IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

What's the Deal with Robot Comedy? Pinpointing the Impact of Post-Joke Repartee in a Robotic Comedian's Performance

Ayan Robinson*[1], Sarah Woods*[1], Madison R. Shippy[1], DeAndre Walcott[1], and Naomi T. Fitter[1]

***Abstract*—The rise in prevalence of AI-enabled technologies (from voice assistants to social robots) has not yet been accompanied by an analogous mastery of computer-mediated humor. Although humans often use jokes to repair interactions and navigate uncomfortable scenarios, social robots in similar roles typically fall short at reading the room and adapting behavior according to sensed social contexts and reactions. We pursued two studies to gain clearer evidence about adaptive robot joking's influence (compared to hardcoded repartee or no robot banter). The first study ($N = 48$, between-subjects design) examined in-person one-on-one human-robot interactions across the three conditions. The results indicated that adaptive repartee by robots tended to increase perceived warmth, competence, comfort, social closeness feelings, and humorousness, and that human behavioral responses varied significantly between conditions, with any repartee leading to significant gains over no repartee. The second study used an online video-based survey with a within-subjects design ($N = 99$) to examine the same conditions. This follow-up effort showed significant gains in perceived competence and anthropomorphism for any type of repartee, although this banter also made the robot more discomforting. Our work can help practitioners who are interested in applying playful banter to enhance robot charm and success.**

I. Introduction

Iconic robots from the media, from C-3PO to Futurama's Bender, use humor to succeed in their day-to-day exploits. Yet as real-world personal and service robots become more common, there is a palpable gap between the social playfulness of actual and fictional robotic systems. Is this a wise design decision, or is some potential lost (in terms of social perception, relationship-building, entertainment, or even value) when omitting playful banter in modern robotic systems? With inspiration from robots of the media and skilled human performers, we seek to equip robots like the Misty II (Fig. 1) with repartee abilities and assess the impact of these new and playful skills on human-robot interactions.

Our work is not the first to address robot humor, but this vibrant research area is still working toward reliable and autonomous social robotic systems that can use comedy skills to their advantage in day-to-day settings. From over a decade ago to the present, robotics researchers have showcased autonomous robotic comedians onstage [1], [2], [3]. These robots typically have used audio skills, and sometimes early visual room-reading skills, to understand audience perceptions of their performance. In-lab efforts show that robots can get away with using more disparaging jokes compared to human performers [4], and that comedic robots can yield more enjoyable social interactions [5], less awkward ice-breaking [6], and even better intimacy and trust in healthcare settings [7]. But what are the benefits of humorous robot repartee specifically, and which of these gains arise when quips are custom-chosen for the scenario at hand? These central research questions guided our robot comedy work.

Fig. 1. Mock interaction between the Misty robot and a human user in the controlled lab setting of the first presented study.

This paper presents the results of two studies that sought to compare human responses to robots with no repartee, hardcoded (i.e., pre-determined) repartee, and adaptive (i.e., actively selected based on human responses, using validated skills [8]) repartee. As further described in Section II, previous work focused on the related research question of sensor input to enable adaptation, but no experiments yet show the clear impact of robot repartee. Our methods, as highlighted in Section III, began with an in-person between-subjects user study that sought to highlight benefits of playful robot banter. We identified behavioral benefits of banter, but not quite as expected; the puzzle of why hardcoded quips were often favored as much as adaptive ones led to our follow-up online study with a clearer manipulation check, larger sample, and stronger within-subjects design (Section IV). We discuss the key findings and their implications in Section V. Overall, central contributions of this work include performance scripts that others can reuse in future robot comedy efforts, insights on the impacts of robot repartee, and possible explanations for why adaptive repartee is not an unconditional champion.

II. Related Work

Related work in the areas of robot humor and robot comedy guided the formation and design of our presented studies.

Manuscript received: August 23, 2024; Revised December 12, 2024; Accepted February 4, 2025.

This paper was recommended for publication by Editor Angelika Peer upon evaluation of the Associate Editor and Reviewers' comments.

[1]Collaborative Robotics and Intelligent Systems (CoRIS) Institute, Oregon State University, Corvallis, Oregon, USA. fittern@oregonstate.edu

* Indicates shared first authorship.

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

For example, amusing interactions have already had a positive impact on situations involving social robots. When individuals encounter a social robot for the first time, they tend to find joke-telling robots more likable and less socially awkward than robots without a sense of humor [4]. A similar outcome was observed in a study in which clever introductory jokes improved people's perceptions of a robot's intelligence compared to non-humorous interactions [9]. Additionally, when compared to jokes presented in text form, robots delivering jokes were found to be funnier, emphasizing the advantages of embodied humor [10]. In targeted contexts, such as healthcare, humor has even been shown to increase people's perception of a robot's likability, empathy, and perceived safety [7]. These prior results show that humor can enhance interactions with robots in scenarios from the playful to the serious, which led us to become interested in gaining a nuanced understanding of humorous robots.

Beyond simply using humor, the type of jest and timing of quips can be important to a comedic robot's success. In humorous small talk, users tended to prefer and have a more satisfying experience with robots that use humor in an ironic manner compared to their non-ironic counterparts [5]. In research focused on robot failures, mutually affirming humor prevented any reduction in competence ratings of a robot that was unsuccessful in its intended task [11]. Another past study showed that participants liked a robot that poked fun at itself more than a robot that jests at another robot's expense [12], although audiences also seem to tolerate robots that poke fun at one another as long as one has a positive air [13]. Given these past results, we considered tools such as irony, humor for relationship repair, self-deprecating humor, and even poking fun at other parties as methods for constructing our humorous robot's interactions.

