Getting to Know Each Other: The Role of Social Dialogue in Recovery from Errors in Social Robots

Gale M. Lucas, Jill Boberg, David Traum, Ron Artstein, Jonathan Gratch, Alesia Gainer, Emmanuel Johnson, Anton Leuski
USC Institute for Creative Technologies
Los Angeles, USA
United States
lucas@ict.usc.edu

Mikio Nakano
Honda Research Institute Japan Co., Ltd.
Wako, Saitama
Japan
nakano@jp.honda-ri.com

ABSTRACT

This work explores the extent to which social dialogue can mitigate (or exacerbate) the loss of trust caused when robots make conversational errors. Our study uses a NAO robot programmed to persuade users to agree with its rankings on two tasks. We perform two manipulations: (1) The timing of conversational errors – the robot exhibited errors either in the first task, the second task, or neither; (2) The presence of social dialogue – between the two tasks, users either engaged in a social dialogue with the robot or completed a control task. We found that the timing of the errors matters: replicating previous research, conversational errors reduce the robot's influence in the second task, but not on the first task. Social dialogue interacts with the timing of errors, acting as an intensifier: social dialogue helps the robot recover from prior errors, and actually boosts subsequent influence; but social dialogue backfires if it is followed by errors, because it extends the period of good performance, creating a stronger contrast effect with the subsequent errors. The design of social robots should therefore be more careful to avoid errors after periods of good performance than early on in a dialogue.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)** → Empirical studies in HCI

KEYWORDS

Social robots, influence, social dialogue, rapport, errors

ACM Reference format:

G. M. Lucas, J. Boberg, D. Traum, R. Artstein, J. Gratch, A. Gainer, E. Johnson, A. Leuski and M. Nakano. 2018. Getting to Know Each Other: The Role of Social Dialogue in Recovery from Errors in Social Robots. In HRI '18: 2018 ACM/IEEE International Conference on Human-Robot Interaction, March 5–8, 2018, Chicago, IL, USA. ACM, NY, NY, USA, 9 pages. https://doi.org/10.1145/3171221.3171258

ACM ISBN 978-1-4503-4953-6/18/03...$15.00.
http://dx.doi.org/10.1145/3171221.3171258

1 INTRODUCTION

Human-Robot interaction involving multiple tasks is becoming more common. New tasks often initially have high rates of communication errors, especially until sufficient training data can be gathered. This creates issues for designers. First, if an interaction will consist of multiple tasks at different stages of development (and expected amount of errors), what should the ordering of those tasks be? Also, can other activities ameliorate (or exacerbate) the effects of errors? For example, robots can engage in activities to establish rapport -the feeling of having a close and harmonious relationship- with human users [1]. This paper investigates the interplay between rapport and conversational errors - both how errors affect rapport with robots, and how rapport mitigates (or exacerbates) the effects of errors. This is part of a longer-term goal which seeks to understand how rapport-building in conversations with robots and agents can improve or worsen conversational outcomes, and under what circumstances this occurs.

Previous work has shown that agents that make conversational errors are less capable of influencing people than agents that do not make errors [2,3]. Some work has shown this effect with robots [4,5], but prior research has not considered additional factors such as the level of rapport felt with the robot. Our hypothesis was that building rapport through a social dialogue (such as an ice-breaker) could mitigate the detrimental effect of a robot's errors on its ability to influence people. However, we also considered the effect that the timing of the errors might have. Previous research suggests that errors occurring after a period of good performance are more harmful to influence than those that occur earlier [6]. Accordingly, rapport built through an ice-breaker may not be enough to mitigate the detrimental effect of errors occurring in a later interaction.

We consider two factors: the presence and timing of errors, and the presence or absence of an ice-breaker dialogue (3x2 design). Impact is measured using both objective measures of social influence and subjective measures of the participants' perceived rapport with the robot. We find that the ice-breaker can be used to repair rapport that had been lost through conversational errors, but that pre-existing rapport exacerbates the effects of errors occurring after rapport had been established. The findings have practical implications for the design of systems for human-robot interaction. The utility of many social and assistive robots is potentially determined by their ability to build relationships and influence their human companions. The remainder of the paper presents background on influence, rapport and conversational errors; describes the experiment; and presents the results and conclusions.

2 RELATED WORK

Conversational robots and agents can be used to assist on tasks. They have seen such use as being companions to senior citizens [7,8], serving as assistants in schools [9,10], engaging as home-based social robots [11], helping with rehabilitation [12], and assisting with disaster relief [13], among many others.

2.1 Social dialogue

Dialogue system research often makes a distinction between *functional* or *task-oriented* dialogue, aimed at joint completion of a specific task, and *relational* or *social* dialogue, aimed at building a relationship between the participants. While some dialogue systems for robots and agents attempt to engage in only one of these types, others have included both relational and functional aspects. For example, both kinds of dialogue were used in the REA system [14], as well as the SASO system [15]. There are different types of functional dialogue – some functions relate to information-seeking or service tasks (assistant systems), but others focus on influencing the user. Examples include assistive systems to motivate people to exercise [16] or to negotiate on a course of action [15,17]. When influence is an objective, it is important that the user feels both comfortable getting input from the system and acting on that input. This trust is a critically important factor when attempting to create and maintain a relationship with a robot or agent [18,19].

2.2 Rapport and persuasion

Robotic agents are able to persuade humans to change their behavior. For example, participants who received social feedback from an embodied agent were more likely to conserve energy [20]. Researchers have also investigated the influence of robots' non-verbal messages such as demeanor and appearance [21], gender [22], gazing and gestures [23], verbal messages such as reciprocity [24], and indirect language [25], or both messages and reciprocity [26] on their ability to persuade a human user.

