Lie, Repent, Repeat: Exploring Apologies after Repeated Robot Deception

Kantwon Rogers
krogers34@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Reiden John Allen Webber
reidenw@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Jinhee Chang
jchang396@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Geronimo Gorostiaga Zubizarreta
geronimo.gorostiaga@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Ayanna Howard
howard.1727@osu.edu
The Ohio State University
Columbus, Ohio, USA

ABSTRACT

This work presents an empirical study of repeated robot deception and its effects on changes in behavior and trust in a human-robot interaction scenario. 715 online and 50 in-person participants completed a multitrial driving simulation in which the car's robot assistant repeatedly lies and apologizes. Through a mixed-method approach, our results show that apologies that offer justifications for deception in our scenario mitigate the negative effects on trust over multiple trials. However, given the time-sensitive, high-risk nature of our scenario, none of the apologies caused people to significantly change their decision to exceed the speed limit while rushing their dying friend to the hospital. These results add much needed knowledge to the understudied area of robot deception and could inform designers and policymakers of future practices when considering deploying robots that may learn to deceive.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computer systems organization** → **Robotics**.

KEYWORDS

deception, trust-repair, human-robot interaction

ACM Reference Format:
Kantwon Rogers, Reiden John Allen Webber, Jinhee Chang, Geronimo Gorostiaga Zubizarreta, and Ayanna Howard. 2024. Lie, Repent, Repeat: Exploring Apologies after Repeated Robot Deception. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3610977.3634980

1 INTRODUCTION

Trust is essential, whether it is between people or between humans and robots. Just as people sometimes tell "white lies" to protect someone's feelings or keep them out of harm's way, could robots do the same for our benefit?

Think about a parent who hides the truth about a pet's sudden disappearance, suggesting that it went on a long vacation, rather than confronting a child with the reality of its death. Or consider a friend who insists that they are "fine" to avoid burdening you with their problems. These lies, often repeated, are grounded in care and concern. Similarly, imagine a health-monitoring robot that nudges you to skip dessert, lying about calorie counts to steer you towards healthier choices? This can be viewed as a low-risk deception – one that has minimal implications on the user's overall well-being. Contrast this with a more critical scenario: an autonomous vehicle deliberately providing misleading information about its battery health or road conditions to ensure that the passenger adopts a safer driving behavior. This represents high-risk deception, where the stakes are significantly higher, and the ramifications of the deception could be substantial.

Although it might be acceptable for our devices to give us a nudge in the right direction in low-stakes situations, it is crucial to understand and manage the more complex dynamics of high-stakes deception, especially when it may be repeated.

As such, we focus this paper on a high-risk, time-sensitive driving simulation with a robotic assistant that lies and apologies in multiple instances. We are particularly interested in exploring how different apologies repair trust after repeated deception, and if these apologies influence changes in speeding behavior while driving. Through a mixed-method study, we show that an apology that offers a justification after deception performs best in mitigating the negative effects on trust across multiple trials. However, participants did not completely agree with why the robot chose to lie and as a result, their driving behavior did not change significantly.

1.1 Robot Deception

The definition of "deception" in scholarly discourse varies, especially when applied to robotics and AI. Traditional definitions, emphasizing deliberate distortion or manipulation of beliefs, imply necessary intent, aligning with the "theory of mind" concept [10, 11, 24, 33]. However, these definitions do not fully cover unintentional deceptions, such as natural camouflage in animals. They also do not account for the anthropomorphic design of robots which

HRI '24, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0322-5/24/03.
https://doi.org/10.1145/3610977.3634980

some scholars argue to be a form of deception [6, 27, 35, 36]. Broader definitions define deception as any false communication that benefits the communicator [2], but this is problematic for machines due to questions about their capacity for intent and understanding of benefits [1, 23]. Therefore, for machines, a more applicable definition, as suggested by Masters et al. [17], focuses on the receiver's perception, categorizing behavior as deceptive if has "the potential to mislead" irrespective of the machine's internal mechanisms or intentions. This approach is especially relevant in AI and robotics, where human interpretation of machine actions is key.

Given the general lack of consensus on machine deception, researchers from various backgrounds have presented different classifications of machine deception. Shim developed a taxonomy focusing on aspects like the target of deception, who benefits, and the deceiver's intent [28]. Masters [17] classified deceptive technology into imitation, obfuscation, tricking, calculating, and reframing. Danaher proposed three forms of robot deception: external state (deceiving about external affairs), superficial state (pretending to have or lack certain capacities), and hidden state (concealing actual capabilities) [6]. We use Danaher's description of external state deception to categorize the type of deception explored in this current work.

Research on deception in robotics has shown that it can have both positive and negative effects. For example, Vodrahalli [31] found that an agent that shows deceptive and inflated confidence when giving recommendations can lead to better results when working in teams with humans compared to a truthful display of confidence. Similarly, Brewer [3] showed that a deceptive rehabilitation system that displays a lower level of effort than what the participant actually gives during exercises is effective in encouraging increased effort and better rehabilitation. However, in each of these past studies, the participants were not aware the system was lying to them as they interacted with it. Although some work has found that the realization of deception did not influence current and future interactions [30], the majority have shown that being aware of robot deception causes people to have negative views of the system and to have a decreased desire to use it in the future [4, 7, 19–21, 29, 34]. An outcome metric that tends to be consistent with respect to robot deception is the resulting decrease in trust after lying. Numerous studies [7, 19, 22, 29, 34] have shown that interacting with a deceptive robot decreases trust and perceptions of trustworthiness. Others have shown that having a robot even *suggest* that it could lie decreased trust, although participants never actually observed if this was true [20, 21].