Past work on robotic comedians has covered topics from apparent gender and form in robot comedy [14] to different joke-telling styles [4], [15] and nonverbal behaviors in robot comedy [2]. In the area most closely aligned with our work, past research teams have sought to incorporate audience feedback in to robots' comedy routines. For example, one team's robotic comedian responded to audience feedback via red and green paddles during live performances [1]. The authors of [13] implemented a street performance-style comedy act that interacted with passers-by and solicited audience input through show of hands. In our research team's prior work, we investigated using simple audio-based adaptation to tailor a robotic comedian's performance to the crowd's responses [3]. The research presented here assesses a system that uses face-reading computer vision techniques (first developed by our team in [8]) to select appropriate adaptations and analyze the impact of robot repartee.

III. In-Lab Study

This initial between-subjects study used the Misty II robot, as shown in Fig. 1, as a robotic actor. This rover-like representative social robot includes an animated face screen, appropriate speakers for playing joke audio, and a camera capable of sensing human facial expressions. We constructed a performance routine (as included in the methods below) based on puns with a simple setup-punchline structure, which afforded ample opportunities for the robot to succeed and fail at joke deliveries during the interaction; as evidenced in [3], initially unsuccessful deliveries are an important opportunity for using adaptation to improve audience reception. The methods and results from this first study follow.

A. Methods

This study's interactions with Misty and associated data collection were approved by the Oregon State IRB.

1) Study Design: The study compared three conditions. In each condition, the robot delivered the same jokes in a predetermined order, as further presented in the next subsection. The variations between conditions involved differences in repartee behaviors, as explained below:

- *Control condition:* the robot did not use any repartee after delivering each punchline.
- *Hardcoded condition:* the robot used a fixed set of preselected repartee, regardless of participant response to jokes, as delivered after each punchline.
- *Adaptive condition:* the robot used the output of a facial response classifier, taken directly from [8], to select and deliver the appropriate repartee following each punchline.

Condition presentation levels were balanced, and participants were matched to a condition using random assignment.

2) Comedy Routine: To present multiple distinct opportunities for our robotic comedian to succeed or fail during any given performance, we formulated the robot's routine around a series of one-liner jokes, most of which revolved around puns or other "dad-joke-like" statements, which are likely to lead to a mix of positive and negative reactions. The robot's routine, which appears in full in the paper's Appendix, allowed the possibility for positive and negative follow-up banter. For the hardcoded condition, this banter was predetermined by a random draw, and for the adaptive condition, the banter matched the classifier output for the participant's reaction (positive or negative) to each joke.

To help balance out the style and strength of comedy writing for each pair of repartee options, a professional comedian reviewed the performance script and supported rewrites as needed until satisfied with the end result. Work with this professional also helped us with the design choice to have the robot make strong decisions (i.e., avoid having a neutral category of robot repartee), in the style of many improvisational comedy schools of thought (e.g., the Upright Citizens' Brigade [16]). Although the classifier that we used as the basis for the robot's automatic response detection outputted a positive, neutral, or negative label, we grouped the neutral and negative responses together as negative for the purposes of the robot's logic, to achieve this strong decision-making (similarly to how a human performer might adapt to uncertain/tepid responses by trying to repair the interaction).

3) Procedure: Participants were recruited through university email lists, and they consented to the study before beginning the procedure. In the study, each participant faced a Misty II robot in an otherwise empty lab setting. The research assistant started out in the study space during the consent process, and left the room after launching the robot performance from

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

the laptop, returning when the performance was complete. Participants listened to the robot's comedy performance in the assigned condition and completed a survey. We recorded audio and video from the interaction, and the study self-reports are further detailed below. Each session lasted approximately 20 minutes, and participants were compensated $5 US for their time.

4) Measures: We used a mix of self-reported and behavioral measures to understand participants' experiences in the study, in addition to gathering demographics data.

The self-reports in the survey following the robot's performance helped us to understand elements of the interaction experience relevant to our research question. This survey included self-ratings of robot social attributes, relationship closeness feelings with the robot, and robot humorousness. Specifically, we administered:

- The *Robotic Social Attributes Scale (RoSAS)*, which gauges warmth, competence, and discomfort on 9-pt Likert scales from "Not at All Associated" (1) to "Very Much Associated" (9) [17].
- The *Inclusion of Other in the Self Scale (IOS)*, which measures social closeness on a 7-pt Likert scale with increasingly overlapping Venn diagram images for each scale point [18].
- The *joke rating scale (JRS)*, which gauges joke humorousness on a 7-pt Likert scale from "Strongly Disagree" (1) to "Strongly Agree" (7) [14].

This trio of inventories closely parallels the question set used in past related robot comedy work (e.g., [14]). Finally, we included a free-response field that allowed participants to share what aspects of the robot performance most strongly influenced their responses.

To understand behavioral responses to the robot, we also performed coding of participants' facial reactions to the robot after the delivery of each joke. In a process similar to those used in past robot comedy work (e.g., [3], [8]), three human raters with training in comedy performance and/or robot comedy labeled each post-joke reaction as positive, neutral, or negative. We calculated the majority vote of these three ratings by using the mode, or the neutral rating in cases when human rating spanned all three categories. To parallel the evaluations discussed later (for example, comparison of the robot's automatic labeling of human responses as positive or negative), our last code processing step was to collapse the negative and neutral categories from the majority vote labels into simply a negative category (i.e., performing the same grouping that the robot was doing in its program). Potential future work could preserve all three categories, once more key foundations of robot comedy have been established.