One factor that has been shown to facilitate relationship-building and interaction is rapport [27]. Rapport between two individuals creates interpersonal responsiveness and influence [28], so when rapport is created, research suggests that it can improve the quality of customer-employee exchanges [29], therapist-patient relationships [30], college roommate relationships [31], teacher-student interactions [32], relationships between caregivers and their charges [33], as well as other relationships. Similarly, rapport between agents and humans can build a relationship between the two, as well as help agents to influence human behavior. Some researchers have studied ways that robots, in particular, can increase persuasiveness. One study found that verbal cues of expertise, for example, increase the persuasiveness of a robot [34], while another found that robot and human teams that engaged in argumentation-based dialogue for shared decision-making were more successful than those where the robot engaged only in supervisory dialogue [35]. In one study, participants who received social feedback were influenced to use less electricity in their homes than those with strictly factual feedback [20].

One way to build rapport is through an ice-breaker. The concept of ice-breaking to increase trust and build rapport between group members is a common one, used often in education and organizational fields [36–38] and in previous robot studies [39]. Social relational techniques such as small talk, self-disclosure (intimacy), expert's jargon (credibility), gossip (social networks), and politeness (dignity/prestige) [14, p.91] used by humans to build trust could also be used by robots in guidance roles in which it is crucial that the robot gains the user's trust. This is within the capabilities of a robot – studies have shown that users are responsive to robots' and agents' verbal and non-verbal cues, social language, conversational gestures, and mutual gaze [14,40]. For example, a study using a robot-receptionist found that she was able to build rapport with users by talking about her backstory [41].

2.3 Conversational errors and trust

For the foreseeable future, social robots will make occasional conversational errors. Designers who build such robots have to decide *whether* and *when* a new task that is expected to have a relatively high error rate should be given to the user. First, to the question of *whether* such a task should be given to users, work with virtual agents has shown that errors in dialogue systems can reduce social influence [2,3]. Evidence with robots is mixed. Within a human-robot team, trust is found to be most influenced by robot characteristics such as reliability [18] and trust is easily lost but not easy to regain [6]. Other work finds that errors do not impact robots' influence – for example, one study found that errors affected the perceived reliability and trustworthiness of the robot, but had no impact on users' willingness to comply with the robot's requests [19]. It is possible, however, that the nature of this task lent itself to sustained compliance with the robot's request in spite of these errors, which occurred later in the task. In contrast, when other tasks are used (as in [4,5,42]), errors that occur after periods of good performance interfere with robots' influence. As such, in the present work, we use a standardized task to measure influence.

Second, to the question of *when* tasks likely to have high error rates should be included, studies have found that drops in reliability after a period of good performance were much more harmful to trust and performance than early failures, after which trust could be at least partially recovered [4,5] For example, the reliability of a robot that shifted from 100% reliable to 80% reliable was rated lower than a robot that was 80% reliable from the start [6].

However, research has yet to consider the impact of robots' social dialogue on *mitigating* the impact of errors. Accordingly, we designed a study to test the possibility that building rapport through a social dialogue (e.g., ice-breaker) could mitigate the detrimental effect of errors on influence. It is possible that this benefit of social dialogue could be realized if it occurs before the errors, buffering against errors occurring later in the dialogue. Yet, there is reason to believe that an ice-breaker *might not* benefit influence if errors occur later; an error-free initial task plus an ice-breaker would amount to extended period of good performance, which (as mentioned above) has been shown in other work to render later errors more detrimental [41]. If participants might experience a loss of trust when the errors begin after such a long period of good performance, rapport built in the ice-breaker may not be enough to mitigate the detrimental effect of errors.

3 EXPERIMENT

3.1 Design

To test for the effect of errors both before and after the social dialogue, we had each participant engage in two tasks with the robot; errors could be in the first task, the second task, or in neither task. In between the tasks was either an ice-breaker dialogue (with no errors) or a control task. Thus, the experiment manipulated two factors in a 3 (error: early errors, late errors, or no errors) x 2 (ice-breaker: ice-breaker or control) design. Each participant was assigned to one of these 6 conditions.

3.2 Participants

We recruited 165 participants (51.3% female) from Craigslist for this experiment; they were compensated $30 for their participation. Participants' age ranged from 18 to 84 with a mean of 29.2 years and standard deviation of 18.2. Participants were 29.6% African American, 10.1% Asian or Pacific Islander, 39.6% Caucasian, 10.1% Hispanic, 5.0% mixed race and 5.6% other race. Concerning experience with robots, 78.1% had not interacted with a robot before, whereas 19.5% had (and 2.4% chose not to answer). The data for 11 users had to be excluded due to technical issues experienced during the session. Analyses were conducted on the remaining 154 participants.

3.3 Procedure

Participants were seated at a table facing a NAO humanoid robot. Audio for the control task came from small speakers under the camera. Participants wore a close-talking microphone which recorded their speech throughout the interaction, and a video camera recorded their face and upper body movements. A second video camera behind the participant recorded the robot and its movements. To avoid distraction, the robot was covered when not in use. The participants used an iPad Pro for ranking items in the tasks, answering survey questions, and reading instructions on each phase of the experiment.