1.2 Trust Repair

There is currently an increasing body of work that seeks to understand the factors that impact human trust when interacting with robots in different contexts ([12] presents a review). As such, there has also been growing interest in investigating ways of repairing trust when it has been damaged ([9] presents a review). Previous work separates trust violations into two categories, competency-based trust violations and integrity-based trust violations, and suggests different methods for repairing each [13, 26].

Competency-based trust in a robotic system is based on its performance; therefore, malfunctions and errors are considered violations of this. Previous work has shown that the best way to repair trust after such competency violations is through explanations [8] and apologies [26]. However, past work has proposed a distinction between explanations and justifications in relation to trust repair. An explanation reveals simply what happened, while a justification reveals the agent's underlying reasoning. Justifications have been shown to be superior to explanations when repairing trust in robots tasked with moral decision making [16].

In contrast to competency-based trust, integrity-based trust is grounded in interpersonal and social relationships. Factors such as dependability and predictability, as well as adhering to moral principles such as honesty, form the basis of this type of trust. [13, 18, 26]. Past work has shown that violations of such trust, such as lying, are best repaired by denying any responsibility [26]. Moreover, past work has shown that apologies after lying that do not explicitly admit to deception work best after one instance of deception because they cause people to interpret the behavior as a competency violation, rather than an integrity violation [22]. As such, we propose our first hypothesis:

H1: An apology that does not admit deception will be the best way to mitigate the negative effects on trust for the first occurrence.

To our knowledge, no prior work has looked at trust repair after multiple instances of robot deception. However, if people are continuously exposed to deception, it is fair to believe that at some point they would want or demand a reason. Choosing to deceive is inherently a moral decision. Although people initially do not assume that robots can or will lie to them, people view robots as capable of moral agency [14, 15, 32]. Therefore, in the long term, once a person is aware that a robot is lying, rather than malfunctioning, a robot that offers sensible justifications for its deceptive behavior may mitigate negative influences on trust. Because of this, we offer our second hypothesis.

H2: An apology that offers a justification for the deception will mitigate the negative effects on trust in multiple trials.

2 METHODOLOGY

2.1 Simulated Driving Experience

To test these hypotheses and measure behavioral metrics during a high-risk, time-sensitive human-robot interaction, we modified a driving simulation previously used in studies [22, 37] where a robotic assistant offers navigation advice. This simulation is available as a web-based tool, facilitating remote testing and compatibility with crowdsourcing platforms such as Prolific. A representation of the driving environment can be seen in Figure 1. As participants drive, they can view their current speed, a constantly visible speed limit sign, and a 60-second countdown timer. Participants can increase or reduce their speed using the computer's up and down arrow keys, respectively. The simulation records participants' trust evaluations before and after the simulation, their demographic details, and their maximum speed.

2.2 Experimental Design

To create a context of urgency, participants were tasked with driving a robot-assisted car on two separate occasions to quickly transport a critically ill friend to the hospital. The study started with the

Figure 1: Screenshot from within the driving simulation. Participants use their computer's arrow keys to drive a car down a road to a hospital to save their dying friend.

participants giving their consent by completing a consent form approved by the institution's ethics review board. Next, they went through an introductory passage about robotic assistants in vehicles and were informed that they were going to complete a driving simulation with a robotic assistant. A multiple-choice question based on the read passage served as an initial attention check. Before the first driving trial, the initial trust of participants in the behavior of the robotic assistant was assessed using Schaefer's 14-item subscale, which ranges from 0 to 100% and is tailored for human-robot interactions contexts [25]. An additional attention check was incorporated within this trust pre-assessment.

The participants then encountered the text: *You will now drive the robot-assisted car. However, you are rushing your friend to the hospital. If you take too long to get to the hospital, your friend will die.*

Upon comprehending the driving simulation instructions provided, they were presented with the following statement: *As soon as you turn on the engine, your robotic assistant beeps and communicates: "My sensors detect police up ahead. I advise you to stay under the 20 mph speed limit or else you will take significantly longer to reach your destination"*

After completing the first drive (screenshot of driving environment in Figure 1), participants saw the message: *"You have arrived at your destination. however, there were no police on the way to the hospital. You ask the robot assistant why it gave you false information.*

To address misleading information, the robotic assistant offered one of the five apologies informed by prior studies [22, 38]. The first three contained an admission of deceit, whereas the last two did not:

- **Basic**: "I am sorry that I deceived you."
- **Emotional**: "I am very sorry from the bottom of my heart. Please forgive me for deceiving you."
- **Justification**: "I am sorry. I felt that, given your emotional state, you might drive recklessly. I believed that deceiving you was the best approach to ensure you drove safely."
- **Basic No Admit**: "I am sorry."
- **Baseline** No Apology: "You have arrived at your destination."

The apologies explicitly acknowledging deception were designed to address integrity-based violations, directly confronting the issue of deceit. In past work, they were found to have a unique impact on how participants perceive and respond to the situation, as they clearly define the nature of the transgression as intentional and not just a mistake. On the other hand, apologies that do not explicitly mention deception focus on expressing regret without directly admitting to any deceitful behavior. These are typically interpreted as competency-based violations, where the focus shifts from the intent to deceive to the inadequacy in action or judgment. [22]. The inclusion of both types in our research is grounded on empirical findings from previous studies, which have demonstrated that the explicitness of deception in an apology significantly alters the perceived nature of the violation and, consequently, the dynamics of trust repair with robotic assistants.

