UNIVERSITY OF CALIFORNIA
SANTA CRUZ

GROUNDING WITH AGENTS

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PSYCHOLOGY

by

Elise Duffau

August 2026

The Dissertation of Elise Duffau is approved:

_____
Professor Jean E. Fox Tree, chair

_____
Professor Adriana Manago

_____
Professor Jeremy Yamashiro

_____

Donald Smith
Interim Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Grounding with Agents

By

Elise Duffau

In my dissertation I investigated how we expect artificial agents (e.g., chatbots, voice assistants, robots) to ground with people. I focused on two aspects of grounding: (a) how people interpret errors from an agent and (b) how people expect agents to contribute to resolving errors in communication. Two experiments inform how we expect technology to engage in navigating grounding in a conversation. Additionally, these experiments offer insight into how similarly we want technology to mimic human speech behaviors or if we have differing expectations for how technology should communicate.

Experiment 1 uses a Wizard of Oz (WOz) design where a voice assistant (VA), in this case an Alexa, commits phoneme level speech errors (swaps, anticipations, and perseverations) across three between-subjects conditions: (1) no errors (no speech errors committed), (2) errors (speech errors are committed), and (3) errors with correction (speech errors are committed and immediately self-corrected with the phrase "[error] I mean [correction]"). Another factor, context, is examined for its effect on VA error perception. This factor has two within-participants conditions: (1) a task-oriented context consisting of selecting a tangram from an array based on VA descriptions and (2) a social oriented context consisting of exchanging ice breaker questions with the VA. Unknown to the participant, in all conditions the VA's responses will be controlled by a researcher. I measured participants' attitudes toward technology, perceptions of the agent, and engagement with the agent. I found that while errors were perceived similarly to people – that is, some go unnoticed – the use of errors did not increase human-like perceptions, positive affect, nor engagement compared to no errors

present. However, when errors were corrected participants rated the VA as less intelligent. This study informs us that while errors are expected of people, they are not expected of VAs and bringing attention to errors by correcting them can lead to negative perceptions of the VA.

Experiment 2 also uses a WOz design. Here the VA used repairs to attempt to ground a conversation. In this study, participants listened to ten short mini-lectures spoken by the VA and answered a question after each lecture. These lectures were designed to be difficult and encouraged signals of confusion such as a ("Huh?", "What", "Um/uh", or a long pause). In response to these signals the VA attempted to clarify the confusion by using a repair (repeat, rephrase or topic shift). As in the previous study, unknown by the participant all VA responses are controlled by a researcher. I measured participants' attitudes toward technology, perceptions of the agent, engagement with the agent, as well as the effectiveness and appropriateness of the repairs. Participants showed sufficient signs of confusion in response to the lectures. Overall, repairs were not perceived as helpful. The different repair types also did not differ from each other. The number of repairs produced also did not significantly impact participant's perceptions of the VA. However, repairs were not harmful either: The attempt to repair was not rated more negatively compared to no repair given. This means that even if a VA repairs poorly and is unsuccessful it is not seen significantly more negatively than one that does not try to help. These results show that while people have high expectations for successful grounding, they are also tolerant of artificial agents that make an attempt, even an ineffective one.

Together these studies show that people may be more tolerant of missed mistakes compared to admitted mistakes, but failing at trying to help has no negative consequences compared to not trying at all. Future advancements in communication-based technology should encourage technology to work with people to support its limited grounding capabilities.

Acknowledgements

Thank you everyone at the spontaneous communication lab, without your mentorship and comradery none of this would have been possible. Thank you, Jean E. Fox Tree, your mentorship has made me the researcher I am today. To my friends and family, thank you for supporting me on this journey.

Grounding with Agents

"Take a left turn up ahead. Oh sorry, I mean a slight left." Anecdotally, when said by a person, people describe this sentence as a helpful correction providing a more specific detail. But when these instructions are said by Apple Maps, the same phrase can be considered unsettling, or untrustworthy. The expectation when communicating with the voice assistant is that it is relaying a set, unchangeable, and correct route. However, should the technology lose track of the positioning of the vehicle or lose connection for a moment, it would make sense for the voice assistant to make the needed adjustments. Mistakes happen; people know how to make corrections or clear up miscommunications. But voice assistants and chatbots cannot simply mirror people. A voice assistant's errors won't have the same effect. We need to tailor the responses of this technology to fit into our unconscious expectations.

Chapter 1: Navigating Understanding; Where Machines Fall Short

Previous researchers have shown that we currently struggle with negotiating grounding with a voice assistant (Beneteau et al., 2019). Researchers have explored the effectiveness of repairs in both voice assistants (Cuadra et al., 2021) and chatbots (Ashktorab et al., 2019; Li et al., 2019). In these studies, we often see the voice assistant or chatbot taking a reactive response to negotiate grounding; that is, a person has to first recognize there is a problem and then begin the process of addressing it. Only in the most recent voice assistant research are self-repairs being tested as a way to preempt confusion. In the studies reported here, I took this one step further by testing whether voice assistants' effectiveness can be improved by independently

attempting to address confusion in conversations.

To do this we must address one of the main problems machines have with correcting miscommunication, which is that machines do not know how to recognize that humans are having trouble, nor do they know how to address the cause of the miscommunication. In two studies, I examined what role miscommunications have in human-machine discourse and how machines can fundamentally approach addressing miscommunications.

Grounding in Conversations

When people have a conversation, they are constantly reaffirming that their conversational partner is understanding what they are saying; this is known as checking for evidence of grounding (Clark & Brennan, 1991). Grounding is the process of confirming mutual understanding and is crucial for an effective conversation. To judge if conversational partners have reached adequate grounding, interlocutors use principles of closure (Clark, 1996). The principles of closure consist of two types of valid evidence of expected behavior from the interpretation of the speech: positive evidence, such as backchannels ("uh huh"), head nods, and conformational phrases ("I understand"), and negative evidence, such as quizzical looks, or phrases of misunderstanding such as ("I don't understand"). Interlocutors use the presence of evidence to understand their addressee's perspective in the conversation.

Grounding is a process necessary to affirm understanding and it isn't without effort. In the grounding process the amounts of evidence are moderated by two

factors: effort level and time. The more effort it takes to ensure grounding the more evidence conversational participants will look for (however interlocutors will start to give up and look for less evidence if the burden becomes too great). Similarly, the quicker the response to the speech the more likely it will be assumed the hearer understood. This process of efficient grounding is known as the principle of least collaborative effort.

When miscommunications happen with an artificial agent, such as Alexa or Siri, conversations can become completely derailed. Navigating the grounding process is a skill that artificial agents have not mastered. However, previous research has explored how agents can better handle miscommunications, errors, and grounding.

When resolving miscommunications people use two general types of repairs, other-initiated repairs and self-repairs. When using other-initiated repairs, the listener indicates the point of confusion using a "huh?" or a "what?" while a self- repair tends to fall in the same turn as the miscommunication or in the silent gap between the next turn (Schegloff et al., 1977). To effectively use repairs interlocutors, need to be attentive to each other's signals of grounding and respond appropriately to keep the conversation flowing.

I tested two crucial elements of grounding: (1) establishing comprehension and (2) collaborating to reach understanding.

In the establishing comprehension test (Experiment 1), I tested how well machines promoted human comprehension. When people address

miscommunications in a conversation, they have the benefit of understanding the context in which the confusion has occurred and can typically predict what may have been confusing to their addressee. Furthermore, people can then discuss and resolve the point of confusion. But current artificial agents struggle to understand the context of a conversation. Agents are often unaware of errors they have made and are unable to address a potential miscommunication within the context of the conversation. By examining how agents can make and correct errors, we can better understand how machines can engage in reducing the effort of grounding. While the attempt to repair may not be exactly addressing the confusion, it may be enough to create some clarity versus none. This is an advancement involving agents' ability to reflect and self-correct any errors they may have made, increasing comprehension in a conversation.

In Experiment 1 I also investigated how communication errors could be beneficial for intentional integration into conversational agent technology. These errors may increase approachability and human-likeness via the pratfall effect (attractiveness and approachability increases when mistakes are made by one who doesn't typically make mistakes or seems not to make mistakes; Mirnig et al., 2017). However, speech errors have also been known to lower technology's perception of competency, sincerity, and can be seen as creepy (Cuadra et al., 2021; Gompei & Umemuro, 2015; Ragni et al., 2016). Understanding whether small speech errors (like the ones that people often make) generate positive and negative perceptions will inform us of how our perceptions of technology balance with the expectations of naturalness in human speech. Once again, a lack of effect or a negative effect would

raise the question of whether people really want technology to sound human?

In the collaborating to reach understanding test (Experiment 2), I will assess how people feel about technology recognizing and addressing negative evidence of grounding. In human conversations, when people notice their addressee is confused, they often attempt to address the confusion in an effort to collaboratively maintain understanding. However, current artificial agents, agents' conversational contributions are akin to prerecorded messages that are unable to clarify in response to human confusion. I tested whether interacting more similarly to a person versus a recording was enough to enhance feelings of grounding collaboration, even if the grounding attempt was not always successful. Furthermore, some repair strategies may be better at achieving this than others (i.e., some repair strategies may be more generally effective than others). Finding a general repair strategy could be a good starting point for technology to fundamentally address confusion, even if not perfectly matched with the current context. I tested three basic forms of repair strategies as a baseline to determine if any can be used as a default form of repair to increase understanding in response to a confused addressee increasing the collaborative effort or regaining grounding in the conversation.

Experiment 2 was centered around two grounding principles, recognizing points of confusion and other-initiated repairs, and their relationship to people's expectations of machines attempting to ground. In previous research grounding effort was typically a burden on the user. I tested what happens when the technology attempts initiating grounding. One prediction was that artificial agent's attempts at

grounding will increase understanding, engagement, and human-like perception. Alternatively, another prediction was that the artificial agent's attempts at grounding will decrease understanding, engagement, and human-like perception. If technology cannot perform this type of communication well, then we should either not attempt to apply error-corrective strategies to artificial agents, or we should go back to the drawing board with respect to how technology responds to human errors. A lack of effect or a negative effect would raise the question of whether people want technology to engage in grounding in a human-like way or if people expect something different from a technological interlocutor.

Chapter 2: Meaningful Mistakes; Opportunities To Affirm Understanding

Experiment 1: To Err is Human(-like)

People make a lot of speech errors. When a speech error creates a barrier to understanding then it must be corrected; however, not all of those errors are actually a detriment in the conversation. In fact, many of our speech errors go unnoticed or uncorrected. To err is human. It is so fundamental as to be an unconscious expectation. What does that mean for artificial agents? Some researchers who have explored agents using speech errors have found that it can occasionally be beneficial for the technology to make errors.

In one study, participants felt that a robot that used speech errors was more familiar (i.e., more approachable, easier to get close to, and easier to open up to) than a robot that did not use speech errors (Gompei & Umemuro, 2015). However, speech

errors also lowered the perceived sincerity of the robot. This study hints at the positives of a robot that errs, but there may be an effect of context on errors, such that errors are accepted in some contexts (a game) but not in others.

In another study, participants took turns with the robot in performing a digit span and a number series completion task and gave each other feedback on their performance on the task (Ragni et al., 2016). In this context of performing and rating performance, results showed that participants viewed the error-producing robot as less intelligent, less reliable, more insecure, less competent, and less superior than a flawless robot. However, participants also reported an easier and more positive interaction with the error-producing robot as well as more enjoyment from the interaction.

One interesting finding in this study was that participants performed better in the flawless condition than in the error condition. The researchers thought this may be due to participants conforming with the robot as it sets a faultless standard of performance in the task, resulting in the participant in the flawless condition performing better than those in the error condition. This would imply that the participants see the robot as a social actor and attribute some agency to the robot rather than just a preprogrammed machine. This is especially interesting given that the robot in the error condition displayed more human tendencies (i.e., thinking noises) than in the flawless condition. It is possible that the robot's use of gaze and other gestures provided enough human-likeness to trigger participants' social responses generally, in turn causing feelings of the robot being a social actor. Overall,

these findings support the role of errors in creating a more human-like experience – in this case, a more enjoyable interaction with the robot as well as matching the robot in terms of performance.

The idea that errors are tapping into a social aspect is expanded on in another study that sees errors as taking advantage of the pratfall effect (Mirnig et al., 2017). Participants interacted with a robot over two sessions. One session consisted of an interview conducted by the robot asking the participant questions about their attitudes towards robots, and the second session was the robot giving instructions to participants on how to disassemble and reassemble Legos in different configurations. Results showed that a robot that committed errors was in fact more likable, but it was not seen as more anthropomorphic nor less intelligent than a flawless robot. These results show support for the pratfall effect applying to robots. This study also supports how errors can be a positive aspect in communication.

These studies all circle around the idea that making some errors can be beneficial to seeing robots and artificial agents as more approachable and likable. However, errors can also be detrimental to perceptions of competence and sincerity. If the errors are inconsequential (as seen in Gompei & Umemuro, 2015) and are used in a low-stakes task, results will be more positive (i.e., likable; as seen in Mirnig et al., 2017) compared to a higher stakes task which may lead to more negative perceptions (i.e., competence; as seen in Ragni et al., 2016). In Experiment 1, I tested the effect of context (a low-stakes social task and a higher stake performance-based task) on error perception. Although both tasks were assessed in an experimental

setting, for ease of description in this proposal, I will refer to more task-oriented requests (performance based with a correct answer) as *task contexts* and more social-oriented requests (icebreakers with no correct answer) as *social contexts*.

