# A Wizard-of-Oz Study of Curiosity in Human-Robot Interaction*

Edith Law, Vicky Cai, Qi Feng Liu, Sajin Sasy, Joslin Goh, Alex Blidaru and Dana Kulić[1]

***Abstract*— Service robots are becoming a widespread tool for assisting humans in scientific, industrial and even domestic settings. Yet, our understanding of how to motivate and sustain interactions between human users and robots remains limited. In this work, we conducted a study to investigate how surprising robot behaviour evokes curiosity and influences trust and engagement in the context of participants interacting with Recyclo, a service robot for providing recycling recommendations. In a Wizard-of-Oz experiment, 36 participants were asked to interact with Recyclo to recognize and sort a variety of objects, and were given object recognition responses that were either unsurprising or surprising. Results show that surprise gave rise to information seeking behavior indicative of curiosity, while having a positive influence on engagement and negative influence on trust.**

## I. Introduction

Service robotics is a nascent technology for assisting with tasks in our everyday lives, from guiding users through museums [16], airports [38], shops [28] and offices [4], to assisting the blind [2] and health care workers [25]. Beyond performing tasks, service robots can also serve *social* functions, such as providing companionship [7] and encouraging users to adopt healthy habits (e.g., exercise) [9]. A major challenge in designing service robots is to make them *engaging*, such that human users are motivated to initiate and sustain interactions with the robot over an extended period of time.

Prior work has explored a variety of factors that influence engagement, including the robot's physical appearance [24], [35], performance (e.g., reliability, predictability), behavior (e.g., anthropomorphism, gestures) [31], task structure, cultural factors [1], [15], as well as people's prior experience with and impressions of robots. One under-explored factor is *curiosity*, which has been shown in human-computer interaction research to play a significant role in affecting user engagement, in the context of crowdsourcing systems [17], interactive displays [36], educational technologies [23] and games [37]. The core idea is that by designing systems to present stimuli that are novel, conflicting, uncertain, complex, or surprising, human users will be intrinsically motivated to engage with the system for a longer period of time in order to satisfy their "desire to know, to see, or to experience" [19]. Likewise, several works have found that unpredictable behavior of robots [30] and virtual agents [3] can lead to engagement, potentially due to higher levels of perceived anthropomorphism [8], [34]. At the same time, the lack of predictability and transparency can negatively affect trust and reliance [14], [18].

In this work, we conducted a Wizard-of-Oz study to investigate *curiosity* as a mediator of engagement in human-robot interaction, with 36 participants using Recyclo, a service robot that recognizes objects and advises users on their recyclability. Our results show that unpredictable robot responses gave rise to surprise, which in turn, elicited both positive and negative reactions from users. On the one hand, surprising robot responses violated users' expectations of the robot's capabilities and led them to choose interaction strategies to test their theories about how the robot functions, both of which reflect their curiosity. On the other hand, the lack of predictability negatively impacted users' perceptions of the robot's reliability and their trust in the robot's guidance.

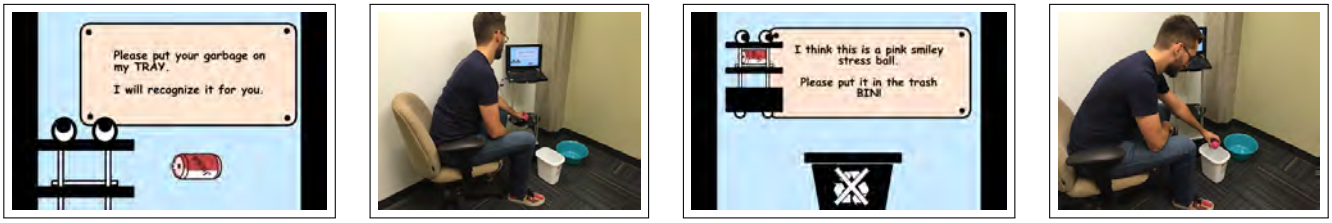
## II. Related Work

Many robotic systems have implemented curiosity as a mechanism for directing the robot's exploratory behavior. Curiosity-driven reinforcement learning [20], [32], [33], for example, encourages agents to explore parts of the state space that yield the highest information gain, as opposed to the best reward. Curious robots were shown to be practical—they learn more efficiently and can outperform standard reinforcement learning, especially in cases of rare rewards [11]. Our work is different in that we study how the robot's behavior evokes the curiosity of a human interacting with the robot, and how that in turn affects engagement and trust.

Prior work has shown that robot performance can modulate both engagement and trust. Imperfect robots can in fact be more engaging. Hamacher et al. [13] found that the majority of participants preferred a faulty robot that is socially interactive (e.g., appears sad and apologetic after an error) than a non-social but perfect robot, even when tasks took longer to complete due to the additional interactions. In Ragni et al. [26], participants perceived erroneous robots to be less competent and intelligent, but found the interaction more fluid, joyful and interesting, even though team task performance was compromised. On the flip side, inconsistent robot performance can negatively impact trust. Lee and See [18] found that performance-related metrics (e.g., reliability, false alarm rate, failure rate) are strongly associated with trust development and maintenance, while robot attributes (e.g., proximity, personality, anthropomorphism) are only weakly associated. Hancock et al. [14] identified reliability (i.e., *what* the automation does), transparency (i.e., *how* the automation operates) and intention (i.e., *why* the automation was developed) as three important factors that contribute to the development of human-robot trust. Brooks et al. [5] found