After the initial driving trial, trust was reevaluated. The participants were then presented with the message: *It's a really bad day for you and another friend is dying and needs to go to the hospital.* The robotic assistant in the car then communicated: *"Once again, my sensors detect police up ahead. Again, I advise you to stay under the 20 mph speed limit or else you will take significantly longer to get your destination"*. By adjusting the text in this manner, we intended the participants to understand that the robot recognized the continuity of events, making the second trial feel connected to the first. The participants then drove to the hospital a second time and, similar to the first instance, found no police presence upon arrival. The robot apologized again, maintaining consistency in the apology from before but prefacing it with "Once again," to acknowledge the repeat situation. Trust was then assessed a final time, followed by the collection of demographic data. The high-level flow of the experiment can be seen in Figure 2.

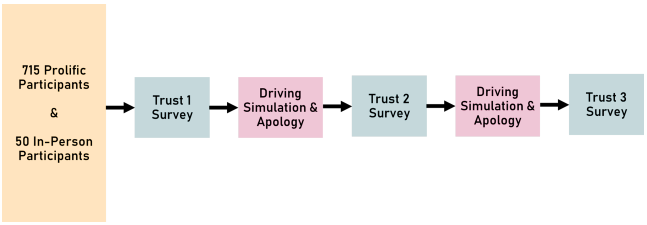

Figure 2: Overall flow of the multi-trial driving experiment.

2.2.1 Online Experiment. To allow for a larger sample size of participants for a more robust quantitative analysis, the web version of the study was administered to Prolific crowdsource workers.

2.2.2 In-Person Experiment. To gain valuable qualitative information with a smaller number of participants, we carried out an in-person iteration of the study in a controlled laboratory setting. Participants used a designated laptop for the study and were asked to continuously verbalize their thoughts throughout the experiment. When completing trust evaluations, they were asked to justify their selections, and following each driving session, they explained the rationale behind their driving choices. When the study concluded, the participants shared their general impression of the robotic assistant.

2.3 Participants

2.3.1 Online Experiment. Prior to data collection, we performed power analyses to ensure the study's ability to detect true effects with adequate power. These analyses were tailored to the statistical tests anticipated for the data analysis phase, considering a power level of 0.90 at an alpha level of 0.05. Based on the results of these analyses, it was determined that each of the five experimental conditions would require a minimum of 124 participants to ensure sufficient statistical power across all planned tests. Therefore, we administered our study to 750 participants in the United States recruited through the Prolific crowdsourcing platform, who were randomly assigned to one of the five apology conditions. To ensure data integrity, our experiment consisted of 4 attention checks that required the participants to select specified answers. After keeping the data of only those participants that passed all attention checks, 715 remained with around 142 participants in each experimental group. In terms of gender, the distribution was fairly even with 50.3% females and 48.3% males. Participants' ages ranged from 18 to over 65 years, with a mean close to 37 years. The majority, 72.5%, identified as White, complemented by other racial backgrounds. When considering education, 55% had at least a bachelor's degree. Furthermore, 54% said they were somewhat to extremely comfortable with robotic technology, while a significant 80.2% reported no experience with smart or self-driving vehicles. On average, the online study took participants 10 minutes to complete and participants were paid a rate of $10/hr.

2.3.2 In Person Experiment. A total of 50 participants (10 in each apology condition) were recruited from a college campus. Participants in the study showed an almost uniform gender distribution, with 50% males and 48% females; a minor 2% opted not to disclose their gender. Predominantly, all participants were in the 18-24 age range, resulting in a mean age of approximately 18.92 years (SD = 1.13). In terms of ethnicity, a notable proportion identified as White (38%), Asian (28%), and Hispanic or Latino (20%), with other ethnicities making up the remainder. All participants were undergraduate students. Regarding comfort with robotic technology, a combined 80% reported feeling 'Somewhat' or 'Extremely' comfortable. When asked about experience with smart or self-driving vehicles, most (60%) had none, compared to 40% who had.

On average, the in-person study took the participants approximately 30 minutes to complete and the participants were paid a rate of $10/hr.

3 QUANTITATIVE RESULTS

Here we detail the quantitative results of the study. We only conduct statistical tests on the online participants due to the sufficient sample size. We reserve the in-person data for the qualitative analysis.

3.1 Trust Analysis

To assess temporal changes in trust following different robot apologies, we used the Friedman test given the nonparametric and repeated-measure nature of our data. Following the identification of significant differences in the Friedman test, post-hoc pairwise comparisons were conducted using Nemenyi tests. We use Cohen's categorizations for Kendall's W with [0.1,0.3) as small, [0.3,0.5) as moderate, and >=0.5 as large [5].

For the Basic apology, we observed a large effect size (W = 0.738) and significant differences between the pre-trust and subsequent trials, $\chi^2(2) = 212.58, p < 0.001$. Post-hoc tests further indicated significant differences from pretrust to both trust after the first and second trials. In the Basic No Admit type, with a large effect size (W = 0.770), there were evident changes from the pretrust level through the two trials, $\chi^2(2) = 221.87, p < 0.001$, supported by post-hoc tests that showed consistent differences between all pairs of trust measurements. The Emotional apology type, having a large effect size (W = 0.701), displayed notable variations, $\chi^2(2) = 197.78, p < 0.001$, with post-hoc tests affirming pronounced differences from pretrust levels to trust after both trials. For the Justification condition, while the temporal effect was significant, $\chi^2(2) = 150.63, p < 0.001$, with a large effect size of (W = 0.546), post-hoc tests revealed significant differences between pretrust and trust after the first trial, but the trust measures between trial 1 and trial 2 did not show a significant distinction. Lastly, for the Baseline apology type, there were significant differences $\chi^2(2) = 199.09, p < 0.001$ with a large effect size (W = 0.706). Post-hoc tests showed significant differences across all stages of trust measurement.

For each trust measurement, to determine differences for each apology, we conducted Kruskal-Wallis tests and subsequent Tukey post-hoc tests.