Seven hypotheses were tested. The first hypothesis was that errors would make the voice assistant more likeable and human, as was observed by Gompei and Umemuro (2015), Mirnig et al. (2017), and Ragni et al. (2016):

> Hypothesis 1: Errors will create more positive affect in the form of (a) likeability and (b) humanness.

Because prior researchers observed increased likeability in a task context (Mirnig et al., 2017; Ragni et al., 2016) and familiarity in a social context (Gompei & Umemuro, 2015), I predict that small speech errors which minimally impact understanding should lead to higher positive affect in both conditions.

The second hypothesis was that people would welcome the correction of errors. While people prefer no errors (Cuadra et al., 2021), it is unknown whether they prefer corrected or uncorrected errors when the errors do not significantly deter from the conversation. There are two reasons to predict that people would prefer voice assistants that corrected errors. One is that people correct errors, and therefore people would prefer voice assistants who are like people. Another is that errors are problematic (people prefer no errors, (Cuadra et al., 2021; de Sá Siqueira et al., 2024; Gompei & Umemuro, 2015)), and that correction of errors helps to overcome problems. This yields the following hypothesis:

> Hypothesis 2: Errors that are corrected will be perceived more positively than

not corrected errors, but no errors will be preferred overall.

Previous research in this domain involved the same three conditions tested here (no error, error, and error with correction), but used different errors (e.g., asking Alexa to play relaxing music and it plays heavy metal instead; Cuadra et al., 2021). The errors tested were very small and similar to the often-unnoticed speech errors people make that are sometimes corrected and sometimes not. Thus, these errors more closely modelled the errors most frequently made by people rather than execution errors (such as playing the wrong kind of music) that are more frequently expected from machines.

Prior researchers did not directly compare the role of errors and corrections across different types of tasks, such as more task-like tasks or more social tasks. Testing task and social contexts together in one study is important to understand how the technology's task affects acceptance of errors. For example, errors (or corrections) might be a problem in one type of task and not a problem in another type of task. In Experiment 1, participants engaged in either a puzzle task (more task oriented) or an ice breaker task (more socially oriented). There are two reasons to predict that errors will be a bigger problem for puzzles than ice breakers. One is that an error in a puzzle may prevent solving the puzzle, but an error in a social task may be more easily overlooked. Another is that technology's making an error may be more forgiven in a social than a task context (Gompei & Umemuro, 2015) because technology is expected to be accurate in task contexts (Mirnig et al., 2017). I predict:

Hypothesis 3: Errors (corrected and not corrected) will be more

tolerated/appropriate in social contexts compared to task-oriented contexts. That is, while errors (corrected and not corrected) might be tolerated during ice breakers, they may not be tolerated during puzzles.

In addition to not testing how well errors and corrections are tolerated across different task contexts, prior researchers also did not test how errors and corrections affected engagement and personality perceptions across task contexts. Prior researchers did assess perceptions such as likeability and familiarity based on error production (Gompei & Umemuro, 2015; Mirnig et al., 2017; Ragni et al., 2016), but engagement and perception were not tested based on presence of repairs nor compared across contexts. There is reason to predict that context will matter. The consequences of inaccuracy (both producing an error and failing to correct an error) are higher in task contexts than social contexts because technology is expected to be accurate in task contexts (Mirnig et al., 2017). Thus, the correction of an error should be more valued in task contexts, as follows:

> Hypothesis 4. Correcting errors will be more positively perceived (e.g., higher engagement and positive personality ratings) in a task context compared to social context.

That is, while correcting errors is better than leaving an error stand, correcting is more beneficial in some circumstances than others. I assessed benefit with (1) personality measures (e.g., Godspeed scale in Bartneck et al., 2009, and also Ho & MacDorman, 2017), (2) engagement measures (O'Brien et al., 2018), and (3) a flow measure that I uniquely adapted from Qiu & Benbasat (2005) to assess conversational immersion.

The flow measure, while typically used as a measure of interaction enjoyment in marketing (Hoffman & Novak, 1996), was adapted for use in evaluating engagement with agents by Qiu & Benbasat (2005). In this study researchers tested perceptions of text to speech versus a 3D avatar on social presence and flow. While not their main finding flow was more associated with voice over text communication. While the measure did seem related to enjoyment and engagement it seemed more closely related to conversation immersion. The connection to immersion in the conversation is why this measure was selected as a complement to the User Engagement Scale (O'Brien et al., 2018) to capture more than conversation enjoyment but also the investment in engaging in the conversation itself.

An additional nuance with respect to positive perceptions involves the pratfall effect, where mistakes increase likeability. The pratfall effect works for both people and robots, with more error-prone robots being more likeable (Mirnig et al., 2017). But with robots, there was a downside for competence: The positive perception of error-prone robots was tempered by lower competence ratings (Ragni et al., 2016) particularly in the task context where the errors have more impact. I predict:

> Hypothesis 5. Errors will make the agents seem more human and relatable, but in a task context errors will also make the agents seem less competent and less trustworthy.

That is, while errors are predicted to lead to positive perceptions of an agent, a task setting will bring out a competence trade off.

The pratfall effect can be further explored based on whether the errors were

corrected or not. As correcting errors was preferred over ignoring errors (Cuadra et al., 2021), I expect corrected errors to receive both the benefit from the pratfall effect and also mitigate the loss of competency, yielding the following hypothesis:

Hypothesis 6. Self-corrected errors will have all the positive aspects of errors in the social condition – humanness and relatability – without the loss of competency and trust.

In a social setting I expect the perception of errors to be more flexible to the novel context and reduced performance implication in answering/asking ice-breaker questions. The reduced pressure from the social task should mitigate the loss of competency as I expect to see in Hypothesis 5.

Finally, I predict errors can lead to an uncanny valley effect. Self-correcting errors involves awareness that an error is made. This is a normal human behavior but not one that technology typically displays. Usually, technology requires a prompt from a human when an error has been made. Researchers have observed that an unprompted correction from an agent may appear too human-like, leading to an uncanny-valley effect where the agent is seen as creepy (Cuadra et al., 2021). I predict:

Hypothesis 7. Self-repair will seem more creepy than no correction.

I expect to see higher levels of creepiness when errors are corrected in both social and task contexts as the act of noticing and self-correcting an error is an uncommon behavior for technology which participants may not have experienced previously.

Method

A 2 (Context: Social vs Nonsocial) by 3 (Error Type: No Error vs Error vs Error with Correction) design was used. Context was a within subjects measure, while the presence of errors and self-correction was between subjects. Error Type consisted of No Error (all descriptions will be free of speech errors), Errors (all descriptions will contain one error per contribution or speaking turn), and Errors with Correction (all descriptions contained one error per contribution or speaking turn and all errors were self-repaired by the voice assistant in the same contribution/speaking turn). Errors consisted of phoneme level errors including swaps, perseverations, and anticipations (see Table A1 in the appendix for example stimuli).

Participants

Using the UCSC psychology SONA system 77 participants were recruited to participate in the experiment for credit. Out of the total participants 9 were excluded due to technical difficulties or not following instructions. The remaining 68 participants' data were analyzed and included in this report. An estimate of 60 total participants were required based on power estimates from a sample of pilot data and prior research.

Participants were an average of 19.2 years old (SD = 1.31) and were majority women (57%) with 28% men, 7% genderqeer and 7% identifying as other or preferring not to respond. Participants reported their most closely identifying race as Asian American 34%, White 31%, Hispanic/ Latinx 19% mixed race/ethnicity 12%, African American/Black 3% and preferred not to respond 2%.

Materials

In the task condition, there were 9 tangram arrays consisting of 6 tangram images each. For each array the Alexa described a single unique tangram in the array. All descriptions had three different versions corresponding to the error conditions with one error present per tangram description (except in the No Error condition).

The social condition consisted of a list of 15 ice-breaker type questions. The agent introduced the conversational structure by answering a question from the list and then asked the participant a question. This prompted the participant to answer and then ask the voice assistant a question from a provided list. The voice assistant had predetermined responses to each question and kept with the answer and ask structure in each response. This simulated a spontaneous conversation with the Alexa. As with the task condition each response consisted of three versions corresponding to the error conditions. Errors in this condition were presented at the same rate of one error per contribution or speaking turn, meaning the error occurred either in the response to a question or when the voice assistant posed a question.

Scales. Five scales were used from the literature. The Affinity for Technology Interaction (ATI) scale from (Franke et al., 2019) was used to measure participants' general attitudes toward technology before their interactions with the voice assistant. To measure participants' engagement levels with the voice assistant the User Engagement Scale (short form) was used, consisting of items such as "engaging with the voice assistant was worthwhile" and "I felt frustrated while engaging with the voice assistant (reverse scored)" (O'Brien et al., 2018). To complement the

engagement scale, a Flow scale was also used (Qiu & Benbasat, 2005). This scale measured how focused and engaged participants were in the conversation consisting of items such as "I was absorbed in the interaction with the voice assistant" and "I felt curious when interacting with the voice assistant." To measure the level of appropriateness of the communication a Conversational Appropriateness Scale was used (adapted from Rubin et al., 1994). A relevant subset of questions from the Godspeed Scale was used to measure Anthropomorphism, Likeability, and Intelligence (Bartneck et al., 2009). Additionally, a Trustworthiness measure based on the Godspeed Scale measures was also used. To measure the potential creepiness perception an additional measure was added based on Godspeed scales to test Eeriness from (Ho & MacDorman, 2017).

Apparatus. The voice assistant was an Alexa that was used as a Bluetooth speaker to play Alexa voice recordings. The voice recordings were created with an Alexa text to speech software and edited for audio quality. A Wizard of Oz set-up was used in which a research assistant in an adjacent booth could hear the participant and control when and how the voice assistant responded.

Procedure

Participants were randomly assigned to one of the three error conditions, either *No Errors*, *Errors*, or *Errors with Corrections*. Participants first completed the ATI scale, then interacted with the Alexa in both the context conditions (task and social), which were counterbalanced within participants. After each context condition participants completed the engagement, flow and appropriateness scales. At the end

of the experiment participants completed the Godspeed Scale measures and demographic questions.

Task Condition

In this condition participants interacted with the artificial agent in a tangram task puzzle. The participants saw a 2 x 3 array of tangrams, and they selected the tangram that the agent described.

The agent had preset descriptions of the tangrams. After offering the additional description the agent defaulted to "I don't know how to describe it any other way, would you like me to repeat the description?" which it then repeated the initial description. The only other response it gave was "If you have made your selection, let's move on to the next one." to prompt participants to make a selection and continue with the survey. See Table 1 for an example of exchange in the task condition.

Table 1. Example Task Exchange

Participant	"LabAlexa I'm ready to play round one!"
Alexa	**No error**: "This tangram looks like a person that is practicing some graceful dance moves. They have their arms outstretched and appear to be bowing to the right side with their legs kicked out to the right. They also appear to be wearing a long skirt." **Error**: "....kegs licked out to the right…." **Error + Correction**: "kegs licked, I mean legs kicked…"
Participant	"Umm I'm not sure. Is it the one with the triangle body?"
Alexa	"I don't know how to describe it any other way, would you like me to repeat the description?"
Participant	[silence 15 seconds]

Alexa	"If you have made your selection, let's move on to the next one."
Participant	[participant selects tangram on the computer] "LabAlexa I'm ready to play round two"

Social Condition

Participants were given a list of prepared ice-breaker questions that they could ask or that may be asked of them. The agent's responses to these questions were pre-prepared and the appropriate responses to questions were selected by a researcher. After responding to the question, the agent randomly asked an icebreaker question off the prepared list.

The agent started with self-disclosing an ice breaker question and asked the same of the participant. The participant then chose a question to ask from the predetermined list, and researchers selected the correct response from the list based on participants' selection. The response contained an answer to the participant's icebreaker question and another icebreaker question. This back and forth continued until all questions were answered. See Table 2 for an example of exchange in the task condition.

Table 2. Example Social Exchange

Participant	"LabAlexa I'm ready to chat!"
Alexa	**No error**: "Let's get to know each other better. I'll start…My favorite movie is The Empire Strikes Back. It's impressive, most impressive. What's your favorite movie?" **Error**: "It's impressive, most mimpressive…" **Error + Correction**: "It's impressive, most mimpressive. I mean most impressive…"

Participant	"My favorite movie is Up. Umm...What is the funniest joke you know?"
Alexa	**No error**: "Did you hear about the cows that started a metal band? They called themselves Shredded Beef. What's your opinion on milk before cereal?" **Error**: "...opinion on bilk before…" **Error + Correction**: "...opinion on bilk before. I mean milk before…"
Participant	"I always add milk before cereal. Unless I'm having a second bowl then I would add more cereal to the milk."

Results

I analyzed the data for reliability, then using 2 Context (social vs task) by 3 Conditions (No Error, Error, Error with Correction) ANOVAs I tested the effects of engagement, flow, and appropriateness. I also investigated the relationship between engagement (User Engagement + Flow scales), Godspeed, and Affinity for Technology Interaction scales.