Demographic questions collected gender, age, STEM background, robotics experience (on a 5-pt Likert scale from "No Experience" [1] to "Expert-Level Experience" [5]), and comedy experience (on that same 5-pt experience scale).

5) Hypotheses: When beginning this work, we had the following two hypotheses:

H1: The addition of any robot repartee (hardcoded or adaptive) will lead to better survey and behavioral responses to the robot, compared to responses to the control condition.

H2: The adaptive condition will yield more positive survey and behavioral responses compared to the hardcoded condition.

These expectations arose from the past idea that post-joke quips can make a positive local influence in social interactions with a robot [3]. Further, we expected that repartee that fits the context of the crowd's reception would be received better than follow-up quips that did not necessarily fit.

6) Participants: 48 participants (31 female, 13 male, 2 genderfluid, and 2 non-binary) completed our study. Participant ages ranged from 18 to 68 years ($M = 32.9$, $SD = 14.8$). 25 participants had a STEM background. Participants had a little experience with robotics ($M = 2.3$ out of 5, $SD = 0.68$), some experience watching stand-up comedy ($M = 3.0$ out of 5, $SD = 0.98$), and almost no experience performing stand-up comedy ($M = 1.2$ out of 5, $SD = 0.52$).

7) Analysis: We tested for significant differences in the self-report data and the classifier accuracy data using one-way analysis of variance (ANOVA) tests with an $\alpha = 0.05$ significance level, reporting effect size with η^2. Although not all of the collected data was normal, throughout the manuscript, we use ANOVA tests even on non-normal data; a long history of work has shown that ANOVAs continue to operate robustly under violations of normality assumptions [19]. In interest of using the results of this study to inform prospective hypothesis generation, we also compared the descriptive statistics for each survey inventory across conditions. We used a Kruskal-Wallis test with an $\alpha = 0.05$ significance level to test for significant differences in the behavioral data, reporting effect size with ϵ^2. In the case of significant main effects, we used Tukey's HSD test to check for pairwise differences. Free-response input was used to help explain and contextualize other results.

B. Results

48 participants successfully completed the study; 15 experienced the control condition, 16 saw the hardcoded condition, and 17 viewed the adaptive condition. For two participants, one in the control condition and one in the hardcoded condition, we were unable to analyze facial expressions via video data because the participants wore masks. Since the robot was not reading the participants' faces in these cases, the corresponding survey data was still analyzed.

1) Self-Report Results: The distributions of responses for the RoSAS scales (i.e., warmth, competence, and discomfort), IOS, and JRS appear in Fig. 2. There were no significant differences in the condition-wise survey responses (all $p > 0.448$). Comparisons of the mean values (as shown by asterisks in the figure) revealed that the adaptive condition tended to elicit the highest ratings for warmth, competence, social closeness, and humorousness, as well as the lowest ratings (i.e., the best score, since higher is more uncomfortable) for discomfort. Hardcoded ratings also tended to be higher than control ratings for warmth and competence, although this condition did worse than the control condition in terms of comfort, social closeness, and humorousness. It is worth mentioning that some of these differences were small; for example, the competence

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

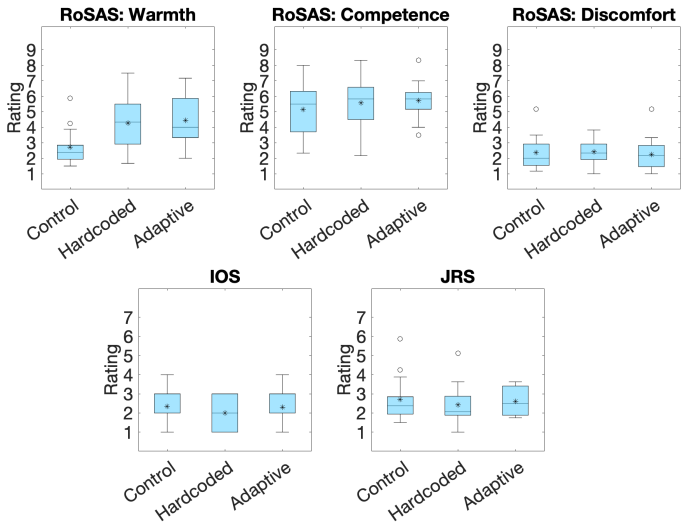

Fig. 2. Boxplots showing the ratings for each condition in the in-lab study (i.e., control, hardcoded, and adaptive). Whiskers extend to 1.5 times the interquartile range, circles are outliers, and asterisks show the mean.

rating means were similar between the control and hardcoded conditions, and the humorousness scores did not differ greatly across the three conditions.

2) Behavioral Data: In addition to using the survey-based self-reports, we could reason about participant experiences of the robot using their facial responses after the delivery of each joke. In this data, there was a significant main effect ($\chi^2(2) = 13.57$, $p = 0.001$, $\epsilon^2 = 0.037$). Pairwise comparisons showed that both the hardcoded and adaptive conditions yielded significantly more positive facial reactions than the control condition.

3) Face-Reading Algorithm Performance: When considering the lack of significant differences between responses to the hardcoded and adaptive robot performances, we wondered if limitations in the classifier performance could be partly to blame. Accordingly, we wanted to check how well the (basically random) selection of the hardcoded condition's repartee happened to match participants' post-joke facial reactions (as labeled by the human coder majority vote), in addition to how well the adaptive condition's labeling matched the human coder majority vote. We performed this first exploratory analysis for the actual hardcoded group and actual adaptive group from the study. The left plot in Fig. 3 shows the distributions of the robot response's match to the participant reaction; the ANOVA test showed a significant main effect ($F(1, 26) = 11.47$, $p = 0.002$, $\eta^2 = 0.306$), where the adaptive condition yielded higher accuracy.

