking about the difference in *what* to explain (i.e., “there should be no difference between what a robot says to explain its behavior and what a person would say to explain the same behavior.”) The second question was asking about the difference in *how* to explain (i.e., “if a person did what the robot did, they should both explain the same behavior in the same way.”)

Thus, to test the responses, we ran a chi-square goodness-of-fit test on the responses to both items independently. The chi-square test for the first question indicated a statistical significance ($\chi^2(6) = 108.58, p < 0.0001$). Post hoc binomial tests with Holm-Bonferroni correction show statistically significant differences between Likert responses -3 ($p < 0.0001$), 1 ($p < 0.0001$), 2 ($p < 0.0001$), 3 ($p < 0.0001$) but not in -2 ($p = 0.06$), -1 ($p = 0.46$) and 0 ($p = 0.17$).

We also ran the tests on the responses to the second question. The chi-square test indicated a statistical significance ($\chi^2(6) = 75, p < 0.0001$). Post hoc comparisons show statistically significant differences in Likert responses -3 ($p < 0.0001$), 1 ($p < 0.0001$), and 3 ($p < 0.01$) but not in -2 ($p = 0.46$), -1 ($p = 0.46$), 0 ($p = 0.26$) and 2 ($p < 0.06$).

Thus, **H6** was partially supported. Approximately half of the participants agreed that the robot should explain the behavior by saying the same thing as a human would (53%, Figure 9, left). But only one-third of participants agreed they should explain in the same way (35%, Figure 9, right). More than one-third of participants who indicated that the robot should explain in a different way than a human may have happened by chance (43% and 42%, respectively).

5.6 H7, H8: Detail, summarization, and follow-up questions

Cronbach’s alpha reports -0.57 on the three summarization questions with the first being reversed, but 0.65 (acceptable with a large sample) with the first question dropped, suggesting the first question may be measuring an independent metric—detailed—and the last two questions for another metric—summarized.

5.6.1 Explanation detail. Similar to the human-robot difference question, we ran a chi-square goodness-of-fit test on the responses to the first question, indicating a statistically significant difference ($\chi^2(6) = 118.07, p < 0.0001$) between responses. Post hoc binomial tests with Holm-Bonferroni correction for pairwise comparisons show statistically significant differences in -3 ($p < 0.0001$), 1 ($p < 0.0001$), 3 ($p < 0.0001$) but not in -2 ($p = 0.99$), -1 ($p = 0.99$), 0 ($p = 0.28$), and 2 ($p = 0.99$). Thus, at least 35% agreed that explanations should be very detailed, shown in Figure 10 (left). However, the fact that there were four Likert scale points not being statistically significant may imply that almost half (47%) of participants may not have been certain about the need for detailed explanations.

Fig. 10. Two summarization aspects (green lines indicate median values). *Detailed*: While 35% participants (1 & 3) agreed explanations should be detailed, almost half of the responses (−2, −1, 0, 2) may happen at random. *Summarized*: 72% participants preferred explanations to be concise.

5.6.2 Explanation summary. We calculated an unweighted average score from the responses to the last two questions to achieve a composite score. A chi-square goodness-of-fit test was performed and indicates a statistically significant difference ($\chi^2(6) = 281.06, p < 0.0001$), and Post hoc binomial tests with Holm-Bonferroni correction show statistically significant differences in all ($p < 0.0001$) but not in 0 ($p = 0.16$). In summary, 72% participants preferred explanations to be concise, with only 11% disagreeing, shown in the right of Figure 10.

5.6.3 Level of verbosity. We performed a multinomial goodness-of-fit test on the responses, which reveals statistical significance ($p < 0.0001$). Post hoc binomial tests with Holm-Bonferroni correction for pairwise comparisons were performed and show significant differences in all levels ($p < 0.0001$). As shown in Figure 7 (right), 75% preferred “Only A Few (1 to 3).”

Thus, **H7** and **H8** were supported. Participants preferred that the robot’s provided explanation be a summary as opposed to being detailed, and they preferred fewer follow-up questions as opposed to more.

5.7 Additional analyses: Explanation content

We coded the open-ended comments on what should be explained and grouped them into multiple categories. An independent coder coded a random sample of 10% of the participant data and the experimenter coded all responses. After merging codes with similar meanings, we achieved a Cohen’s κ value of 0.78, considered substantial agreement between raters by Reference [34]. Figure 12 shows those conditions in which at least six participants (approximately 10% of participants in a given condition) endorsed the coded category. Figure 11 shows the data without conditions.

Common to all conditions except for the Look with Headshake condition (31% participants), around half of the participants (39% to 59%) wanted the robot to explain **why it failed** to pass the cup (pink bars in Figure 12): No Cue without Headshake: 59%; No Cue with Headshake: 44%; Look without Headshake: 44%; Look & Point without Headshake: 44%; Look & Point with Headshake 39%.

For the Look with Headshake condition (top middle in Figure 12), people were more specific, with 46% participants asking **why the robot disobeyed** them by shaking its head (purple in Figure 12). This perception of defiance also held for 34% of the participants who were in the No Cue with Headshake condition (top left in Figure 12).

Perceptions of the robot acting defiantly were not shown for the Look & Point condition, because arm movement likely indicated that the robot obeyed, but 36% in the Look & Point without Headshake (22) and 25% in the Look & Point with Headshake (15) conditions were confused about **arm movement intention** (cyan in Figure 12) and only six participants in total (Look & Point

Fig. 11. The most wanted robot explanations (top categories regarding explanation content). The top one is why the robot failed to hand the cup.