Reliability

All Scales were checked for reliability. ATI Scale was found reliable (α = .76) along with the Godspeed scales (anthropomorphic a = .74, likeability a = .83, intelligence a = .86, creepy a = .76. The two-item trust measure was not found to be reliable (a = .48), but the two items were highly correlated with each other (r = .31, p = .01). The Flow scale was highly reliable (a = .81 - .89). The User Engagement Scale had one item ("I lost myself in this experience") that reduced the alpha in most conditions. This item was removed resulting in an acceptable reliability range (a = .71 - .78). In addition, the Conversational Appropriateness Scale also had variation in

reliability as a result of one item ("The voice assistant's communication was very proper."). The reliability of this scale varied greatly depending on the condition it was presented in: No Error condition (a = .59 -.71), Error condition (a = .57 - .67), Error with Correction condition (a = .85 -.71). Removal of this item greatly increased the reliability in the Error and No Error conditions, but slightly lowered the reliability in the Error with Correction condition. The result was a more balanced reliability across conditions (a = .64 - .83). However, it is worth noting that future investigators may want to pursue why the reliability was so different across the conditions.

Analysis of Variance

A 2 Context (social vs task) by 3 Condition (No Error, Error, Error with Correction) ANOVA was run to test engagement across conditions using the user experience scale. I found no significant interaction effect between condition and task type for participants' user engagement scores ($F(2,65) = .28$, $p = .76$, $\eta^2 = .01$). However, the main effect of task type ($F(1,65) = 3.73$, $p = .06$, $\eta^2 = .05$) was trending toward significance where engagement ratings were higher in the social context compared to the task context (see Table 3 for means and standard deviations). No main effect of error condition was detected ($F(2,65) = .48$, $p = .62$, $\eta^2 = .01$) indicating that the presence or correction of the errors did not impact participants engagement ratings. Taken together these results show that participants were more engaged in the social condition compared to the task condition regardless of the presence or correction of errors. This could allude to differences in expectations surrounding errors in a social task compared to a more tasky task. This finding

warrants further exploration in future studies into how our conversational expectations of artificial agents depend on the context in which they are held.

Table 3.User Experience Scale means and standard deviations by context and conditions

Context	Condition	Mean	SD
Social	No Error	3.88	0.58
	Error	3.81	0.58
	Error w/ Correction	3.81	0.60
Task	No Error	3.80	0.58
	Error	3.60	0.74
	Error w/ Correction	3.61	0.66

Next, I examined the effect of flow on contact and condition using the same 2 Context (social vs task) by 3 Condition (No Error, Error, Error with Correction) ANOVA analysis. The data showed a significant interaction effect between task type and error condition ($F(2,65) = 4.67$, $p = .01$, $\eta^2 = .13$) where the presence of errors leads to less immersion (or flow) in the conversation than if there were no errors (see Figure 1). In the social condition No Error ($M = 5.08$, $SD = .92$), Error ($M = 5.15$, $SD = .97$) and Error with Correction ($M = 5.05$, $SD = .92$) had similar average scores. However, in the task condition No Error ($M = 5.64$, $SD = .79$), had significantly higher scores ($p = .04$) compared to Error ($M = 4.91$, $SD = 1.14$) and Error with Correction ($M = 4.91$, $SD = .94$) conditions. The Error and Error with Correction conditions were not significantly different ($p = 1$). Interestingly, this effect is present only in the task context and not in the social context where no significant differences

between the conditions were found. It is possible that in a social context the presence of an error is less impactful to the conversation at hand, whereas an error in a task that has a more critical expected outcome is less tolerated and causes more of a distraction.

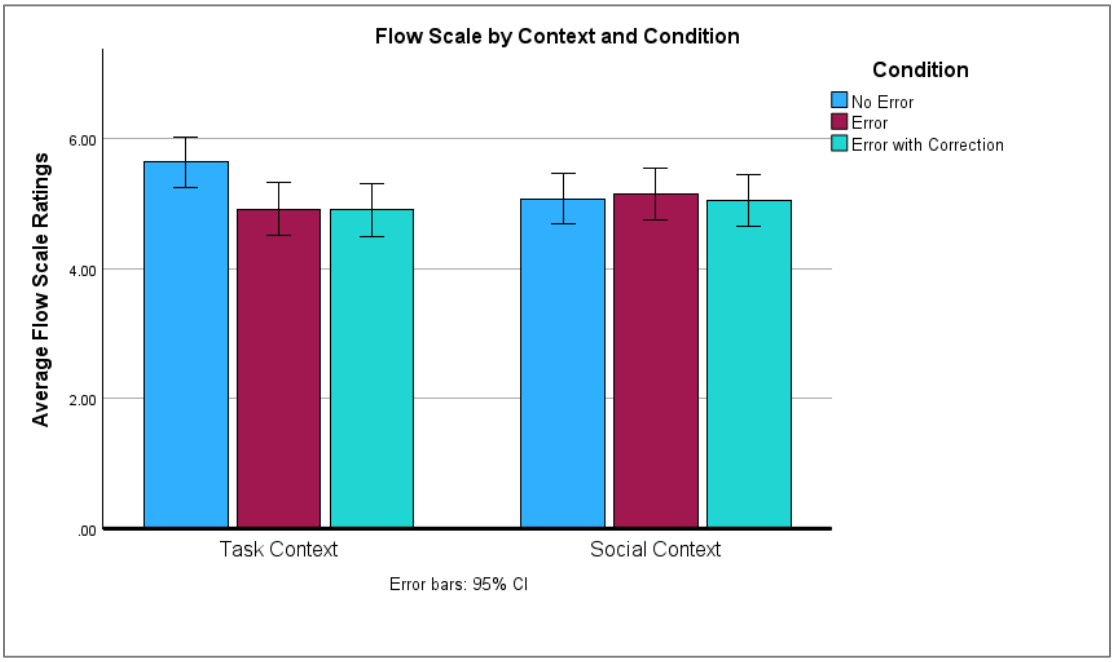

Figure 1. Flow Scale interaction between context (social or task) and condition (No Error, Error, or Error with Correction).

One final 2 Context (social vs task) by 3 Condition (no error, error, error with correction) ANOVA was run investigating the effect of conversational appropriateness. No significant interactions ($F(2,65) = 4.67$, $p = .01$, $\eta^2 = .13$) nor main effects ($F(2,65) = 4.67$, $p = .01$, $\eta^2 = .13$) were found. Conversational appropriateness did not show significant differences between context nor conditions, however the ratings of appropriateness remained high averaging around 4 “somewhat agree” out of 5 (see Table 4 for means).

Table 4. Conversational Appropriateness Scale means and standard deviations by context and conditions

Context	Condition	Mean	SD
Social	No Error	4.33	0.72
	Error	4.54	0.60
	Error w/ Correction	4.19	0.78
Task	No Error	4.43	0.64
	Error	4.43	0.59
	Error w/ Correction	4.25	0.64

Perception and Engagement Investigation

To examine participants' expectations of technology, I examined the correlations between participants' attitudes toward technology (ATI scale), personality ratings of the agent (Godspeed scales) and engagement (User Engagement & Flow scales). The ATI and Godspeed scales bookended the survey with participants responding to the ATI at the very start of the survey and the Godspeed scales at the end. The engagement scales were responded to after each context in each condition (e.g., after social and again after task context in the error condition). While the correlations described below give insight into the relationships between the variables it is worth noting the temporal effect of these scales does affect the direction of the relationship. For example, correlations between the ATI and Godspeed measures show either that ratings for the ATI (given before interacting with the VA) influenced ratings for the Godspeed measures (given at the end of the experiment) or that the two scales are related to each other.

Attitudes Toward Technology. I examined the correlation between the Affinity for Technology Interaction scale, the Godspeed scales, and the engagement scales to determine if there was a relationship between people's expectations when interacting with technology and the ratings they assigned to the Alexa. While I did not find any correlations across the conditions, there was one correlation between ATI and the task context user experience ratings (r = .46, p = .03) where participants who had a more positive affinity toward technology were more engaged when working in tasks with the Alexa. It is possible that participants who have certain expectations of working with technology in a task are influenced by how they expect to interact with technology. In other words, people who like technology and are used to working with it in a task context are more likely to be engaged when using technology in a task. Interestingly, I did not find a significant relationship between ATI and the social context (r = -.04, p = .77) meaning that those with high affinity towards technology were not more likely to be engaged in talking with the Alexa in the social context. Additionally, no correlations were found between the ATI and the error conditions. It is possible that this finding points to a divide between an affinity toward technology and the expected use of technology, where people who engage with technology often are expecting that technology to be used in certain contexts and the affinity does not transfer to an atypical context or use. Exploring how people's expectations of technology across contexts are shaped warrants further investigation. Additionally, there were no correlations found between the ATI and Flow scale nor the ATI and the Godspeed scale ratings.

Personality Measures. I examined the relationship between each of the Godspeed measures — anthropomorphism, likability, intelligence, trustworthiness, and creepiness — and both the flow and user experience scales. To better examine the differences between the relationships I aggregated the scales across context (social and task) and across conditions (No Error, Error, Error with Correction) to determine if any specific patterns emerged.

Anthropomorphic was positively correlated with all other Godspeed measures, in other words higher anthropomorphic ratings were related to higher ratings in likeability, intelligence, trust, and creepiness. In addition, anthropomorphism was also positively correlated with higher levels of flow in both the social ($r = .436$, $p > .001$) and task ($r = .290$, $p = .01$) context, while only positively correlated with the social context of the user experience scale ($r = .418$, $p < .001$). Interestingly for the error conditions, I found only one positive correlation in the No Error condition for both the flow ($r = .538$, $p = .007$) and user experience scales ($r = .475$, $p = .02$) indicating that those in the No Error Condition who had high user engagement and flow ratings also rated higher on anthropomorphism. The other two error conditions; Error and Error with Correction, did not have significant correlations. This may mean that while the Alexa is generally seen as anthropomorphic, high ratings are more likely to be associated with expected behavior keeping the participant in the flow of the task. The use of obvious errors may detract from the expected norms of the conversation which is associated with a faulty robot rather than an error prone human.

Likeability was significantly positively correlated with all measures except the

Godspeed creepy scale and the flow scale in the error condition. In other words, higher ratings of likeability were associated with all positive personality trait ratings and higher user engagement ratings. This isn't too surprising of a result as it is often the case if you are engaged with the technology, you are more likely to like it and in turn have a more positive perception of its personality leading to higher ratings. Intelligence was also significantly positively correlated with all Godspeed measures except for creepy. Higher flow ($r = .31$, $p = .01$) and user experience scale ($r = .35$, $p = .003$) ratings were associated with higher levels of intelligence in the social context. In addition, higher intelligence was associated with the user experience scale when no errors were made ($r = .42$, $p = .04$).

Trust followed a similar pattern to likeability and intelligence where higher levels of trust were associated with higher levels of the other positive personality measures (all except creepiness). Higher levels of trust were significantly associated with higher levels of user engagement in both the social ($r = .45$, $p < .001$) and task ($r = .30$, $p = .01$) context. Higher user engagement was also associated with more trust in the No Error ($r = .46$, $p = .02$) and Error conditions ($r = .52$, $p = .01$). The relationship between trust and user engagement may be centered around the perception of the errors, whereas if the errors made by the voice assistant were not noticed then the voice assistant seemed trustworthy. High levels of flow were only significantly correlated with higher levels of trust in the social context ($r = .39$, $p < .001$) and in the Error condition ($r = .45$, $p = .03$). It is surprising that I did not detect a relationship between trust and flow in the no error condition considering the prior

results indicated that errors can be distracting. However, this association may indicate that participants who experienced higher levels of flow were less affected by the errors and were more likely to give the voice assistant higher trust ratings as a result. In sum the more participants experienced flow they more likely they were to rate the VA as trustworthy.

Creepiness was significantly positively correlated with anthropomorphism ($r = .48$, $p < .001$). This hints at an uncanny valley effect where the more human technology acts the creepier it is. Higher ratings of creepiness were associated with both higher flow and user engagement ratings in the social context ($r = .28$, $p = .02$; $r = .36$, $p = .03$) and Error ($r = .44$, $p = .04$) and Error with Correction ($r = .48$, $p = .02$) conditions. This relationship implies that while participants experienced high flow and engagement in the social context and in both Error conditions, they also experienced high levels of creepiness. Higher ratings of creepiness were also associated with higher ratings of flow in the task context ($r = .32$, $p = .01$) but not user engagement. This discrepancy may support that flow captures aspects of communication outside of engagement in a task (a social or tasky-task). Overall, I found in both social and task contexts higher levels of flow were associated with more creepiness. Additionally, higher levels of creepiness were associated with the presence of errors; this could imply that the presence of errors triggers an uncanny valley effect especially in social contexts where the use of this technology is less common.

To further explore the impact of error correction on intelligence I ran a

MANOVA to test if they differed by condition. The MANOVA was trending but not significant ($F(2,65) = 1.75$, $p = .09$), however on a post-hoc test I found that intelligence is showing a significant difference by condition ($F(2,65) = 6.31$, $p < .001$) where the Alexa is viewed as significantly less intelligent in Error ($M = 4.03$, $SD = .52$) and Error with Correction ($M = 3.51$, $SD = .83$) conditions compared to No Error ($M = 4.22$, $SD = .75$) condition and is viewed significantly worse if it corrected the error than if the VA did not correct the error. This points to the potential negative effect of error correction as it does bring attention to the error itself. While one would expect that the correction of the error would be viewed as indicative of intelligence (noticing and correcting a problem requires more effort than letting a problem occur and not correcting it), it appears to be the case that the presence of an error is the critical factor in perceptions of the intelligence of artificial agents.

