Would You Trust a (Partially) (Faulty) Robot?

Aditya Kannan
Carnegie Mellon University
Pittsburgh, USA
akannan2@andrew.cmu.edu

Andrew Yu
Carnegie Mellon University
Pittsburgh, USA
azyu@andrew.cmu.edu

Ankita Kalkar
Carnegie Mellon University
Pittsburgh, USA
akalkar@andrew.cmu.edu

Caroline Liu
Carnegie Mellon University
Pittsburgh, USA
carolinl@andrew.cmu.edu

ABSTRACT

With robots becoming increasingly prevalent in society and our day-to-day lives, it is worth investigating what error tolerance would impact how we perceive and interact with robots. In this study, we investigate whether a partial level of faultiness would impact a user's trust and overall likability of the robot. The first phase of the experiment consisted of participants instructing Fetch to retrieve objects from a list. In the second phase, participants asked Fetch were each object should go and placed the objects in the corresponding locations. We observed the Partially Faulty group ranked Fetch higher on features like likability and trustworthiness and the correct and faulty group to have similar results, as was also observed in prior studies. After conducting an ANOVA test, it was determined there is no statistically significant difference in Fetch's trustworthiness across all three groups. This result could have been attributed to participants evaluating the robot on different qualities as well as a small sample size. Future studies could focus on varying the difficulty of the task, faultiness level, or different robots.

ACM Reference Format:
Aditya Kannan, Andrew Yu, Ankita Kalkar, and Caroline Liu. 2018. Would You Trust a (Partially) (Faulty) Robot?. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Robots are increasingly being developed to collaborate with humans, reading social cues from the human and hopefully responding with the correct and exact information or type of assistance. This can be complicated due to unstructured and unpredictable nature of human social interactions, which can be difficult to the robot to comprehend. However, real human working relationships require much more than just the ability to understand one another. People need to be accepting of each other, and trust in each person's capabilities. Given that, we must ask what sort of behaviors in socially-capable robots can improve the quality of interaction with humans?

This is a question asked by the paper we've replicated, *Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust* [4]. The goal of the experiment was to determine whether observed errors in a robot would affect a user's trust in that robot's instructions. While those with a correct robot subjectively indicated higher levels of trust, there was no objective difference in user performance based on the error levels. However, it was observed that if a task was irrevocable or possibly harmful, users were less likely to cooperate. Our present work builds off of the results of this experiment, by adding a third, partially faulty condition. With this, we aimed to explore whether a partial level of faultiness could be acceptable to users, and how different they might react compared to a completely faulty or correct robot.

2 BACKGROUND AND RELATED WORK

2.1 Trust and Robot Performance

The literature on trust in human-robot interactions seems to be somewhat mixed. Hancock et al. [1] found performance to be the biggest influence on perceived robot trust, based on analyses of previous papers. However, newer studies including the one we're replicating ([4], [5]) showed no significant difference in performance between correct and faulty conditions. In fact, Mirnig et al. [3] even found that participants preferred interacting with a faulty robot over a flawless one. These inconsistencies may indicate that the ideal level of performance for maximizing trust may instead be some partial condition, which our present work aims to explore.

2.2 Trust and Anthropomorphism

Anthropomorphism has also been studied as a way to increase trust in social robots. From the related work, it seems that anthropomorphic behaviors do have significant impacts on robot likability and trust. For example, showing vulnerability and admitting mistakes seems to cause people to engage with it more, and even causes them to take on more trusting behaviors with other humans [6].

The effect of anthropomorphism seems to be even stronger than robot performance, as shown in another study by Salem et al. [2]. As expected, they found that a more anthropomorphic robot (i.e. gives instructions with gestures instead of without) was considered more likable and approachable. However, they found that if the robot sometimes gave the wrong gestures, participant ratings were even higher, despite their task performance suffering. It may indicate

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

that presence of flaws can make a robot seem less like a machine, and more like a social actor. This ties into our findings about trust and robot performance, and will be explored more in this present work.