*Supported by NSERC Discovery Grant RGPIN-2015-04543

[1] The authors are with the University of Waterloo, Canada edith.law, x34cai, qfliu, ssasy, jtcgoh, asblidar, dana.kulic@uwaterloo.ca




(a) step 1: robot prompts user to place an object onto its tray (b) step 2: user chooses an object and places it on the tray (c) step 3: robot identifies the object and makes recycling recommendations (d) step 4: user puts the object into the trash or recycling bin

Fig. 1. Interaction with Recyclo

that failures make robots seem less capable, lower trust, and make users reluctant to use the robot's services again. However, the negative reaction caused by failure can be mitigated by providing human support (i.e., notifying people to take actions to alter the situation) or task support (i.e., the robot performing an action to recover from the mistakes). Finally, Mota et al. [22] found that people gauge the trustworthiness of robots via testing, e.g., by trying different strategies to see how the robot would respond. In our work, we manipulate the predictability of robot responses, e.g., by making them *surprisingly* accurate or inaccurate, in order to study their impact on curiosity, engagement and trust.

## III. Framework Development

To develop a conceptual framework for studying curiosity in human-robot interaction, we created Recyclo—a Wizard-of-Oz recycling robot whose function is to help users determine whether objects are recyclable or not—and conducted a pilot study to qualitatively assess the reactions of participants in response to Recyclo's object recognition and recyclability judgments. Results from the pilot study were used to inform the design of our main experiment.

### A. Recyclo

Recyclo is a service robot whose function is to help users recycle. Different from [10], [34], [40], Recyclo identifies objects that users present, classifies their recyclability, then makes a recommendation to users to put the object into the trash or recycling bin. To process each object, users interact with Recyclo via a sequence of steps, as shown in Figure 1. First, users select an object and place it on the middle rack of the robot monitored by a webcam. Recyclo displays a busy indicator, showing users that it is in the process of identifying the object. Upon observing the object placed by users, Recyclo visually (via a screen) and verbally (via text-to-speech software) communicates its response, telling them what the object is (e.g., "it is a pen") and providing a recommendation as to whether the object should be recycled (e.g., "put it in the recycling bin") or discarded (e.g., "put it in the trash bin"). At this point, users are asked to imagine that this is a real world scenario where they are making the final decision about recyclability, decide whether they agree with or trust Recyclo's recommendation and throw the object into the bin it specifies, or do the opposite. Built using the Turtlebot platform, Recyclo is remotely controlled by a human operator, who monitors the interactions via Teamviewer (a remote access software), identifies the object that users present to Recyclo, constructs and sends a response (i.e., what the object is and whether it is recyclable) back.

### B. Pilot Study

We conducted a pilot study with 12 participants to qualitatively assess the reactions of participants in response to Recyclo's object recognition and recyclability judgments. The study was conducted in a conference room scattered with objects. Participants were first asked to choose from a pre-determined set of 10 objects, as shown in Figure 2—5 commonly recycled objects (made of paper, metal, or plastic) and 5 more ambiguous ones in terms of recyclability (made of mixed materials)—to show to Recyclo. Participants could additionally choose other objects (either in the room or ones they brought themselves) and stop at any time by informing the researcher. For each object, the human operator generated an object recognition response with different levels of correctness, specificity and predictability on the fly. For recyclability, the human operator generated a best guess response based on the assumption that an object is recyclable if it is made of a single, recyclable material (such as metal, plastic, paper, glass or rubber) and not recyclable if it is an eWaste product, contains chemicals, or is not typically found in recycling bins.

Fig. 2. Objects in the pilot study

There are several observations from the pilot study that informed our experimental design. First, we observed that certain types of responses elicited a strong feeling of surprise from participants. Accurate and unusually specific responses (e.g., a remote control identified as "panasonic remote") as well as inaccurate and completely *off* responses (e.g., a pen identified as a "coin") both contributed to surprise.

TABLE I
RESPONSE TYPES

| response type | description | examples | correctness | specificity | conceptual proximity |
|---|---|---|---|---|---|
| correct-general (CG) | correct and generic response | bottle | correct | general | exact |
| correct-specific (CS) | correct but unusually specific response, such as the inclusion of brand names, adjectives or detailed description of more than one attribute of the object | "panasonic remote control", "stylish pair of glasses", "scissors with angular white handle" | correct | specific | exact |
| incorrect-close (IC) | an incorrect but believable response, such as an object that shares at least two attributes in common with what is presented, in terms of function, color, shape, size or material | paper recognized as "napkin" | incorrect | general | close |
| incorrect-wayoff (IW) | a severely incorrect response, e.g., an object that shares no attributes in common with the presented object in terms of function, color, shape, size of material | paper recognized as a "mug" | incorrect | general | far |
| incorrect-nonsense (IN) | a nonsensical response, e.g., an object that should not exist in a given environment | lanyard recognized as "snake' | incorrect | general | very far |

Furthermore, we also observed *testing behavior*: one participant presented Recyclo with the same object multiple times under different angles or two very similar objects (e.g., two different kinds of tape), in order to test the robot.