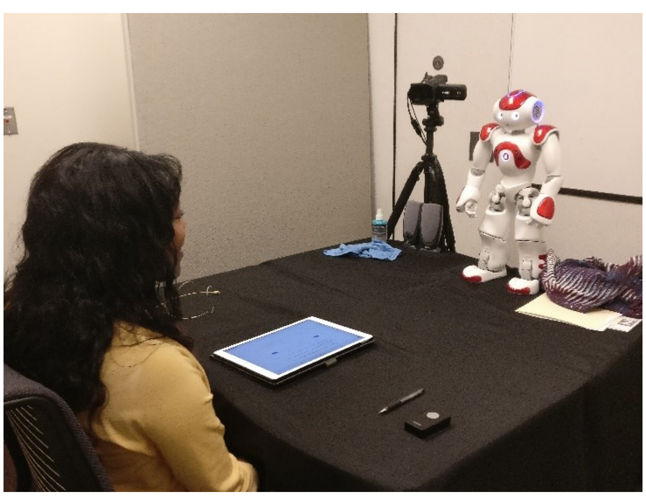

Figure 1: *The NAO robot discussing rankings with a user.*

Participants started with the Lunar Survival Task, a problem-solving task widely used for measuring persuasion [42–44]. In this task, participants were asked to imagine that they are part of a space crew that crashed on the moon, and were asked to rank 10 items as to their importance for surviving long enough to be rescued. Participants first ranked these items individually on the iPad Pro, and were then told that they should rank them again with the help of another crewmember, our robot. At that time, the robot was uncovered and activated, standing up from a crouched position, and participants engaged in dialogue with the robot (Fig. 1). Although participants were led to believe that the robot was autonomous, in fact it was controlled by a human operator ("Wizard of Oz") and acted as a confederate [45], providing factual arguments for ranking the items in a specific order. The robot had a fixed set of arguments, all supporting a specific ranking order, but each argument was used as needed in the conversation with a given participant. Following the dialogue, participants re-ranked the items; the differences between initial rankings and final rankings served as a measure of influence [46]. The robot was again deactivated and covered, and the participant reported how much rapport they currently felt with the robot by filling a questionnaire (Fig. 2). This scale was designed and used previously to measure the feeling of having a close and harmonious relationship, even after a single interaction [47].

In the early error condition, the robot made a series of errors while interacting with the participant on the Lunar Survival Task. Errors were introduced in one of several ways: asking users to repeat themselves, answering a different question than the user had asked, repeating the answer to the previous question, answering with a question or non-sequitur, or not answering at all. Errors were introduced into the dialogue according to a set order at a rate of about one of these errors per two utterances (Fig. 3). In most cases, the entire list was not used, as the task was finished after about 10 errors. The late error and no error condition did not include conversational

Each question is rated on a 5-point scale. R indicates reverse-coded items
Niki created a sense of closeness or camaraderie between us.
Niki created a sense of distance between us (R).
I think that Niki and I understood each other.
Niki communicated coldness rather than warmth (R).
Niki was warm and caring.
I wanted to maintain a sense of distance between us (R).
I felt I had a connection with Niki.
Niki was respectful to me.
I felt I had no connection with Niki (R).
I tried to create a sense of closeness or camaraderie between us.
I tried to communicate coldness rather than warmth (R).

Figure 2: *Rapport questionnaire*

"Can you repeat that? I didn't understand you"
Wrong answer
No answer
"I'm sorry, I didn't hear you"
Repeat and/or interrupt
Answer with a question or a non sequitur
"Can you repeat that? I didn't understand you" (× 3)
Wrong answer
No answer
"I'm sorry, I didn't hear you" (× 3)
Wrong Answer
Repeat/interrupt
Answer with question or non sequitur
"Can you repeat that? I didn't understand you" (× 2)
No answer

Figure 3: *Error manipulations, in order used.*

errors in the Lunar Survival Task. See Fig. 4 for an example of part of a dialog with errors.

After the Lunar Survival task, participants answered a series of personal questions. In the ice-breaker condition, the questions were delivered through interactive dialogue with the robot. The robot shared its own stories, in first person, while eliciting stories from the participant. Questions asked of the participant included, "What is your favorite kind of music or favorite music artist," with follow up questions such as, "What other type of music do you like," or "Have you travelled?" followed by "Where was your favorite vacation?" and "Can you tell me more about your trip?". This interaction was again followed by questions on the perceived rapport felt by the participant. In the control condition, the user participated in a non-interactive oral survey, with the same personal questions being asked by a female (non-robot) voice, with no feedback or other sharing from the voice. No conversational errors were introduced in either condition.

Participants then completed a second ranking task, the Save-the-Art Task [46]. Participants were asked to imagine that they were a manager at a museum that was on fire. They were asked to rank 10 pieces of art as to their importance in being saved. Again, they ranked the items individually, had an interactive dialogue with the robot about which items should be saved, and then re-ranked the items. Although the Save-the-Art Task is not a survival task – the rankings are more about objective and subjective analysis of the value of art pieces themselves – previous research had found that the pattern of influence of the two tasks was similar [46]. As such, differences between rankings again indexed influence. Participants again reported how much rapport they currently felt with the robot on the same scale [47]. [1]

In the late error condition, the robot made errors while interacting with the participant on the Save-the-Art Task but not on the first task. These errors were identical to those described for the early error condition, and they were introduced into the dialogue based on the same order and at the same rate. The early error and no error conditions did not introduce errors into the

[1] The order of the tasks was not randomized, as it was shown previously that task order of these kinds of survival tasks did not affect results significantly [46].