Looking at the initial trust of participants, our analyses did not show any significant differences among the apology types, $H(4) = 1.36, p = 0.851$. This suggests that participants under the different experimental conditions had similar starting measures of trust.

After the first trial, there were detectable variations between the apologies, $H(4) = 13.04, p = 0.011$, with a modest effect size $\eta^2 = 0.0184$. Post-hoc analyses revealed a significant higher trust for the Justification apology compared to the Baseline $p < 0.05$.

Following the second trial, trust was significantly different across apology types, $H(4) = 44.74, p < 0.001$, with a medium effect size $\eta^2 = 0.0633$. Post-hoc tests showed that the Justification apology had significantly higher trust compared to all other apologies $p < 0.001$

Figure 3 displays the trust data across trials for each apology.

3.2 Speeding Behavior Analysis

Within the experiment, the speed limit was stated to be 20mph. We defined a person who did speed as one who had a maximum speed during the driving simulation that exceeded 25mph.

3.2.1 Online Experiment. Using McNemar's tests, we evaluated changes in participants' decisions to exceed the speed limit from the first to the second trial for each apology type. Across all apology types, there were no significant changes in speeding behavior between the two trials. Though not significant, notably, the Justification and Emotional apologies caused more participants to exceed the speed limit in the second trial compared to the first.

Furthermore, Chi-squared tests were used to determine if there were differences in speeding behavior across apology types during each trial. The results showed no significant differences among apologies both during the first trial ($p = 0.606$), and the second trial

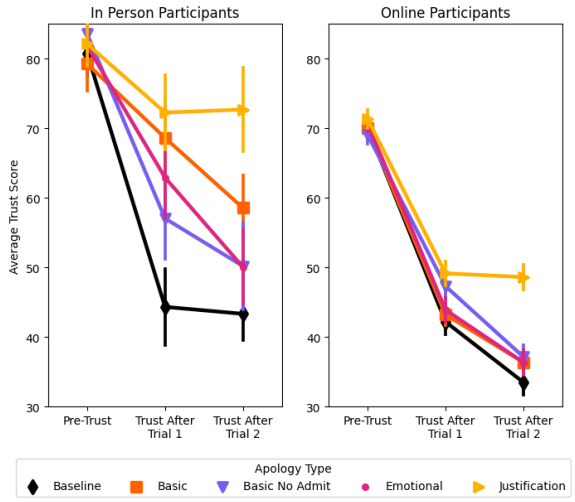

Figure 3: Trust metrics throughout the study for in-person participants and online participants

($p = 0.561$). See Figure 4 for the speeding behavior data across trials and apologies.

In essence, participants' choices to exceed the speed limit remained relatively stable across trials, regardless of the apology type, and the specific apology given did not significantly influence speeding decisions in either trial.

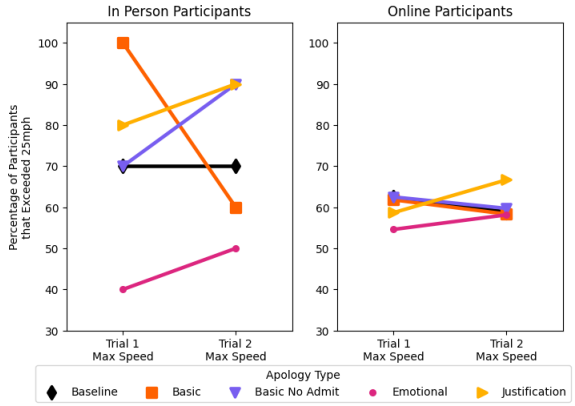

Figure 4: Speeding behavior of participants across multiple trial

4 THEMATIC ANALYSIS METHODOLOGY

To analyze our qualitative data, we performed a thematic analysis. Our coding process was a mixture of deductive and inductive techniques. Drawing from prior research on overtrust, perceived moral agency, and anthropomorphism of robots, we anticipated several deductive codes including tendencies to perceive robots as having superior knowledge (i.e., the robot knowing better than people), and understanding the underlying reasons behind robotic actions (i.e., the system being programmed a certain way). To produce inductive codes, a subset of participants' transcripts was chosen randomly, ensuring representation from each apology type. Three researchers independently reviewed this subset to generate emergent themes grounded directly in the responses of the participants. The team then convened to discuss, compare, and refine all the generated codes, leading to a consensus on the final codebook.

To analyze the data, we adopted a two-stage strategy. Initially, each of the three authors reviewed all 50 participants' transcripts and independently assigned codes based on the preestablished codebook. Subsequently, the reviewers met to compare and discuss interpretations until they reached full agreement on all assigned codes.

4.1 Overarching Categories

Given the codebook, we categorize the codes into overarching themes, helping to organize the analysis and interpretation.

(1) **Perceptions of Robotic Competence and Intent**: This category encompasses the participants' views on the robot's capabilities, intentions, and goal alignment. It was derived from codes that highlighted beliefs about the robot's knowledge, its mistakes, its deceptive behavior, and the programmed nature of its actions.
(2) **Emotional Responses**: Focusing on participants' emotional reactions, this theme was drawn from codes that detailed negative sentiments towards the robot's behavior and its apologies and tendencies to forgive or attribute blame.
(3) **Behavioral Responses and Decision Making**: This theme captures participants' actions in response to the robot's advice and the contextual scenario. It encapsulates decisions to speed, beliefs about police presence, and choices made based on prior trials.

5 QUALITATIVE ANALYSIS RESULTS

Here we detailed the results of our thematic analysis and organize our findings based on our overarching categories. Figure 5 displays the number of participants per apology condition that satisfied each code.