Fig. 12. The most wanted robot explanations across conditions (most frequent categories regarding explanation content across conditions). An explanation for why the robot failed to pass the cup is common in all conditions. Please see Section 5.7 for other categories.

with Headshake) explicitly expressed a desire for the robot to say **that it could not reach the cup**.

For No Cue execution conditions, **why it dropped** the small toy robot that Baxter was holding onto the desk (sky-blue in Figure 12) and **why it shook its head** (brown) were mostly questioned in the No Cue without Headshake condition (46%, 28) and the No Cue with Headshake condition (26%, 16), respectively. With an additional cue like head-turning or arm movement but not Headshake, the percentage of participants who wanted to know why the small toy robot is dropped decreased to 18% (11) in the Look without Headshake condition and disappeared in the Look & Point without Headshake condition. With Headshake (No Cue with Headshake,

Fig. 13. Most frequent coded responses for why participants want the robot to explain (i.e., explanation reasoning). Two of them are endorsed by more than 20% of participants: The robot should explain because its behavior does not meet their expectations, and to confirm the robot will do the task, will do it correctly, and whether it is capable of finishing the task.

Look with Headshake, Look & Point with Headshake), no participants wanted to know why the robot dropped the small toy robot Baxter was holding onto the desk. Except for the No Cue with Headshake condition where Headshake was the only motion cue, no participants wanted to know why the robot shook its head.

For general explanations, around 10% participants wanted explanations for the robot's **actions**, mostly current actions, in each condition (No Cue without Headshake: 13%, 8; No Cue with Headshake: 15%, 9; Look without Headshake: 10%, 6; Look with Headshake: 18%, 11; Look & Point with Headshake: 20%, 12), except for the Look & Point without Headshake condition, where explanations for the arm movement intention were more desired. Being related, participants in the Look without Headshake (30%, 18) and Look & Point without Headshake (10%, 6) conditions wanted explanations for the robot's **intended actions**. Last, when there were no motion cues (No Cue without Headshake) or just head turning (Look without Headshake), 10% (6) and 18% (11) wanted **confirmation** of the participant's request from the robot, respectively.

5.8 Additional analyses: Reasoning behind explanation content

Similarly, we also coded participant comments on why the robot should explain. This question was asked immediately after the explanation content question.

Figure 13 shows the top 23 most frequently coded responses to the “Why the robot should explain” open-ended question, that appeared more than six times, i.e., 10% of participants. Figure 14 shows the same data but across conditions.

Fig. 14. Top codes for explanation reasoning across conditions. Please see Section 5.8 for a detailed analysis.

Ninety-eight (26.8%) participants wanted the robot to explain, because the handover failure did **not meet their expectations**. While there were around 20 cases for No Cue conditions (both without Headshake and with Headshake) and the Look condition with Headshake, there are only around 10 for the Look condition without Headshake and both Look & Point conditions.

By comparing the No Cue and Look conditions, we found that participants expected the robot to turn its head towards the cup (Look without Headshake) but without a Headshake, which explains why the count for the Look without Headshake condition is lower. The takeaway here is that **participants set expectations of successful handovers from the robot after their request, and the robot should have explained when it could not meet those expectations.**

By comparing the Look and Look & Point conditions, the additional arm movement does not lead to any change when there is no headshake, but the count reduces by half with a headshake. However, the reason why more participants want the robot to explain itself is to fix the robot or the second composite code, revealing how problems are perceived. The takeaway here is that **when the robot cannot complete the task yet exhibits some unclear behaviors without explanation, participants interpreted them as problems and that the robot needed to be fixed.**²

For other provided reasons, the differences are not large across conditions, usually within 5–10, so we will not discuss them per condition below.

As seen in Figure 13, the second most frequent response by 81 participants (22.1%) was a composite one: *Confirm* | *Correct* | *Capable*, indicating that the robot should explain to confirm it will do the task, will do it correctly, and whether it is capable of finishing the task. Around 51 participants (13.9%) expressed general reasoning: robot explanation helps them better understand the robot. Interestingly, in the fourth composite response: *Fix* | *Troubleshoot* | *Help* | *Get Fixed*, 42 participants (11.5%) expressed interest in solving the problem of the robot, either by themselves or by contacting the manufacturer of the robot. Related to this, we found 21 participants (5.7%) who stated that if the robot explained that the participant was the cause of the problem that they would like to adjust their own actions to help the robot complete its tasks, which are coded as *Human-Correction* | *-Correctness*.

Due to the open-ended nature of the question, all other themes found in the responses were fragmented and limited to less than 10% of participants. Some interesting reasons included understanding the decision-making process of the robot (34 participants, 9.3%) and being able to solve future problems after understanding current problems (21 participants, 5.7%). If these options are given explicitly, e.g., in a forced-choice question, then more participants may choose them.

6 REPLICATING THE STUDY TO VERIFY RESULTS

To validate the robustness of the results, we conducted a strict replication of the study. We launched the study again on MTurk, following the identical procedures reported above. The original study was conducted in August 2019 (“2019 Study”), and the replication was conducted 15 months later in November 2020 (“2020 Study”). Due to the ongoing COVID-19 pandemic, our ability to conduct in-person human subjects studies has been limited as in many other research labs. We thus were unable to replicate the study in a laboratory setting.