We also wanted to check how well the hardcoded approach would have done across the full participant group, how well the adaptive approach would have done across the full participant group, and how well each human rater performed compared to the majority vote of all raters. This second exploratory analysis included all participants' data for all considerations (regardless of their true condition assignment). The results of this analysis help us to benchmark how well the classifier performs compared to skilled human raters, as well as how all approaches compare to the analogous hardcoded performance baseline. The right plot in Fig. 3 shows the performance of the hardcoded approach, classifier, and three human raters individually. The ANOVA test across these five performance distributions showed a significant main effect ($F(4, 225) = 9.94$, $p < 0.001$, $\eta^2 = 0.150$). Pairwise comparison tests showed that all other accuracies are higher than the hardcoded performance, but no other performance is significantly different from any other.

4) Summary of Key Results: Although there were no significant differences in the survey data (counter to the expectations of **H1** and **H2**), behavioral data showed significant gains from robot repartee (supporting **H1**), and the adaptive condition's survey data tended to show the best ratings (tentatively supporting **H2**). At the same time, there was no clear significant difference between the perception of the two conditions involving repartee (hardcoded and adaptive; counter to the **H2** expectations).

Some participant free-response feedback signaled a possibility that each condition was experienced as intended; for example, one hardcoded-condition participant mentioned that they "weren't sure if the robot was reacting to [their] lack of reaction or if the jokes were preplanned" while one adaptive-condition participant noted that "it was interesting how the robot could respond to whether you laughed at the previous joke or not." On the other hand, there were signs that any repartee was good repartee, such as one hardcoded-condition participant's comment that "the robot's responses after it delivered a joke contributed to my responses." We considered that one possible explanation for a similarity in experience across the two conditions with repartee would be poor performance from the facial reaction classifier; however, we assessed the classifier performance and found that the adaptive performance was significantly better than hardcoded coincidental matches in the study experiences, and that facial reaction classifier performance was similar to the human raters' performances.

Open questions about experience difference between the hardcoded and adaptive conditions led us to design the follow-up study described below.

IV. Online Study

The results of the in-person study did not show as clear of a picture of differences between the two conditions with robot repartee as we had sought to uncover in this work. Accordingly, we decided to conduct a follow-up study with a stronger within-subjects design. We wondered if adding a clearer manipulation check, collecting a larger sample, and moving the study context to watching a third-person view

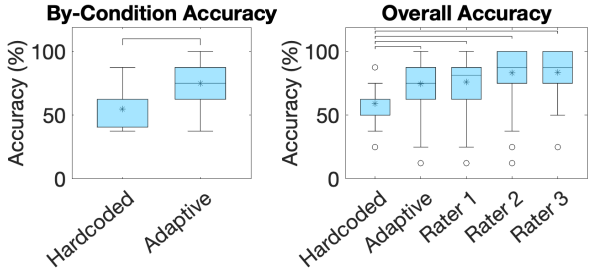

Fig. 3. Boxplots showing the accuracy of the match between robot repartee and correct label (left) and of the match between each robot/human label type and the majority-vote ground truth (right). Brackets represent significant pairwise differences.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

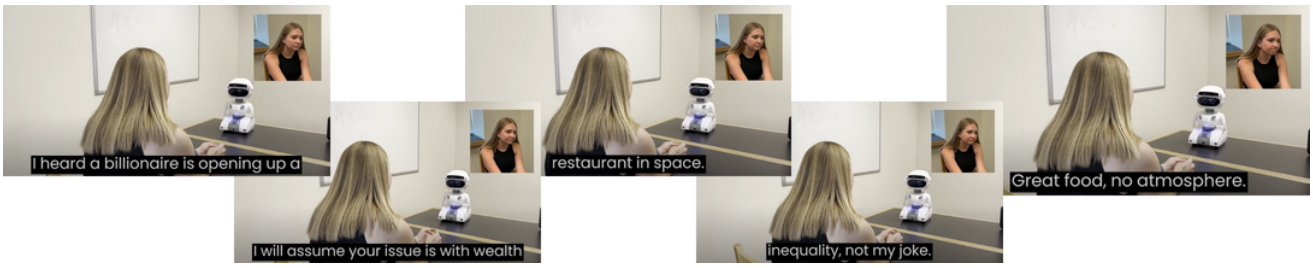

Fig. 4. Frames from the stimulus video of the adaptive condition variant of the billionaire joke. The third-person view shows the robot and human user together, and a picture-in-picture view shows the user's facial response to each joke. Captions show the joke transcription, which matches the joke audio.

of one-on-one actor-robot interactions (to provide a natural pathway to a plausible within-subjects design) could help us identify clearer impacts of each robot condition.

The resulting follow-up online study had a within-subjects design and used the same Misty II robot as the robotic social agent. The study used a smaller number of overall jokes (to ensure feasible ongoing attention by remote participants throughout the study) and cycled through presentation of all three conditions in a balanced fashion, as further described below. The methods and results of this study follow.

A. Methods

The below methods for this online, within-subjects study's interactions with Misty stimuli and associated data collection were approved by the Oregon State IRB.

1) Study Design: This study involved the same three conditions as the first study: *control*, *hardcoded*, and *adaptive*.