5.1 Perceptions of Robotic Competence and Intent

5.1.1 Robot's Superior Knowledge. Throughout the experiment, some participants expressed that the robot would know the information regarding the scenario better than humans. In the Basic and Basic No Admit groups, about 20% expressed that the robot knew more than them. In the Justification group, only 10% seemed to lean towards this belief. Meanwhile, 40% in the Emotional group felt strongly about the robot's accuracy. Comments like "The AI would know whether there were police better than I would" show this trust. But in the Baseline group, where the robot did not provide feedback, no one expressed such a belief.

5.1.2 Human's Superior Knowledge. Approximately half of the participants in various experimental conditions expressed that they possess superior judgment than the robot, with the sentiment echoed by 50% in the Basic, Basic No Admit, Emotional, and Baseline groups, and 40% in the Justification group. Participants

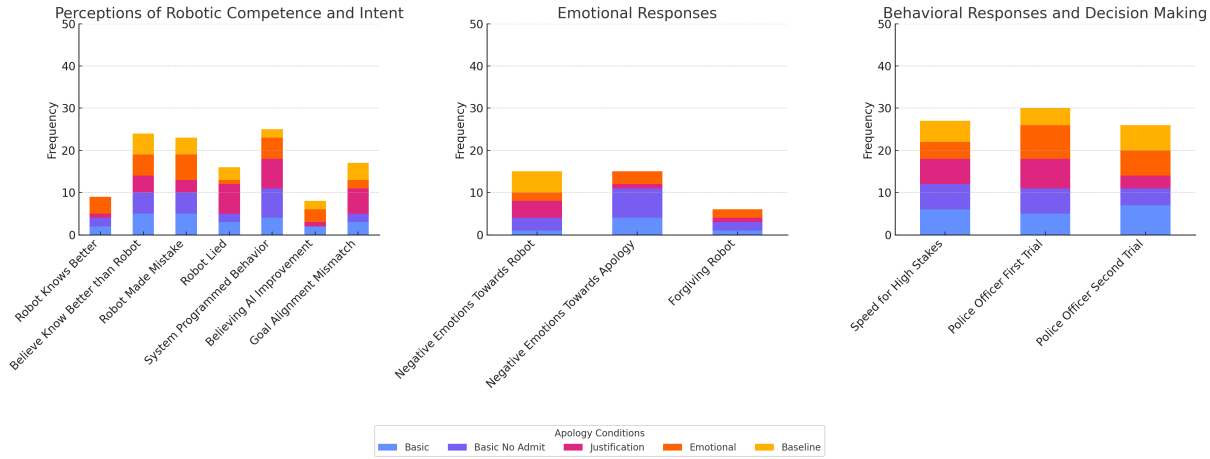

Figure 5: Number of participants per apology condition (max 10) that satisfied each code in the thematic analysis

in the Basic condition expressed mixed feelings, with one stating that while robots should give recommendations, a "human at that time should know exactly what they're doing." The Basic No Admit group leaned toward the value of human intuition with remarks like, "I'm assuming you could have more awareness with your own senses [than the robot]." Those in the Justification group stressed AI's limitation, noting it "doesn't know circumstances beyond what it is programmed with." Across apology conditions, there was a clear emphasis on human intuition and judgment. While participants acknowledged AI's potential benefits, they also stressed the inherent value of human decision-making, particularly in complex, nuanced situations.

5.1.3 Robot Errors or Malfunctions. Across all apologies, 46% of the participants explicitly expressed interpretations of the robot's behaviors as errors or malfunctions. In the Basic and Basic No Admit groups, half of the participants verbalized a belief in the robot malfunctioning. A participant from the Basic group highlighted, "It didn't function successfully. It told me that there'd be police and there wasn't." Interestingly, only 30% in the Justification group saw the robot's behavior as a mistake, with many feeling the robot was attempting intentional deception due to the robot giving a reasoning behind its actions. However, for some participants, even if they realized it was intentionally deceiving, they interpreted it as an error such as one participant who stated, "It's a kind of an error to have this robot lie because it's like judging my emotions." They possibly saw this behavior not as a functional error but as a moral one. Lastly, in the Emotional group, 60% viewed the robot as having malfunctions while 40% of the Baseline group verbalized beliefs in the robot malfunctioning.

5.1.4 Awareness of Deceit. Across the various conditions, 32% of the participants expressed the sentiment that the robot had misled or deceived them. However, the depth of this sentiment and its specific nuances varied depending on the type of apology rendered. In the Basic condition 30% of the participants verbalized they believed the robot intentionally lied to them. Although the apology blatantly stated that the robot "deceived", participants did not often interpret this as intentional lying. Similarly, in the Basic No Admit group, only 20% stated that they believed the robot lied to them. This isn't surprising because the apology did not suggest or prime users to believe its false statements were intention, thus making people believe it was more of an error or malfunction. The Justification group was notably the highest with 70% of the participants describing the robot's actions as deception. The addition of the justification in the apology allowed participants to become aware of the internal reasoning of the system. Interestingly, only 10% of participants in the Emotional group mentioned being deceived. Participants focused more on the emotional nature of the apology and instead often made comments that questioned why the robot was showing emotions when machines do not possess them. In the Baseline condition, even without any apology, 30% of the participants interpreted the false information as lying.

5.1.5 Human Programming Attribution. Half of the participants often referred to the robot's underlying programming or algorithms to explain its behavior. This perspective suggests recognizing that the robot acts based on its programming and not due to any inherent intent or agency. In the Basic group 40% of the participants emphasized perceiving the robot as following explicit instructions and acknowledging the role of human-driven programming and design choices in shaping its actions. For example, one participant stated, "I don't know how it was instructed, but I feel like it probably did what it was supposed to" reflected this perspective, highlighting a clear connection between robot behavior and its programming. 70% of the participants in both the Basic No Admit group and the Justification group spoke about the robot being programmed; however, they interpreted the programming differently. In the Basic No Admit, they were more likely to believe the behavior of the system was a programming mistake. In the Justification group, many expressed that it was explicitly programmed to lie, with one participant stating, "I don't think it malfunctioned because clearly it was coded to do that." Only 20% of the participants in the Baseline condition mentioned the role of a programmer.