We aimed to recruit an identically sized sample as the 2019 Study from MTurk following the same procedure. For the 2020 Study, MTurk workers who participated in the 2019 Study were excluded from participation. While reviewing the responses to the attention-check questions, we also found that 38 participants gave suspect answers to the question asking about explanation content. Specifically, they provided a full or partial copy of the definition of what a robot is and how it works from top web search returns including Wikipedia. We decided to exclude these participants’ responses, as the answer was non-responsive and suggested either the person was not paying close

²Due to the open-ended nature, we did not find the same evidence in the 2020 Study.

Fig. 15. Top categories regarding explanation content (replication results from the 2020 Study). The top one remains unchanged.

attention or a bot may have been filling out the survey. Given this change to our attention-check questions, we went back to the data from the 2019 Study, as we described before its analysis.

After employing the same screening procedures as the 2019 Study, including removing participants who gave suspect answers noted above, we trimmed the dataset to provide an equal number of participants in each of the six experimental conditions and to match the size of the 2019 Study sample. Due to this process, we again had data from $N = 366$ participants, which is included in subsequent analyses.

The demographics of the final sample for the 2020 Study are very similar to the 2019 Study, with 206 males, 157 females, 1 participant who preferred not to say, and 2 transgender people; with ages ranging from 20–69, $M = 36$, $median = 33$, $skewness = 1.05$. 101 participants (28%) agreed with the statement, “I have experience with robots,” 191 disagreed (52%), and 74 (20%) responded that they neither agreed nor disagreed.

After following identical analysis procedures as with the 2019 Study data, we found that the 2020 Study replicated all of the 2019 Study findings except for the Unexpectedness score of the Look & Point with Headshake condition. As seen in the top right subfigure in both Figures 22 and Figure 23, we no longer found any significant difference (was $p < 0.01$ for the 2019 Study) between the mean estimate of responses and 0 (used to code neutral responses).

We placed all of the figures for the 2020 Study and the 2019 Study side-by-side in the Appendix for easy comparison and to document the few minor statistical significance changes in findings regarding Hypotheses H1–H8 between the two studies.

6.1 Additional analyses: Explanation content

As shown in Figures 15 and 16, we were able to draw roughly the same conclusions for explanation content as in the 2019 Study (Section 5.7).

The top explanation that the participants explicitly wanted the robot to provide remained **why it failed** to pass the cup (40%). The same as in the 2019 Study, and this finding was common to all conditions, with a maximum difference of nine participants (No Cue with Headshake vs. Look & Point with Headshake): No Cue without Headshake: 45%; No Cue with Headshake: 47%; Look without Headshake: 33%; Look with Headshake: 42%; Look & Point without Headshake: 48%; Look & Point with Headshake, 33%.

Regarding the explanation for **why the robot disobeyed** by shaking its head, similar to the 2019 Study, it only appeared for the No Cue with Headshake and Look with Headshake conditions.

Fig. 16. Most frequent categories regarding explanation content across conditions (replication results from the 2020 Study). Still, the explanation for why the robot failed to pass the cup is common in all conditions. Please see Section 6.1 for other categories.

With the data about compliance above, the conclusion of our previous analysis still holds: “Perceptions of the robot acting defiantly were not shown for the Look & Point condition, because arm movement likely indicated that the robot obeyed.” There were still participants who explicitly expressed that they would like explanations about **the arm movement intention**, although the number dropped from 22 and 15 participants to 12 and 8, respectively.

For No Cue without Headshake condition, **why it dropped** the small toy robot that Baxter was holding onto the desk remained to be mostly questioned (19, 32%). While **Why shook head** is still the most questioned for No Cue with Headshake condition, its number has dropped from 16 to 10 participants. One new finding is that in the Look with Headshake condition, 11 participants (18%) were confused about why the robot shook its head. The conclusions given in the second-to-last paragraph of Section 5.7 still largely hold.

For general explanations, there were around 10% of participants who wanted explanations for the robot’s **actions**. These participants were roughly evenly distributed in most conditions, but not in the No Cue with Headshake and Look with Headshake conditions (Figure 16). This differs from the 2019 Study in which participants did not want an explanation for the robot’s actions in only the Look & Point without Headshake condition (Figure 12).

For **confirmation**, there were still approximately the same number of participants who wanted a confirmation of their request from the robot: no motion cues (No Cue without Headshake, 12 participants in the 2020 Study vs. 6 in the 2019 Study) or just head turning (Look without Headshake, 14 in the 2020 Study vs. 11 in the 2019 Study).

6.2 Additional analyses: Reasoning behind explanation content

As shown in Figure 17, that the robot did **not meet their expectations** remains to be the top reason behind what to explain, which increased from 98 participants (26.8%) to 113 (30.9%). Across

Fig. 17. Most frequent coded responses for explanation reasoning (replication results from the 2020 Study). The top two remain unchanged and are still endorsed by more than 20% participants. For the changes from the 2019 Study, please see Section 6.2.

conditions (Figure 18), there were still only around 15 cases in the Look without Headshake and more than 20 cases in the No Cue and Look with Headshake conditions. The conclusions given in the fourth paragraph in Section 5.8 still hold.

The composite code “Confirm | Correct | Capable” was still the second-most frequent reason, with roughly the same number of participants in the 2019 Study. Forty-seven vs. 27 participants expressed that they wanted to get confirmation from the robot out of an explanation. The codes “Better Understand” and “Fix | Troubleshoot | Help | Get Fixed” dropped 5.4% from 40 participants to 20. Due to the open-ended nature, we did not find the same conclusion earlier about getting the robot fixed. The frequency values for other codes from the 2019 Study are within around 10 participants each.

