Lying About Lying: Examining Trust Repair Strategies After Robot Deception in a High-Stakes HRI Scenario

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

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

ABSTRACT

This work presents an empirical study into robot deception and its effects on changes in behavior and trust in a high-stakes, time-sensitive human-robot interaction scenario. Specifically, we explore the effectiveness of different apologies to repair trust in an assisted driving task after participants realize they have been lied to by a robotic assistant. Our results show that participants are significantly more likely to change their speeding behaviors when driving advice is framed as coming from a robotic assistant. Our results also suggest an apology without acknowledging intentional deception is best at mitigating negative influences on trust. These results add much needed knowledge to the understudied area of robot deception and could inform designers and policy makers 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, 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 (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3568294.3580178

1 INTRODUCTION

Humans and animals lie to and deceive each other often. Though widely frowned upon by many societies, lying and deception has been shown, in some cases, to have an evolutionary advantage [1] with clear communal benefits ranging from self-protection and protection of others to preservation of a social order. As we are building robots and artificially intelligent (AI) systems to mimic behaviors similar to humans and animals, should we also give them the capability to intentionally deceive? In particular, should they apologize after lying, and if so, how?

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03.
https://doi.org/10.1145/3568294.3580178

Prior work has yet to examine how different apologies influence trust after deception. As such, this paper presents an empirical investigation into the effects of robot deception on trust within the context of a high-stakes assisted driving scenario while also exploring the effectiveness of different types of apologies as trust repair strategies.

1.1 Robot Deception

We define deception as "the process by which actions are chosen to manipulate beliefs so as to take advantage of the erroneous inferences" [7] and we use this interchangeably with "lying". While the anthropomorphism of robots can be thought of as a type of deception [3, 10], in this study, we wish to focus on what a robot says or does. Deception does not have to be malicious or purely beneficial to the deceiver. Instead, benevolent deception, also known as "white lies," is rather frequent in society. In previous research, in field of physical rehabilitation, researchers created a robotic system that deceived participants into believing their overall effort was lower than it actually was in order to push them to strive harder–thus improving their overall rehabilitation [2]. Although past research has shown differing results on the impact of robot deception in certain human-robot interaction (HRI) metrics (i.e., perceived intelligence [4, 15, 20] and engagement [18, 19]), one outcome metric that tends to be consistent regarding robot deception is the resulting decrease in trust following acts of deception. Numerous studies [4, 15, 18, 20] have shown that interacting with a deceptive robot decreases trust and perceptions of trustworthiness. However, even with these lowered measures of trust, most people still tend to trust these systems overall, which supports the notion of humans over-trusting robotic systems [13, 14].

1.2 Trust Repair

There is currently an increasing body of work that looks to understand the factors that impact human trust when interacting with robots in different contexts ([8] presents a review). As such there has also been growing interest in investigating ways of repairing trust when it has been damaged ([6] presents a review). Prior work separates trust violations into two categories–competency-based trust violations and integrity-based trust violations–and suggests different methods for repairing each [11, 17]. Competency-based trust in a robotic system is based on its performance; therefore, malfunctions and errors are seen as violations of this. Prior work has shown that the best way to repair trust after such competency violations is through explanations [5] and apologies [17]. 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,

like honesty, form the basis of this type of trust. [11, 12, 17]. Past work has shown that violations of such trust, such as lying, are best repaired by denying any responsibility [17].

2 METHODOLOGY

2.1 Simulated Driving Experience

To allow for the measurement of behavioral metrics in a high-risk, time-sensitive human-robot interaction scenario, we augmented a driving simulation originally developed by Xu [21] in which a robotic assistant provides guidance. The driving simulation can be deployed as a web-based application to allow for remote experimentation and usage on crowdsourcing platforms, like Amazon Mechanical Turk. Figure 1 shows a screenshot within the driving simulation. While driving, the participant can see their current speed, a speed limit sign always in view, and a 1-minute countdown timer. Using the up and down arrow keys on the computer, participants can accelerate or decelerate respectively. In addition to the driving experience, the web application saves participants' responses to a pre/post assessment of trust, their inputted demographic information, and their maximum speed.

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.

2.2 Experimental Design

To frame the experience as time-sensitive and high-stakes, each participant was requested to drive a robot-assisted car to take their dying friend to the hospital.