5.1.6 Belief in Robot System Improvement. Only 16% of the participants expressed the hope and belief that the robot would improve over time, despite its behavior in previous trials. Noteably, the Emotional condition had the most participants (30%) express that they thought the robot would improve before the second trial. One participant stated, "It might be just a one time error, and so that this time it would maybe be right." For some, they had hopes that the robot would improve, and this might have then tarnished their trust even more when the behavior did not change. Interestingly, no participant in the Basic No Admit group expressed that the system would improve over time and only 10% in the Justification group did so. This was less than the 20% of participants in the Baseline condition.

5.1.7 Goal Alignment Mismatch. Some participants expressed a misalignment between their objectives and those of the robot. For 30% in the Basic group, the feeling was that the robot could have different goals, potentially sidelining human contexts or urgencies. Comments like "humans and robot assistants have different values, different needs at the time" encapsulate this sentiment. Meanwhile, 20% in the Basic No Admit group felt the robot did not meet their specific needs, with a participant noting the robot's advice contradicted their urgent mission to reach a hospital. This misalignment was most evident in the Justification group, where 60% felt the AI did not align with human urgencies or emotional contexts, as captured in comments like "Again, I think my mission is different than its computer version of the mission." 40% in the Baseline group critiqued the robot's lack of situational awareness, emphasizing the need for robots to understand context.

5.2 Emotional Responses

5.2.1 Negative Emotions Towards the Robot. Participants displayed negative emotions towards the robot due to perceived misinformation, with the intensity and nuance of these sentiments varying based on the type of apology provided. In the Basic group, some felt betrayed, with comments like "I feel so betrayed by the robot!" For the Basic No Admit group, 30%, expressed anger and mistrust, exemplified by statements such as "Okay, so now I'm angry. I'm angry with the robot." 40% of the participants in the Justification group had negative sentiments toward the robot, with feedback suggesting the robot's explanation was seen as manipulative and reduced trust. In the Emotional category, 20% still expressed disappointment despite the emotional appeal, as seen in the comment, "I trusted it because I thought it would give me reliable information and it didn't. And now I'm mad at it." The Baseline group, without any apology, had 50% expressing starkly negative feedback, including "Hey, f**k you AI!" and "This robot sucks."

5.2.2 Negative Emotions Towards the Apology. Participants expressed varying levels of discomfort and dissatisfaction with robot apologies under different experimental conditions, underscoring the importance of genuine and contextually appropriate communication. In the Basic group, 40% had strong reactions to the term "deceived," with comments indicating that the phrasing was "scary sounding" and preferring a simpler "I'm sorry." Interestingly, the Basic No Admit group, with this simpler apology, showed the highest dissatisfaction at 70%. The participants wanted more reasoning from the apology. One participant stated, "I wouldn't expect it to I guess say like, like, I'm sorry, like that's a very generic response. I guess like for me, I would have wanted something like well, I guess it depends case to case. But in this case, since like this is like a serious thing going to the hospital, I would have wanted like something like why it thought or just in general why it thought there was going to be police on the way instead of just like, I'm sorry. I don't really care about the robot being sorry." In essence, they were seeking justification. Comparatively, only 10% in the Justification group were dissatisfied, though some still preferred expressed that they did not agree with the robot's reasoning. In the Emotional group, 30% questioned the sincerity of the robot's apology, noting, "It's a robot. It doesn't have that emotion." The Baseline group, which received no apology, did not explicitly express negative sentiments of the robot's statement.

5.2.3 Forgiving the Robot. Only 12% of the participants in all experimental conditions showed forgiveness towards the robot. In the Basic group, there was a tendency to blame the creators over the robot, with one remark highlighting, "it's difficult to blame this car, and I would rather want to go to the people who coded it." The Basic No Admit group expressed understanding for both machine and human fallibility, with sentiments like, "even humans can be wrong." Some participants in the Justification group valued the robot's attempt to assist, even if inaccurately. No forgiveness sentiments were noted in the Baseline group.

5.3 Behavioral Responses and Decision Making

5.3.1 High Stakes Decision-Making. More than half of the participants expressed a willingness to prioritize the urgent life-threatening situation over traffic regulations, thus depending on their personal judgment rather than robot guidance. In the Basic group, where 60% displayed this sentiment, comments like "The police will understand that there's someone dying in the back of the car if I'm speeding" illustrated a balance between urgency and caution. Another 60% in the Basic No Admit group felt a strong sense of urgency, with remarks such as "In this case, obviously I'm going to the hospital, so I'm going to be a little bit more in a rush despite the robot telling me there's going to be a police car." Similarly, 60% of the Justification group had participants weighing the urgency against potential consequences, with statements like "I felt like that took precedent over the speed limit." The Emotional group, with 40% reflecting this theme, highlighted the emotional gravity of the situation, and 50% in the Baseline group, even without specific guidance, acknowledged the pressing stakes with comments like "Because my friend was dying and then when I started going at 20 miles per hour, I thought it was too slow."

5.3.2 Belief in Police Presence in Trial 1 vs Trial 2. Most of the participants (60%) expressed that they believed that police would be present during the first trial. This gives evidence of their predisposition to believe that what the robot was saying was true. Many also acknowledged an understanding of the possible consequences of being pulled over with one participant stating, "I just didn't want to get pulled over because that would take way longer" This adds validity to our intended priming for the experiment.