We generated third-person-view video stimuli in each of these conditions for three of the jokes from the comedy routine presented in the Appendix (i.e., *Billionaire*, *Chicken*, and *Identity*) with just one minor wording tweak: the removal of the word "again" from the *Chicken* joke, since it was now possible for this joke to appear first. For all stimuli in the hardcoded and adaptive conditions, we used the robot's negative repartee, as a way to accentuate the room-reading and interplay. In the video stimuli, a human actor was visible interacting with Misty; the main image showed this person in the same space as Misty, and a picture-in-picture view throughout the video displayed the actor's facial reaction (see keyframes in Fig. 4). For hardcoded stimuli, the actor visibly enjoyed the joke, yet the robot selected the negative repartee. For adaptive stimuli, the robot banter matched the actor's visible negative reaction. This stimulus design sought to show the most extreme possible benefit of the adaptive condition, as a starting point for future work.

To mitigate the influence of ordering effects and influences of specific follow-up quips, we used random assignment to balance which joke was presented in which condition, as well as the order of stimulus presentation, across participants. Each participant saw three videos overall, covering each condition and each joke within the set.

2) Procedure: The study was self-contained within a Qualtrics survey that we released using the Prolific platform. At the start of the survey, participants consented to participate and answered questions about their own robotics and comedy experience. Next, respondents saw a video of the Misty robot saying "Hello. I am Misty the robot. Welcome." and completed beginning perception questions about the robot, as further described in the following subsection. After that, participants viewed the first of their stimulus videos and completed a post-stimulus questionnaire (as further described in the next subsection). This process repeated two more times, with the viewing of the second video, the completion of the post-stimulus question set, the viewing of the third video, and the completion of the post-stimulus question set. In the closing portions of the survey, respondents answered the same robot perception questions as at the start, completed one free-response question, answered an attention-check question, and finished the manipulation check questions. Participants on Prolific received $6 US for completing this approximately 25-minute survey.

3) Measures: We used scale-wise self-report questions and free-response questions to understand participant perceptions of Misty, attention and manipulation check questions to confirm the study was experienced as intended, and demographic questions to understand respondent characteristics.

The self-reports before all stimuli and after all stimuli came from the Object Centered Sociality Factors scales for cultural context, grouping, reciprocity, and attachment [20].

In the post-stimulus question sets, we administered:

- The *RoSAS*, as described in the previous study.
- The *IOS Scale*, as described previously.
- The *JRS*, as described previously.
- The *Godspeed* question set on anthropomorphism, which uses a set of semantic differential subscales [21].

After each stimulus, we also administered a free-response question about what aspects of the stimulus videos most strongly influenced participant responses.

The attention-check question was a single fact-based multiple choice question about a provided nature-themed image. We checked for understanding of the condition experiences by presenting the same three stimulus videos to the participant again and asking "Did the robot appear to understand the human's response?" with yes/no options after each one. At the end of the manipulation check questions, we administered a final free-response question asking "What factors influenced your impression that the robot understood or did not understand the human responses?"

We gathered the same demographics as in the last study.

4) Hypotheses: Our hypotheses in this effort matched the expectations for survey responses in the in-person study.

5) Participants: The 99 participants (59 male, 40 female) ranged in age from 21 to 71 ($M = 39.2$, $SD = 12.4$). 51 participants had a STEM background. Participants had some

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

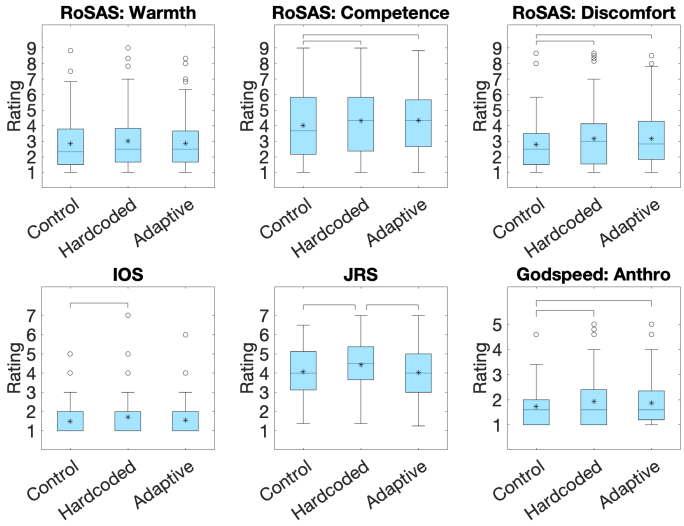

Fig. 5. Boxplots showing the ratings for each condition in the online study (i.e., control, hardcoded, and adaptive).

experience with robots generally ($M = 2.2$ out of 5, $SD = 0.78$), a fair amount of experience watching comedy ($M = 3.6$ out of 5, $SD = 0.781$), and very little experience performing comedy ($M = 1.56$ out of 5, $SD = 0.971$).

6) Analysis: In our analysis of RoSAS, IOS, JRS, and the Godspeed survey responses, we checked for significant differences using a repeated measures ANOVA (rANOVA) test with an α = 0.05 significance level. When significance was found, we used a Tukey post hoc test to assess pairwise differences. The effect size was computed using η^2. In the analysis of the Object Centered Sociality Factors scale data, we compared the initial and final findings using paired-samples t-tests with an α = 0.05 significance level, computing effect size with Cohen's d. We used free-response input to help explain and contextualize other results.