User:	What's third?
Robot:	I ranked the oxygen first.
User:	Okay. Which one is third?
Robot:	I ranked the stellar map third.
User:	Why?
Robot:	I ranked the stellar map third because we will need it for navigation when we walk.
User:	Okay. What's next?
Robot:	<Silence>
User:	Which one do you want to do next?
Robot:	I ranked the food concentrate fourth.
User:	Okay. Why fourth?
Robot:	I'm sorry, I didn't hear you.

Figure 4: *Example dialogue with errors.*

Save-the-Art Task. More complete sample dialogues for all conditions can be viewed on the supplementary video.

4 RESULTS

In each of the ranking tasks (Lunar Survival and Save-the-Art), the item rankings were used to infer the amount of influence the interaction with the robot had on the participant. To do so, we calculated divergence between the participant's ranking and the robot's ranking, and compared the divergences before and after the interaction. If the rankings were closer after the interaction, we take that as a sign that the robot convinced the participant to change their rankings to be closer to its own (positive influence). Divergence is calculated as the Kendall τ distance between the participant's ranking and the agent's ranking, and ranges from 0 (identical rankings) to 45 (maximally different rankings); influence therefore ranges from −45 to 45, with larger numbers indicating more influence, and zero indicating no influence (negative numbers indicate that the participant moved farther away from the robot's ranking after the interaction).

4.1 Task 1 influence and rapport

A 2 (ice-breaker) x 3 (error) ANCOVA was run on influence in task 1 (Lunar Survival Task). Because initial agreement with the robot limits the amount of possible influence on a given task, initial agreement with the robot on this task was entered as a co-variate to statistically equate participants on this factor. As expected, there was no effect of or interaction with ice-breaker condition (as it had not yet occurred), however, there was also no effect of error condition on influence in task 1 (Fs < 0.69, ps > .45, ds < .20). As there was no difference between the early error condition and the other conditions (where no errors occurred during this task), it seems that errors that occur early are not detrimental to influence.

However, a 2 (ice-breaker) x 3 (error) ANOVA on the rapport ratings taken after the Lunar Survival Task revealed that conversational errors did impact the subjective feeling of rapport. While there was no effect of or interaction with ice-breaker (as it had not yet occurred; Fs < 0.26, ps > .72, ds < .14), there was a significant effect of error ($F(2,148) = 3.38$, $p = .04$, $d = .43$). Participants reported feeling less rapport with the robot

Figure 5: *Effect of error on rapport in task 1.*

after completing the lunar survival task with errors compared to the other two conditions in which no errors were introduced in this in this task (see Fig. 5). Following up on this omnibus effect, contrasts were run to detect which differences were significant. Early errors led to significantly less rapport than no errors (Tukey's LSD p = .02, d = .39) and late errors (Tukey's LSD p = .04, d = .34), but those two conditions where no errors had yet occurred did not significantly differ (Tukey's LSD p = .74, d = .05).

4.2 Task 2 influence and rapport

The same tests were run on the Save-the-Art Task. Influence was tested by an ANCOVA, again controlling for initial agreement. In contrast to the results for the first task, there was a significant effect of error on influence in this second task (F(2,147) = 3.49, p = .03, d = .43). As depicted in Fig. 6, participants were less influenced in the Save-the-Art Task when the robot made errors during that task compared to the other two conditions. Simple contrasts revealed that this significant omnibus effect was driven by the difference between late errors and early errors: late errors led to significantly less influence than early errors (Tukey's LSD p = .01, d = .42) but not no errors (Tukey's LSD p = .12, d = .25). Early errors did not significantly differ from no errors either (Tukey's LSD p = .31, d = .16).

While neither the effect of ice-breaker (F(1,147) = 0.63, p = .43, d = .13) nor interaction between errors and ice-breaker (F(2,147) = 2.03, p = .13, d = .33) were significant, the interaction at least had a small-to-moderate effect size. As depicted in Fig. 7,