7 DISCUSSION

Although the data only partially supported H1: The Look without Headshake and the Look & Point without Headshake conditions were considered to be unexpected by participants; we found that outside of H1 on the unexpectedness metric, adding additional non-verbal cues to increase causal information was not always helpful for meeting participant expectations. Comparing vertically in Figure 22 for the 2019 Study and Figure 23 for the 2020 Study, the addition of the Headshake without any verbal explanation made the robot’s behavior unexpected (Look with Headshake vs. Look without Headshake in the 2019 Study and the 2020 Study), equally unexpected (No Cue without Headshake vs. No Cue with Headshake) or more unexpected only in the 2019 Study (Look & Point without Headshake vs. Look & Point with Headshake).

Fig. 18. Top codes for explanation reasoning across conditions (replication results from the 2020 Study). Please see Section 6.2 for a detailed analysis.

Although the addition of head-turning in the Look condition did decrease the unexpectedness ratings over the No Cue condition in the 2019 Study but not in the 2020 Study (compare the left two columns of Figure 22), the inclusion of arm motion in the Look & Point condition failed to

make the robot's behavior less unexpected than the Look condition—the unexpectedness either increased in the 2019 Study (compare the right two columns of Figure 22) or was not significantly significant in the 2020 Study (compare the right two columns of Figure 23).

In contrast to H2, we found that in both the 2019 Study and the 2020 Study, increasing causal information by adding more non-verbal motion cues did not increase the need for robot explanations. Rather, the need was always desired regardless of what non-verbal cues were being used (partially confirms H1). Together with the explanation content responses, more non-verbal cues such as head shaking or arm movement did not seem to decrease the need, but simply led participants to seek additional specific explanations for those cues.

The assumption that the addition of a headshake would lead to less need for explanation, as hypothesized in H3, did not hold. In the comments on explanation content, participants reported being confused about the robot's headshaking behavior: Instead of perceiving the robot as unable to take the cup, the headshake was considered as a sign that the robot was disobeying many of the participants (the most frequent category in the 2019 Study and second-most in the 2020 Study), when additional arm movement was not present. This confusion may explain why H3 was not supported. Even though we limited our recruitment on MTurk to people living in the United States, cultural difference, especially concerning the meaning of a headshake, may be another factor affecting the perception.

Confusion also existed for the Look & Point execution conditions, with some participants not understanding the intention of the robot's arm movement. This suggests that non-verbal behaviors such as head-shaking and artificial yet unpredictable human non-verbal behaviors such as repeated reaching arm movement without explanation may fall short of being clear methods for trying to provide *in situ* motion-based implicit explanations.

For the properties of robot explicit explanations, participants wanted robots to explain as unexpected things happened. However, we saw explicit requests for a running commentary of the robot's behavior by 4 participants in the 2019 Study and 3 participants in the 2020 Study. Similarly, while participants preferred the robot to get their attention by looking at them, 30 participants in the 2019 Study and 19 in the 2020 Study explicitly opted for the robot to address them specifically.

Surprisingly, results showed that there was no desired difference between robot and human explanations, rejecting H6. However, most non-agreement responses to the difference between robot and human explanation items were not of statistical significance (Figure 46 for the 2019 Study and Figure 47 for the 2020 Study), nor were most participants' responses to the detailed aspect in summarization (Figure 49 for the 2019 Study and Figure 50 for the 2020 Study). Also, the distributions shown in Figure 46 for the 2019 Study and Figure 46 for the 2020 Study are very similar, suggesting that people may not differentiate what and how robots explain. Additional investigation, with multiple questions to improve reliability, would be needed to lend more clarity to these findings. In addition to the arm and head movement, (i.e., non-verbal motion cues), future work should also include exploration to learn if eye gaze or facial expression would be unexpected and how they would affect perceptions of robot explanation. Table 3 show a summary of hypothesis support.

8 LIMITATIONS AND FUTURE WORK

In the experiment, we focused on robot explanations for failures that occur shortly after a person's request, using the example of handover interaction. This timing is important, because early failures prevent the whole process from happening and early failures have been shown to decrease a person's trust in a robot system [16]. While seemingly simple, the handover interaction was chosen, as it is expected to be a frequent interaction among humans working in physical proximity to robots. However, due to the timing of our work and the global pandemic, no in-person or physical interaction was involved in this study (i.e., transferring the object to participants). Further

Table 3. Summary of Evidence or Lack Thereof for Hypotheses
(Includes Data from Both the 2019 Study and the 2020 Study)

Hypothesis	Result
H1. Robot behaviors need to be explained. In general, robot behavior will be considered unexpected to people, and there will be a desire for it to be explained.	<i>Partially supported.</i> Results show that the robot should explain even when participants rank its behavior as neither expected nor unexpected (i.e., neutral) in the Look without Headshake and Look & Point without Headshake conditions.
H2. As more causal information about robot behavior is provided, there will be less need for an explanation from the robot.	<i>Not supported.</i> The perceived need for explanation did not drop with more causal information.
H3. Adding a headshake to the robot's explanation will result in less need for an explanation.	<i>Not supported.</i> The addition of the Headshake cue did not result in less perceived need for explanation.
H4. Explanations offered at multiple points in time will be desirable.	<i>Not fully supported.</i> More participants (53%) preferred <i>in situ</i> explanations, fewer (18%) preferred <i>a priori</i> , yet no statistical difference was found for explanations in the end.
H5. Engagement prior to providing an explanation will be important.	<i>Supported.</i>
H6. Similarity to human explanations will be expected.	<i>Partially supported.</i> Half of the participants agreed that robots and humans should say the same thing, but only one-third of participants agreed they should explain in the same way.
H7. Summarization will be preferred.	<i>Supported.</i>
H8. Fewer number of follow-ups will be preferred.	<i>Supported.</i>

investigation is needed for physical proximity and contact between humans and robots, and for different contexts and tasks other than handovers, likely using in-person studies, although this work has demonstrated that consistent results can be attained through online studies.