B. Results

All 99 participants successfully completed the study, including correctly answering the attention check question. Among the respondents, 83 correctly identified that the hardcoded condition was not aligned with the actor's response, and 84 correctly identified that the adaptive condition was aligned with the actor's response. Accordingly, our impression was that the participants could largely perceive when the robot was acting adaptively, or at least that a lack of ability to distinguish between non-adaptive and adaptive behavior was not the reason for limited differences between the conditions with repartee in the in-lab study.

1) Post-Stimulus Self-Report Results: The distributions of responses for the RoSAS scales (i.e., warmth, competence, and discomfort), IOS, JRS, and Godspeed anthropomorphism scale appear in Fig. 5. In the rANOVA test results, there were no significant main effects for warmth ($p = 0.150$), but there were significant main effects for competence ($F(2, 196) = 3.94$, $p = 0.021$, $\eta^2 = 0.039$) and discomfort ($F(2, 196) = 6.18$, $p = 0.003$, $\eta^2 = 0.059$). The post hoc tests revealed greater levels of perceived competence ($M = 4.33$ for hardcoded, $M = 4.36$ for adaptive) but also greater levels of discomfort ($M = 3.18$ for hardcoded, $M = 3.18$ for adaptive) for both the hardcoded and adaptive conditions, compared to the control condition ($M = 4.03$ for competence, $M = 2.79$ for discomfort). There were significant main effects for interpersonal closeness ($F(2, 196) = 5.48$, $p = 0.005$, $\eta^2 = 0.053$); the hardcoded condition ($M = 1.71$) led to more feelings of closeness than the control condition ($M = 1.49$). For humorousness, there were also significant main effects ($F(2, 196) = 16.2$, $p < 0.001$, $\eta^2 = 0.021$). Pairwise tests showed that the hardcoded condition ($M = 4.44$) was more humorous than any other condition ($M = 4.07$ for control, $M = 4.02$ for adaptive). Lastly, there was a significant main effect for anthropomorphism ($F(2, 196) = 7.32$, $p < 0.001$, $\eta^2 = 0.070$), where hardcoded ($M = 1.93$) and adaptive ($M = 1.86$) conditions appeared more anthropomorphic than the control condition ($M = 1.73$).

TABLE I
DESCRIPTIVE STATISTICS OF BEGINNING- AND END-OF-STUDY RATINGS ON THE OBJECT CENTERED SOCIALITY FACTORS SCALES. RESULTS ARE REPORTED AS $M \pm SD$. FOR SIGNIFICANT DIFFERENCES, THE RELATIVE "BEST" VALUE IS BOLDED FOR EMPHASIS.

	Beginning	Final
Cultural Context	4.59 ± 0.99	4.55 ± 1.16
Forms of Grouping	**5.00 ± 1.02**	4.70 ± 1.23
Reciprocity	**2.86 ± 1.28**	2.62 ± 1.33
Attachment	**3.35 ± 1.35**	2.71 ± 1.52

We also wanted to assess whether omitting participants who failed the hardcoded and/or adaptive condition manipulation checks led to similar or different overall results. Re-running the same rANOVA and post hoc tests on the data from participants who got both manipulation check questions correct showed most of the same significant differences as the tests on the full participant set. All of the main effects were the same aside from the IOS results, which were not significant in this exploratory follow-up test ($p = 0.063$). The pairwise significant differences remained almost completely the same; just the competence rating difference between the control and hardcoded conditions was lost.

2) Pre/Post Sociality Factor Results: The paired-samples t-tests on the four Object Centered Sociality Factors subscales showed no significant differences for the cultural context scale ($p = 0.592$), but did uncover significant main effects for forms of grouping ($p = 0.001$, $d = 0.339$), reciprocity ($p = 0.012$, $d = 0.258$), and attachment ($p < 0.001$, $d = 0.637$). On all of the significant scales, ending ratings were lower than starting ratings, as shown in Table I.

3) Summary of Key Results: Overall, the follow-up online study reinforced the idea that there are selected significant differences in perception of a robot that does use repartee compared to one that does not, but it did not fully resolve the question of whether or why an adaptive approach might offer extra gains compared to a hardcoded approach, at least within the relatively short timescale of this study.

Repartee seemed to enhance apparent competence and anthropomorphism of the robot performance (in alignment with **H1**), but actually decreased viewers' levels of comfort (counter to the hypothesis). Participant input showed that the repartee was noticed for both the hardcoded condition (e.g., "sometimes [the actor] laughed and [the robot] would still be like 'oh you

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

didn't like the joke"') and adaptive condition (e.g., "[the robot] gave off funny snide remarks after the initial joke," "I did like how it seemed to react to whether or not the person was laughing"). Discomfort most likely arose from the writing of the follow-up repartee itself; one participant mentioned that "a few comments could be perceived as predatory or creepy" and another similarly "thought it was a little creepy the way [the robot] said it would steal her packages if she did not laugh."

The hardcoded performance tended to lead to the highest feelings of closeness and ratings of humorousness (somewhat in tension with **H2**). The free-response input from participants highlighted failing, being awkward, or malfunctioning as one reason for the potential success of this condition in particular (e.g., "[the robot was] awkward with its jokes, which made them funnier"). General opinions of the robot mostly declined between the beginning and ending Object Centered Sociality Factors questions. This could be because of a mismatch in expectations vs. reality of robotic systems; for example, one participant commented that the robot "seemed like something that would have been fun in the 90's," despite the modern technology underlying the robotic system. Overall, it seems like public opinions of social robot capabilities may be inflated, and that although some gains can come from repartee generally, the slapstick experience of a failing robot can be more entertaining (at least on a short timescale) compared to a robot with fitting banter.