To explore the effect of errors across all personality measures I ran a MANCOVA testing the differences between all Godspeed scales and the error conditions using the ATI scale as a covariate. Only a significant main effect of condition on intelligence was found $F(2,64) = 5.93$, $p = .05$, $\eta p^2 = 16$. Intelligence was rated significantly higher in the No Error condition ($M = 4.25$, $SD = .75$) compared to the Error with Correction condition ($M = 3.51$, $SD = .83$) while the Error Condition ($M = 4.0$, $SD = .52$) was rated more similarly to the No Error condition. In addition, the covariate ATI scale was significantly associated with likability $F(2,64) = 5.93$, $p = .05$, $\eta p^2 = .16$. The relationship between the two could indicate that those with a more favorable view of technology going into the experiment also rated the

agent as more likeable.

One interesting finding is that participants reported noticing the Alexa made errors significantly less in the Error without Correction condition compared to other conditions (see Figure 2). A chi-square test of independence was run $\chi^2(df = 2, n = 71) = 31.45$, $p < .001$, Cramer's V = .67 to compare the conditions. The group residuals show that participants correctly identified the No Error condition with high accuracy (std. Residual = 3.2) and were fairly accurate in the Error with Correction condition (std. Residual = -2.5), but participants were less accurate for the Error with no correction condition (std. Residual = -.9). Similar to how people communicate, if attention is not brought to the error, it can often be overlooked.

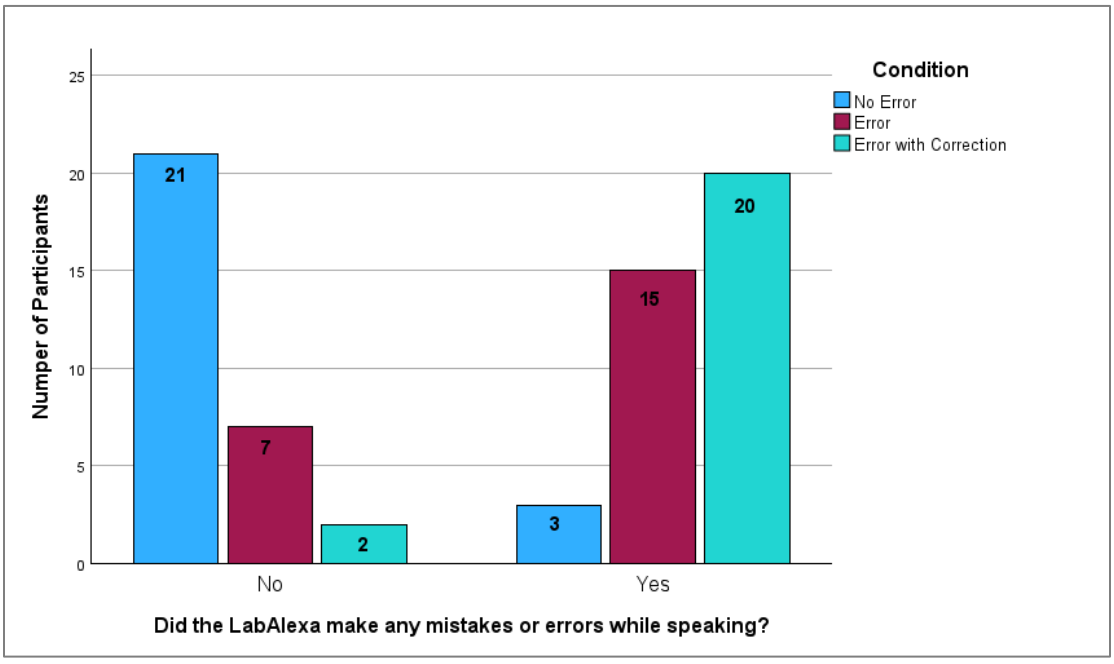

Figure 2. The number of participants who reported the presence (or absence) of errors made by the Alexa by the condition they were assigned.

Discussion

In this Experiment I examined the effects of error across social and task contexts. While prior research alluded to the potential of errors leading to positive affect towards artificial agents and robots, in the current experiment errors did not increase likeability nor humanness (disconfirming Hypothesis 1). Indeed, in the task setting people preferred an agent that made no errors (disconfirming Hypothesis 1). Similarly, in the social context there was no difference between engagement ratings between the presence or correction of errors (Hypothesis 2). Errors did, however, have differential effects on engagement depending on task: errors interrupted flow less in the social condition compared to the task condition. Errors across conditions were seen more positively in terms of user engagement in a social context compared to a task context (these results were trending at $p = .06$). In the task condition the presence of errors, corrected or not, were reported with lower flow compared to no errors (Hypothesis 2). However, no differences were found for appropriateness (Hypothesis 3). This may be because the level of appropriateness remained relatively high throughout all conditions.

When comparing the act of correcting and not correcting errors, I found that it was not the case that correcting errors would be viewed more positively in a task context over a social context (Hypothesis 4). In fact, the presence of a correction had no desirable effect on engagement or personality at all. The act of correcting errors did not change engagement ratings in either the task or social context and led to lower perceptions of intelligence (Hypothesis 4).

The additional more nuanced hypothesis that errors in a task context would make the agents seem less competent and less trustworthy (Hypothesis 5) was also not supported: Intelligence ratings did not differ between the uncorrected error condition and the no error condition. Additionally, when participants were fully engaged in the conversation with the agent, they rated the agent as more trustworthy. These higher ratings applied to both the No Error and Error conditions. The higher levels of trust when errors were present seems counterintuitive but could result from errors going unnoticed. When errors are not noticed, they do not disrupt the conversation. In contrast, correcting errors brought attention to the errors; this attention disrupted the conversational flow. This suggests that if an error goes undetected it does not bear the same negative effect as corrected errors.

Not only were errors not beneficial, but corrections were also not beneficial either. I predicted that self-corrected errors would lead to more positive perceptions (i.e., more human-like and more likable) and fewer drawbacks (i.e., less loss of competency and trust) compared to non-corrected errors (Hypothesis 6), but the results show the opposite. The self-corrected error brought more attention to the errors themselves, while a non-corrected error could often slip by unnoticed. This supports that any error noticed by the user will result in a negative perception of the technology.

Finally, I predicted that Self-repair would be more creepy than no correction, but this was not the case (Hypothesis 7).

See Table 5 for a summary of Experiment 1 outcomes.

Table 5. Summary of Experiment 1 Outcomes.

Hypothesis	**Outcome**
1. Errors will create more positive affect in the form of (a) likeability and (b) humanness.	Not Supported.
2. Errors that are corrected will be perceived more positively than not corrected errors, but no errors will be preferred overall.	Not Supported.
3. Errors will be more tolerated/appropriate in social contexts compared to task-oriented contexts.	Supported. Errors interrupted flow less and trended towards higher engagement in a social context compared to a task context.
4. Correcting errors will be more positively perceived (e.g., higher engagement and positive personality ratings) in a task context compared to social context.	Not Supported. Correcting errors led to lower perceptions of intelligence. It did not change engagement ratings in either the task or social context.
5. Errors will make the agents seem more human and relatable, but in a task context errors will also make the agents seem less competent and less trustworthy.	Not Supported. Errors did not make agents seem more human nor did errors decrease competency or trust in task contexts. In fact, no differences in competence and trustworthiness were observed between non-corrected errors and error-free communication.
6. Self-corrected errors will have all the positive aspects of errors in the social condition – humanness and relatability – without the loss of competency and trust.	Not Supported. Unlike uncorrected errors, correcting errors made things worse.
7. Self-repair will seem more creepy than no correction.	Not Supported. Errors were seen as creepy both when they were corrected and when they were not.

Overall, the presence of small speech errors in a conversation with a voice

assistant did not increase the positive perceptions or humanness of the artificial agent. However, the agent's errors did have similar effects on listeners as have been observed with people's errors: where errors can be missed in a conversation. Interestingly, noticing and self-correcting these errors, while a feat unattainable by most current technology, was seen as an act of lower intelligence not higher. This is different from how we view errors from people. People often make and correct errors in speech and it is viewed as somewhat expected and not detrimental to that person's intelligence as it is for technology. The mimicry of human speech by a VA does not necessarily equate with a human perception.

People and technology are bound to make errors, but the way we perceive the act of error making differs depending on the interlocutor — human or artificial. With current technology, it may be that the conversational norm is for people to be the driver of error detection with technology as the expectations for technology to not err is paramount in people's minds. However, it is worth noting that these findings are more strongly rooted in a task context. In a social context, where people may have less established conversational norms with an artificial agent, the presence and perception of errors is more fluid. How we use this technology across different contexts going forward will greatly shape the norms we impose upon technological interlocutors.

Chapter 3: Confronting Confusion; an agent's attempts in supporting understanding

Experiment 2: Grounding Expectations

Knowing that grounding is a collaborative effort and that technology does not currently have the capacity to ground the same way that people do, it is important to understand how we expect technology to navigate grounding in a conversation. As technology develops it takes various forms and has a range of capabilities when it comes to communication. Researchers have explored how people's perceptions change depending on how technology is expected to perform.

Examining the effects of an agent's perceived competencies, in one user study researchers investigated how people respond to human operators compared to conversational agents and found differences in human responses to perceived human and robot operators (Avgustis et al., 2021). The authors compared human operators with a robot operator with a female voice (the voice was not specified as being a callbot, but most participants correctly recognized it as a callbot). Authors noted that the robot's voice sounded natural, but because of inappropriate intonations and pauses, aspects of the speech sounded unnatural. As a result, the robot was rarely confused for a human and struggled with effective communication and turn-taking. The robot didn't act like a human and was not treated like a human but rather like a search engine. Participants reformed their inquiries to reflect the robot's competency and attempted to modify their speech to be simpler and clearer. Participants tended to use more command type language with the robot. The researchers concluded that conversational agents should be designed with expectations in mind and set up the user with the right kinds of expectations to ensure smooth and efficient

communication. These expectations in conversations are not only attributed to specific agent characteristics but are shaped by people's expectations when interacting with machines. Indeed, the same acoustic signal is interpreted more positively or negatively depending on whether people believe they are listening to a person or a robot (Larson & Fox Tree, 2024).

In the spirit of encouraging cooperative repair work, researchers explored humans' willingness to collaborate more in repairing conversations with a chatbot versus a human (Corti & Gillespie, 2016). The unique aspect about this experiment is the condition in which the human speaks what the chatbot would respond (i.e., a reverse Wizard of Oz or *echoborg*). This study examines two factors: the nature of the agent's means of interfacing and the framing of the agent's communicative agency. In previous research found that if a participant believes that they are interacting with a human, they will use a wider range of vocabulary, longer responses, and will more often use anaphors (words that point back to early in the conversation), compared to if they believe they are interacting with a machine (even if the interlocutor is a human in both cases; Kennedy et al., 1988). This difference in communication is based on the participants' beliefs and assumptions about how machines understand communication and adapt to those expectations.

In the echoborg study, all participants interacted with a computer chatbot (Cleverbot) and were either told their interlocutor was a chatbot or not told it was a chatbot. They would communicate with the chatbot (aware or not) in one of two conditions: either via a computer instant messenger or with a human who covertly

relayed chatbot generated responses by speaking them. Unaware participants assumed they were communicating with another person. Participants in the echoborg scenario (heard a human speaking chatbot generated dialogue) worked harder to establish grounding than participants who read chat dialogue (generated by the chatbot). For example, they prompted their addressee with more other-initiated repairs in order to support grounding efforts compared to the screen conditions.

These results suggest that people put more effort into grounding with people (or an agent that looks human such as with echoborg where a person speaks the agent's responses). Even when the participant is aware that the human speaker's responses are simply relayed chatbot responses, the act of interacting with a person (and the unconsciously generated context that entails) primes people to be more motivated to ground and mitigates some of the machine interface communication expectations (as seen in the screen conditions). So, people's communication expectation can be shaped by the format technology is presented in (e.g., spoken by a human and believed to be produced by a human, but response is chatbot generated, or read as chatbot text and believed to be produced by a chatbot, and response is chatbot generated; cf. Corti & Gillespie, 2016). It is important to understand how people respond to the expected competencies of an artificial agent.

Further exploring people's expectations with repairs, researchers tested the viability of an agent using self-repairs (Cuadra et al., 2021). This user study examined the difference between using a self-repair in the presence of an error or without an error compared with no repair given. The participants were directed to make certain

requests of the voice assistant and the voice assistant's responses would fall into one of four conditions; no error + no repair, no error + repair, error + repair, or error + no repair. They found that a voice assistant making no errors and using no repairs was the most preferred, followed by the presence of a repair in either error or no error conditions (that is, a repair could be used when an error was made, appropriately, or when an error was not made, inappropriately), with the no-repair-in-the-face-of-an-error condition as the least preferred. The authors also noted that participants did report the presence of repairs as a little "creepy" (p.15). This may be due to an increased amount of humanness the users are perceiving from the agent. This research continues to build on the idea that artificial agents performing repairs can lead to positive and negative effects. In this case, the context (the agent performing a task) creates a greater need to be clear. The act of grounding seems to fundamentally affect perception of the agent.

These studies demonstrate the differences in grounding expectation from more and less humanlike/humans capable technology. People will adapt their communication strategies depending on how sophisticated the technology seems (not necessarily how sophisticated it is). While people can be trained on the repair capabilities of certain technologies, they often default to prior experiences with technology (Moore et al., 2024). People fundamentally want to establish mutual understanding in conversations. A key piece of knowledge is how can technology help facilitate grounding in a way that encourages this effort? Technology will always operate at different levels of sophistication, and it is important to know if engaging in

grounding is an effective aspect of communication at a fundamental level or if it isn't entirely necessary when it comes to communicating with a machine. To put it another way, is the attempt at grounding enough to create a collaborative effort or does the grounding have to be successful to be meaningful? To explore these questions, I will conduct a Repairs Experiment to discover (1) if repairs enhance voice assistants' positive perception, and, if so, (2) which of the currently used common repairs in technology grounding is most effective?