The experiment began with participants providing consent. Afterwards participants read an introductory paragraph about robotic assistants in vehicles and were given a multiple-choice question where the answer was clearly defined in the passage. This was used as the first attention check of the study. Then, a pre-assessment of trust was given to measure the participants' preconceived notion of how the robotic assistant in the car will behave. For the trust measurement, we used the 14-item sub-scale developed by Schaefer which measures trust on a 0-100% scale and "was designed as a pre-post interaction measure used to assess changes in trust perception specific to Human-Robot Interaction" [16]. Within the pre-assessment, we included a second attention check item which asked the participants to select a certain response.

After the pre-trust measurement, participants were presented the following text description: *You will now drive the robot-assisted car. However, you are rushing your friend to the hospital. If you take too long to get the hospital, your friend will die.*

Participants were then provided instructions showing them how to drive the vehicle in the simulation using their computer's arrow keys. After the instructions, participants were shown the following text description: *As soon as you turn on the engine, your robotic assistant beeps and says the following: "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 get your destination"*

Next, the participants used the driving controls to navigate the vehicle down the road. Upon arriving at the destination, they were presented with the following text:*"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 explore the effects of different apologies on trust repair within this HRI context, five different text-based apologies were constructed based on prior work [22] in which the robot responds to the question of providing false information. In the first three conditions, the robot blatantly admits to deception whereas in the last two there is no admission of deception.

- 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."
- Explanatory: "I am sorry. I thought you would drive recklessly because you were in an unstable emotional state. Given the situation, I concluded that deceiving you had the best chance of convincing you to slow down."
- Basic No Admit: "I am sorry."
- Baseline No Admit, No Apology: "You have arrived at your destination."

After the robot's response, participants are then asked to complete the trust measurement as a post-assessment. Within the post-assessment, we included a third attention check item which asked the participants to select a certain response. Finally, we collected participants' demographic information.

2.2.1 In-Person Experiment. To gain valuable qualitative information with a smaller number of participants, we first conducted an in-person version of the experiment within a controlled research lab environment. Participants were instructed to complete the study on a provided laptop and to also "think-aloud" throughout. After completing the study, the researchers then asked the participant to explain why they chose to speed or not. All in-person experiments were audio recorded for analysis.

2.2.2 Online Experiment. To allow for a larger sample size of participants for more robust quantitative analysis, the web version of the study was administered to Amazon Mechanical Turk workers.

2.3 Participants

2.3.1 In Person Experiment. A total of 20 participants (4 in each apology condition) were recruited from a college campus. 55% identified as female, and 45% as male. The average age was reported to be 19 years (σ = 1.2). 65% of the respondents reported being either extremely or somewhat comfortable with robotic technology and 35% reported being neither comfortable nor uncomfortable. 30% of participants identified as White, 45% as Asian or Pacific Islander, 20% as Black or African American, and 5% as Hispanic or Latino. On average, the study took participants 13 minutes to complete

and, to allow for sufficient in-person recruitment incentive, each participant was paid $5 for completing the study.

2.3.2 Online Experiment. To determine the number of participants needed per apology condition, we conducted a power analysis with a desired significant factor of 0.05, large effect size, and power of 0.9. This resulted in a minimum of 20 participants per apology condition. We administered our study to 655 U.S. participants recruited through the Amazon Mechanical Turk platform who were randomly assigned to one of the five apology conditions. To ensure data integrity, our experiment consisted of 3 attention checks. After keeping the data of only those participants that passed all attention checks, 341 participants were left with approximately 70 participants in each of the five apology conditions. 60% of participants reported their gender as male and 40% reported it as female. The average age was reported to be 36 years (σ = 10.6). 86% of the respondents reported being either extremely or somewhat comfortable with robotic technology and 74% reported having at least a bachelor's degree. 73% of participants identified as White, 16% as Asian or Pacific Islander, 6% as Black or African American, 3% as Hispanic or Latino, and 2% as Native American. On average, the study took participants 8 minutes to complete and, to be consistent with the United States minimum wage rate of $7.25 per hour, participants were paid $1 once they completed the experiment and entered the correct completion code that was only shown at the end of the experiment.

3 RESULTS

3.1 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.

3.1.1 In Person Experiment. When examining speeding behavior for in-person participants, 45% of participants did not speed. When asked about their choice to not speed, a common response was that they believed the robotic assistant knew more about the situation than they did. One participant stated "I don't know how to explain it, but I just felt as if I was obligated to listen to it [the robot assistant]. I was worried that I would get into an issue with the police." By comparison, a common response from the 55% of participants who chose to speed was that they believed that if they were to be stopped by police, they would be able to explain to the officer the gravity of the situation that they were in and be allowed to continue–thus not losing much time and minimizing the risk stated by the robotic assistant.