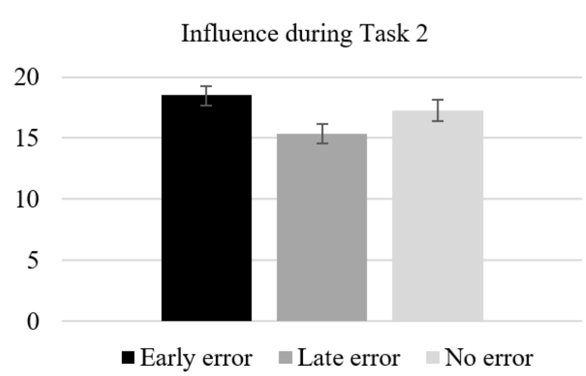

Figure 6: *Effect of error on influence in task 2.*

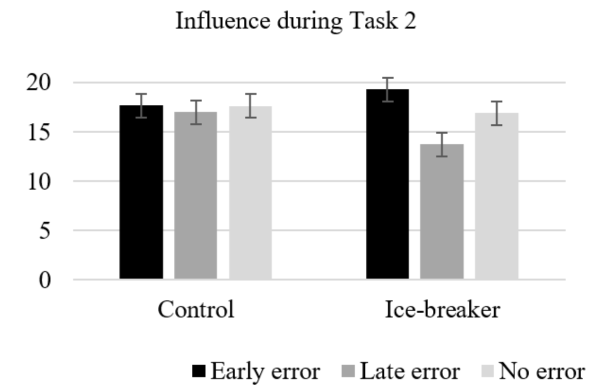

Figure 7: *Effect of error and ice-breaker on influence in task 2.*

the aforementioned effect of error was driven by the condition in which participants engaged in an ice-breaker with the robot between tasks 1 and 2. If anything, errors after the ice-breaker hurt influence, but having the robot make early errors followed by a recovery ice-breaker helps restore influence in the second task. Following errors in the first task up with an ice-breaker brings influence back on par with (and appears to even exceed) the condition in which no errors were introduced at all. Follow-up simple contrasts confirm that the ice-breaker condition drove the aforementioned effects of errors. While there were no significant differences based on errors in the absence of the ice-breaker (control condition; ps > .62), there were when the ice-breaker was introduced. In this case, late errors reduced influence significantly compared to early errors (Tukey's LSD p = .003) and shows the same trend compared to no errors (Tukey's LSD p = .09), but the difference between early errors and no errors was not significant (Tukey's LSD p = .20).

We find a similar pattern with an ANOVA on rapport ratings after the Save-the-Art Task. There was a significant effect of error on rapport felt during this second task (F(2,148) = 7.21, p = .001, d = .63). As can be seen in Fig. 8, participants felt less rapport when the robot made errors during this second task compared to the other two conditions. Simple contrasts revealed that late errors significantly reduced rapport compared to early errors (Tukey's LSD p = .04) and no errors (Tukey's LSD p < .001). Moreover, analyses revealed a trend for no errors to be better than early errors for rapport felt during this second task (Tukey's LSD p = .07).

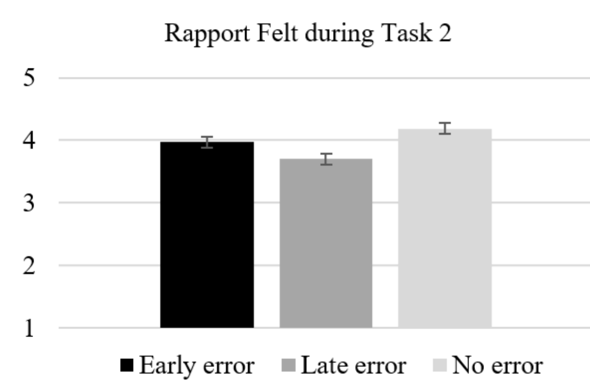

Figure 8: *Effect of error on rapport in task 2.*

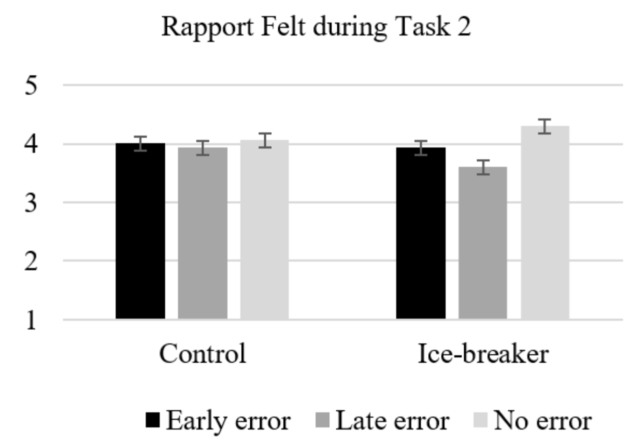

Figure 9: *Effect of error and ice-breaker on rapport in task 2.*

While neither the effect of ice-breaker ($F(1,148) = 0.02$, $p = .88$, $d = .02$), nor interaction between errors and ice-breaker ($F(2,148) = 1.96$, $p = .14$, $d = .34$) were significant, the interaction at least had a small-to-moderate effect size. As with influence, errors after the ice-breaker were most detrimental to rapport. Yet, in contrast to influence, having no errors and an ice-breaker was the best for rapport (see Fig. 9). Again, follow-up simple contrasts confirm that the ice-breaker condition drove the effects of errors. In the absence of the ice-breaker, differences between error conditions were not significant (ps > .16). In contrast, with the ice-breaker included, no errors are significantly better for rapport than late errors (Tukey's LSD $p > .001$, $d = .61$) and early errors (Tukey's LSD $p = .045$, $d = .32$), and early errors are also marginally better for rapport than late errors (Tukey's LSD $p = .08$, $d = .29$).

4.3 Correlations between influence and rapport

To determine whether perceived rapport relates to influence, we also examined the relationship between rapport and influence across the tasks. Given that errors have an effect on perceived rapport but not on influence during the Lunar Survival Task, we found a disconnect between rapport and influence for this task: they were uncorrelated ($r(152) = .04$, $p = .65$). Despite this disconnect between rapport on the first task, rapport built over time does relate to influence. For the second, Save-the-Art Task, rapport significantly relates to influence ($r(152) = .18$, $p = .02$). Considering change over time in rapport and in influence, we found a similar effect. The extent of the increase in rapport from the first task to the second task significantly correlates with the increase in influence across tasks ($r(152) = .23$, $p = .004$). Unexpectedly though, rapport during the first, lunar survival task does not correlate with influence in the second, save-the-art task ($r(152) = .04$, $p = .67$), nor does it correlate with change in influence from the first task to the second task ($r(152) =.01$, $p = .91$). While building rapport during the session was related to greater influence in the second task, having high rapport early does not result in greater influence later.