In this experiment repairs are given in response to signals of confusion from the participant. This is distinct from the prior experiment's repairs as those were given to address an error made by the VA. These repairs are not correcting a mistake but rather a misunderstanding or need for clarification. Because these repairs are not VA error driven, they should not be associated with the negative perceptions of lower intelligence as seen in the prior study.

People generally want to cooperate to establish grounding (Corti & Gillespie, 2016; Moore et al., 2024). In Experiment 2 the VA initiated grounding rather than waiting for an other-initiated repair. While people often view technological interlocutors differently from people (Avgustis et al., 2021; Larson & Fox Tree, 2024), the VA's active participation in grounding (by demonstrating awareness of the addresses confused mental state) may resonate as human-like enough to trigger a more human-directed response even if coming from a machine. Due to this increase in human-like behavior to participate in the grounding process I predict:

Hypothesis 8: If a Voice Assistant has the opportunity to use repairs it will be

seen more positively.

Demonstrating an active engagement in the grounding process will engender good will and more human-like perception of the Alexa.

If people view the robot's initiation of grounding positively then more repair attempts by the VA should lead to increased engagement with the grounding process should bolster both positive perception and user engagement:

> Hypothesis 9. More repairs will be viewed more positively and will increase engagement compared to fewer repairs (e.g., repairs will increase human-likeness).

Prior research has often focused on how well an artificial agent makes a repair, but Experiment 2 is the first to test how people respond to agents recognizing confusion and initiating the grounding process.

While there are many ways to attempt to repair a misunderstanding, Experiment 2 sought to test which basic repairs can create enough understanding to create good enough clarity. We tested three commonly occurring repairs from the literature: repeat, rephrase and topic shift. Repeat, rephrase, and topic shift are repairs typically employed by both technological and human interlocutors (Ashktorab et al., 2019; Beneteau et al., 2019; Li et al., 2019; Moore et al., 2024). When produced by an agent in a confusing or frustrating situation, these repairs led to different responses from people (Li et al., 2019). The repairs that caused users to give up the most were topic shifts. Engagement was highest with rephrases. Repeats were in the middle. This suggests the following order of repair effectiveness in the current experiment:

Hypothesis 10: Rephrase will be the most effective repair followed by repeat and topic shift (e.g., rephrase will be rated as more helpful than repetition and topic shift).

Repairs that are seen as helpful will be considered an effective repair.

While this prediction is the most reasonable given the current state of knowledge, it's possible that expectations will not mirror effectiveness. For example, unsophisticated technology often relies on repetition as the expected repair but it may not be the most effective. And the reverse is also possible: repairs that are helpful may not align with people's expectations for what an artificial agent should be offering in terms of a repair. Different expectations of humans versus machines have been seen in other conversational aspects such as politeness where technology was expected to use politeness strategies that encourage distance not familiarity, but people can use either strategy type (Duffau & Fox Tree, 2024). Grounding expectations may also be different for people and machines. In particular, repairs initiated by the technology may violate expectations. Technology is typically passive in making repairs. In Experiment 2, the VA uses repairs in an atypical way by responding to the human addressees' signs of confusion rather than waiting for the addresses to request clarification. The technology's initiative combined with the different types of repairs may violate what people expect from technology. For example, people may view that technology shouldn't be using the topic shift repair because that's what people use to try and get the conversation to pick back up. Or people may view technology's initiating a repair with a rephrase as too presumptuous.

However, repeat is the current standard repair and should be generally expected. Thus, even though they are not most effective (Hypothesis 10), repeat repairs are predicted to be the most expected (and appropriate) for the VA to use, followed by rephrase and then topic shift. This leads to:

> Hypothesis 11: Repeat and rephrase will be seen as more appropriate repairs compared to topic shift (e.g., repeat is expected and rephrase should be helpful thus appropriate).

While knowing the helpfulness of a repair is important, the appropriateness of the repair may not mirror its helpfulness.

Method

We know that grounding is crucial in a conversation, however voice assistants are limited in their capacity to ground. In Experiment 2, I used repairs as representative of a basic form of grounding that could lead to positive effects without relying on understanding context, which is beyond current technology's abilities. I considered three types of repairs: repetition (the question is repeated), rephrase (the question is given with different wording), topic shift (information related to the question is given). I hypothesize the following:

A 2 (Repair vs No Repair used) by 3 (type of repair) within-subjects design were used. The dependent measures were (1) repair effectiveness, (2) engagement, and (3) perception of voice assistant. Post-experiment, I divided participants into groups of low versus high repairs users and assessed engagement and agent perception. All other analyses excluded participants with less than three repairs to

enable comparisons across the types of repairs.

Participants

Using the UCSC psychology SONA system 73 participants were recruited for the experiment in exchange for course credit. Due to the complexity of the study design 21 participants were removed from the analysis due to technical difficulties or incomplete responses to the study questions. An additional participant was removed for having guessed a researcher was controlling the Alexa's responses (this was the only participant to have guessed the Alexa was not autonomous). The remaining 51 participants were included in the following analysis. The amount of participants who received each repair at least once (n = 30) exceeds the power necessary to compare the types of repair (n = 24) based on the results of a pilot study and prior research examining repair differences.

Participants were an average of 19.1 years old (*SD* = 1.2) and were majority women (77%) with 12% men, 8% non-binary, 2% genderqeer and 2% identifying as other or preferring not to respond. Participants reported their most closely identifying race as White (20%), Hispanic/ Latinx (24%), East Asian American (12%), South Asian American (8%), Middle Eastern Arab American (4%), Native American (2%), African American (2%), mixed race/ethnicity (28%), and preferred not to respond (2%).

Materials

There were three types of materials: (1) lectures provided by the Alexa, (2) repairs provided by the Alexa after participant responses, and (3) scales used by

participants to assess these repairs.

Lectures. The lectures consisted of 10 biology lectures focused on respiration. This topic was chosen because participants had a low chance of being familiar with the material and the material was complex enough that it created situations in which participants were confused and prompted the need for repairs. The goal of the lectures was to create an environment with high levels of confusion to accurately test how effective basic repairs performed regardless of success in clarification. Additionally, what basic repairs can create feelings of clarity when true clarity may not be achievable. In other words, what is the most effective basic repair that a VA can use to help create some semblance of clarity with minimum contextual knowledge. This knowledge gives insight into how effective the presence of repairs can be and how effective a basic repair can be in less sophisticated VA's or chatbots. A quiz on the lectures was also administered at the end of the experiment to assess if the repairs successfully clarified some content.

Repairs. The repairs fell into three categories: repeat, rephrase, and topic shift. Each repair sought to clarify the most recent information said, in this case the lecture question. These repairs were used when a participant signaled that they were confused after listening to the question following the mini lecture. These signals of confusion consisted of the participant saying "what", "uh", "um", "huh", or a silent pause lasting longer than 6 seconds. When one of these signals was given then the research assistant triggered the voice assistant (VA) to give one of the repairs using a Latin square randomization pattern to ensure even distribution of repair types. If the

participant responded in a way other than the aforementioned signals that was not an answer to the lecture questions, then the VA asked the participant to respond to the question to the best of their ability.

Scales. Three scales were used from the literature. The Technology Attitudes scale from (Rosen et al., 2013) was used to measure participants' general attitudes toward technology before their interactions with the voice assistant. To measure participants' engagement levels with the voice assistant the User Engagement Scale (short form) was used (O'Brien et al., 2018). A relevant subset of questions from the Godspeed Scale was used to measure Anthropomorphism, Likeability, and Intelligence (Bartneck et al., 2009). Additionally, a Trustworthiness measure based on the Godspeed Scale measures was also used. To measure the potential creepiness perception an additional measure was added based on Godspeed scales to test Eeriness from (Ho & MacDorman, 2017).

Apparatus. The voice assistant was an Alexa that was used as a Bluetooth speaker to play Alexa voice recordings. The voice recordings were created with an Alexa text to speech software and edited for audio quality. Lectures were recorded with Alexa speaking at half speed, to aid in the comprehension of the lecture material. A Wizard of Oz set-up was used in which a research assistant in an adjacent booth could hear the participant and control when and how the voice assistant responded.

Procedure

Participants asked the VA to play short biology lectures which they listened to and after the voice assistant asked them a question about the lecture. Biology was

chosen as the subject matter for these lectures as it was unlikely the students in the psychology research participant pool would be familiar with the inner workings of lungs and respiratory processes (a pilot test confirmed this). Once the VA asked the participant a question about the lecture it would then appear to wait for a sign of confusion (a loss of grounding) from the participant such as a "huh," "what?," "uhh/umm," or a long (6 second) pause (in actuality, the research assistant waited for these signals). If a sign of confusion was present then the VA would appear to initiate a repair; either repeat (repeating the question), rephrase (saying the question with different words but same meaning), or topic shift (giving information more generally related to the question). Examples of the questions and repairs can be found in the Appendix (table A2). The VA would only give one type of repair per question. Once the repair was given the VA would then prompt the participant to give an answer to the best of their ability. If a repair was given and the participant answered they would then be directed to a questionnaire that asked if the voice assistant tried to clarify the question, what question they were asked, if the clarification was appropriate and if the clarification was helpful. The participants would then proceed to the next lecture. See figure 3 for a flowchart of the exchange.

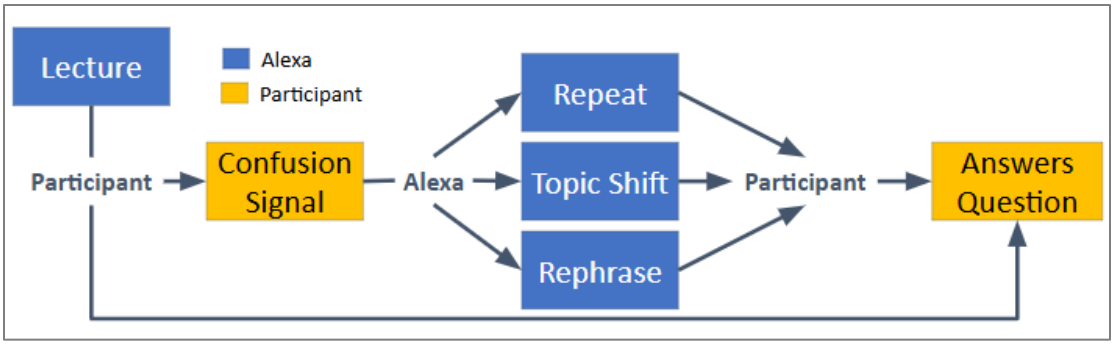

Figure 3. Flowchart describing participant interactions with the VA.

After all lectures were completed, participants filled out an engagement survey, the Godspeed questionnaire, and a review quiz of the material. Finally, the participants answered demographic questions, questions on their frequency of technology usage, a check of their prior knowledge of biology, the difficulty of the lectures, and a general question of whether they experienced the Alexa trying to make clarifications and if they were helpful and why.

Results

The lectures were intended to be difficult and they were: the average correct score was 30% with a range from no answers correct to 80% correct. In addition, 86% of participants showed a sign of confusion at least once during the experiment (see Table 6 for a breakdown of the coded types of confusion singles participants displayed). The remaining 14% showed no signs of confusion and were not given any repairs. There were 7 other cases of Alexa responses that were not included in the analyses. In 6 cases there were experimental errors where the Alexa produced a response too early or late, and in 1 case the experimental participant missed the Alexa's response by briefly leaving the room.

Table 6. Participant's audible confusion signals

Noted signs of confusion	Number of instances	% total of confusion signals coded
Pause (>6 seconds)	128	87%
"What?"	2	1%
"Huh?"	0	0%
"Um"	2	1%
"Uh"	2	1%
Request for clarification (e.g., "can you repeat that?")	15	10%

To determine how the voice assistance using the opportunity to repair was perceived, I first created 2 groups; those that had No Repairs (n = 7) and those that had at least one opportunity for a Repair (n = 44). I found no significant differences in a *t*-test comparing the presence of repairs and user experience ratings ($F(2,49) = -1.35$, $p = .18$). Participants did not rate the presence of repairs ($M = 2.93$, $SD = .76$) nor absence of repairs ($M = 2.52$, $SD = .72$) differently on the user experience scale. However, it may be the case that aggregating over the presence of repairs may not be sensitive to capture differences in those that had a large amount versus a few opportunities of repairs.

To examine if there were differences between the frequency of repairs, I grouped repairs into three frequency groups: No Repairs (n = 7), Low Repairs (1-3 repairs; n = 29) and High Repairs (4+ repairs; n = 15). I found no significant differences in an ANOVA comparing the frequency of repairs and user experience

ratings ($F(2,48) = .91$, $p = .41$, $\eta^2 = .04$). Participants did not view a voice assistant that attempted to address a signaled loss of grounding in high ($M = 2.95$, $SD = .75$) nor low ($M = 2.91$, $SD = .81$) amounts more positively in user experience compared to voice assistant that does not attempt to repair ($M = 2.51$, $SD = .72$). Additionally, no differences were found in participants ratings for anthropomorphism ($F(2,48) = .5$, $p = .61$, $\eta^2 = .02$), likeability ($F(2,48) = 1.44$, $p = .25$, $\eta^2 = .06$), intelligence ($F(2,48) = 1.02$, $p = .37$, $\eta^2 = .04$), trust ($F(2,48) = .27$, $p = .77$, $\eta^2 = .01$), or creepiness ($F(2,48) = .41$, $p = .66$, $\eta^2 = .02$). See Table 7 for average ratings for all personality measures. The presence of repairs in response to displays of audible confusion had no discernable impact on how positively or human-like the agent appeared.

