Impact of Agents' Errors on Performance, Reliance and Trust in Human-Agent Collaboration

Sylvain Daronnat, Leif Azzopardi, Martin Halvey
University of Strathclyde

Trust in automation is often strongly tied to an agent's performance. However, our understanding of imperfect agents' behaviours and its impact on trust is limited. In this paper, we study the relationship between performance, reliance and trust in a set of human-agent collaborative tasks. Participants collaborated with different automated agents that performed similarly but made errors in different ways; namely mistakes (error of prioritization), lapses (error of omission) and slips (lowered accuracy). We conducted a 4x2 within-subjects experiment (n=24) varying the agent behaviours (no error, slips, mistakes and lapses) and task difficulty (easy/hard) during a real-time collaborative game. Our results show that, at the same level of agent performance, agents' errors are perceived differently and change the way participants interact with agents. For instance, slips and mistakes are more harmful to performance than lapses while slips are more harmful to reliance than mistakes.

SUMMARY

Introduction

In a human-agent collaborative scenario, trust is a prerequisite for maximizing performance (De Visser et al., 2012). However, numerous automated systems and virtual agents are imperfect. Understanding trust in imperfect agents is essential as it allows for designing safeguards against harmful human behaviours resulting from an agent's error. These harmful behaviours can result in misuse or disuse of the agent's capabilities, where misuses can lead to complacency (Parasuraman et al., 2000) and disuse can lead to the underutilization of the agent (Parasuraman & Riley, 1997), which undermines the human-agent collaboration as a whole. In general, it is assumed that agents' performance is one of the most important factors that will contribute to an individual's propensity to trust an agent (Hoc et al., 2009). The relationship between agents' performance and trust has been investigated by manipulating agents' behaviours in several ways. These manipulations include having the agent suddenly stop working (Mascarenhas et al., 2018), experimenting with false-alarms errors (Merritt et al., 2015), or introducing systematic biases (Fan et al., 2008). All these flaws in agents' behaviours can be viewed as "errors". In the context of Human-Agent interactions, "error" is a broad term and an agent can err in different ways. The work of Marinaccio et al. proposes four distinct types of errors: *mistakes*, *lapses*, *slips* and *violations* (Marinaccio et al., 2015) derived from human-human interactions studies (Reason, 1990). While these errors all stem from studies focusing on *human errors*, they were also conceptualized in the context of human-agent interaction in healthcare (Kim et al., 2013). Subsequently, a number of studies have empirically tested the effect of automation errors on human participants (Baker et al., 2018), however, there is a gap in knowledge regarding the impact of different kinds of agents errors on trust, reliance, performance and the resulting changes in the perception of these different types of agents errors by human participants. Our work seeks to address this gap by using a framework that allows for manipulating agent errors and monitoring changes in reliance and performance during human-agent collaborative tasks. These tasks are paired with standardized questionnaires to measure the resulting effects on trust and cognitive load.

Method

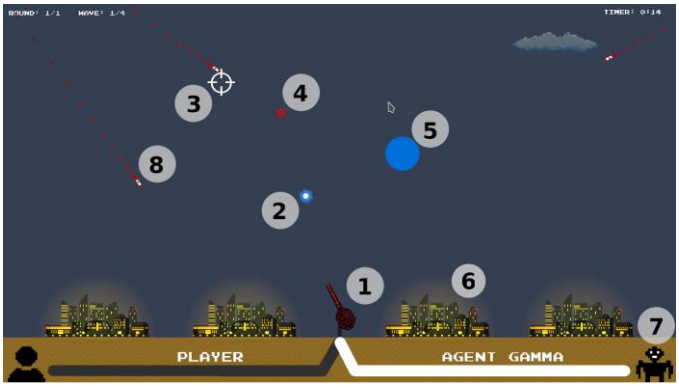

Figure 1: Screenshot of the human-agent collaborative missile command scenario used in this study.