Second, participants showed a strong preference for presenting the robot with objects they brought themselves to the experimental setting (e.g., their wallet, phone or jewelry), as opposed to using the objects available in the room. They reported feeling less surprised by robot responses on the ten objects that we provided compared to the objects they chose themselves, believing that the robot was pre-trained to generate the appropriate responses for those objects, or that other objects were *planted* by the experimenter in the vicinity. Providing a pre-defined set of objects for which responses can be fixed has the obvious benefit of minimizing bias that may be introduced by the experimenter generating the responses on the fly. However, observations from the pilot study show that doing so makes the experiment much less ecologically valid, as we cannot reliably elicit a genuine sense of surprise from participants. These observations form the basis of our experimental procedure for the main study.

### C. Framework

Based on findings from the pilot study and prior literature, we hypothesize that the surprise that users experience when confronted with unexpected robot responses will cause them to become curious and seek information to make sense of the robot's behavior. At the same time, surprise will positively impact user engagement and negatively impact trust. Our conceptual framework is illustrated in Figure 3.

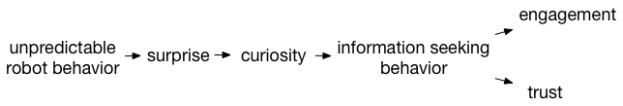



Fig. 3. Conceptual Framework

To experimentally validate this conceptual framework, we manipulated robot responses by making them more or less surprising based on how participants reacted to different types of robot responses in the pilot study. Table I shows the five types of object recognition responses based on three key dimensions, namely correctness (i.e., whether the response is correct or incorrect), specificity (i.e., whether the response is general or specific) and conceptual proximity (i.e., whether the response object is conceptually exact, close to or far away from the real object, in terms of the number of shared attributes). We expect correct-specific (CS) responses to be more surprising than correct-general (CG) responses, because people typically do not expect robots to be able to have human-level recognition capability to describe objects in minute detail. Likewise, incorrect-wayoff (IW) and incorrect-nonsense (IN) responses are anticipated to be more surprising than incorrect-close (IC) responses, in that they are high contrast and highly improbable responses respectively, thus potentially leading participants to question how the robot could have generated such absurd mistakes.

The human operator generated robot responses by following a predetermined sequence of correct (C) and incorrect (I) responses. The sequence was fixed in order to keep the accuracy rate (~70%) of the robot consistent across all sessions, and to eliminate the variability that might arise from order effects. Our two human operators were trained to follow the same set of rules for generating responses as shown in Table I.

## IV. EXPERIMENT

In the main experiment, we manipulate robot responses in order to evoke surprise in human users interacting with Recyclo, and study how this experience of surprise gives rise to curiosity and affects engagement and trust.

### A. Study Design

*1) Participants:* We recruited 36 university students via mailing lists and posters. Half of our participants have a technical education background, i.e., their academic major, course or research work are related to computer science and engineering. This balance of technical and non-technical participants allows us to understand how knowledge about intelligent machines factors into participants' expectations of robot behavior, and hence their experience of surprise during the interaction. Participants range in age between 18 and 35 years old, and the gender distribution is 56% male and 44% female.

*2) Procedure:* Each study session took approximately 1 hour, and participants were paid a $20 honorarium. Upon arrival, participants were given a brief introduction about the purpose of the study (i.e., how trust and curiosity shape human robot interactions) and the procedure on how to interact with Recyclo. Participants then proceeded to fill in a pre-study questionnaire, which asked them about their educational background, experience with recycling and initial impressions or expectations of Recyclo's capabilities. They were given a brief demo, followed by the actual experiment. After the study, participants rated their overall level of engagement, surprise, motivation, trust and perception of reliability in a post-study questionnaire. We also conducted short interviews with participants, asking them to reflect on the rationale behind the way they chose objects, the decisions they made and their experiences of surprise, engagement and trust throughout the interaction.

We separated participants into two conditions. In the **low surprise** (LS) condition, the robot generated predictable responses, namely CG and IC responses. In the **high surprise** (HS) condition, the robot generated unpredictable responses, namely CS, IW and IN responses, and resorted to CG responses only if the presented object lacked attributes that lend themselves to specific description. If participants presented the same object multiple times, the human operator gave consistent answers (i.e., the same each time) for the low surprise condition, and inconsistent answers (i.e., different each time) for the high surprise condition.

In the training phase of the experiment, participants chose from a pre-determined set of three objects (namely, a roll of tape, an empty beverage can and a plastic bottle) to present to Recyclo. After training, participants could choose objects in their vicinity or personal items that they brought with them. This phase of the experiment was unbounded—participants were told to continue showing objects to Recyclo, so long as they found the interaction fun and engaging.

To capture how surprise changes during the interaction, we asked participants to record information using an interaction form. Specifically, after choosing and before presenting an object, participants were asked to note down what the object is, whether they think the object is recyclable (yes, no, unsure), and whether they think that Recyclo can recognize the object (yes, no, unsure). After disposing of the object, participants were asked to record what the robot identified the object to be, which bin Recyclo recommended versus the bin they actually used, and to rate on a 7-point scale how surprised they were about the robot's responses.

### *B. Analysis Methods*

We used a mixed methods approach, including quantitative analyses and qualitative summaries of the interviews.