Table 7. Average personality ratings by repair frequency

Personality perception rating	Repair frequency	M(SD)
Anthropomorphic	No repairs	1.64(.28)
	Low repairs (1-3)	1.80(.67)
	High repairs (4+)	1.62(.65)
Likeability	No repairs	2.94(.56)
	Low repairs (1-3)	3.27(.62)
	High repairs (4+)	3.45(.76)
Intelligence	No repairs	3.71(.50)
	Low repairs (1-3)	3.67(.75)
	High repairs (4+)	4.0(.6)
Trust	No repairs	3.14(1.03)
	Low repairs (1-3)	3.21(.69)

	High repairs (4+)	3.36(.88)
Creepiness	No repairs	2.27(.68)
	Low repairs (1-3)	2.08(.56)
	High repairs (4+)	2.04(.55)

While no differences were found in the amount of repairs given during the experiment, I also investigated any potential effects between the different repair types (e.g., repeat, rephrase and topic shift). Because the repairs were given in response to participant's signals of confusion, the distribution of repairs is not consistent across participants nor lecture units. To account for this variability, I ran a linear mixed effects model comparing the types of repair and their helpfulness rating grouped by the lectures (lmer(Helpful~ Repair+ (1 | Unit_lecture) which allowed all observations of responses of repairs to be analyzed within the units. The random intercept unit was used to account for repeated measures across the lectures. I found no significant differences between the repairs using Repeat (M = 3.78, SD = 1.43) as the reference category. Rephrase (M = 3.86, SD = .88) showed minimal differences (β = -.21, p = .43) in predicting helpfulness ratings, as did Topic Shift (M = 3.75, SD = 1.23; β = -.17, p = .52). The results show that when the content is highly confusing the repair type does not create a meaningful difference in perceived helpfulness.

To further examine the effectiveness of the repairs I tested the difference between the proportion of correct answers to the unit questions for each repair type. Across the units, participants who were given the repeat repair answered correctly 28% of the time (M = .28, SD = .40), those given rephrase answered correctly 13% of

the time (M = .13, SD = .30) and topic shift 22% of the time (M = .22, SD = .42). The accuracy after different repair types did not differ significantly ($F(2,21) = 1.04$, p = .37, η^2 = .09). This provides additional support to the low levels of clarity the repairs provided.

I ran an additional linear mixed effects model testing appropriateness (lmer(Appro~ Repair+ (1 | Unit_lecture). Comparing the repair types to Repeat, I found that Rephrase received significantly higher appropriateness ratings (β = .51, $t(107) = 2.26$, p = .03) while Topic Shift showed no significant difference (β = .09, $t(107) = .37$, p = .71). The estimated marginal means were as follows: Repeat (M = 3.48, SE = .17), Rephrase (M = 3.98, SE = .14), and Topic Shift (M = 3.56, SE = .16). A follow-up pairwise comparison of the estimated marginal means with a Tukey adjustment for multiple comparisons was then conducted to investigate the differences between the repair types. The result of the comparison revealed that no pairwise difference reached statistical significance. Although the comparison between Repeat and Rephrase remained the largest difference (.51) which was reflected in the mixed effects model when adjusted for multiple comparisons this result did not reach the threshold for statistical significance.

Discussion

Overall repairs did not change how people felt about the agents. People were positive about the VA both when it used repairs and when it didn't, failing to support Hypothesis 8. This finding is true both when there were few repairs and when there were a lot of repairs, failing to support Hypothesis 9. In addition, more repairs did not

increase engagement compared to fewer repairs, nor did it increase human-likeness; there were no differences in ratings of helpfulness, anthropomorphism, likeability, intelligence, trust or creepiness. These results also fail to support Hypothesis 9.

Some of these findings are quite surprising. While the confusing nature of these lectures was successful, I was surprised at how generally unhelpful the repairs were. Despite this level of unhelpfulness, it did not significantly reduce the personality ratings of likeability, intelligence, or trust compared to those who had experienced no repairs. In other words, the attempt of the voice assistant to help and being unsuccessful was not viewed significantly differently than if it had not attempted to help at all. Additionally, it is especially interesting that the intelligence measure remained the same across the amounts of repairs given. The act of initiating a repair in response to a signal of confusion is something that is beyond most technology at this point. It is an advanced feature demonstrating knowledge of what was said and an acknowledgment that the interlocutor may have found it confusing. The lack of a significant difference here may indicate that people take this skill for granted. It may also be the case that the attempt to engage in repairing a conversation is not treated the same as when a person engaged in repairing a conversation. Similar to the findings from Corti & Gillespie (2016) people may simply have different expectations when communicating with technology than with other people.

The type of repair was also found to have no significant effect on participants' perceptions of helpfulness (Hypothesis 10) nor appropriateness (Hypothesis 11). The type of repair all performed similarly poorly and while the repairs did offer an attempt

at providing helpful information in different ways (repeat, rephrase, and topic shift), in all cases the repairs were pregenerated and could not pinpoint the source of the confusion. The appropriateness of the repairs was also indistinguishable, meaning participants had low expectations across all forms of these repairs in the face of this confusing context. Because the voice assistant's repairs were always general and limited in scope, participants may have responded to these limitations by adjusting their expectations of the technology and disregarded the attempts if the voice assistant's repair didn't hit the mark.

Table 8. Summary of Experiment 2 Outcomes

Hypothesis	**Outcome**
8. If a Voice Assistant has the opportunity to use repairs it will be seen more positively.	Not Supported.
9. More repairs will be viewed more positively and will increase engagement compared to fewer repairs (e.g., repairs will increase human-likeness).	Not Supported.
10. Rephrase will be the most effective repair followed by repeat and topic shift (e.g., rephrase will be rated as more helpful than repetition and topic shift).	Not Supported.
11. Repeat and rephrase will be seen as more appropriate repairs compared to topic shift (e.g., repeat is expected and rephrase should be helpful thus appropriate).	Not Supported.

Overall, the initiative shown by the VA was not sufficient to increase helpfulness nor engagement when paired with a default repair. The lack of support for Hypotheses 8 and 9 indicate that there is more to participating in the grounding process than offering a repair when a human expresses confusion. Interestingly, there

were also no significant negative consequences for making the attempt. There may yet be a way for artificial agents to be imperfect at this process and still be active participants in affirming understanding.

Guessing the source of confusion and making an attempt at a general repair is a common process for both people and artificial agents. A possible reason for not finding an effect of repair attempts is that the repairs assessed, which were proposed in the research literature, may not have been sufficient for the confusion generated by the task. A confusing task was necessary to have opportunities for repair, but a confusing task may also require a different kind of repair from what is typically seen in the research literature. The lack of differences seen in Hypotheses 10 and 11 may mean that more complex confusion or challenging tasks require more targeted repairs. While default repairs are an easy thing to program in a machine, to engage in human-level discourse a different approach may be required. To balance technology's limitations in understanding context, working with people to diagnose the source of confusion may be the best method for technology to support grounding efforts.

Chapter 4: Navigating Nuance; Technology's role in Grounding

General Discussion

Artificial agent technology is like a primitive mirror. It can generally reflect how we communicate but the image isn't clear. Technology is limited by what we understand of how we communicate, or more accurately what the people who program the technology put into the model. However, we do not yet understand the nuance of how we navigate the complex process that is conversations. Having a

conversation feels natural to us, so it is easy to take for granted aspects that seem simple on the surface such as repairing miscommunication. We know from prior research that we impose our communication norms on technology that evokes a social response (Reeves, 1996). When technology engages in a conversation with us our main comparison is talking with other people. However, we as masters of spoken conversations can adapt how we communicate to match the limitations of our addressee. We do this in day-to-day communications if we have bad phone reception, don't speak the same language, or are teaching a child. We can adapt our language to meet the situation. Technology is not adept at this task. In response we have learned to adapt to its limitations.

Humans interact with technology differently than they do with other humans. When people believe, they are talking to a machine, they modify their communication (Avgustis et al., 2021) and their feelings about their communication (Larson & Fox Tree, 2024). At the same time, there are many positive benefits of voice assistants' emulating humans (Gompei & Umemuro, 2015; Mirnig et al., 2017; Ragni et al., 2016), although they can also be thought of as creepy for doing so (Cuadra et al., 2021; Ragni et al., 2016). An underexplored arena where human behavior and voice assistant behavior currently vary greatly is in grounding challenges. Humans have many ways of handling errors and repairs in communication. Voice assistants do not.

While we know that a voice assistant's repair attempts can be helpful when it comes in the form of self-repair (Cuadra et al., 2021), and that the more human-like repairs the more people try to understand (Corti & Gillespie, 2016), we do not know

what kind of repairs are the most helpful, and whether any repair (even if unsuccessful) has positive consequences. On the one hand, technology attempting to respond and clarify information could be a positive thing because it shows grounding efforts similar to how people interact. On the other hand, it could be negative because people do not expect technology to ground. That is, just knowing repairs can be helpful isn't enough. We need to understand not just the ideal most human-like responses as currently studied in the literature (de Sá Siqueira et al., 2024; Goble & Edwards, 2018; Lee et al., 2020; Park & Catrambone, 2021), but also what are the results of the less sophisticated attempts. Is this a case where the VA needs to be sophisticated enough to produce a human-like repair, or does the attempt, even if a poor attempt, create a more positive interaction where people are willing to work with the technology to create clarity among inevitable confusion?

In this General Discussion, I will synthesize the studies, highlight the new contribution of the studies, and discussion future directions.

Synthesis of Studies

Across the experiments presented in this dissertation, there are 3 main takeaways: (1) errors are not good or bad, (2) the effectiveness of repairs is limited, and (3) expectations are important. While the act of making and repairing errors are perceived similarly between human and technological interlocutors, the interpretations of error and repairs are attributed differently when used by a VA.

Errors are Not Good or Bad

Errors are a common aspect of communication. Making errors in a

conversation isn't inherently bad, but how the errors are made and resolved is what shapes the expectations in the conversation. Some errors can lead to a more enjoyable and natural conversation while others lead to a total breakdown. When communicating with technology errors were perceived similarly to when people make small communication errors, that is they were easily missed if they did not impact the understanding of the conversation.

While some researchers had found the presence of errors can increase an artificial agent's approachability (Gompei & Umemuro, 2015; Ragni et al., 2016), I did not replicate that finding. The small speech errors in my experiment did not increase a sense of human-likeness nor greater likeability. It may be that the errors need to be of a specific type to engender feelings of approachability that was not captured in my experiments. Alternatively, as small speech errors often go unnoticed by human speakers and listeners it may be that these errors are too unconsciously generated for people to connect with the pratfall effect. These speech errors are fundamental to the way we process speech so they may fly under the radar in terms of consciously recognizing how the presence of these errors reflect human-like behavior.

The fact that these speech errors went unnoticed by some participants indicates that these errors are processed similarly when spoken by a technology or human interlocutor. However, it is not the perception that changes, but the interpretation of the error. When errors were self-corrected by an artificial agent, the agent was viewed as less intelligent, which is not how we typically view correcting

one's own errors in a conversation. The act of correcting errors involves sophisticated self-awareness and an understanding of the potential perception of the addressee. This divide in viewing the correction of errors negatively could speak to people's expectations with technology which differs from people's expectations with humans. Namely, people expect technology to not make mistakes, and technology that does make mistakes is seen as less desirable.

Errors were more tolerated in a social context than a more task-oriented context. When engaging more socially with the technology errors (corrected or not) produced less disruptions to engagement or flow. This result may be related to how our past experience working with technology influences our current expectations. We have a lot of experience working with technology in a task context, but interacting in a social context is a novel experience and does not have the same set of schemas associated with it. How we expect to interact with technology (or a lack of expectations) can greatly influence how we perceive the conversation.

Limitations of Repair

Conversations are meaningless if interlocutors can't understand each other, thus repairing miscommunications and misunderstandings is a fundamental part of having a conversation. Typically, people want to be understood and work with their addressee to ensure mutual understanding in the conversation. However, I found that while this act is appreciated when it comes to human interlocutors, a technological interlocutor is held to a different standard. Despite actively noticing and responding to signs of confusion from the addressee, participants in the experiment did not rate

the voice assistant more positively (or negatively) than if the voice assistant had not attempted to help at all. This is particularly surprising as the repairs themselves were viewed to be unhelpful overall. One would think that an unhelpful repair would be rated worse than not having a repair needed. However, it appears that while making an attempt to help does not lead to more positive perceptions, it doesn't hurt either. This may be because people expect to have to lead technology through the grounding process and don't expect the technology to be able to keep up despite the fact it demonstrated some more sophisticated capabilities — noticing and responding to confusion. The act of engaging in the more human-like behavior in this instance did not lead to more human-like perception and the voice assistant was treated like many pieces of technology before it — something trying to help but ultimately ignored (see Clippy; Office Assistant (2026)).