In addition, while we aimed to have our findings generalize to the general public, preferences are inherently subjective, influenced by culture, and individual differences exist. Thus, adaptive explanation to individuals is of interest and needs more investigation but is out of the scope of this work. As we have only used the Baxter robot with a slightly smiling face and moderate human-likeness, future work could further investigate how different robot designs impact participant expectations for robot explanation.

Knowing people's preferences for robot explanations paves the way to explanation generation. In our parallel work [26], we proposed such algorithms to generate shallow explanations for only a few follow-up questions, informed by the summarization and verbosity preferences. Currently, we are investigating the effects of verbal explanation mixed or not mixed with non-verbal projection mapping [27] cues, informed by the engagement and timing preferences, i.e., addressing humans *in situ*.

9 CONCLUSIONS

We have investigated desired robot explanations when coupled with non-verbal motion cues. Results suggest that robot explanations are needed even when non-verbal cues are present, in most cases where the robot is unable to perform a task. The robot needs to get the attention of the person and then concisely explain why it failed and its other behaviors *in situ*, in a similar way to what humans would expect other humans to do. We are able to make the same conclusion in a strict replication study after 15 months, providing stronger evidence for our findings.

APPENDICES

A INTERNAL CONSISTENCY (2019 STUDY VS. 2020 STUDY)

Table 4. Explanation Measure Items

Unexpectedness (Cronbach's $\alpha = 0.80$ vs. 0.70)

1. *I found the robot's behavior confusing.*
 2. *The robot's behavior matched what I expected.* (Reversed)
 3. *The robot's behavior surprised me.*
-

Need (Cronbach's $\alpha = 0.74$ vs. 0.65)

1. *I want the robot to explain its behavior.*
 2. *The robot should not explain anything about its behavior.* (Reversed)
-

Human-robot Difference (Cronbach's $\alpha = 0.49$ vs. 0.60)

1. *There should be no difference between what a robot says to explain its behavior and what a person would say to explain the same behavior.*
 2. *If a person did what the robot did, they should both explain the same behavior in the same way.*
-

Summarization (Cronbach's $\alpha = -0.57$ vs. -0.77 ; 0.65 vs. 0.56 if Q1 is dropped)

1. *The robot should give a very detailed explanation.* (Reversed)
 2. *The robot should concisely explain its behavior.*
 3. *The robot should give a summary about its behavior before giving more detail.*
-

Likert items are coded as -3 (Strongly Disagree), -2 (Disagree), -1 (Moderately Disagree), 0 (Neutral), 1 (Moderately Agree), 2 (Agree), and 3 (Strongly Agree).

Cronbach's α for Need dropped from 0.74 to 0.65. As shown in Table 4. The reason is possibly that the second Need question is phrased as a double negative and difficult to understand. To answer, it becomes "I (strongly/moderately) disagree that the robot should not explain anything about its behavior." We plotted the responses to both questions and found that they do not belong to the same distribution. Specifically, responses to the second question are not normally distributed. We thus analyzed the first Need question and present the results in Appendix C.

B UNEXPECTEDNESS

B.1 Unexpectedness (interaction effect): No change

In the 2019 Study, no interaction effect was suggested by a two-way between-subjects factorial ANOVA. In the replication (2020 Study), the same ANOVA test suggested there might be an interaction effect ($p < 0.001$) but post hoc pairwise comparisons using Tukey's test with Holm-Bonferroni correction did not find a statistical significance between No Cue conditions (leftmost column in Figure 23).

Fig. 19. Interaction plot of the unexpectedness responses (original result from the 2019 Study).

Fig. 20. Interaction plot of the unexpectedness responses (replication result from the 2020 Study).

Fig. 21. Interaction plot of the unexpectedness responses (combined result).

In terms of main effects in the 2019 Study, a two-way between-subjects factorial ANOVA revealed 2 statistically significant main effects in the original experiment for both *Headshake* ($F(1, 360) = 15.30, p < 0.001$) and *Execution Type* ($F(2, 360) = 11.98, p < 0.0001$).

In the 2020 Study, main effects were found again for *Headshake* ($F(2, 360) = 7.07, p < 0.001$) and *Execution Type* ($F(2, 360) = 11.90, p < 0.0001$).

B.2 Unexpectedness (bar chart)

Comparing Figure 22 and Figure 23, Look & Point with Headshake (top right) changed for $H_0 : \mu = 0$, meaning participants neither disagreed nor agreed that the Look & Point with Headshake is unexpected. This statistically significant change led to pairwise comparison changes for the condition ($H_0 : \mu_i = \mu_j$), in addition to pairwise comparison changes for the conditions in the left two columns.

Fig. 22. The distribution of Unexpectedness responses with median lines (original result from the 2019 Study).

Fig. 23. The distribution of Unexpectedness responses with median lines (replication result from the 2020 Study).

Fig. 24. The distribution of Unexpectedness responses with median lines (combined result).

B.3 Unexpectedness (box plot)

Fig. 25. Boxplot of Unexpectedness responses (original result from the 2019 Study).

Fig. 26. Boxplot of Unexpectedness responses (replication result from the 2020 Study).