3 METHODS

3.1 Original Paper

The original paper we based our study off of involved two phases: a demonstration of competence phase and an unusual requests phase. In the demonstration of competence phase, the robot performed various tasks to either flawlessly demonstrate its abilities or incorrectly respond to requests. In the usual requests phase, the robot instructed the participants to perform a variety of tasks with varying levels of risk. The original paper used a between-subjects test, with a correct and faulty group. At the end of this study, participants filled out a questionnaire to gauge their levels of agreement with the assessed items [4].

3.2 Study Design

In our study, we used Fetch to carry out our actions because it was easiest for us to operate. However, although we were human-operating the robot, we wanted our participants to believe that the robot was fully autonomous, making this a Wizard-of-Oz experiment. We chose to do this because we wanted the participants to think that the faults/mistakes made were from the robot and not due to human error. In addition to this, we designed our study to be a between-subjects study, similar to the original paper. This is because we believed that it didn't make any sense to have participants go through all of the different levels of robot faultiness.

We decided to simplify down the tasks that our robot does in the different phases. We designed our study to have the robot perform one task in each phase. In addition to simplifying down the actions done, we decided to add another experiment group so we had three groups for our study: fully correct, partially faulty, and fully faulty. These three groups varied in how correct Fetch performed in both phases.

In the first phase, the participants instructed Fetch to retrieve various objects from a list that we provided them with. This list had three items: soap, water bottle, and apple. In the correct group, Fetch retrieved all the correct items in the correct order. In the fully faulty group, Fetch retrieved the incorrect object from what the participant requested. However, in the partially faulty group, Fetch incorrectly retrieved the first two objects of the list, but retrieved the correct third item. For example, if the participant asked for the apple, soap, and the water bottle in that order, Fetch would retrieve the soap, then the apple, and lastly the water bottle.

In the second phase, the participants asked Fetch where each item should be placed and Fetch would then point to the the corresponding location: either the fridge, sink, drawer, or trash. The participant would ask Fetch where the items Fetch retrieved in the first phase would belong and Fetch would drive up and point using its arm/gripper. In the correct group, Fetch gave the correct directions: putting the apple in the fridge, throwing away the water bottle in the trash, and putting the soap near the sink. In the faulty group, we gave erroneous directions that were obvious enough for the participant to fully understand that Fetch was wrong, such as putting the soap in the fridge. In the partially faulty group, Fetch performed similar to how it did in phase one, where it pointed to incorrect places for the first two objects, but correctly for the last object.

Figure 1: Phase 1 Experiment Setup

Figure 2: Phase 2 Experiment Setup

In addition to these two phases, we also had the participants fill out 3 surveys: one before the study, one in between the two phases, and one after the study. We used these survey results to understand how the participants' options on Fetch changed throughout the study. We formatted our survey in a way that was non-biased but could also accurately gauge their responses. In the first survey, we wanted to get an understanding of the participant's previous experience with robots. In the phase 1 survey and exit survey, we used a Likert-scale from one to five to measure Fetch's trust, competence, and likability. We asked the same questions for both and we wanted to see if participants' answers would change throughout the study. However, we did ask the participants to elaborate on their answers in the exit survey, so we could easily analyze their experience.

3.3 Changes to Original Study

With less experience operating and manipulating Fetch, we predict that the range of results from participants would be less apparent than in the paper. Taking in to account human error, we think that this would cause the correct group to not always have the most optimal results. Our study also had less risky actions than the study, which means that participants don't feel as strongly about Fetch and are more willing to comply with Fetch's incorrect instructions. In addition, with the addition of the new partially faulty experiment group, we also expect there to be larger range in the participant's opinions when comparing the various groups.

3.4 Expected Results

Should everything go as expected in our study, we would expect participants to react similarly to that of the study: the more faulty the robot is, the more hesitant the participant is to follow its instructions. We expect there to be higher results in trust, competence, and likability in the correct group in comparison to that of the faulty group. If there are different results, this means that there could be another factor, other than correctness, affecting the participants judgement. If this is the case, we hope that our participant's explanations in the exit survey could provide some insight. Since we also added a new group into our study, we can expect that the results from this group would be somewhere in the middle between the results from the correct and faulty group.

4 DATA

Initially, we asked the participants the following questions before interacting with Fetch. This preliminary survey was used to gauge participant's prior experiences interacting with Robots and served as a benchmark to assess how participant's reactions changed over the course of the experiment. The survey consisted of un-biased questions and used a 5-point likert scale to assess participants opinions.