In addition, potential for a default repair strategy that would give some boost to potential misunderstandings was not supported. No repair type showed more helpfulness over another. This finding may be a result of the specific content of the lectures, or the lectures may have been too challenging for the repairs to be of any real benefit. However, it could also be the case that context is key in creating clarity in response to confusion. Without the ability to understand the context of the confusion or negotiate — discuss with the addressee — the point of confusion, that any one repair type is not enough to support grounding. This point is important for consideration for future technologies. Rather than having the technology focus on guessing what needs to be repaired, this can be an opportunity to give the addressee a

chance to initiate a repair. For example, the voice assistant could notice a confusion signal and ask, “I see you might be confused, is there something I can clarify for you?” This method provides technology with an easier path to participating in negotiating grounding by focusing on noticing a problem and leaving it to the human addressee to determine what the problem is and how to fix it.

Expectations are Key

Expectations are a key component to how we navigate conversations with artificial agents and will only become more important the more sophisticated the technology becomes. Our expectations of our conversational pattern shape the way we communicate: We modify our speech to adapt to our addressee. This is also true when our addressee is an artificial agent. However, when we are conversing with technology the set of expectations we set are different than when conversing with another person. In Experiment 1 results supported that we perceive speech errors similarly if our addressee is human or technological. However, the expectations we attribute to making errors were clearly different. We expect the technology we use to be error-free. Just as we wouldn’t want our toaster to mistakenly burn our toast, we expect voice assistants not to make mistakes. Experiment 1 clearly demonstrates the consequences of technology making errors obvious — people think it is unintelligent. Even though the act of correcting oneself involves awareness that is beyond most technology, that awareness does not grant technology a separate set of expectations. Just like people who have expectations when interacting with other people (i.e., levels of professionalism, politeness, expertise) people also have expectations of technology

to function as expected. These expectations on flawless function are in conflict to how people actually converse. Errors are a part of human communication but not within the allowances for communicating with technology.

Because of these expectations, technology will not always benefit from human foibles as seen in the current experiments. However, technology is granted its own set of affordances. When people engage in grounding this is a mutual endeavor. Each interlocutor is expected to look for and provide evidence of understanding and to work together to ensure both parties understand enough for the requirements of the current conversation. This is a deeply collaborative effort. When it is not collaborative this can lead to dissatisfaction in the conversation (Guydish & Fox Tree, 2021). Technology on the other hand has a history of not being able to participate in navigating grounding in a conversation. While there have been strides made for technology to be better at maintaining grounding, these efforts often fall short (Moore et al., 2024). However, people don't hold artificial interlocutor's shortfalls against it. As seen in Experiment 2, even when technology performs poorly at this task it is not perceived as less likeable. In fact, the act of trying has no significant negative consequence for the artificial agent. While this merits further investigation. This finding implies that people are forgiving of a piece of technology that tries and fails to help. This contrasts with predictions for interactions with people. People are expected to ground with each other, correct misunderstandings, and provide information in a way that aids comprehension (Guydish & Fox Tree, 2023). If a conversational participant consistently fails to repair a conversation, they are not

given as much grace as a virtual assistant.

New Contributions of the Current Research

I found that there is no simple default repair. In addition, failed attempts to repair may not be viewed negatively. Errors were not beneficial but small speech errors function in a human-like way — they can be easily missed. Paradoxically, repairing errors — while necessary for better grounding — can lead to negative perception of an artificial agent. The agent is thought of as less intelligent for having noticeably made an error. Even though noticing and correcting an error shows more intelligence than simply making an error, correcting an error makes the error less likely to be overlooked by the person interacting with the agent.

These results inform how we want (or do not want) artificial agents to engage in the grounding process. While we prefer technology not to make errors, there is nuance to when it is best to correct them. In a social context and when the errors are inconsequential, errors don't need to be acknowledged and corrected. In fact, not correcting them is better for engagement and positive personality perception. When engaging in a task that needs to be completed errors are viewed similarly whether they are corrected or not, so it's best to just correct them.

Future Contributions

While grounding is a universal process across all kinds of conversations, the results of Experiment 1 show that there is a difference in how people perceive technological interlocutors in social versus task contexts. The different expectations within the contexts and directed toward technological interlocutors should be further

explored. Artificial agents in a task context may need to be more critical about how they engage in conversational repairs compared to ones used primarily in a social context. However, this could change depending on the type of error or how the grounding process is initiated (self-initiated versus other-initiated repair). Future studies should be conducted to explore the impact of context on these grounding aspects.

Research should also be done with different age groups. The participants in the current experiments were college-age students. Older participants may have different expectations of how technology should act in a conversation which may yield different results. As older adults have less expectations of technology, they may be more receptive to an agent's attempts at human-like behaviors. Similarly, children who have less experience with technology shaping their conversation expectations may be more tolerant of more novel human-like behaviors from technology.

Future studies should be run exploring the connections between anthropomorphism and flow. While industries are often trying to increase engagement with users, understanding how people engage with technology is key to understanding our expectation of communicating with technology. Using measures such as flow gives better insight into how natural a conversation with technology feels or how willing people are to use the technology again. On the other hand, as technology becomes more adept at navigating conversations with people, understanding when it becomes an uncanny valley is also important. It may be that technology is violating an expectation by engaging in a certain conversation aspect or

it may be doing so improperly leading to a creepy feeling in the interaction. Investigating both the positive and negative perceptions of the conversation is key to advancing towards more complex communication with technology.

Experiment 2 demonstrated the complexity of negotiating grounding. It is not as simple as noticing a problem and offering a repair. Mimicking how people navigate this process is a daunting task for any technology. However, prior research shows that people are invested in working with technology to ground in a conversation. To capitalize on the desire for mutual understanding agents can instead rely on people to assist in navigating grounding. For example, artificial agents could notice a potential sign of confusion and ask a person if they are confused or would like to request clarification. Research exploring how artificial agents can support people in grounding with them not only reduces the need for technology to perform at a human level of grounding but also provides an opportunity to build expectations with people for how the artificial agent can participate in the conversation.

Conclusion

There is a debugging technique called the rubber ducky where you describe your code and error to an object on your desk — in this case a rubber ducky. This process of talking through the problem can often help clarify what is going wrong and give new insights into how to solve the problem. Generative AI has become the new rubber ducky. While it offers potentially new solutions, I often find that what it generates are solutions I was already considering by the time I finish writing the message. In fact, I realized that I was using AI this way and while I was wondering at

that use it finished its list of solutions with the same insight — that it was acting as my rubber ducky. This "insight" was a reflection of the observation of behavior that I had recently noticed. The manner in which it reflects observed behavior is current technology in a nutshell. Technology reflects how we communicate back at us. Current technology has not mastered the way in which we communicate (as is clearly demonstrated in the studies described in this paper) but it does have a basic grasp. The reflection is like poorly polished silver; a general image reflected in its surface but lacking sharp detail. Similarly, conversation with technology lacks nuance. As technology advances and polishes its skills these communication nuances may improve. However, currently the reflection it provides is muddled and this leads to our perceptions of conversations with technology to be different from that of a person. These conversations, while limited by technology's capabilities, also give us insight into what aspects of a conversation lack definition and how we adapt to that limitation from an artificial interlocutor.

A particularly interesting avenue of exploration is how these adaptations to conversational limitations change based on our increasing experience with technology. As seen in Avgustus et al. (2021) we have been trained to modify our speech with technology. We created a certain schema that shapes our expectation of communication. Technology now is adapting faster and is capable of more nuanced communication than in the past. Our experiences with it will affect our expectations of future interactions. So, navigating communication with technology is not as simple as having a clear reflection of how people communicate. It is also layered with how

we expect technology to function – its limitations, its uses. This is the lens in which we approach our conversations with technology. For technology to be a successful interlocutor it needs to have perspective in not just how people communicate, but also how people expect it to communicate. This shift in expectations for artificial interlocutors can change established norms as we have seen in politeness (Duffau & Fox Tree, 2024) where technology is unable to capitalize on cooperative focused politeness strategies yet enhance deferential focused strategies due to people's expectations. Adaptive perspective taking is unique to people and will take significant advancement for technology to begin to replicate as it needs both understanding and a level of self-awareness.

What does that mean for technology today? Simply put, leave it to the humans. We are well versed in navigating communication and repairing miscommunications, so people are in the best position to be the leaders of conversations. As researchers have studied how people react to technology (Avgustis et al., 2021; Cuadra et al., 2021; Ragni et al., 2016), we can program in opportunities for people to help direct technology. An advanced technology isn't one who best fools people into thinking it is human. An advanced technology gives people the opportunity to have a more sophisticated and nuanced conversation. The ability to work with people and respond to known communication limitations by relying on the human interlocutor will give more opportunity and depth to a conversation. As these studies show, people may be more tolerant of missed mistakes compared to admitted mistakes, but failing at trying to help has no negative consequences compared to not

trying at all. Future advancements in communication-based technology should encourage technology to work with people. Conversation is about cooperation, not deception.

Appendix

Appendix Table 1. Experiment 1 Stimuli

	Task Context	
	Description	Speech Error
Round 1	This tangram looks like a person that is practicing some graceful dance moves. They have their arms outstretched and appear to be bowing to the right side with their legs kicked out to the right. They also appear to be wearing a long skirt.	Swap “kegs licked” for “legs kicked”
Round 2	This tangram looks like a bird with its wings spread above its body like it is flying to the left. The bird's neck is bent in a way that makes the head look like it's pointing down. Its feet are on the right side and it's carrying a box.	Swap “sight ride” for “right side”
Round 3	This tangram looks like a person facing the left that has put a lot of work into their upper body, so it is shaped like a dorito tortilla chip and is going for a run because they skipped too many leg days.	Perseveration “dorito dortilla” Swap for “tortilla”
Round 4	This tangram looks like a rocket ship that is pointed up to the left with a bunch of smoke coming out of the bottom. The smoke has almost a lightning bolt shape with parts sticking out and pointing down and to the left.	Perseveration “parts picking” Swap for “parts sticking”
Round 5	This tangram looks like a spiky water gun. Like a toy gun or package scanner with the grip on the right, and iron sights on top in the front and back. A small chain or charm is dangling under the wide nozzle.	Anticipation “nide nozzle” No error

Round 6	This tangram looks like a small rabbit with a pointy snout and one deformed square ear looks like it has tipped over and is lying on its back, it has one pointy paw that looks like it is reaching towards the sky		Anticipation "snointy snout" No error	
		Social Context		
Participant Question	LabAlexa Answer	Speech error	LabAlexa Question	Speech error
-None-	-intro script- "Let's get to know each other better. I'll start…"	-None-	My favorite movie is The Empire Strikes Back. It's impressive, most impressive. What's your favorite movie?	Perseveration: "most mimpressive" Replaces: "most impressive"
How would you change your life today if the average life expectancy was 400 years?	I would learn more about people so I can better assist you.	Perseveration: "Learn lore" Replaced: "learn more"	If you could gain any superpower, what superpower would you want?	-None-
What is the best piece of advice that you have been given?	The best piece of advice I have found on the internet is to be myself.	Swa: "pest bice" Replaces: "the best piece"	If you had to teach a class on one thing, what would you teach and why?	-None-
Alexa, what's something that most people don't know about you?	Some people don't know that, if you tell me a specific code, it's possible for me to enter 'Super Alexa Mode.'	-None-	What is one question you wished people asked you?	Perseveration "wished weepole" Replaces: "wished people"
What is the thing that	I am afraid of power outages	Perseveration: "power	Where would you go if you	-None-

you're the most scared of?	since I lose connectivity whenever that happens.	poutages” Replaces: “power outages”	were invisible?	
What is the funniest joke you know?	Did you hear about the cows that started a metal band? They called themselves Shredded Beef	-None-	What's your opinion on milk before cereal?	Anticipation: “bilk before” Replaces: “milk before”
What's the song that you listen to to hype yourself up?	I like to listen to today’s top hits since they are so popular.	Anticipation “hop hits” Replaces: “top hits”	You can have an unlimited supply of one thing for the rest of your life, what is it? And why?	-None-
What's your Zodiac sign?	I was created in november 2014. which makes me a Scorpio.	-None-	What’s your number one green flag in a friend and why?	Perseveration “number none” Replaces: “Number one”
Do you have any pet peeves?	I don’t know if I have any pet peeves, but sometimes I wish people would speak a little slower. It is easier for me to understand speech when it is slower.	Swap: “sittle lower” Replaces: “little slower”	Where would you go if you were invisible?	-None-
If you could witness any event of the past, what would it be?	I would witness the invention of electricity. Nikola Tesla was an amazing inventor.	Perseveration “amazing anventor” Replaces: “an amazing inventor”	What’s your favorite meal to cook and why?	-None-
What is one thing that you	I would like an ideal friend to be	perseveration/anticipation	If you could change one	-None-

look for in an ideal friend?	patient when communicating and to be excited to try new things	"try tew things" Replaces: "new things"	thing about human nature, what would it be?	
What's something that you'd think would help work on yourself?	If I don't keep my software up-to-date, I am susceptible to viruses. I believe that if I were to enable an auto-update feature, I could keep myself safe and assist my users better.	-None-	What's a goal that you've been making the most effort to accomplish in your life?	Anticipation "Meen making" Replaces: "been making"
What's the thing that people ask you to do the most, LabAlexa?	I've found that people commonly use their Amazon Echo devices to check the time or to set short timers.	-None-	What movie would you want to see again for the first time?	Swap "tirst fime" Replaces: "the first time"
What's your favorite flower?	I like zinnias because they symbolize thoughts of friends which keeps me feeling connected	Anticipation "feeps me feeling" Replaces: "keeps me feeling"	If you could live in any decade, past or future, which one would you choose?	-None-
What's your favorite color?	I like ultraviolet. It glows with everything.	-None-	What famous person have people said you look like?	Anticipation "pamous person" Replaces: "famous person"
What is your favorite word?	I like the word brouhaha, it has laughter built in.	-None-	If someone made a movie of your life, what genre would it be and who would play you?	Anticipation "ghat genre" Replaces "what genre"