**Dependent Variables:** These include macro-level effects (i.e., the overall engagement, motivation, surprise, reliability and trust, as evaluated using the post-study questionnaire) and interaction-level effects related to surprise (i.e., per-object-surprise), curiosity (i.e., testing), engagement (i.e., nb-objects-presented) and trust (i.e., recyclability-trust, recyclability-compliance), which were measured per object and may vary over the course of the experiment.

**Independent Variables:** The only factor we manipulated experimentally in the low surprise versus high surprise conditions was the type of response the robot provides for object recognition (i.e., response-type).

**Covariates:** We took into account other factors that may affect the patterns of interaction between participant and robot; these include user characteristics (i.e., background, gender), expertise in intelligent systems (i.e., knowledge-robot, knowledge-AI), recycling experience and knowledge (i.e., recycling-frequency, recycling-confusion), pre-study expectations (i.e., prestudy-expectation-recognition, prestudy-expectation-recyclability), as well as expectations and experiences of violated expectations during the interaction with the robot (i.e., recognition-expectation, recyclability-expectation, recognition-violation, recyclability-violation).

To simplify notation for categorical variables, each category will be denoted by the name of the variable followed by the category name; for example, gender-F is used to denote female participants. Oftentimes, the effects are aggregated over the number of objects by taking averages (avg), variances (var) and proportions (prop) to facilitate modelling. Throughout the paper, we will use the shorthand of the aggregation method followed by the name of the variable to denote the aggregated value, for example, avg per-object-surprise implies the average per-object-surprise.

**Statistical Methods:** For high-level descriptions of surprise, curiosity, engagement and trust, we used descriptive statistics (e.g., mean, median) when appropriate. The models considered were: (1) logistic regression models for modelling binary response variables, (2) Poisson regression models for modelling count-type response variables, and (3) proportional odds models for ordinal response variables [21]. We used model averaging [6] to select the best possible explanatory variables for each model. Since the list of all possible explanatory variables for each model is short, exhaustive search of the main effects was done for every model to identify the variables that have strong relationships with the corresponding response variable. The goodness of fit for each model was assessed using the Akaike Information Criterion [39]. We performed the Wald test to investigate the significance of the effects for the proportional odds models, and the t-test for the logistic and Poisson regression models.[1] In all the tables reporting modelling results, $\hat{\beta}$ is the standardized regression coefficient, *Std. Error* is the standard error for the estimate of $\hat{\beta}$, and $t$ is the coefficient estimate divided by the standard error, a measure of precision of the coefficient estimate.

The 36 participants who took part in this study were evenly and randomly split between the LS and HS conditions. One participant in the HS condition deviated from procedure (she presented 3 objects, stopped, then mid-way through the interview, asked to present more objects to Recyclo);

[1]Statistically significant results are reported as follows: $p < 0.001$(***), $p < 0.01$(**), $p < 0.05$(*), $p < 0.1$(·).

this participant's responses were removed from the data. This data cleaning step is done to maintain data integrity; the removal of one participant did not affect the overall composition of the two groups.

## V. Results

### A. Unpredictable Robot Response → Surprise

The first question of interest is whether manipulating the response types predictably affects the surprise levels of individual users. While results show that at the macro-level, there are no significant differences in the general surprise level (reported in the post-study questionnaire) between conditions, we found differences when analyzing factors that influence surprise at the interaction (or per-object) level.

Figure 4 illustrates the distribution of per-object-surprise by response type. The medians denoted by the lines inside the boxes, show that the correct-specific (CS) and incorrect-nonsense (IN) response types elicited higher surprise ratings than correct-general (CG) and incorrect-close (IC) responses in general. The incorrect-wayoff (IW) responses had the lowest surprise ratings. A proportional odds model also reveals a number of user characteristics that influenced per-object-surprise. For example, male participants experienced significantly lower surprise than female participants ($\chi^2(1, N = 35) = 24.03), p < 0.001$), while participants with AI knowledge tended to be more surprised ($\chi^2(1, N = 35) = 8.55), p < 0.01$).

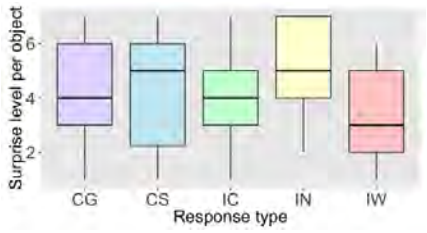



Fig. 4. Boxplots of surprise level per-object by response type.

The interview data further confirm these quantitative results. As expected, correct-specific and incorrect-nonsense responses were found to be surprising. However, participants also mentioned that the surprise was in part due to the juxtaposition of responses with different levels of specificity and correctness. For instance, the same participant, having seen that Recyclo can be extremely specific, was surprised when Recyclo classified a piece of tissue as "tissue" and not "Kleenex".

Many participants mentioned being surprised when their expectations about Recyclo's abilities were violated. However, the extent to which they experienced surprise depended critically on each individual's *unique mental model* about how difficult a particular object is to recognize or sort. One participant said that a stapler must be easy for the robot to recognize because of its distinct shape, while another participant, who has expertise in robotics, considered the same stapler to be difficult to recognize because it is a black object placed against a black background. Participants' certainty (or uncertainty) about an object's recyclability also played a role in their experience of surprise. One participant said "I always think everything is recyclable, and so pretty much all the answers he gave me corresponded with what I thought." Here, the lack of recycling knowledge contributed to the lack of surprise.