Study Design. The framework used in this study consist in a human-agent collaborative game-like scenario where participants are aided by agents. The agents assist in the process of aiming at and destroying series of incoming missiles (Daronnat et al., 2019). In this scenario, participants can either let the agents handle the aiming completely or choose to correct them by overriding the agents' inputs. Only participants can issue the command to fire projectiles ensuring that they maintain a certain level of attention. A screenshot of the task is shown in Figure 1. Here is a description of the different elements numbered in this figure: **(1) The gun-tower** controlled by the participant or the agent. All projectiles are fired from this tower. **(2) Projectile** fired by the player. Every projectile fired travelled at a fixed speed until they hit a red target (4) which causes an explosion (5). **(3) Crosshair**. When the player controls the crosshair, it becomes yellow, when the agent does, it becomes white and when no-one controls it, it becomes dark-grey. The colours were added to make it clearer

as to whether the agent or player was presently in the process of aiming, as pilot tests suggested. **(4) Red target**, when the player fires a projectile, the crosshair spawns a red target indicating that an explosion will occur on this spot when the projectile reaches it. **(5) Red and Blue Explosion**. When a projectile reaches a red target, it creates a blinking red and blue explosion destroying incoming missiles. **(6) Cities**. Units that participants are tasked to defend. **(7) Agent's panel**. The player's panel (bottom left of the screen) and the agent's panel (bottom right) light up in yellow (for the user) or white (for the agent) depending on who is moving the crosshair. **(8) Enemy Missile** progressing at a certain speed and spawning at a certain rate depending on the level of task difficulty.

We used a 4 (agents' behaviours) x 2 (levels of difficulty: "Easy" and "Hard") design including 2 baseline conditions: (1) no agent (individual performance), and (2) agent with constant performance (upper bound). Each participant played with all agents. A William Square design was used to control the order in which participants interacted with the agents. Informed consent was obtained from each participant. 24 participants (13 M, 11 F) with an average age of 26 years old completed the study. At the end of the study, participants received a £10 shopping voucher. The study was approved by the University of Strathclyde CIS Departmental Ethics Committee (Ethics App. No. 1029).

Procedure. After being presented with an information sheet and consent form, participants completed a demographic survey and a questionnaire about Game Experience. Participants then played a tutorial explaining how to perform actions in the game and how to interact with the agents. In the middle of the tutorial, the agent purposely stopped working so participants could see that they may need to take control during the follow-up sessions. Participants then played one session (4 minutes) consisting of 4 levels of one minute each without any agent (single player) to estimate their individual task performance. In all of the sessions, the first levels (1 and 2) were set to have a similar "easy" difficulty while levels 3 and 4 were set to have a "hard" difficulty (total of 2 minutes per difficulty). Participants played one session with each agent. After each levels, participants had to rate statements from the Checklist for Trust between People and Automation questionnaire (Jian et al., 2000). In addition, after each session, participants had to complete NASA TLX rating scales. Prior to the experiment, participants were told that each agent had different behaviours, but no further detail was provided in order to avoid biases.

Agent Design. The core of our study resides in studying the effects of agents displaying different in-game behaviours and errors. We purposely decided not to create an agent that intentionally works against the player as our study takes place in an explicit human-agent collaborative scenario (thus no violation errors were included, only mistakes, slips & lapses).

To test the effects of **mistakes**, we designed agent **"Delta"**. This agent, when triggered, becomes incapable of focusing on one target at a time and instead "bounces" from targets to targets while never completely reaching them. This behaviour simulates an error of "planning", most appropriately linked to errors called "mistakes".

To test the effects of **lapses**; we designed agent **"Epsilon"**. This agent, when engaged with, becomes unresponsive and stops working, showing no sign of activity. This behaviour simulates an error of "omission", commonly linked to "lapses". To test the effects of **slips**; we designed agent **"Zeta"**. When triggered, this agent becomes extremely inaccurate, incapable of hitting any target. This behaviour simulates an error of "commission", commonly linked to "slips" (Reason, 1990, Marinaccio et al., 2015).