Fig. 27. Boxplot of Unexpectedness responses (combined result).

C NEED (QUESTION 1): SAME CONCLUSIONS

The conclusion about the need to explain still holds. To analyze the results, we used the same significance tests as the ones in Section 5.2.

Fig. 28. The distribution of Need (questions 1) responses with median lines (original result from the 2019 Study).

C.1 Results for 2019 Study (Figure 28)

A between-subjects factorial ANOVA did not show a statistically significant interaction between *Headshake* and *Execution Type*, but found a statistically significant main effects for *Headshake* ($F(1, 360) = 11.10, p < 0.001$) but not *Execution Type* for explanation scores.

Before we conducted pairwise comparisons across conditions, we used the ANOVA model to calculate estimated marginal means of all conditions and performed multiple comparisons with Holm-Bonferroni correction [30] to test whether these means significantly deviate from 0 ($H_0 : \mu = 0$). Results showed statistical significance across all conditions:

Fig. 29. The distribution of Need (questions 1) responses with median lines (replication result from the 2020 Study).

Fig. 30. The distribution of Need (questions 1) responses with median lines (combined result).

- Without headshake:
 - Look & Point: $0.54 \pm 0.20, p < 0.0001$
 - Look: $0.41 \pm 0.20, p < 0.001$
 - No Cue: $0.89 \pm 0.20, p < 0.0001$

- With headshake:
 - Look & Point: $1.13 \pm 0.20, p < 0.0001$
 - Look: $1.10 \pm 0.20, p < 0.0001$
 - No Cue: $1.21 \pm 0.20, p < 0.0001$

Pairwise comparisons were conducted with Tukey's test ($H_0 : \mu_i = \mu_j$), not revealing any statistically significant differences between the Execution Type conditions.

C.2 Results for 2020 Study (Figure 29)

A between-subjects factorial ANOVA did not show a statistically significant interaction between *Headshake* and *Execution Type*, but found statistically significant main effects for *Headshake* ($F(1, 360) = 4.96, p < 0.05$) and *Execution Type* ($F(2, 360) = 4.01, p < 0.05$) for explanation scores.

Before we conducted pairwise comparisons across conditions, we used the ANOVA model to calculate estimated marginal means of all conditions and performed multiple comparisons with Holm-Bonferroni correction [30] to test whether these means significantly deviate from 0 ($H_0 : \mu = 0$). Results showed statistical significance across all conditions ($p < 0.0001$):

- Without headshake:
 - Look & Point: $0.87 \pm 0.19, p < 0.0001$
 - Look: $0.93 \pm 0.19, p < 0.001$
 - No Cue: $1.07 \pm 0.19, p < 0.0001$
- With headshake:
 - Look & Point: $0.77 \pm 0.19, p < 0.0001$
 - Look: $1.67 \pm 0.19, p < 0.0001$
 - No Cue: $1.46 \pm 0.19, p < 0.0001$

Pairwise comparisons were conducted with Tukey's test ($H_0 : \mu_i = \mu_j$), revealing only a $p < 0.05$ statistically significant difference between Look with Headshake and Look & Point with Headshake.

C.3 Results for both experiments combined (Figure 30)

A between-subjects factorial ANOVA did not show a statistically significant interaction between *Headshake* and *Execution Type*, but found statistically significant main effects of *Headshake* ($F(1, 360) = 16.10, p < 0.0001$) but not *Execution Type* for explanation scores.

Before we conducted pairwise comparisons across conditions, we used the ANOVA model to calculate estimated marginal means of all conditions and performed multiple comparisons with Holm-Bonferroni correction [30] to test whether these means significantly deviate from 0 ($H_0 : \mu = 0$). Results showed statistical significance across all conditions ($p < 0.001$):

- Without headshake:
 - Look & Point: $0.75 \pm 0.13, p < 0.0001$
 - Look: $0.68 \pm 0.14, p < 0.001$
 - No Cue: $1.00 \pm 0.14, p < 0.0001$
- With headshake:
 - Look & Point: $0.96 \pm 0.13, p < 0.0001$
 - Look: $1.37 \pm 0.13, p < 0.0001$
 - No Cue: $1.35 \pm 0.13, p < 0.0001$

Pairwise comparisons were conducted with Tukey's test ($H_0 : \mu_i = \mu_j$), not revealing any statistically significant differences between the Execution Type conditions.

D NEED: STATISTICAL SIGNIFICANCE REMAINS THE SAME

D.1 Need (interaction effect): No change, still no interaction effect

No interaction effect was found in neither the original experiment nor the replication.

For main effects, there were no changes in terms of whether statistical significance was found and just changes significant p values. The main effect of *Headshake* was changed from $F(1, 360) = 14.99, p < 0.0001$ to $F(1, 360) = 5.32, p < 0.05$. The main effect of *Execution Type* was changed from $F(1, 360) = 3.31, p < 0.05$ to $F(1, 360) = 2.77, p < 0.05$.

Fig. 31. Interaction plot of the unexpectedness responses (original result from the 2019 Study).

Fig. 32. Interaction plot of the Need responses (replication result from the 2020 Study).

Fig. 33. Interaction plot of the Need responses (combined result).

D.2 Need (bar chart): Same Conclusions

Fig. 34. The distribution of Need responses with median lines (original result from the 2019 Study).

Fig. 35. The distribution of Need responses with median lines (replication result from the 2020 Study).

Fig. 36. The distribution of Need responses with median lines (combined result).