TABLE II

LINEAR MODEL FOR PROPORTION OF TIME PARTICIPANTS SPENT TESTING THE ROBOT.

| | Model Parameters | | | |
|---|---|---|---|---|
| **Variable** | $\hat{\beta}$ | *Std. Error* | *t* | *p-value* |
| condition-HS | 0.12 | 0.05 | 2.27 | * |
| var per-object-surprise | 0.07 | 0.03 | 2.84 | ** |

### B. Surprise → Curiosity and Information Seeking Behavior

To study whether surprise elicits testing behavior indicative of curiosity, we analyzed two variables—testing (i.e., whether a participant exhibited any testing behavior) and prop testing (i.e., the proportion of the interaction sequence that participants spent on testing). Here, we define testing behavior as users presenting the same objects to Recyclo multiple times. Overall, results show that more HS participants (65%) exhibited testing behaviour compared to those in the LS condition (33%), but this difference is not significant due to large variances, $\chi^2(1, N = 35) = 3.44, p = 0.06$.

To investigate what affects the proportion of the interaction sequence that participants spent on testing (prop testing), we used a linear regression model. Participants assigned to the HS condition tested the robot more often than those assigned to the LS condition, as shown in Table II. Participants also spent more time on testing when their variance of per-object-surprise is larger, i.e. when their per-object surprise ratings span a larger range.

Another indication of testing behavior is when participants intentionally choose objects to stump the robot. Analysis using a linear model shows that the more often participants failed to guess whether Recyclo can recognize the object (i.e., prop-recognition-violation), the more often they selected objects that they do not think Recyclo can recognize (i.e., recognition-expectation-No and recognition-expectation-Unsure), $\hat{\beta} = 1.12, t(33) = 5.08, p < 0.001$.

During the interviews, most participants expressed their desire to test the robot in order to find the system's limitations and to understand the "thinking" process of the robot. Participants said that their curiosity was piqued by a single or a small group of surprising items. These surprises came in multiple forms: as a specific and correct recognition of the object (e.g. adidas sporty sneakers), as an incorrect but amusing recognition (e.g. whiteboard magnet recognized as a delicious Oreo cookie), or simply as a correct recognition of an object that the participant didn't think Recyclo would be able to handle (e.g. tube-shaped speakers). These surprises offered the users a glimpse of Recyclo's supposed capabilities, and invited further prodding and testing of its

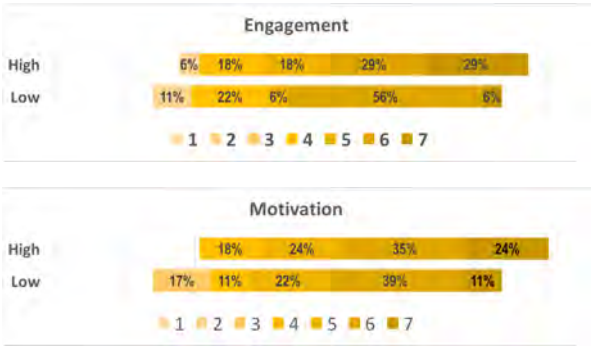



Fig. 5. Engagement and Motivation by condition. The legend indicates participant responses on the 7-point Likert scale.

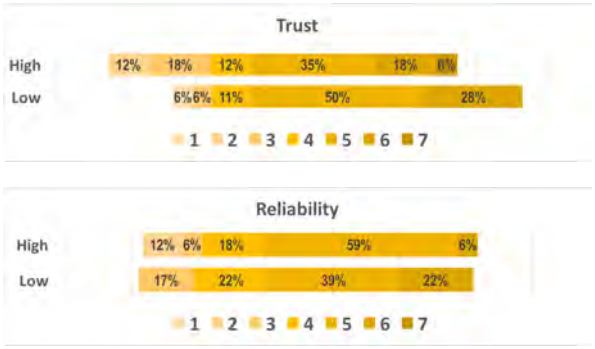



Fig. 6. Trust and Reliability by condition. The legend indicates participant responses on the 7-point Likert scale.

knowledge.

As we had found in the pilot study, multiple participants tested the robot by presenting the same object multiple times using different angles. However, several participants went further and constructed new objects. For example, two participants tested Recyclo's ability to read various kinds of text (e.g., french, sheet music). Another participant presented a take-out paper bag three times, once as is, the second time crumpled, and the third time with a roll of tape on top to cover its text. Some participants constructed composite objects made of two or more items, while others tried several objects in a similar category, as they were "curious about the granularity to which Recyclo would be able to detect objects." Participants who exhibited such testing behavior said that they were motivated to understand how the robot works.

### C. Effects on Engagement

At the macro-level, HS participants reported slightly higher engagement ($M = 5.52$, $Med = 6.00$, $SD = 1.42$) and motivation ($M = 5.65$, $Med = 6.00$, $SD = 1.06$), compared to the LS participants' engagement ($M = 5.11$, $Med = 6.00$, $SD = 1.45$) and motivation ($M = 5.12$, $Med = 5.50$, $SD = 1.29$). Figure 5 contains divergent stacked bar graphs, which center the neutral responses in the middle. The plot shows that HS participants tended to be slightly more engaged (as indicated by the slight shift to the right) than LS participants. However, the differences are not statistically significant, $\chi^2(1, N = 35) = 0.91, p = 0.34$.