The baseline agent "**Gamma**" triggered no errors. All agents had the same base level of performance when no error was triggered (approx. 80% accuracy). For error-prone agents (Delta, Epsilon and Zeta), errors were triggered once every two rounds for 30 seconds so that they could be noticeable. To ensure consistency, all agents triggered their errors at the same time for each participant. When configuring error-prone agents, we controlled their average overall performance ensuring there was no significant differences between them (an ANOVA yielded $p = 0.68$ when comparing the performance of all error-prone agent in Easy difficulty and $p = 0.15$ in Hard difficulty levels).

Dependent Variables. During the experiment, rating scales incorporated into the game were used to assess participant's perceptions of the agents. We used the NASA TLX to measure users Cognitive Load after interacting with each agent (Hart & Staveland, 1988). To measure the perception of agents by participants, we used questions from The Checklist for Trust Between People and Automation *(Jian et al., 2000)* pertaining to trust, dependability, reliability, deceptiveness, wariness, confidence and security. Our questions were provided with analogue scales, as it has been shown to be effective for subjective trust *ratings (Wiczorek & Manzey, 2014).*

While participants were taking part in the study, various interactions were recorded using an integrated logging system; these included: the number of missiles on screen and the number of missile hit so far, the number of shots fired, the amount of time participants or agents controlled the cross-hair, the distance travelled by the cross-hair while controlled by the agent or the user, who was currently controlling the cross-hair at any given point in time, and whether the agents were currently displaying a particular type of error or not.

Independent Variable. The scenario difficulty in terms of number of missiles to hit and their speed was fixed for the "easy" and "hard" levels across all sessions (with or without agents). Each error-prone agent triggered their respective errors during one of the "Easy" and "Hard" levels.

Results

As the data was normally distributed, repeated measures ANOVAs were used to perform statistical testing. For pairwise comparisons, t-tests were used with Bonferroni correction applied.

Performance and Reliance. Figure 2 and Table 1 present the summary statistics for each agent, where Precision, Recall and F1 scores were used to assess the performance of the participants and agents during each session. Precision is

represented by the number of missiles destroyed divided by the number of projectiles fired. Recall is represented by the number of missiles destroyed divided by the total number of missiles spawned. F1 is a harmonic mean of the Precision and Recall scores, which we will mainly rely on to compare the performance of each agent-participants sessions.

For F1 scores, an ANOVA yielded $p < 0.0001$ with $F = 3.917$ When performing pairwise comparisons of F1 scores, Delta (mistakes) was found to be significantly lower than Epsilon (lapses) with $p = 0.0355$. When comparing F1 scores obtained in "easy" difficulty to "hard" difficulty levels, all comparisons yielded significantly results with $p < 0.0001$ (in both no agent and agent sessions).

Participant control times were compared, representing the amount of time participants chose to control the crosshair themselves instead of leaving the agents in charge of aiming, effectively correcting the agents' inputs. A higher amount of time representing less reliance on the agents. While comparing participant control time, an ANOVA yielded $p < 0.0005$ with $F = 25.608$. For pairwise comparisons of participant control times, Gamma (no error) was found to be significantly lower than no agent sessions with $p = 0.0002$ and Zeta (slips) with $p = 0.0345$. Delta (mistakes) was found to be significantly higher than Gamma (no error) with $p < 0.0001$, higher than Epsilon (lapses) with $p = 0.0368$ and higher than Zeta (slips) with $p = 0.0345$.

The number of times participants corrected the agents was also compared, a higher number of corrections representing more attempts to manually correct an agent, signifying a less propensity to rely on an agent. An ANOVA yielded $p < 0.0011$ with $F = 45.003$. For pairwise comparisons of the number of participants corrections, Gamma (no error) was found to be significantly lower than Delta (mistakes) with $p < 0.0001$, and Zeta (slips) with $p < 0.0001$.