There were no changes for $H_0 : \mu = 0$, meaning participants still agreed that the robot should explain across all conditions, we found statistically significant differences between two pairs in pairwise comparisons—still remaining to agree. They were between Look with Headshake and Look & Point with Headshake ($p < 0.01$) and Look with Headshake and Look without Headshake ($p < 0.05$).

D.3 Need (box plot)

Fig. 37. Boxplot of the Need for explanation responses (original result from the 2019 Study).

Fig. 38. Boxplot of the Need for explanation responses (replication result from the 2020 Study).

Fig. 39. Boxplot of the Need for explanation responses (combined result).

E EXPECTED PROPERTIES

E.1 Explanation timing and verbosity preferences: Conclusions remain the same

Timing: “*a priori*” changed from *** to n.s. This has not changed our conclusion that participants wanted robots to explain *in situ*, not at the end of the task. Verbosity: results remained the same.

Fig. 40. Timing and verbosity preferences (original result from the 2019 Study).

Fig. 41. Timing and verbosity preferences (replication result from the 2020 Study).

Fig. 42. Timing and verbosity preferences (combined result).

E.2 Engagement importance/preference: Conclusions remain the same

Importance: no change in terms of statistical significance and significance levels. Preference: the conclusion remains the same but fewer participants chose “Other.”

Fig. 43. Engagement (Look: Look at me, Raise: Raise volume, Other: Other (Please elaborate)). [original result from the 2019 Study].

Fig. 44. Engagement (Look: Look at me, Raise: Raise volume, Other: Other (Please elaborate)). [replication result from the 2020 Study].

Fig. 45. Engagement (Look: Look at me, Raise: Raise volume, Other: Other (Please elaborate)). [combined result].

E.3 Similarity to human explanation: Largely remained the same; the same conclusion

What: Scale point 2 moved from **** to ***, 3 from **** to *. How: scale point -1 moved from n.s. to *, 2 from n.s. to ****, 3 from ** to n.s.

Fig. 46. Robot vs. human explanations in what and how (green line indicates median). [original result from the 2019 Study].

Fig. 47. Robot vs. human explanations in what and how (green line indicates median). [replication result from the 2020 Study].

Fig. 48. Robot vs. human explanations in what and how (green line indicates median). [combined result].

E.4 Detail, summarization: Same conclusions

Detailed: scale point -2 moved from n.s. to **, 2 from n.s. to *, 3 from **** to n.s. Summarized: scale point moved 3 from **** to **.

Fig. 49. Two summarization aspects (green lines indicate median values). [original result from the 2019 Study].

Fig. 50. Two summarization aspects (green lines indicate median values). [replication result from the 2020 Study].

Fig. 51. Two summarization aspects (green lines indicate median values). [combined result].

REFERENCES

- [1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 582:1–582:18.
- [2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.
- [3] Henny Admoni, Thomas Weng, Bradley Hayes, and Brian Scassellati. 2016. Robot nonverbal behavior improves task performance in difficult collaborations. In *Proceedings of the 11th ACM/IEEE International Conference on Human-robot Interaction*. IEEE, 51–58.
- [4] Dan Amir and Ofra Amir. 2018. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi-agent Systems (AAMAS'18)*. 1168–1176.
- [5] Ofra Amir, Finale Doshi-Velez, and David Sarne. 2018. Agent strategy summarization. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi-agent Systems (AAAMS'18)*. International Foundation for Autonomous Agents and Multiagent Systems, 1203–1207.
- [6] Sule Anjomshoa, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi-agent Systems (AAMAS'19)*. 1078–1088.
- [7] W. A. Bainbridge, J. Hart, E. S. Kim, and B. Scassellati. 2008. The effect of presence on human-robot interaction. In *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'08)*. IEEE, 701–706.
- [8] Gisela Böhm and Hans-Rüdiger Pfister. 2015. How people explain their own and others' behavior: A theory of lay causal explanations. *Front. Psychol.* 6 (2015), 139.
- [9] Daniel J. Brooks, Momotaz Begum, and Holly A. Yanco. 2016. Analysis of reactions towards failures and recovery strategies for autonomous robots. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'16)*. IEEE, 487–492.
- [10] Elizabeth Cha, Yunkyung Kim, Terrence Fong, Maja J. Mataric, et al. 2018. A survey of nonverbal signaling methods for non-humanoid robots. *Found. Trends® Robot.* 6, 4 (2018), 211–323. Retrieved from https://www.lizcha.com/publications/ft_2018.pdf.