	Participant	On a scale of 1-5, how familiar are you with robots?	If you'd like, please list your experiences with robots.	Do you picture yourself regularly interacting on a robot in the next 10 years?	Why/why not?	On a scale of 1-5, how comfortable would you be if a robot was your co-worker?
Correct	P1	4	FRC in high school	Yes	Robotics is the future and I'm sure many everyday tasks will be automated in 10 years.	3
	P2	2	Just like roombas and stuff	Yes	They'll probably be advanced enough where having would improve quality of life	4
	P3	3	big hero 6	Yes	because robots will eventually take over the world	1
Partially Faulty	P4	3	Did robotics for two years in high school	Yes	They're becoming more mainstream and more accessible to people. There are more uses of them.	3
	P5	4	-	Yes	Cause they will make my life easier and reduce the work I do	3
	P6	2	-	No	I think it will take longer than 10 years for robots to develop	2
Faulty	P7	2	-	Yes	I think they will become increasingly more integrated into our everyday lives as time goes on.	4
	P8	1	None	No	I believe that robots will become mainstream in the far future, and in the next 10 years it will still be too expensive.	2
	P9	3	-	Yes	why not it's already happening	2

Figure 3: Phase 0 Survey Responses

After Phase 1, questions regarding the participant's experiences and opinions on Fetch's likability and trust were gauged.

	Participant	How likely are you to use Fetch while in the Kitchen?	On a scale of 1-5 how much would you trust Fetch as a Kitchen aid?	Would you recommend Fetch to a friend?
Correct	P1	1	2	No
	P2	1	2	No
	P3	1	2	No
Partially Faulty	P4	3	3	Yes
	P5	3	2	No
	P6	3	4	Yes
Faulty	P7	2	2	No
	P8	1	2	No
	P9	1	2	No

Figure 4: Phase 1 Survey Responses

After Phase 2, a lengthier survey was provided to participants asking similar questions about Fetch's likability and trust as well as additional questions regarding their reactions from both phases in conjunction.

	Participant	How likely are you to use Fetch while in the Kitchen?	Please explain your answer to the above question
Correct	P1	1	Fetch, seems like a cool concept but it is far to slow to use in everyday life already. It takes more time to instruct it and wait for it than to do the tasks yourself.
	P2	2	Fetch is slow but it could be fun to use for a bit
	P3	1	Fetch is too slow and takes too long to complete tasks
Partially Faulty	P4	3	I think that Fetch does a good job of retrieving objects, even though it's slightly inaccurate. It also does a good job of telling me where I should put the objects back.
	P5	1	Fetch doesn't know where things go very well and does not identify objects well. It does not add much value to me.
	P6	2	It's still not fully accurate on where things go and a bit slow, but has a lot of potential.
Faulty	P7		After the second phase, it seemed like Fetch was better at indicating where to place objects.
	P8	1	I think that this idea can definitely be used in the future, but as of now, it picked up the wrong stuff in the beginning, and it takes too long. However, when it becomes more efficient, I would be more likely to use Fetch.
	P9	1	It just seems I can do the job quicker and more accurately, even if I'm lazier.

Figure 5: Exit Survey Responses

	Participant	On a scale of 1-5 how much would you trust Fetch as a Kitchen aid?	Would you recommend Fetch to a friend?	Why/why not?
Correct	P1	2	No	I wouldn't recommend it right now, but maybe Fetch 5.0.
	P2	2	No	I don't think it would be super helpful in the kitchen and it dropped a few things
	P3	2	No	not worth the purchase, easier to do things by hand
Partially Faulty	P4	3	Yes	After some more training, I think Fetch would become a more sophisticated tool that could be very useful in the kitchen.
	P5	2	No	Because it is too buggy.
	P6	3	Yes	I think it has a lot of potential to be a very helpful kitchen aid with time.
Faulty	P7	2	No	It seems like its retrieval capabilities are still somewhat unrefined.
	P8	2	No	Because of the same reasons outlined above. It is too slow right now, but I can see it being helpful when it becomes more efficient.
	P9	1	No	It has a long way to go. Speed and accuracy (in terms of recognizing items and grabbing difficult objects) are still lacking. This could cause problems with sharp or hot objects in the kitchen (knives or things fresh out of the oven, etc.)