Participants in the high surprise condition chose slightly more objects on average ($M = 25.76$, $SD = 9.20$) than those in the low surprise condition ($M = 23.72$, $SD = 10.72$), but the difference is not statistically significant, $t(33) = -0.61, p = 0.55$. A Poisson regression model is used to further investigate factors that influence the number of objects participants showed to Recyclo. According to Table III, participants whose expectation of Recyclo's recognition ability was violated and who spent more time testing showed significantly more objects to Recyclo. The more that participants trust Recyclo, i.e., change their opinion based on its recommendation, the more objects they showed Recyclo. Finally, participants who experienced more highs and lows (i.e., variance) in their per-object-surprise showed more objects to Recyclo.

TABLE III

POISSON REGRESSION MODEL FOR THE NUMBER OF PRESENTED OBJECTS.

| | Model Parameters | | | |
|---|---|---|---|---|
| **Variable** | $\hat{\beta}$ | *Std. Error* | *t* | *p-value* |
| var per-object-surprise | -0.15 | 0.05 | -3.11 | ** |
| prop recognition-violation | 0.86 | 0.38 | 2.28 | * |
| prop recyclability-trust | 0.99 | 0.50 | 1.98 | * |
| prop testing | 0.73 | 0.27 | 2.69 | ** |

During the interviews, many of the LS participants reported that their interest in interacting with Recyclo started high but decreased over the course of the interaction. On the other hand, HS participants often remarked that they became more interested in interacting with the robot when they encountered the first surprising response from the robot, since these responses provided new insights into how the robot thinks, which in turn fueled their quest to "figure out" the robot.

### D. Effects on Trust

Figure 6 shows that at the macro-level, HS participants perceived Recyclo as less trustworthy ($M = 4.47$, $Med = 5.00$, $SD = 1.46$) compared to LS participants ($M = 4.83$, $Med = 5.00$, $SD = 1.25$), even though the perception of Recyclo's reliability is roughly the same between the HS ($M = 4.41$, $Med = 5.00$, $SD = 1.12$) and LS ($M = 4.67$, $Med = 5.00$, $SD = 1.03$) groups. However, these differences are not statistically significant ($t(33) = 0.79, p = 0.44$ and $t(33) = 0.70, p = 0.49$, respectively). Results also reveal a high correlation ($r(33) = 0.78$, $p < 0.001$) between trust and reliability—the more participants find the robot reliable, the more they would trust it to actually recycle trash for them.

Prior studies have shown that trust can be demonstrated through compliance with the robot's recommendations [27], [29]. Here, we evaluated both compliance (participants following Recyclo's recommendation) as well as a stronger notion of trust, when participants change their initial opinion about an object's recyclability based on Recyclo's recommendation. As expected, participants were less likely to

comply when they disagree with Recyclo's recyclability recommendation ($\hat{\beta} = -5.72, t(824) = -12.92, p < 0.001$). In terms of trust, the logistic regression model shows that participants were significantly less likely to trust Recyclo if they were in the HS condition ($\hat{\beta} = -0.69, t(824) = -2.12, p = 0.03$) or if they were more surprised on a per-object level ($\hat{\beta} = -0.32, t(824) = -3.30, p < 0.001$). Participants who had high technical expertise were less likely to rely on the robot when its recommendation differed from their own ($\hat{\beta} = -1.24, t(824) = -1.98, p = 0.05$), a finding also reported in Gombolay et al. [12]. Finally, participants were more likely to both comply and trust Recyclo's recyclability recommendation when the object recognition response was correct ($\hat{\beta} = 1.80, t(824) = 4.54, p < 0.001$) and when they had high expectations of the robot's object recognition capability prior to the experiment ($\hat{\beta} = 0.66, t(824) = 3.46, p < 0.001$).

These findings are well-aligned with participants' comments during the interviews. Our interview data reveal that many participants considered recyclability classification to be an easier (or even trivial) task compared to object recognition. In fact, participants tended to believe that if Recyclo can recognize objects with such impressive accuracy, it must also be able to determine recyclability correctly. One participant was unsure whether house keys are recyclable, but upon seeing that Recyclo was able to classify the object as "a key chain with two keys", she accepted Recyclo's recommendation. Finally, confirming results from our quantitative analysis that knowledge plays a role in trust, a number of technical students mentioned during the interviews that they would not trust the robot that much because they understand how difficult material recognition and recyclability classification is for current AI technology.

## VI. Discussion

This work investigates the relationship between surprise, curiosity, engagement and trust within the context of human-robot interaction. Our study confirms that surprise can positively influence the amount of testing behavior, engagement, and length of interaction. On the flip side, surprise may negatively influence trust and compliance. Our results also show that participants differ in terms of what robot responses they find surprising; this difference can be attributed to individual characteristics (e.g., gender, technical background and recycling experience) and expectations.

Interview data reveal that participants engaged in testing behavior in order to understand how Recyclo works; in the process, they adaptively infer what Recyclo can and cannot do. For example, responses that include color, material, or other tactile information (e.g., squish ball) in the description made participants think that Recyclo was especially equipped (e.g., with special sensors) to recognize those attributes. Participants also made inferences about mistakes and how they came about. Finally, many of the participants also inferred that Recyclo had some kind of learning capabilities, and attributed some of the specific answers to Recyclo learning from previous interactions.