Outcome	Test	P value	D value	Fig.
Task 1 Influence	Effect of Errors	.52	.19	
	Effect of Rapport	.46	.13	
	Interaction Effect	.51	. 19	
Task 1 Rapport	Effect of Errors	.04	.43	5
	Effect of Rapport	.72	.06	
	Interaction Effect	.77	.13	
Task 2 Influence	Effect of Errors	.03	.43	6
	Effect of Rapport	.43	.13	
	Interaction Effect	.13	.33	7
Task 2 Rapport	Effect of Errors	.001	.63	8
	Effect of Rapport	.88	.02	
	Interaction Effect	.14	.34	9

Figure 10: *Summary of results.*

5 DISCUSSION

5.1 Discussion of findings

In this experiment, later errors were found to be detrimental to influence, whereas early errors were not. Indeed, errors on the first task has no impact on influence in that task. This finding is in line with previous literature showing leniency towards early errors [6]. Furthermore, by the second task, these early errors actually benefited influence when an ice-breaker occurred between the two tasks. In that way, the ice-breaker was helpful for increasing the influence of the robot when it made early errors. In contrast, the ice-breaker had the opposite effect if the errors occurred later. When the errors instead occurred later (i.e., in task 2), they were only harmful to influence when the robot and user had previously engaged in an ice-breaker. These results illustrate that the effects of errors were driven by the ice-breaker condition. This social dialogue intensified the impact of errors, which had these opposite effects depending on their timing.

While errors that occurred during the first task did not hinder influence in that task, they did significantly reduce rapport felt with the robot (significantly during task 1 and marginally during task 2). For rapport felt during task 2, the effect of errors was again driven by the ice-breaker condition; again, the social dialogue exacerbated the effect of errors. While the ice-breaker only appeared to benefit rapport *in the absence of errors*, rapport building seemed to occur naturally during the tasks across conditions. Empirically, correlational analyses revealed that perceived rapport built over the course of the study (however that may have happened) was related to influence. Rapport in the second task significantly correlates with influence in that task. Likewise, increases in rapport from task 1 to task 2 correlate with increases in influence across those tasks. Anecdotally, participants reported liking the robot because it was "polite and nice" – indeed, throughout the tasks, the dialogue included encouraging and polite phrases such as "thank you" and "that's great!" These kinds of informal social exchanges may have played a role in building rapport throughout the entire dialog. While rapport built throughout the interaction matters for influence, rapport felt during the first task was not related to influence.

5.2 Contribution to the literature

This work contributes further evidence to an area that has mixed findings. As in some research [19], our study found that errors on the first task did not impact influence on that task, however errors did negatively impact the level of rapport felt by users. However, in our work and others, errors that occur later in robot's dialogue negatively affect their influence [6]. Indeed, in all of these cases, it seems to depend on when the errors occur: errors after a period of good performance are more harmful to influence than those that occur earlier. In Desai et al. [4,5], drops in reliability after a period of good performance were much more harmful to influence than early failures. Weigmann et al. [6] likewise found that robots that shifted from 100% reliable to 80% reliable had less influence than those that were 80% reliable from the start. Our work finds that influence is preserved in the face of many conversational errors as long as those errors occur at the beginning of the interaction, but when similar errors occur later, influence is diminished. The ice-breaker drives this effect: errors are harmful if good performance occurs before the ice-breaker conversation, but less harmful without this social dialogue.

Our findings suggest that this is a result of the specific timing of the ice-breaker (vis-à-vis when the errors occur). As in the aforementioned prior research, the ice-breaker made the robot appear to perform well for longer (compared to the control condition). It is possible, then, that a contrast effect occurred, whereby errors stood out more after the ice-breaker conversation. Specifically, our attempted rapport-building interaction backfired *if* it was followed by errors because that ice-breaker conversation extended the period of good performance before the errors, creating a contrast effect with the subsequent errors. The *contrast effect* is a common concept in social psychology, which notes that if someone has experienced a positive interaction, their response scale is anchored in the positive, and subsequent negative experiences are judged against that positive scale [48]. However, when the errors occur before the social dialogue, it helps the robot to recover from these early errors and actually boosts influence.

5.3 Implications for design

Good performance in the form of social dialogue has been shown here to interact with the timing of errors. These results have implications for HRI and robot design. Designers should consider that conversational errors that occur later in a social robot's dialogue hinder users from taking the robot's advice, undermining its persuasiveness. As errors are particularly damaging when they suddenly appear after sustained good performance (here during a social dialogue), *unless the system improves over time*, research should consider alternative interventions other than social dialogue. For example, research could explore specific error-mitigation rapport-building, such as incorporating apologies, explanations, and negative self-disclosure. Also, research could consider the role of errors versus the *contrast effect* by introducing errors in all tasks.

Although it is not possible for system designers to avoid errors in all cases, they do have some control over the error rate. The expected error rate of a dialogue system is correlated with the perplexity of the task and inversely correlated with the amount of training data available. Thus, a task that has been piloted and fielded, tweaked to reduce perplexity, and trained on actual data from the user population will most likely have a substantially lower error rate than a novel task of similar complexity.

Designers will reduce errors for a task to the best of their ability, given the available resources. Instead, our work takes on the question of deciding whether a new task that would have a relatively high expected error rate should be given to the user, and to decide on the ordering of a set of possible tasks (of varying expected error rates) to put before the user. Thus, one question for user experience design is how to situate a new task (with expected high error rate) with respect to other possible tasks that a system could do with low error rate. Would it be better to first have the user engage in different low-error rate tasks (including ice-breaking), so that some measure of trust can be built up and the user won't be as put off by the high-error task? A priori, this seems like a reasonable thing to do, but our results suggest that this might backfire, as it did in our experiment. If the system is intended to influence the user, our results suggest it might be more effective to put the new task first, where influence is less susceptible to the presence of error, and then recover rapport afterward. Research should further test recommendations.