- [11] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. 2020. The emerging landscape of explainable automated planning & decision making. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI'20)*. 4803–4811.
- [12] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*. 156–163.
- [13] Angelos Chatzimpampas, Rafael M. Martins, Ilir Jusufi, and Andreas Kerren. 2020. A survey of surveys on the use of visualization for interpreting machine learning models. *Inf. Visual.* 19, 3 (2020), 207–233.
- [14] Sachin Chitta, Ioan Sucan, and Steve Cousins. 2012. MoveIt! *IEEE Robot. Autom. Mag.* 19, 1 (2012), 18–19. Retrieved from <https://moveit.ros.org>.
- [15] M. M. de Graaf and Bertram F. Malle. 2017. How people explain action (and autonomous intelligent systems should too). In *Proceedings of the AAAI Fall Symposium on Artificial Intelligence for Human-robot Interaction*. AAAI, 19–26.
- [16] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction*. IEEE, 251–258.
- [17] Robert F. DeVellis. 2016. *Scale Development: Theory and Applications*. Vol. 26. SAGE.
- [18] Anca Dragan and Siddhartha Srinivasa. 2014. Familiarization to robot motion. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*. ACM, 366–373.
- [19] Anca D. Dragan, Kenton C. T. Lee, and Siddhartha S. Srinivasa. 2013. Legibility and predictability of robot motion. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction*. IEEE, 301–308.
- [20] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Meth.* 39, 2 (2007), 175–191. Retrieved from <http://www.gpower.hhu.de/>.
- [21] Naomi T. Fitter and Katherine J. Kuchenbecker. 2016. Designing and assessing expressive open-source faces for the Baxter robot. In *Proceedings of the International Conference on Social Robotics*. 340–350. Retrieved from <https://github.com/nfitter/BaxterFaces>.
- [22] Cliff Fitzgerald. 2013. Developing Baxter. In *Proceedings of the IEEE Conference on Technologies for Practical Robot Applications (TePRA'13)*. IEEE, 1–6.
- [23] Cheryl D. Fryar, Deanna Kruszan-Moran, Qiuping Gu, and Cynthia L. Ogden. 2018. Mean body weight, weight, waist circumference, and body mass index among adults: United States, 1999–2000 through 2015–2016. *Nat. Health Stat. Rep.* 122 (2018), 1–16.
- [24] Alison Gopnik. 2000. Explanation as orgasm and the drive for causal knowledge: The function, evolution, and phenomenology of the theory formation system. In *Explanation and Cognition*, Frank C. Keil and Robert Andrew Wilson (Eds.). The MIT Press, 299–323.
- [25] Alison Gopnik and Laura Elizabeth Schulz. 2007. *Causal Learning: Psychology, Philosophy, and Computation*. Oxford University Press.
- [26] Zhao Han, Daniel Giger, Jordan Allspaw, Michael S. Lee, Henny Admoni, and Holly A. Yanco. 2021. Building the foundation of robot explanation generation using behavior trees. *ACM Trans. Hum.-robot Interact.* 10, 3 (2021).
- [27] Zhao Han, Alexander Wilkinson, Jenna Parrillo, Jordan Allspaw, and Holly A. Yanco. 2020. Projection mapping implementation: Enabling direct externalization of perception results and action intent to improve robot explainability. In *Proceedings of the AAAI Fall Symposium on the Artificial Intelligence for Human-robot Interaction (AI-HRI'20)*.
- [28] Zhao Han and Holly Yanco. 2019. The effects of proactive release behaviors during human-robot handovers. In *Proceedings of the 14th ACM/IEEE International Conference on Human-robot Interaction (HRI'19)*. IEEE, 440–448.
- [29] Bradley Hayes and Julie A. Shah. 2017. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*. ACM, 303–312.
- [30] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* 6, 2 (1979), 65–70.
- [31] Keith James Holyoak and Robert G. Morrison. 2012. *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press.
- [32] Barbara Koslowski. 1996. *Theory and Evidence: The Development of Scientific Reasoning*. The MIT Press.
- [33] Minae Kwon, Sandy H. Huang, and Anca D. Dragan. 2018. Expressing robot incapability. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*. ACM, 87–95.
- [34] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174.
- [35] Ask Media Group LLC. 2019. How Long Is the Average Human Arm? Retrieved from <https://www.reference.com/science/long-average-human-arm-62c7536c5e56f385>.
- [36] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in Cognitive Sciences* 10, 10 (2006), 464–470.

- [37] Bertram F. Malle. 2006. *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. The MIT Press.
- [38] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (2019), 1–38.
- [39] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 279–288.
- [40] Daniel E. Moerman. 2002. *Meaning, Medicine, and the “Placebo Effect.”* Vol. 28. Cambridge University Press, Cambridge, MA.
- [41] AJung Moon, Daniel M. Troniak, Brian Gleeson, Matthew K. X. J. Pan, Minhua Zheng, Benjamin A. Blumer, Karon MacLean, and Elizabeth A. Croft. 2014. Meet me where I’m gazing: How shared attention gaze affects human-robot handover timing. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*. ACM, 334–341.
- [42] Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F. Malle. 2018. What is human-like? Decomposing robots’ human-like appearance using the Anthropomorphic roBOT (ABOT) database. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*. 105–113.
- [43] Stela H. Seo, Denise Geiskkovitch, Masayuki Nakane, Corey King, and James E. Young. 2015. Poor thing! Would you feel sorry for a simulated robot?: A comparison of empathy toward a physical and a simulated robot. In *Proceedings of the 10th ACM/IEEE International Conference on Human-robot Interaction*. ACM, 125–132.
- [44] Aaquib Tabrez, Matthew B. Luebbbers, and Bradley Hayes. 2020. A survey of mental modeling techniques in human-robot teaming. *Curr. Robot. Rep.* 1, 4 (2020), 1–9.
- [45] Sam Thellman, Annika Silvervarg, and Tom Ziemke. 2017. Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Front. Psychol.* 8 (2017), 1662.
- [46] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2020. Explainable agents through social cues: A review. *arXiv preprint arXiv:2003.05251* (2020).
- [47] Henry M. Wellman, Anne K. Hickling, and Carolyn A. Schult. 1997. Young children’s psychological, physical, and biological explanations. *New Direct. Child Adoles. Dev.* 1997, 75 (1997), 7–26.
- [48] Fabio Massimo Zanzotto. 2019. Human-in-the-loop artificial intelligence. *J. Artif. Intell. Res.* 64 (2019), 243–252.

Received July 2020; revised January 2021; accepted April 2021