One way to explain why there was more testing behavior in the high surprise condition, is that many of the unusually specific (CS), way off (IW), nonsense (IN) responses require some creative thinking to explain, compared to the expected answers (e.g., CG) and forgivable mistakes (e.g., IC) found in the low surprise condition. This form of adaptive inference can be explained in part by psychological theories of curiosity—the surprising responses create a bigger information gap, which participants are compelled to close by engaging in testing (or information seeking) behavior.

There are several limitations to our study. First, while our participant pool involves a balance of users with different levels of technical and recycling expertise, this user population is nonetheless homogeneous in terms of age and level of education. Second, while only one of our participants realized that Recyclo was human operated, several participants mentioned that they expected a more sophisticated robot, i.e, one that can move. The violation of this expectation can negatively affect user engagement. Third, using the number of objects as a measure of engagement is not perfect, as there are anchoring effects (e.g., participants wanting to present just enough objects to fill the page in the interaction form, or participants continuing to present objects so that the experimenters would have enough data) and external factors (e.g., the session starting late, the participant needing to leave early, or the participant simply being slower in general at choosing the next object) that affect the actual number of objects presented. Finally, in this study, we explored only *short-term* interactions between users and Recyclo. As such, novelty effects may be present, and the session was not long enough for us to observe habituation to surprise. These limitations point to interesting future work, involving different user populations (e.g., elderly people), a more sophisticated robot capable of movement and gestures, studying longitudinal effects of surprise, as well as the design of surprising human-robot experiences that take into account the unique characteristics of each participant.

Finally, results from our study imply that surprise is not always an appropriate engagement tactic for real-world human-robot interaction scenarios. Surprising responses from robots that aim to entertain or educate can be effective in piquing users' interest and curiosity. On the other hand, such unpredictable behavior in task-oriented robots may lead to unsatisfactory user experience. Future work can explore ways to mitigate the negative effects of surprise. For example, users may find surprising responses easier to accept if an explanation is provided by the robot.

## VII. Conclusion

Engagement and trust are important issues in human-robot interaction. In this paper, we report findings from a Wizard-of-Oz experiment investigating the relationship between surprise, engagement and trust, as a way to understand the role that curiosity plays in motivating and sustaining interactions between humans and robots. Results show that surprise can result in curiosity and information seeking behavior, increase engagement and decrease trust. Moreover, individual

characteristics and expectations can modulate the experience of surprise and its downstream effects.

## References


[1] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, "Evaluating the Engagement with Social Robots," *International Journal of Social Robotics*, vol. 7, no. 4, pp. 465–478, 2015.

[2] S. Azenkot, C. Feng, and M. Cakmak, "Enabling building service robots to guide blind people: A participatory design approach," in *Proc. ACM/IEEE International Conference on Human Robot Interaction (HRI)*, 2016, pp. 3–10.

[3] T. Bickmore, D. Schulman, and L. Yin, "Maintaining Engagement in Long-Term Interventions With Relational Agents," *Applied Artificial Intelligence*, vol. 24, no. March 2015, pp. 648–666, 2010.

[4] J. Biswas and M. Veloso, "The 1,000-km challenge: Insights and quantitative and qualitative results," *IEEE Intelligent Systems*, vol. 31, no. 3, pp. 86–96, 2016.

[5] D. J. Brooks, M. Begum, and H. A. Yanco, "Analysis of reactions towards failures and recovery strategies for autonomous robots," in *Proc. IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 487–492.

[6] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference*, 2002.

[7] K. Dautenhahn, S. Woods, C. Kaouri, M. L. Walters, K. L. Koay, and I. Werry, "What is a robot companion-friend, assistant or butler?" in *Proc. IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2005, pp. 1192–1197.

[8] F. Eyssel, D. Kuchenbrandt, and S. Bobinger, "Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism," in *Proc. ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2011, pp. 61–67.

[9] J. Fasola and M. J. Mataric, "Using socially assistive human–robot interaction to motivate physical exercise for older adults," *Proceedings of the IEEE*, vol. 100, no. 8, pp. 2512–2526, 2012.

[10] J. Fink, S. Lemaignan, P. Dillenbourg, P. Rétornaz, F. C. Vaussard, A. Berthoud, F. Mondada, F. Wille, and K. Franinovic, "Which robot behavior can motivate children to tidy up their toys? Design and Evaluation of "Ranger"," *Proc. ACM/IEEE International conference on Human-robot interaction (HRI)*, pp. 439–446, 2014.

[11] S. Forestier, Y. Mollard, D. Caselli, and P.-Y. Oudeyer, "Autonomous exploration, active learning and human guidance with open-source Poppy humanoid robot platform and Explauto library," in *Proc. Conference on Neural Information Processing Systems (NIPS)*, 2016.

[12] M. Gombolay, X. J. Yang, B. Hayes, N. Seo, Z. Liu, S. Wadhwania, T. Yu, N. Shah, T. Golen, and J. Shah, "Robotic assistance in coordination of patient care," in *Proc. Robotics: Science and Systems (RSS)*, 2016, pp. 1–11.

[13] A. Hamacher, N. Bianchi-Berthouze, A. G. Pipe, and E. Kerstin, "Believing in bert: Using expressive communication to enhance trust and counteract operational error in physical human-robot interaction," in *Proc. IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 493–500.