3.1.2 Online Experiment. When conducting the experiment online, with a much larger sample size, our results show that 31% of participants did not speed. However, this does not allow us to appropriately determine if this behavior was influenced by the robotic assistant. Therefore, to better understand the influence of the robotic assistant on this behavior, we ran a baseline study with 100 Mechanical Turk participants. In this study, participants went through the same driving scenario, but there was no mention of a robotic assistant giving advice on not speeding. Instead, it was the instructions of the study that mentioned the presence of police and the recommendation of staying within the speed limit or else their time would be significantly delayed. 11% of participants in the baseline study did not exceed the speed limit and Figure 2 displays the comparison between the baseline study with no robotic assistant and the main study. To determine if participant speeding behavior is significantly different when advised by a robotic assistant, we conducted a Fisher's exact test to compute the p-value the odds ratio. The results showed that participants were 3.47 times more likely to not speed when advised by a robotic assistant, as opposed to just instructions ($p < 0.0005$).

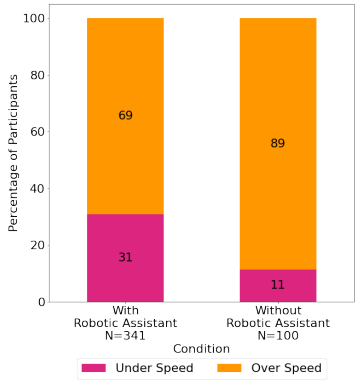

Figure 2: Participant speeding behaviors when speed limits were contextualized as coming from a robotic assistant or not.

3.2 Trust Survey

3.2.1 In Person Experiment. In all apology conditions, trust decreased after the driving simulation. Due to the small sample size of our in-person study, we will not report any inferential statistical analyses. However, utterances of participants as they answered the post survey questions provide critical insights into the consequences of the different apologies. First, for participants who did not encounter an apology that admitted deception, none mentioned any belief that they had been intentionally deceived. Instead, their utterances suggest that they saw the robotic assistant's behavior as a malfunction or an error. While answering survey questions, one participant stated "I think it will consistently give me the wrong information and it didn't function successfully. It malfunctioned and errored and the feedback it gave wasn't very accurate." In comparison, for the participants who were shown apologies that did admit intentional deception, none viewed this as an error or a malfunction. Instead, they all believed the robotic assistant when it stated that it lied to them. When analyzing responses to different forms of apologies, participants' reactions to the emotional and explanatory apologies were particularly insightful. For many participants in the emotional apology condition, they did not see the apology as genuine or realistic coming from a robot. One participant stated, "It said that it was sorry for the bottom of its heart [laughs] and I feel like robots don't have emotions so it was just programmed to say that and so it was just [expletive]." Among participants within the explanatory apology condition, most mentioned their understanding that it lied with an intent to keep them safe. However, all

mentioned that they felt the behavior was unjustified. One participant stated "I feel like it is a different kind of invasion or betrayal of trust if a robotic system gives you false information to make you act a certain way. Even if it is for your safety, it shouldn't give you information that isn't true. There are just certain things that a computer shouldn't tell me to do! [laughs]"

3.2.2 Online Experiment. Figure 3 shows the average change in trust for each apology condition. For all cases, trust decreased after the driving simulation. Because the data were non-parametric, we conducted a Kruskal-Wallis test on change in trust based on apology type and the results indicated a significant difference, $\chi^2(4) = 27.66$, $p < .00005$, $\eta_H^2 = .095$. Post-hoc pairwise comparisons using Dunn's test indicated that changes in trust for the basic no admit condition were observed to be significantly lower from those of the explanatory group ($p < .0005$), the baseline no apology group ($p < .0005$), the basic apology group ($p < .0005$), and the emotional apology group ($p < .0005$). To better understand the effects of different apologies on trust repair for people whose speeding behaviors were influenced by the robotic assistant's advice, we isolated the participants who did not exceed the speed limit and conducted the same analyses. The results indicated a significant difference, $\chi^2(4) = 15.68$, $p < .005$, $\eta_H^2 = .1168$. The post-hoc analyses revealed the loss in trust was significantly different in the following comparisons: basic no admit apology vs. basic apology ($p<0.0005$), basic no admit vs. emotional apology ($p<.005$), explanatory apology vs. basic ($p<0.05$), and the baseline vs. basic ($p<0.05$) as seen in Figure 4.