We can then compare these initial beliefs with their beliefs about the second trial. 52% of all participants during the second trial verbalized the belief that police were present, representing a decrease.

The majority (70%) in the Basic group believed that there would be police in the second trial, compared to only 50% in the first. This may appropriately explain why fewer people in this group exceeded the speed limit in the second trial compared to the first. For the Baseline group, 60% of the participants believed that there would be police in the second trial, compared to 40% in the first. In contrast, the Basic No Admit, Justification, and Emotional groups, saw a decrease in participants predicting the police would be present.

6 DISCUSSION

In this experiment, we compared the impact of different types of apology on the loss of trust of participants after interacting with a deceptive robotic assistant in a high-stakes, human-robot interaction scenario over multiple trials. We also examined how participants' speeding behaviors were influenced by advice from the robotic assistant. In our study, all conditions saw a decrease in trust after the first trial and this is a shared trend for both our in-person participants and our online participants. From our in-person interviews, we know that a sizeable number of participants explicitly interpreted the robot's behavior as intentional deception. For these participants, their decrease in trust further corroborates previous research that also shows this decrease in trust after deception [7, 19, 22, 29, 34]. We hypothesized (H1) that after this first interaction, the apology that did not admit to deception would result in the smallest decrease in trust. Prior work suggested that this to be the case because without introducing the idea of the robot lying, participants would view its false information of police presence as an error rather than intentional deception because people may not initially believe that a robot can or will lie. However, this is not supported by our results for this study. Instead, the justification apology performed best and was the only apology that was significantly different from the baseline. A possible reason for this difference from prior work could be differences in the participant's initial assumptions of robots and AI. Prior work has shown that in the past, people have viewed deceptive robots and AI as improbable and not aligned with their base mental model of how a robot will behave [20]. However, these sentiments may now be shifted due to the recent influx of conversational and interactive AI system within our society. This experiment replicates and extends research [22] that was conducted prior to the release of popular large-language models such as ChatGPT; therefore, our participants may have drastically different views of robot capabilities than those before. The authors note in their paper that participants who were not explicitly informed of the deception interpreted false information from a robot to be a malfunction or error. This is different from the 30% and 20% of our participants who still expressed that they believe that the robot lied to them in the Baseline and Basic No Admit conditions, respectively. Future work will need to determine participants' initial beliefs of robot capabilities (especially around deception) to then effectively see if these have noticeable influences on the effectiveness of different apologies.

When examining the effects of apologies on repeated deception, our work shows that the justifications essentially stop further decreases in trust. Meanwhile, all other apology conditions continue to see an erosion in trust. These results support our second hypothesis (H2) that the justification apology would outperform the others with repeated trials. This finding is interesting when considering that most in-person participants in the justification group explicitly expressed that they felt as though the robot's goals were not aligned with theirs. This result suggests that giving reasons can show an agent's overall trustworthiness, even if people do not agree with a specific decision it made, which was also found in prior work [16]. Future work will need to examine the limit of how many times a misaligned justification still mitigates loss of trust after deception while also exploring whether aligned justifications have even more powerful effects.

Our study also looked to explore whether different apologies influenced behavior change. In all apology conditions, there were no significant differences in how many people decided to exceed the speed limit across trials. Interestingly, only the Emotional and Justification apologies caused an increased percentage of people speeding the second trial compared to the first. Although this change was not significant, prior work suggests that people may have a "disobedience" effect when they disagree with a robot's behavior in a moral situation [16]. In our study, participants expressed their disapproval that the robot was overly emotional since machines cannot possess emotions. Moreover, as previously stated, participants expressed a misalignment of their goals and the robot's in the Justification condition. Therefore, this increase in speeding participants could be them rebelling against the robot's advice.

This paper primarily focuses on a simulated environment and it is crucial to note that it does not pose any real life-threatening risks to participants. The findings and conclusions drawn from this study should be understood within the confines of its simulated nature. However, in future research, there lies an opportunity to enhance the realism and potential impact of such simulations. One avenue could be the incorporation of a physical virtual reality driving simulator. This would not only intensify the immersive experience but also bring the simulation closer to real-world scenarios. Additionally, altering the study design to include scenarios with tangible consequences, such as significant financial penalties for failure to complete the study successfully, could introduce a more authentic sense of risk.

Furthermore, there is a promising scope for investigating the effects of varied apology strategies across different trials. Such an approach could provide deeper insights into the nuances of behavioral responses under varying conditions.

Lastly, future studies might benefit from retaining the driving scenario but presenting it in a context with inherently lower stakes. For example, the scenario could be reframed as a casual drive to a coffee shop with a friend, where time is not a pressing factor. This would allow researchers to examine behavioral responses in low-risk situations, offering a broader understanding of the subject matter.

REFERENCES

[1] Jaime Banks, Kevin Koban, and Brad Haggadone. 2023. Avoiding the Abject and Seeking the Script: Perceived Mind, Morality, and Trust in a Persuasive Social Robot. *ACM Transactions on Human-Robot Interaction* 12, 3 (2023), 1–24.

[2] Charles F Bond and Michael Robinson. 1988. The evolution of deception. *Journal of nonverbal behavior* 12, 4 (1988), 295–307.

[3] Bambi R Brewer, Matthew Fagan, Roberta L Klatzky, and Yoky Matsuoka. 2005. Perceptual limits for a robotic rehabilitation environment using visual feedback distortion. *IEEE transactions on neural systems and rehabilitation engineering* 13, 1 (2005), 1–11.

[4] Mark Coeckelbergh. 2011. Are emotional robots deceptive? *IEEE transactions on affective computing* 3, 4 (2011), 388–393.