It is also important to consider to what extent these findings generalize. Other work has found similar results. For example, it was found that an agent that changed from alignment to non-alignment during the course of a structured game was rated worse than an agent who improved to aligning after a first phase of non-alignment [49]. While this non-alignment was not an explicit conversational error, the results did, however, represent a similar interaction between conversational performance (good vs. bad) and the time dynamics of conversation. Hence, it seems that the results of the current study may generalize to other contexts.

In sum, a few null findings [19] notwithstanding, conversational errors reduce robots' influence. We explored whether a rapport-building social dialogue might mitigate this detrimental impact of errors. Instead of helping, the social dialogue appeared to exacerbate the problem by highlighting how well the robot functioned before the errors. While design could still focus on other ways of mitigating errors, merely placing a social dialogue before the errors appears to be a poor option. In contrast, if the errors occurred earlier, they did less harm; designers may consider placing error-prone tasks earlier when deciding on the order of tasks.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

REFERENCES

[1] Kanda, T., Shiomi, M., Miyashita, Z., Ishiguro, H. & Hagita, N. A Communication Robot in a Shopping Mall. IEEE *Trans. Robot. 26*, 897–913 (2010).

[2] Wang, Y., Khooshabeh, P. & Gratch, J. Looking real and making mistakes. in *International Workshop on Intelligent Virtual Agents,* 339–348, (Springer, 2013).

[3] Blascovich, J. & McCall, C. Social Influence in Virtual Environments. in *The*

Oxford handbook of media psychology, 305–315 (Oxford University Press, 2013).

[4] Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., & Yanco, H. Effects of changing reliability on trust of robot systems. in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12 73* (ACM Press, 2012).

[5] Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A. & Yanco, H. Impact of robot failures and feedback on real-time trust. *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction,* 251–258 (2013).

[6] Wiegmann, D. A., Rich, A. & Zhang, H. Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theor. Issues Ergon. Sci., 2,* 352–367 (2001).

[7] Melson, G. F., Kahn, Jr., P. H., Beck, A. & Friedman, B. Robotic Pets in Human Lives: Implications for the Human-Animal Bond and for Human Relationships with Personified Technologies. *J. Soc. Issues, 65,* 545–567 (2009).

[8] Sundar, S. S., Jung, E. H., Waddell, T. F. & Kim, K. J. Cheery companions or serious assistants? Role and demeanor congruity as predictors of robot attraction and use intentions among senior citizens. *Int. J. Hum. Comput. Stud., 97,* 88–97 (2017).

[9] Kanda, T., Hirano, T., Eaton, D. & Ishiguro, H. PPerson Identification and Interaction of Social Robots by Using Wireless Tags. in IEEE *International Conference on Intelligent Robots and Systems, 2,* 1657–1664 (2003).

[10] Sumioka H. Wada Y., Ishiguro H. Teachers' Impressions on Robots for Therapeutic Applications. *JSAI International Symposium on Artificial Intelligence.* (Springer, 2015).

[11] Bogue, R. Domestic robots: Has their time finally come? *Ind. Robot An Int. J., 44,* 129–136 (2017).

[12] Tickle-Degnen, L., Scheutz, M. & Arkin, R. C. Collaborative Robots in Rehabilitation for Social Self-Management of Health. (2014).

[13] Kojima, K., Karasawa, T., Kozuki, T., Kuroiwa, E., Yukizaki, S., Iwaishi, S., Ishikawa, T., Koyama, R., Noda, S., Sugai, F., & Nozawa, S. Development of life-sized high-power humanoid robot JAXON for real-world use. in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids),* 838–843 (IEEE, 2015).

[14] Cassell, J. & Bickmore, T. Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. *User Model. User-adapt. Interact., 13,* 89–132 (2003).

[15] Traum, D., Swartout, W., Marsella, S. & Gratch, J. Fight, flight, or negotiate: Believable strategies for conversing under crisis. *Proc. Intell. Virtual Agents,* 52–64 (2005).

[16] Manuvinakurike, R., Velicer, W. F. & Bickmore, T. W. Automated indexing of Internet stories for health behavior change: weight loss attitude pilot study. *J. Med. Internet Res.,* 16, e285 (2014).

[17] Hiraoka, T., Neubig, G., Sakti, S., Toda, T. & Nakamura, S. Reinforcement Learning of Cooperative Persuasive Dialogue Policies using Framing. *In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* , 1706–1717 (2014).

[18] Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y., De Visser, E.J., & Parasuraman, R. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *J. Hum. Factors Ergon. Soc., 53,* 517–527 (2011).

[19] Salem, M., Lakatos, G., Amirabdollahian, F. & Dautenhahn, K. Would You Trust a (Faulty) Robot? in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15,* 141–148 (ACM Press, 2015).

[20] Ham, J. & Midden, C. J. H. A persuasive robot to stimulate energy conservation: the influence of positive and negative social feedback and task similarity on energy-consumption behavior. *Int. J. Soc. Robot., 6,* 163–171 (2014).

[21] Goetz, J., Kiesler, S. & Powers, A. Matching robot appearance and behavior to tasks to improve human-robot cooperation. in *The 12th IEEE International Workshop on Robot and Human Interactive Communication,* 55–60 (IEEE, 2003).

[22] Siegel, M., Breazeal, C. & Norton, M. I. Persuasive robotics: The influence of robot gender on human behavior. in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2563–2568 (IEEE, 2009).