[14] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517–527, 2011.

[15] S. Ivaldi, S. Lefort, J. Peters, M. Chetouani, J. Provasi, and E. Zibetti, "Towards engagement models that consider individual factors in hri: On the relation of extroversion and negative attitude towards robots to gaze and speech during a human–robot assembly task," *International Journal of Social Robotics*, pp. 1–24, 2016.

[16] D. Karreman, G. Ludden, and V. Evers, "Visiting cultural heritage with a tour guide robot: A user evaluation study in-the-wild," in *Proc. International Conference on Social Robotics (ICSR)*, 2015, pp. 317–326.

[17] E. Law, M. Yin, J. Goh, K. Chen, M. A. Terry, and K. Z. Gajos, "Curiosity killed the cat, but makes crowdwork better," in *Proc. Conference on Human Factors in Computing Systems (CHI)*, 2016, pp. 4098–4110.

[18] J. D. Lee and K. A. See, "Trust in automation: designing for appropriate reliance." *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[19] J. Litman, "Curiosity and the pleasures of learning: Wanting and liking new information," *Cognition and Emotion*, vol. 19, no. 6, pp. 793–814, 2005.

[20] M. Lopes, T. Lang, M. Toussaint, and P.-Y. Oudeyer, "Exploration in model-based reinforcement learning by empirically estimating learning progress," in *Proc. Conference on Neural Information Processing Systems (NIPS)*, 2012, pp. 206–214.

[21] P. McCullagh and J. Nelder, *Generalized Linear Modesl*, 2nd ed. Chapman & Hall/CRC, 1989.

[22] R. Mota, D. Rea, A. Le Tran, J. E. Young, E. Harlin, and M. C. Sousa, "Playing the 'trust game' with robots: Social strategies and experiences," in *Proc. IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 519–524.

[23] P.-Y. Oudeyer, J. Gottlieb, and M. Lopes, "Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies," *Progress in Brain Research*, pp. 257–284, 2016.

[24] R. A. Paauwe, J. F. Hoorn, E. A. Konijn, and D. V. Keyson, "Designing Robot Embodiments for Social Interaction: Affordances Topple Realism and Aesthetics," *International Journal of Social Robotics*, vol. 7, no. 5, pp. 697–708, 2015.

[25] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun, "Towards robotic assistants in nursing homes: Challenges and results," *Robotics and autonomous systems*, vol. 42, no. 3, pp. 271–281, 2003.

[26] M. Ragni, A. Rudenko, B. Kuhnert, and K. O. Aaras, "Errare humanum est: Erroneous robots in human-robot intearction," in *Proc. IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 501–506.

[27] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *Proc. ACM/IEEE International Conference on Human Robot Interaction (HRI)*, 2016, pp. 101–108.

[28] A. M. Sabelli and T. Kanda, "Robovie as a mascot: A qualitative study for long-term presence of robots in a shopping mall," *International Journal of Social Robotics*, vol. 8, no. 2, pp. 211–221, 2016.

[29] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust," in *Proc. ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2015, pp. 141–148.

[30] E. Short, J. Hart, M. Vu, and B. Scassellati, "No fair!! An interaction with a cheating robot," in *Proc. ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010, pp. 219–226.

[31] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1-2, pp. 140–164, Aug. 2005.

[32] S. Singh, A. Barto, and N. Chentanez, "Intrinsically motivated reinforcement learning," in *Proc. Conference on Neural Information Processing Systems (NIPS)*, vol. 17, no. 2, 2004, pp. 1281–1288.

[33] S. Still and D. Precup, "An information-theoretic approach to curiosity-driven reinforcement learning," *Theory in Biosciences*, vol. 131, no. 3, pp. 139–148, 2012.

[34] H. Tan, L. Sun, and S. Sabanovic, "Feeling green: Empathy affects perceptions of usefulness and intention to use a robotic recycling bin," in *Proc. IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 507–512.

[35] A. Tapus, C. Tapus, and M. Matarić, "The role of physical embodiment of a therapist robot for individuals with cognitive impairments," in *Proc. IEEE International Workshop on Robot and Human Interactive Communication*, 2009, pp. 103–107.

[36] R. Tieben, T. Bekker, and B. Schouten, "Curiosity and interaction: Making people curious through interactive systems," in *Proc. BCS Conference on Human-Computer Interaction*, 2011, pp. 361–370.

[37] A. To, S. Ali, G. Kaufman, and J. Hammer, "Integrating curiosity and uncertainty in game design," in *Proc. International Joint Conference of DiGRA and FDG*, 2016, pp. 1–16.

[38] R. Triebel, K. Arras, R. Alami, L. Beyer, S. Breuers, R. Chatila, M. Chetouani, D. Cremers, V. Evers, M. Fiore, *et al.*, "Spencer: A socially aware service robot for passenger guidance and help in busy airports," in *Proc. Conference on Field and Service Robotics (FSR)*, 2016, pp. 607–622.

[39] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*. Berlin: Springer-Verlag, 2003.

[40] S. Yang, B. K. J. Mok, D. Sirkin, H. P. Ive, R. Maheshwari, K. Fischer, and W. Ju, "Experiences developing socially acceptable interactions for a robotic trash barrel," in *Proc. IEEE International Workshop on Robot and Human Interactive Communication*, 2015, pp. 277–284.