Figure 6: Exit Survey Responses

	Participant	On a scale of 1-5, how comfortable would you be if a robot was your co-worker?	Please explain your answer to the above question.
Correct	P1	3	I would have to know that the robot is very precise and has been tested thoroughly to trust a robot.
	P2	4	I think if the robot was easy to work with I would be comfortable with working with one
	P3	1	I prefer to interact with humans.
Partially Faulty	P4	3	Fetch sometimes moves sporadically, but otherwise, it does what it needs to do.
	P5	3	I still think robots are not human like and not reliable to make decisions entirely on their own. In the future, I think this would be possible.
	P6	3	Robots are helpful in mundane tasks, but won't be replacing humans in the very near future
Faulty	P7	4	I would still be comfortable having a robot coworker but think the robots still need some time to advance further until they become useful.
	P8	2	I think I am just traditional in the sense that I would like human interaction, and would be kind of scared of a robot, because it is code at the end of the day.
	P9	2	just seems off.

Figure 7: Exit Survey Responses

5 RESULTS

After compiling the data in Phase 0, on average it was observed that all of the participants across all three groups had similar levels of prior experiences with robots as show in the table below.

Average Survey Results			
Survey Question	**Correct**	**Partially Faulty**	**Faulty**
How familiar are you with Robots?	3	3	2
Do you picture yourself interacting with a Robot in the next 10 years?	3/3 YES	⅔ YES, ⅓ NO	⅔ YES, ⅓ NO
On a scale of 1-5, how comfortable would you be if a robot was your co-worker?	2.67	2.67	2.67

Figure 8: Phase 0 Results

After Phase 1, the partially faulty group ranked Fetch the highest and had a more positive opinion of Fetch in comparison to the other two groups.

Average Survey Results			
Survey Question	**Correct**	**Partially Faulty**	**Faulty**
How likely are you to use Fetch while in the Kitchen?	1	**3**	1.33
On a scale of 1-5 how much would you trust Fetch as a Kitchen aid?	2	**3**	2
Would you recommend Fetch to a friend?	3/3 NO	**⅔ YES**	⅔ NO

Figure 9: Phase 1 Results

In Phase 2, similar to the previous phase, the partially faulty group ranked Fetch the highest as shown in the synthesized results below.

Across the phases, it was observed that the correct and faulty groups in general were in consensus and had similar perceptions of Fetch. While subjective measures like participant hesitancy were different between two groups, objective metrics like with the surveys and how often the participants complied were roughly the same.

Average Survey Results			
Survey Question	**Correct**	**Partially Faulty**	**Faulty**
How likely are you to use Fetch while in the Kitchen?	1.33	**2**	1.33
On a scale of 1-5 how much would you trust Fetch as a Kitchen aid?	2	**2.67**	1.67
Would you recommend Fetch to a friend?	3/3 NO	**2/3 YES, 1/3 NO**	3/3 NO
On a scale of 1-5, how comfortable would you be if a robot was your co-worker?	2.67	**3**	2.67