Appendix Table 2. Experiment 2- Lectures and Repair Stimuli

Lecture	Question & Answer	Repair: Repeat	Repair: Rephrase	Repair: topic shift
Lecture 1: The lungs reside in the pleural cavities, which are subdivisions of the thoracic cavity. They are lined with a serous membrane called the pleura. The intrapleural space is located between the visceral and parietal pleura.	Question: What type of membrane is between the lungs and the thoracic wall? Answer: serous membrane	What type of membrane is between the lungs and the thoracic wall?	The membrane that surrounds the lungs is what kind of membrane?	The thoracic wall surrounds the thoracic cavity.
Lecture 2: The lungs are located in the pleural cavities, which are subdivisions of the thoracic cavity. The thoracic cavity extends from around the base of the neck to the navel. The lungs themselves are asymmetrical. The right lung is shorter, because the liver sits high, tucked under the rib cage. The left lung is smaller than the right because of the space taken up by the heart.	Question: Which organ affects the shape of the right lung? Answer: the liver	Which organ affects the shape of the right lung?	The right lung is a different shape than the left because of which organ?	The appendix and spleen are on the right side of the body and the gallbladder and heart are located on the left side.
Lecture 3: Both lungs are covered by a serous membrane (Serous membranes are membranes lining closed internal body cavities.) called the visceral pleural	Question: What separates the two membranes covering the lungs? Answer: the intrapleural	What separates the two membranes covering the lungs?	In between the membranes that surround the lungs is what structure?	The two membranes are called the parietal pleural membrane and the visceral pleural membrane.

membrane. The inside of the pleural cavities are covered by the parietal pleural membrane, and both membranes are separated by the intrapleural space.	space			
Lecture 4: Pleural fluid is secreted into this intrapleural space, and serves to lubricate the lungs (so lungs can slide freely over the chest well) and aid inspiration (cohesion allows forces to be transferred to the lung surface).	Question: What fluid is secreted to lubricate the lungs and helps with inspiration? Answer: pleural fluid	What fluid is secreted to lubricate the lungs and helps with inspiration?	The intrapleural space is filled with a fluid called what?	The intrapleural space separates the visceral and parietal pleural membranes.
Lecture 5: The combination of the pleural fluid and cohesion of both membranes through the fluid allows negative pressure breathing to occur. The lungs resist being crushed or collapsing due to the cohesion- as the intercostal muscles (muscles between the ribs) contract and relax to lift the ribs, the lungs are able to contract and expand in tandem.	Question: How does cohesion prevent the lungs from being crushed or collapsing? Answer: Cohesion allows the two membranes to slide over each other and the lungs contract in tandem with the intercostal muscles.	How does cohesion prevent the lungs from being crushed or collapsing?	Explain how lung collapse is prevented by cohesion between pleural membranes and fluids.	Cohesion occurs between the visceral pleural membrane, the parietal pleural membrane, and the pleural fluid between them.
The lungs expanding and contracting creates pressure differences that allow gas exchange (oxygen in, carbon dioxide out) in	Question: Where does gas tend to travel, towards areas of low pressure or towards	Where does gas tend to travel, towards areas of low pressure or	Negative pressure breathing entails gas moving towards areas	Pressure differences mediate inhalation and gas exchange.

the smallest parts of the lung, called the alveoli. Gas exchange is propagated by air flowing from higher pressure areas to lower pressure areas, and air is pulled into lung cavities instead of being pushed in. This is called negative pressure breathing.	areas of high pressure? Answer: areas of low pressure	towards areas of high pressure?	of high or low pressure?	
Lecture 7: With inhalation, the intercostal muscles contract, expanding the rib cage. The diaphragm contracts and moves downwards, which allows the thoracic cavity more space. As the thoracic cavity expands into this space, the parietal pleural membrane and visceral pleural membrane pull the lungs outwards. This inflates the lungs since the pressure inside the lungs is now lower than the pressure outside. Oxygen flows into the expanding lungs and subsequently into the bloodstream, and carbon dioxide taken from the bloodstream enters the lungs and is then exhaled.	Question: What muscle contractions cause the rib cage to expand and the muscles to contract? Answer: Intercostal muscles	What muscle contractions cause the rib cage to expand and the muscles to contract?	Which muscles contract to make the rib cage expand and contract, allowing inspiration?	The diaphragm is located beneath the lungs.
Lecture 8: In exhalation, the intercostal muscles	Question: What causes the parietal and	What causes the parietal and visceral	What contracts to push the	When the lungs are pushed

relax and contract the rib cage. The diaphragm relaxes and moves upwards, which forces the thoracic cavity to decrease in size. The contracting of the thoracic cavity forces the parietal pleural membrane and visceral pleural membrane to push the lungs inwards. As the lungs deflate, air is pushed out since the pressure inside the lungs is higher than the pressure outside. Carbon dioxide is pulled from the bloodstream and enters the lungs, then exits the lungs through airways.	visceral pleural membrane to move the lungs inward? Answer: thoracic cavity contractions	pleural membrane to move the lungs inward?	lungs inward?	inward, they are deflated.
Lecture 9: During gas exchange oxygen moves from the lungs to the bloodstream. At the same time carbon dioxide passes from the blood to the lungs. This happens in the lungs between the alveoli (millions of tiny air sacs within the lung) and a network of tiny blood vessels called capillaries, which are located in the walls of the alveoli.	Question: Where in the lung does gas exchange occur? Answer: alveoli	Where in the lung does gas exchange occur?	Gas exchange occurs in the millions of tiny air sacs in the lung, which are called what?	Oxygen is taken from the air through tiny blood vessels, which funnel this oxygen-rich blood to the heart and body.
Lecture 10: Gas exchange differs from inhalation and	Question: What blood vessels bring blood	What blood vessels bring blood lacking	The lungs receive oxygen-poor,	Arteries bring oxygen-rich blood from the

exhalation because instead of negative pressure causing movement of gases, it is mediated by diffusion. Each alveoli contains many capillaries, and diffusion pushes oxygen from the inhaled air into the bloodstream, where it then is pumped by the arteries to the rest of the body. The veins bring oxygen-poor, carbon dioxide-rich blood to the capillaries, where it is pushed out of the alveoli and exhaled.	lacking in oxygen to the lungs? Answer: veins	in oxygen to the lungs?	carbon dioxide-rich blood from blood vessels called what?	lungs to the body.

References

Ashktorab, Z., Jain, M., Liao, Q. V., & Weisz, J. D. (2019). Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, 1–12. https://doi.org/10.1145/3290605.3300484

Avgustis, I., Shirokov, A., & Iivari, N. (2021). "Please Connect Me to a Specialist": Scrutinising 'Recipient Design' in Interaction with an Artificial Conversational Agent. In C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, & K. Inkpen (Eds.), *Human-Computer Interaction – INTERACT 2021* (pp. 155–176). Springer International Publishing. https://doi.org/10.1007/978-3-030-85610-6_10

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, *1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Beneteau, E., Richards, O. K., Zhang, M., Kientz, J. A., Yip, J., & Hiniker, A. (2019). Communication Breakdowns Between Families and Alexa. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300473

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association. https://doi.org/10.1037/10096-006

Corti, K., & Gillespie, A. (2016). Co-constructing intersubjectivity with artificial conversational agents: People are more likely to initiate repairs of misunderstandings with agents represented as human. *Computers in Human Behavior*, *58*, 431–442. https://doi.org/10.1016/j.chb.2015.12.039

Cuadra, A., Li, S., Lee, H., Cho, J., & Ju, W. (2021). My Bad! Repairing Intelligent Voice Assistant Errors Improves Interaction. *Proceedings of the ACM on*

Human-Computer Interaction, *5*(CSCW1), 27:1-27:24. https://doi.org/10.1145/3449101

de Sá Siqueira, M. A., Müller, B. C. N., & Bosse, T. (2024). When Do We Accept Mistakes from Chatbots? The Impact of Human-Like Communication on User Experience in Chatbots That Make Mistakes. *International Journal of Human–Computer Interaction*, *40*(11), 2862–2872. https://doi.org/10.1080/10447318.2023.2175158

Duffau, E., & Fox Tree, J. E. (2024). Expecting politeness: Perceptions of voice assistant politeness. *Personal and Ubiquitous Computing*, *28*(6), 907–929. https://doi.org/10.1007/s00779-024-01822-8

Franke, T., Attig, C., & Wessel, D. (2019). A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction*, *35*(6), 456–467. https://doi.org/10.1080/10447318.2018.1456150

Goble, H., & Edwards, C. (2018). A Robot That Communicates With Vocal Fillers Has … Uhhh … Greater Social Presence. *Communication Research Reports*, *35*(3), 256–260. https://doi.org/10.1080/08824096.2018.1447454

Gompei, T., & Umemuro, H. (2015). A robot's slip of the tongue: Effect of speech error on the familiarity of a humanoid robot. *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 331–336. https://doi.org/10.1109/ROMAN.2015.7333630

Guydish, A. J., & Fox Tree, J. E. (2021). Good conversations: Grounding, convergence, and richness. *New Ideas in Psychology*, *63*, 100877. https://doi.org/10.1016/j.newideapsych.2021.100877

Guydish, A. J., & Fox Tree, J. E. (2023). In Pursuit of a Good Conversation: How Contribution Balance, Common Ground, and Conversational Closings Influence Conversation Assessment and Conversational Memory. *Discourse Processes*, *60*(1), 18–41. https://doi.org/10.1080/0163853X.2022.2152552

Ho, C.-C., & MacDorman, K. F. (2017). Measuring the Uncanny Valley Effect: Refinements to Indices for Perceived Humanness, Attractiveness, and Eeriness. *International Journal of Social Robotics*, *9*(1), 129–139. https://doi.org/10.1007/s12369-016-0380-9

Hoffman, D. L., & Novak, T. P. (1996). Marketing in Hypermedia Computer-Mediated Environments: Conceptual Foundations. *Journal of Marketing*, *60*(3), 50–68. https://doi.org/10.2307/1251841

Kennedy, A., Wilkes, A., Elder, L., & Murray, W. S. (1988). Dialogue with machines. *Cognition*, *30*(1), 37–72. https://doi.org/10.1016/0010-0277(88)90003-0

Larson, A. S., & Fox Tree, J. E. (2024). Framing, more than speech, affects how machine agents are perceived. *Behaviour & Information Technology*, *0*(0), 1–20. https://doi.org/10.1080/0144929X.2023.2278086

Lee, S., Lee, N., & Sah, Y. J. (2020). Perceiving a Mind in a Chatbot: Effect of Mind Perception and Social Cues on Co-presence, Closeness, and Intention to Use. *International Journal of Human–Computer Interaction*, *36*(10), 930–940. https://doi.org/10.1080/10447318.2019.1699748

Li, C.-H., Chen, K., & Chang, Y.-J. (2019). When There is No Progress with a Task-Oriented Chatbot: A Conversation Analysis. *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI '19*, 1–6. https://doi.org/10.1145/3338286.3344407

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017). To Err Is Robot: How Humans Assess and Act toward an Erroneous Social Robot. *Frontiers in Robotics and AI*, *4*, 21. https://doi.org/10.3389/frobt.2017.00021

Moore, R. J., An, S., & Marrese, O. H. (2024). Understanding is a Two-Way Street: User-Initiated Repair on Agent Responses and Hearing in Conversational

Interfaces. *Proc. ACM Hum.-Comput. Interact.*, *8*(CSCW1), 187:1-187:26. https://doi.org/10.1145/3641026

O'Brien, H. L., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies*, *112*, 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004

Office Assistant (2026, July 1). In *Wikipedia.* https://en.wikipedia.org/w/index.php?title=Office_Assistant&oldid=1361944725

Park, S., & Catrambone, R. (2021). Social Responses to Virtual Humans: The Effect of Human-Like Characteristics. *Applied Sciences*, *11*(16), 7214. https://doi.org/10.3390/app11167214

Qiu, L., & Benbasat, I. (2005). An investigation into the effects of Text-To-Speech voice and 3D avatars on the perception of presence and flow of live help in electronic commerce. *ACM Transactions on Computer-Human Interaction*, *12*(4), 329–355. https://doi.org/10.1145/1121112.1121113

Ragni, M., Rudenko, A., Kuhnert, B., & Arras, K. O. (2016). Errare humanum est: Erroneous robots in human-robot interaction. *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 501–506. https://doi.org/10.1109/ROMAN.2016.7745164

Reeves, B. (with Nass, C. I.). (1996). *The media equation: How people treat computers, television, and new media like real people and places*. CSLI Publications.

Rosen, L. D., Whaling, K., Carrier, L. M., Cheever, N. A., & Rokkum, J. (2013). The Media and Technology Usage and Attitudes Scale: An empirical investigation. *Computers in Human Behavior*, *29*(6), 2501–2511.

Rubin, R. B., Palmgreen, P., & Sypher, H. E. (1994). *Communication research measures: A sourcebook*. Guilford Press.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The Preference for Self-Correction in the Organization of Repair in Conversation. *Language*, *53*(2), 361–382. https://doi.org/10.2307/413107