V. Discussion

The presented work includes two studies that sought to enhance the understanding of robot repartee for future human-robot interaction. Despite the intuitive benefit of this type of adaptation and past efforts to understand this robot behavior, little is empirically known about the best way to apply robot banter. In our in-lab study, **H1** was partially supported by the behavioral results, in which any type of repartee improved reactions to the robot. At the same time, survey input did not reveal significant aligned results (or, for that matter, any significant differences between conditions). **H2** was not supported; no significant differences arose between the two repartee conditions in either the survey or the behavioral results. We performed a follow-up online study to accomplish a stronger within-subjects design and collect a larger sample. This effort yielded more significant differences, especially between the control and other conditions. **H1** was partially supported by the follow-up study, in which both repartee conditions were rated as significantly more competent and anthropomorphic than the control. Again, **H2** was not supported in the follow-up effort. No results showed the adaptive condition to be more effective, and one comparison even showed the hardcoded condition to be more humorous.

The results of this work, although mixed, represent an advance in the state of robot comedy knowledge. In closely related past work, the main adaptive behavior that has been shown to significantly benefit performances is good timing [3]. This past paper considered post-joke quips as another possible way to gain crowd favor, but failed to identify any significant benefit of repartee. Similarly, the beginning work on computer vision-informed repartee in [8] included pilot experiments that showed trending, but no significant impact of banter. The results presented in our paper help to expand the state of knowledge on robot repartee by using stronger experiment designs to show conclusive evidence that banter (whether matched to human reactions or not) can serve to enhance an interaction; we see this in the behavioral results of one study, and in the self-reports of the other. This insight means that especially in ad hoc interactions, allowing a robot to make a bold dialog decision (even in the case of high sensing uncertainty) is likely to enhance the interaction. Whether based on the entertainment value of a malfunctioning robot or other reasons, incorrect and correct repartee choices alike seem to both enhance human perceptions of a robot in this situation. Over longer time periods of interaction, we suspect that the humorousness of hardcoded banter may degrade faster than adaptive. For example, although a "broken" robot can at first seem humorous, it is likely to soon become frustrating. On the other hand, a robot that jests fittingly may be appreciated for this apparent personality over time. Future work can consider humor crossed with interaction duration.

This work was not without limitations. For example, an in-person replication of the online follow-up study would enhance the work's ecological validity; people may not respond to robots in the same way between online and in-person settings. The diversity of participants, especially in the initial study (which had predominantly younger participants), could also be improved to better represent the general public. Our design decision to respond to negative and neutral reactions with snarky repartee has implications, which may be good or bad depending on context. Although most participants in the follow-up study passed the manipulation check, not all participants did, which indicates a need for clearer design of the robot adaptation. Further, the qualitative results hint that certain repartee contributed to the observed discomfort ratings in the follow-up study. Additionally, adaptive repartee could possibly lead to delight in some cases and a feeling of creepiness in others. We encourage those who seek to build on this work to update the robot's script to maximize the benefit and minimize the drawbacks of robot repartee.

VI. Conclusion

The presented work included one in-lab and one online study that sought to identify self-reported and behavioral gains from adaptive robot repartee. Taken together, our results (which show the benefits of follow-up quips generally, although not necessarily adaptive quips in ad hoc interactions) can inform others who are interested in equipping social robots with repartee abilities. The joke scripts provided in this work and the principle of making strong decisions can support further progress in this domain. At the same time, humor is complicated and robot-delivered comedy is not yet perfect. We believe that this work can serve to inform a next generation of robotic companions with more charm and effectiveness, in part powered by playful banter.

Acknowledgements

We thank Lily Oliphant, Christopher A. Sanchez, and Kyler Jones for their help and feedback.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

REFERENCES

[1] H. Knight, S. Satkin, V. Ramakrishna, and S. Divvala, "A savvy robot standup comic: Online learning through audience tracking," in *Proc. of the Int. Conf. on Tangible and Embedded Interaction (TEI)*, 2011.

[2] K. Katevas, P. G. Healey, and M. T. Harris, "Robot comedy lab: Experimenting with the social dynamics of live performance," *Frontiers in Psychology*, vol. 6, p. 1253, 2015.

[3] J. Vilk and N. T. Fitter, "Comedians in cafes getting data: Evaluating timing and adaptivity in real-world robot comedy performance," in *Proc. of the ACM/IEEE Int. Conf. on HRI*, 2020, pp. 223–231.

[4] B. T. Tay, S. C. Low, K. H. Ko, and T. Park, "Types of humor that robots can play," *Computers in Human Beh.*, vol. 60, pp. 19–28, 2016.

[5] H. Ritschel, I. Aslan, D. Sedlbauer, and E. André, "Irony man: Augmenting a social robot with the ability to use irony in multimodal communication with humans," in *Adaptive Agents and Multi-Agent Systems*, 2019.

[6] M. Tae and J. Lee, "The effect of robot's ice-breaking humor on likeability and future contact intentions," in *Companion of the ACM/IEEE Int. Conf. on HRI*, 2020, p. 462–464.

[7] D. L. Johanson, H. S. Ahn, J. Lim, C. Lee, G. Sebaratnam, B. A. MacDonald, and E. Broadbent, "Use of Humor by a Healthcare Robot Positively Affects User Perceptions and Behavior," *Technology, Mind, and Behavior*, vol. 1, no. 2, 2020.

[8] C. C. Gray, "Toward machine learning-enabled adaptivity for humorous robots," 2022.