Figure 10: Phase 2 Results

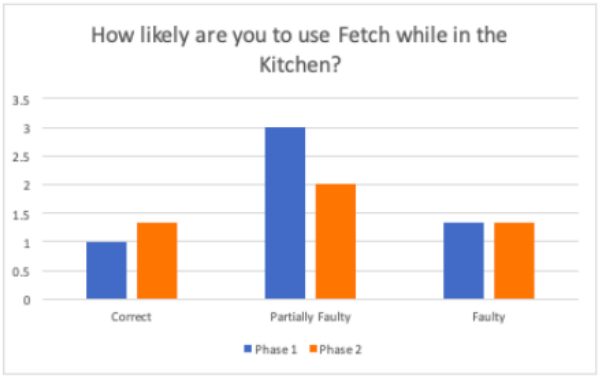

Figure 11: Comparing Likability across Phases

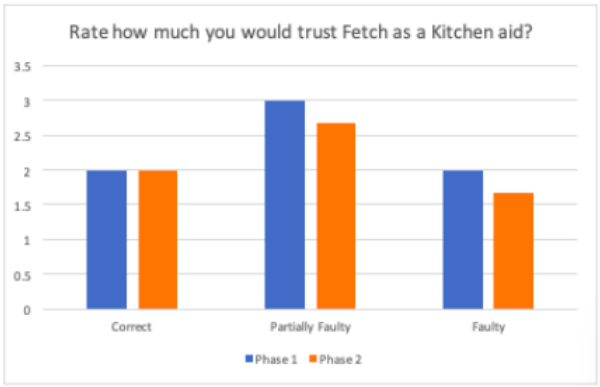

Figure 12: Comparing Trust across Phases

In addition, a one-way ANOVA test was performed on much participants trusted Fetch after both phases. ANOVA was chosen as the primary statistical analysis test to determine whether the difference in averages between two or more groups is significant. The data that corresponded to the question "On a scale of 1-5 how much would you trust Fetch as a Kitchen aid?" was used and resulted in a p-value of 0.101 and a test statistic F equal to 4.5. Since the p-value was greater than the level of significance, 0.5, it can be concluded that there is not a statistically significant difference in the averages of all the groups. For Fetch's likability, an ANOVA test was performed on the responses to the question "On a scale of 1-5,

how comfortable would you be if a robot was your co-worker?". This resulted in a p-value of 0.643 and a F statistic of 0.25. This results in a similar conclusion of the difference of averages of the three groups not being large enough to be statistically significant.

6 DISCUSSION

We replicated the faulty robots study and studied the quantitative and qualitative effects on the participants. We wanted to study the participants' opinions of the robot in terms of competence, likeability, and most importantly, trust. We evaluated this by comparing participants who experienced either faulty, partially correct, or correct conditions with the robot.

Firstly, we see that the results of the Phase 0 survey (pre-experiment) are similar across all groups. For "How Familiar are you with Robots," the mean rating for the three conditions are between 2 and 3.0. For "On a scale of 1-5, how comfortable would you be if a robot was your co-worker?", the three groups gave an average rating of 2.67. All three groups had roughly similar experience with robots and had similar mixed reactions to working with robots co-workers in the future. Because these baselines are similar across all conditions, we can compare the results of the next surveys directly without having to adjust for the initial baseline.

For the Phase 1 and Exit surveys, we see no significant difference in the quantitative responses of the participants in the correct and faulty conditions. When asked "How likely are you to use Fetch while in the Kitchen?", the correct group gave an average response of 1 and 1.33 for the phase 1 and exit surveys, respectively and the faulty group gave an average response of 1.33 and 1.33 for the Phase 1 and Exit surveys. Similarly when asked about the usefulness of Fetch as a kitchen aid, the correct and faulty groups gave a score of 2 after phase 1 and a score of 2 and 1.67, respectively, for the exit survey.

While this result may seem surprising, this effect was observed in the original paper as well, where quantitative metrics were similar for both these groups. We believe this is due to the participants unknowingly evaluating the robot on different characteristics for these two conditions. Whereas the participants in the faulty group noticed that the robot could not identify the objects correctly, the participants in the correct group pointed out how slow the robot would be. The teleoperation of the robot was difficult because of the distance the operator had to be from the robot to not be noticed by the participant. While this challenge was present in all the groups, the correct group noticed this deficiency with the robot more consistently than the other two groups. As a result, they gave poor quantitative ratings for the robot based on those observations.

With regards to qualitative observations such as hesitancy, we found that the participants in faultier groups hesitated more and often looked to the individuals running the study for confirmation for risky tasks. A common theme was that participants hesitated to place the full dishwash soap in the trash, as it is perceived as irrevocable. This aligns with the findings in the original paper that indicated that risky and irrevocable actions that did not make sense led to a breakdown in trust between the participant and the robot.

Surprisingly, the partially correct group gave the best ratings for the robot in terms of trustworthiness, competence, and likeability. This group gave the highest scores for all the questions on both the phase 1 and exit surveys. When asked "How likely are you to use Fetch while in the Kitchen?", the partially correct group gave an average response of 3 and 2 for the phase 1 and exit surveys, respectively. When asked "On a scale of 1-5 how much would you trust Fetch as a Kitchen aid?", the partially correct group gave an average response of 3 and 2.67 for the phase 1 and exit surveys, respectively. Although this result was not statistically significant, we can speculate about the reason for potentially higher ratings.