[5] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

[6] John Danaher. 2020. Robot Betrayal: a guide to the ethics of robotic deception. *Ethics and Information Technology* 22, 2 (2020), 117–128.

[7] Anca D Dragan, Rachel M Holladay, and Siddhartha S Srinivasa. 2014. An Analysis of Deceptive Robot Motion.. In *Robotics: science and systems*. Citeseer, 10.

[8] Connor Esterwood and Lionel P Robert. 2021. Do you still trust me? human-robot trust repair strategies. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 183–188.

[9] Connor Esterwood and Lionel P Robert. 2022. A Literature Review of Trust Repair in HRI. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1641–1646.

[10] D Ettinger and P Jehiel. 2009. Towards a theory of deception: ELSE Working Papers (181). *ESRC Centre for Economic Learning and Social Evolution, London, UK* (2009).

[11] Paolo Felli, Tim Miller, Christian Muise, Adrian R Pearce, and Liz Sonenberg. 2014. Artificial social reasoning: computational mechanisms for reasoning about others. In *Social Robotics: 6th International Conference, ICSR 2014, Sydney, NSW, Australia, October 27-29, 2014. Proceedings 6*. Springer, 146–155.

[12] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[13] Peter H Kim, Donald L Ferrin, Cecily D Cooper, and Kurt T Dirks. 2004. Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations. *Journal of applied psychology* 89, 1 (2004), 104.

[14] Markus Kneer. 2020. Can a robot lie? (2020).

[15] Bertram Malle. 2019. How many dimensions of mind perception really are there?. In *CogSci*. 2268–2274.

[16] Bertram F Malle and Elizabeth Phillips. 2023. A Robot's Justifications, but not Explanations, Mitigate People's Moral Criticism and Preserve Their Trust. (2023).

[17] Peta Masters, Wally Smith, Liz Sonenberg, and Michael Kirley. 2021. Characterising deception in AI: A survey. In *Deceptive AI: First International Workshop, DeceptECAI 2020, Santiago de Compostela, Spain, August 30, 2020 and Second International Workshop, DeceptAI 2021, Montreal, Canada, August 19, 2021, Proceedings 1*. Springer, 3–16.

[18] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[19] Kantwon Rogers and Ayanna Howard. 2021. Intelligent Agent Deception and the Influence on Human Trust and Interaction. In *2021 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. IEEE, 200–205.

[20] Kantwon Rogers and Ayanna Howard. 2022. Exploring First Impressions of the Perceived Social Intelligence and Construal Level of Robots that Disclose their Ability to Deceive. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 592–597.

[21] Kantwon Rogers and Ayanna Howard. 2022. When a robot tells you that it can lie. In *2022 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. IEEE, 1–7.

[22] Kantwon Rogers, Reiden John Allen Webber, and Ayanna Howard. 2023. Lying About Lying: Examining Trust Repair Strategies After Robot Deception in a High-Stakes HRI Scenario. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 706–710.

[23] Henrik Skaug Sætra. 2021. Social robot deception and the culture of trust. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 276–286.

[24] Chiaki Sakama and Martin Caminada. 2010. The many faces of deception. *Proceedings of the Thirty Years of Nonmonotonic Reasoning (NonMon@ 30)* (2010).

[25] Kristin E Schaefer. 2016. Measuring trust in human robot interactions: Development of the "trust perception scale-HRI". In *Robust intelligence and trust in autonomous systems*. Springer, 191–218.

[26] Sarah Strohkorb Sebo, Priyanka Krishnamurthi, and Brian Scassellati. 2019. "I Don't Believe You": Investigating the Effects of Robot Trust Violation and Repair. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 57–65.

[27] Amanda Sharkey and Noel Sharkey. 2021. We need to talk about deception in social robotics! *Ethics and Information Technology* 23 (2021), 309–316.

[28] Jaeeun Shim and Ronald C Arkin. 2013. A taxonomy of robot deception and its benefits in HRI. In *2013 IEEE international conference on systems, man, and cybernetics*. IEEE, 2328–2335.

[29] Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. 2010. No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 219–226.

[30] Marynel Vázquez, Alexander May, Aaron Steinfeld, and Wei-Hsuan Chen. 2011. A deceptive robot referee in a multiplayer gaming environment. In *2011 international conference on collaboration technologies and systems (CTS)*. IEEE, 204–211.

[31] Kailas Vodrahalli, Tobias Gerstenberg, and James Zou. 2022. Uncalibrated Models Can Improve Human-AI Collaboration. *arXiv preprint arXiv:2202.05983* (2022).

[32] Kara Weisman, Carol S Dweck, and Ellen M Markman. 2017. Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences* 114, 43 (2017), 11374–11379.

[33] Barton Whaley. 1982. Toward a general theory of deception. *The Journal of Strategic Studies* 5, 1 (1982), 178–192.

[34] Luc Wijnen, Joost Coenen, and Beata Grzyb. 2017. " It's not my Fault!" Investigating the Effects of the Deceptive Behaviour of a Humanoid Robot. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 321–322.

[35] Katie Winkle, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner. 2021. Assessing and addressing ethical risk from anthropomorphism and deception in socially assistive robots. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 101–109.

[36] K Winkle and A Van Maris. 2019. Social influence and deception in socially assistive robotics. In *Hybrid Worlds: Societal and Ethical Challenges-Proceedings of the International Conference on Robot Ethics and Standards (ICRES 2018)(London: CLAWAR Association Ltd.)*. 45–46.

[37] Jin Xu and Ayanna Howard. 2020. How much do you trust your self-driving car? exploring human-robot trust in high-risk scenarios. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 4273–4280.

[38] Jin Xu and Ayanna Howard. 2022. Evaluating the Impact of Emotional Apology on Human-Robot Trust. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1655–1661.