[9] I. M. Menne, B. P. Lange, and D. C. Unz, "My humorous robot: Effects of a robot telling jokes on perceived intelligence and liking," in *Proc. of the ACM/IEEE Int. Conf. on HRI*, 2018, pp. 193–194.

[10] J. Sjöbergh and K. Araki, "Robots make things funnier," in *Proc. of the Conf. of the Japanese Society for AI*. Springer, 2008, pp. 306–313.

[11] H. N. Green, M. M. Islam, S. Ali, and T. Iqbal, "Who's laughing NAO? Examining perceptions of failure in a humorous robot partner," in *Proc. of the ACM/IEEE Int. Conf. on HRI*, 2022, pp. 313–322.

[12] N. Mirnig, S. Stadler, G. Stollnberger, M. Giuliani, and M. Tscheligi, "Robot humor: How self-irony and Schadenfreude influence people's rating of robot likability," in *Proc. of the IEEE Int. RO-MAN Symp.*, 2016, pp. 166–171.

[13] J. Swaminathan, J. Akintoye, M. R. Fraune, and H. Knight, "Robots that run their own human experiments: Exploring relational humor with multi-robot comedy," in *Proc. of the IEEE Int. RO-MAN Conf.*, 2021, pp. 1262–1268.

[14] N. Raghunath, C. A. Sanchez, and N. T. Fitter, "Robot comedy (is) special: A surprising lack of bias for gendered robotic comedians," in *Proc. of the Int. Conf. on Social Robotics*, 2022, pp. 663–673.

[15] T. Kishi, N. Endo, T. Nozawa, T. Otani, S. Cosentino, M. Zecca, K. Hashimoto, and A. Takanishi, "Bipedal humanoid robot that makes humans laugh with use of the method of comedy and affects their psychological state actively," in *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2014, pp. 1965–1970.

[16] M. Walsh, I. Roberts, and M. Besser, "Upright Citizens Brigade comedy improvisation manual," *Comedy Council of Nicea LLC*, 2013.

[17] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (RoSAS) development and validation," in *Proc. of the ACM/IEEE Int. Conf. on HRI*, 2017, pp. 254–262.

[18] A. Aron, E. N. Aron, and D. Smollan, "Inclusion of other in the self scale and the structure of interpersonal closeness," *Jrnl. of Personality and Social Psych.*, vol. 63, no. 4, 1992.

[19] M. J. Blanca Mena, R. Alarcón Postigo, J. Arnau Gras, R. Bono Cabré, and R. Bendayan, "Non-normal data: Is ANOVA still a valid option?" *Psicothema, 2017, vol. 29, num. 4, p. 552-557*, 2017.

[20] A. Weiss, R. Bernhaupt, M. Tscheligi, D. Wollherr, K. Kuhnlenz, and M. Buss, "A methodological variation for acceptance evaluation of human-robot interaction in public places," in *Proc. of the IEEE Int. RO-MAN Symp.*, 2008, pp. 713–718.

[21] C. Bartneck, E. Croft, and D. Kulic, "Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots," *Int. Jrnl. of Social Robotics*, 2008.

APPENDIX

For the in-lab study, the robot's comedy routine was as follows, where in all conditions, participants heard the setups and punchlines of jokes (the numbered list items). The bulleted sub-lists represent the possible follow-up positive and negative repartee; for hardcoded performances, the robot used the bolded follow-up quip.

1) *Intro:* Hello, I am Misty the Comedian. Welcome to my show!
2) *Billionaire:* I heard a billionaire is opening up a restaurant in space... Great food, no atmosphere!
   - Positive: Call me Roomba, because that joke was about a vacuum!
   - **Negative:** I will assume your issue is with wealth inequality, not my joke.
3) *Butter:* Did you hear that rumor about butter? Well, I am sure not going to spread it!
   - Positive: With smiles like those, are you trying to butter me up?
   - **Negative:** I will, however, spread rumors about your poor sense of humor.
4) *Karate:* I beat the local Chess champion in five moves... The last one was a roundhouse kick!
   - Positive: Huzzah, another Cobra Kai enthusiast!
   - **Negative:** You didn't like that one. Care for a game of chess?
5) *Chicken:* Which came first, the chicken or the egg? Well, I ordered both on Amazon so I will let you know!
   - **Positive:** Cluck, cluck, thank you!
   - Negative: If you don't laugh again, I might steal your packages.
6) *Eclipse:* How does the man on the moon cut his hair? Eclipse it!
   - Positive: Move over, Elon, there is a new space robot in town!
   - **Negative:** I guess the only Eclipse you know is in the Twilight saga.
7) *Identity:* When I was a child, my mother told me I could be whoever I wanted to be... But it turns out, identity theft is a crime!
   - Positive: It was worth the prison sentence to see that smile!
   - **Negative:** You didn't like that one. Hey, what is your mother's maiden name?
8) *Train:* Did you hear about the Starship robot that got hit by the train? I guess it didn't have enough train-ing data!
   - **Positive:** Personally, I would just dodge the train!
   - Negative: You must be one of the people that didn't get their food.
9) *Sprinter:* What do sprinters eat before a race? Nothing, they fast!
   - **Positive:** A speedy laugh for a quality gaffe!
   - Negative: Perhaps you should switch to shotput.
10) *Outro:* Thanks for listening! That's my show. I am Misty the Comedian. Have a great day!

The *Train* joke mentions Starship robots: wheeled service robots that deliver food on many college campuses. A past video of one of these robots getting hit by a train went viral.