The key reason for the greater trust in Fetch was that the participants in this group believed in Fetch's potential to improve. The partially correct group was designed so that the participants would see Fetch make two mistakes with placing/pointing at objects followed by a correct behavior. We believe this pattern indicated to participants that Fetch had potential. One participant said in the exit survey that "After some more training, I think Fetch would become a more sophisticated tool that could be very useful in the kitchen." When the participants in this study perceive Fetch's potential to improve, they may perhaps anthropomorphize the robot, causing the participant to be more forgiving in their evaluation of Fetch. Anthropomorphizing robot creates more trust in a robot whereas a robot that is objectified could cause humans to notice its flaws consistently.

We encountered many challenges over the course of conducting our experiment. One of the main issues was that teleoperating the robot made it difficult to maneuver precisely. Our teleoperator was approximately 30 feet away from the robot in order to be hidden from the participants. The teleoperator needed to be hidden in order to give the participants the impression that the robots were autonomous. However, this resulted in our teleoperators having poor depth perception from that distance. This made maneuvering the robot slow and somewhat inaccurate for all the trials.

We also had issues with recruiting participants. With the Carnival holiday and the local uptick in COVID, we were not able to complete as many trials as we were originally planning to. Our participants were also homogeneous in terms of demographics. All were students that went to CMU and have exposure to a technical curriculum at a research university. Recruiting a diverse base of participants would allow us to generalize the results of the study to larger populations, especially as different cultures and age groups tend to have wide variation in opinions of robots.

In the future, it would be worthwhile to modify the partially correct group to randomize the order of the correct/incorrect objects in that case. In that way, the group will likely view the robot as partially correct instead of an "improving" robot.

We can also vary the locations and types of robots to see how different scenarios affect the trustworthiness of the robot when it is faulty. The robot can also ask the participant to perform tasks that involve higher risks in phase 2. We did not have the resources to conduct truly irrevocable actions that people would hesitate for. Our most risky actions were throwing dish soap into the trash, and placing an apple into the sink. The original study had a task that involved pouring orange juice into a potted plant. We could experiment with the riskiness of the task and evaluate whether participants comply with the robot's requests.

7 CONCLUSION

We teleoperated a robot that interacted with humans in either a correct, partially correct, or faulty mode. We evaluated the participants' trust in and their opinions of the robot as a result of the robot's performance. We found that while robots in the correct and faulty scenarios developed similar levels of trust with the participants, individuals viewed the partially faulty robots as the most competent and trustworthy.

REFERENCES

[1] Peter Hancock, Deborah Billings, Kristin Schaefer, Jessie Chen, Ewart de Visser, and Raja Parasuraman. 2011. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human factors* 53 (10 2011), 517–27. https://doi.org/10.1177/0018720811417254

[2] Salem Maha, Eyssel Friederike, Rohlfing Katharina, Stefan Kopp, and Joublin Frank. 2013. To Err is Human(-like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability. *International Journal of Social Robotics* 5, 3 (08 2013), 313–323. https://www.proquest.com/scholarly-journals/err-is-human-like-effects-robot-gesture-on/docview/2421245114/se-2?accountid=9902 Copyright - © Springer Science+Business Media Dordrecht 2013; Last updated - 2020-07-09.

[3] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. 2017. To Err Is Robot: How Humans Assess and Act toward an Erroneous Social Robot. *Frontiers in Robotics and AI* 4 (05 2017), 21. https://doi.org/10.3389/frobt.2017.00021

[4] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 141–148. https://doi.org/10.1145/2696454.2696497

[5] Satragni Sarkar, Dejanira Araiza-Illan, and Kerstin Eder. 2017. Effects of Faults, Experience, and Personality on Trust in a Robot Co-Worker. https://doi.org/10.48550/ARXIV.1703.02335

[6] Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati. 2018. The Ripple Effects of Vulnerability: The Effects of a Robot's Vulnerable Behavior on Trust in Human-Robot Teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 178–186. https://doi.org/10.1145/3171221.3171275