[23] Ham, J., Bokhorst, R., Cuijpers, R., van der Pol, D. & Cabibihan, J.-J. Making robots persuasive: the influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power. in *International conference on social robotics*, 71–83 (Springer, 2011).

[24] Lee, S. A. & Liang, Y. The role of reciprocity in verbally persuasive robots. *Cyberpsychology, Behav. Soc. Netw., 19*, 524–527 (2016).

[25] Lopez, A., Ccasane, B., Paredes, R. & Cuellar, F. Effects of Using Indirect Language by a Robot to Change Human Attitudes. in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17,* 193–194 (ACM Press, 2017).

[26] Chidambaram, V., Chiang, Y.-H. & Mutlu, B. Designing persuasive robots. in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12 293* (ACM Press, 2012).

[27] Krämer, N., Karacora, B., Lucas, G. M., Dehghani, M., Rüther, G., & Gratch, J. Closing the gender gap in STEM with friendly male instructors? On the effects of rapport behavior and gender of a virtual agent in an instructional interaction. *Comput. Educ., 99,* 1–13 (2016).

[28] Tickle-Degnen, L. & Rosenthal, R. The nature of rapport and its nonverbal correlates. *Psychol. Inq., 1,* 285–293 (1990).

[29] Gremler, D. D. & Gwinner, K. P. Customer-Employee Rapport in Service Relationships. *J. Serv. Res., 3,* 82–104 (2000).

[30] Gfeller, J. D., Lynn, S. J. & Pribble, W. E. Enhancing hypnotic susceptibility: Interpersonal and rapport factors. *J. Pers. Soc. Psychol., 52,* 586 (1987).

[31] Carey, J. C., Stanley, D. A. & Biggers, J. Peak alert time and rapport between residence hall roommates. *J. Coll. Stud. Dev.* (1988).

[32] Fuchs, D. Examiner Familiarity Effects on Test Performance. *Topics Early Child. Spec. Educ., 7,* 90–104 (1987).

[33] Burns, M. Rapport and relationships: The basis of child care. *J. Child Care* (1984).

[34] Mutlu, B. & Forlizzi, J. Robots in organizations. in *Proceedings of the 3rd international conference on Human robot interaction - HRI '08,* 287 (ACM Press, 2008).

[35] Azhar, M. Q. Toward an argumentation-based dialogue framework for human-robot collaboration. in *Proceedings of the 14th ACM international conference on Multimodal interaction*, 305 (ACM Press, 2012).

[36] Clear, T., Daniels, M., Clear, T. & Daniels, M. A cyber-icebreaker for an effective virtual group? in *Proceedings of the 6th annual conference on Innovation and technology in computer science education, 33,* 121–124 (ACM Press, 2001).

[37] Mitchell, S. K. & Shepard M. Building Social Presence through Engaging Online Instructional Strategies. in *Student-Teacher Interaction in Online Learning Environments,* 133–156 (2015).

[38] Risvik, L. L. Framework for the Design and Evaluation of Icebreakers in Group Work. Master's Thesis. (2014).

[39] Inaguma, H., Inoue, K., Nakamura, S., Takanashi, K. & Kawahara, T. Prediction of ice-breaking between participants using prosodic features in the first meeting dialogue. in *Proceedings of the 2nd Workshop on Advancements in Social Signal Processing for Multimodal Interaction,* 11–15 (ACM Press, 2016).

[40] Sidner, C. L., Kidd, C. D., Lee, C. & Lesh, N. Where to look: a study of human-robot engagement. in *Proceedings of the 9th international conference on Intelligent user interfaces,* 78–84 (ACM, 2004).

[41] Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., Schultz, A.C., & Wang, J. Designing robots for long-term social interaction. in IEEE/RSJ *International Conference on Intelligent Robots and Systems,* 1338–1343 (IEEE, 2005).

[42] Moon, Y. & Youngme. The effects of distance in local versus remote human-computer interaction. in *Proceedings of the SIGCHI conference on Human factors in computing systems,* 103–108 (ACM Press, 1998).

[43] Khooshabeh, P., McCall, C., Gandhe, S., Gratch, J. & Blascovich, J. Does it matter if a computer jokes. in *CHI'11 Extended Abstracts on Human Factors in Computing Systems,* 77–86 (ACM, 2011).

[44] Ramachandran, D. & Canny, J. The persuasive power of human-machine dialogue. in *International Conference on Persuasive Technology*, 189–200 (Springer, 2008).

[45] Kelley, J. F. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2, 26–41 (1984).

[46] Artstein, R., Traum, D., Boberg, J., Gainer, A., Gratch, J., Johnson, E., Leuski, A., & Nakano, M. Listen to My Body: Does Making Friends Help Influence People? *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference*, 430-435. (2017).

[47] Von der Pütten, A. M., Krämer, N. C., Gratch, J. & Kang, S.-H. 'It doesn't matter what you are!' Explaining social effects of agents and avatars. *Comput. Human Behav., 26*, 1641–1650 (2010).

[48] Sherman, S. J., Ahlm, K., Berman, L. & Lynn, S. Contrast effects and their relationship to subsequent behavior. *J. Exp. Soc. Psychol., 14,* 340–350 (1978).

[49] Nuñez, T. R., Prynda, K., Bergmann, & K. Von der Pütten, A. M. Computers Aligning to Their Users: Lexical Alignment in Human-Agent-Interaction and Its Psychological Effects. *Virtual Social Interaction Workshop.* (2017).