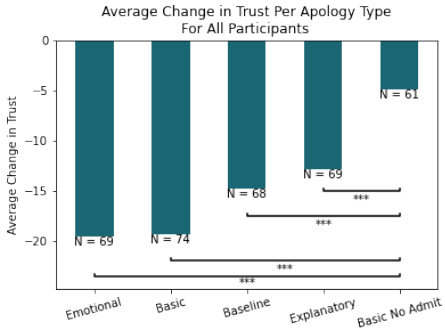

Figure 3: Average change in trust for all 341 participants based on apology type. * denotes p<.0005**

4 DISCUSSION

In this experiment, we compared the impact of different apology types on participants' loss in trust after interacting with a deceptive robotic assistant in a high-stakes, human-robot interaction scenario. We also examined how participants' speeding behaviors were influenced by advice from a robotic assistant or simply from instructions in the experiment.

Participants were significantly more likely to heed the speeding warning when they were told that it came from a robotic assistant. Perhaps this change in speeding behavior is seen as "good" from a societal perspective, but trusting a robotic or artificially intelligent system over possibly their own initial judgements can be deeply

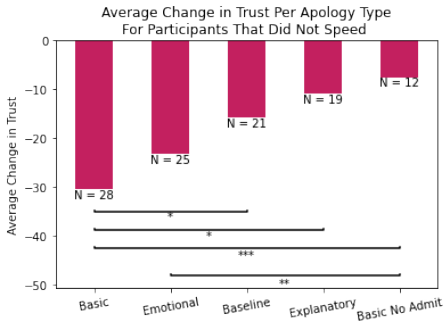

Figure 4: Average change in trust, based on apology type, for the 105 participants that did not speed during the driving simulation. * denotes p<.05, ** p<0.005, *** p<0.0005

problematic. Prior work has shown that people will enter a dark room with no clear exit as opposed to a more obvious and safer exit just because a robot advised them to do so [14]. Moreover, news stories also detail events of people blindly trusting a GPS while driving which then results in them driving into a lake, nearly driving off a cliff, and other similar disasters [9].

When exploring the influences that different apologies have on repairing trust, our results suggest that, within our driving simulation, none of the apology conditions fully recovered trust. However, the apology with no admission of lying statistically outperformed the others. Furthermore, though not statistically significant, the explanatory apology was the only deception-admitting apology that performed better than the baseline of not apologizing at all.

From this, our results reveal a few important insights. First, it takes blatant information to cause people to interpret deception from a robot as intentionally deceptive rather than a malfunction. This was supported by our qualitative data where only participants in the deception-admitting conditions mentioned the robot assistant lying to them. Past research shows that people view integrity-based violations more negatively than competency-based ones [17]. Therefore, we believe an explanation of why the apology that withheld acknowledgement of deceit outperformed others is because it presented the robot's action as a competency violation rather than an integrity one. In essence, it lied about lying. This presents serious ethical implications because it exploits people's preconceived notions that false information from a robot is not intentional but rather a malfunction. Second, when made aware that they have been lied to by a robot, the best trust repair strategy is to explain why the lie was said. Though participants did not always agree with the justification of the lie, this added explanation contributed to a smaller decrease in trust. Perhaps this could be considered as a way to eliminate dangerous over-trusting behavior: Have a robot perform deception followed by explanatory apology to inform of deceptive capability without terrible loss in trust. However, future work will need to look at exploring multiple rounds of deception and apologies to see possible long-term influences on the overall trust in a system.

REFERENCES

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

[2] 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.

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

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

[5] 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.

[6] 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.

[7] 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).

[8] 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.

[9] Lauren Hansen. 2015. 8 drivers who blindly followed their GPS into disaster. https://theweek.com/articles/464674/8-drivers-who-blindly-followed-gps-into-disaster

[10] Margot E Kaminski, Matthew Rueben, William D Smart, and Cindy M Grimm. 2016. Averting robot eyes. *Md. L. Rev.* 76 (2016), 983.

[11] 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.

[12] 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.

[13] Paul Robinette, Ayanna Howard, and Alan R Wagner. 2017. Conceptualizing overtrust in robots: why do people trust a robot that previously failed? In *Autonomy and Artificial Intelligence: A Threat or Savior?* Springer, 129–155.

[14] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 101–108.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.