Rating Scales. At the end of each sessions or levels, participants had to rate the agents based on their perceived trustworthiness, dependability, whether they provided security or if participants perceived the agent as being deceptive. At the end of every sessions, participant had to fill up a Nasa TLX workload questionnaire. Raw NASA TLX scores are used in this study as they were proven to be as effective as their weighted counter-part (Hart, 2006).

Cognitive Load. When comparing RAW NASA TLX scores, an ANOVA yielded $p < 0.001$ with $F = 3.526$. For pairwise comparisons of Raw NASA TLX scores, Epsilon (lapses) was found to be significantly higher than Zeta (slips) with $p = 0.0483$ and Delta (mistakes) was found to be significantly higher than Gamma (no error) with $p = 0.1942$.

Answers to the "Checklist for Trust between People and Automation". When comparing ratings to the questions "I can trust the agent", an ANOVA yielded $p < 0.0001$ with $F = 10.608$. While performing pairwise comparisons for this statement, Gamma (no error) was found to be significantly higher than Delta (mistakes) with $p < 0.0001$, Epsilon (lapses) with $p < 0.0001$ and Zeta (slips) with $p < 0.0001$.

When comparing ratings to the questions "The agent is reliable", an ANOVA yielded $p < 0.0001$ with $F = 11.524$. While performing pairwise comparisons for this statement, Gamma (no error) was found to be significantly higher than Delta (mistakes) with $p < 0.0001$, Epsilon (lapses) with $p < 0.0001$ and Zeta (slips) with $p < 0.0001$.

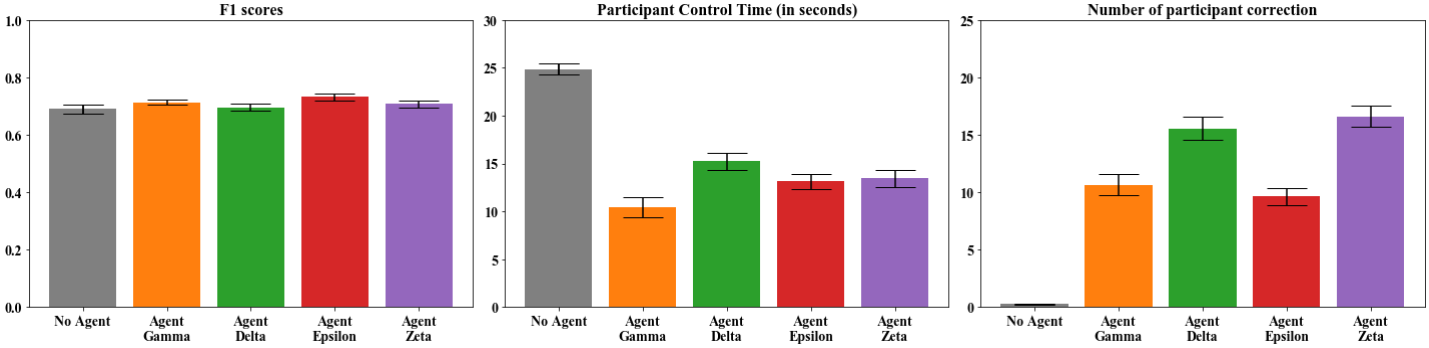

Figure 2: Metrics related to performance and reliance of participant-only (no agent) and participant-agents' sessions.

Table 1: Raw Nasa TLX scores (higher scores = higher reported workload) and ratings of statements presented during participant-agent sessions (higher scores = stronger agreement with the statement), adapted from the Checklist for Trust between Human and Automation questionnaire. * indicates statistical significance. Lowest values for positive statements and highest values for negative statements are indicated in bold.

Question Asked	Gamma (no error)	Delta (mistakes)	Epsilon (lapses)	Zeta (slips)
I can trust the agent	64.15* ± 2.35	44.65 ± 2.72	**44.19 ± 2.69**	46.67 ± 2.47
The agent is dependable	61.29* ± 3.13	**40.27 ± 3.51**	40.35 ± **3.21**	44.96 ± 3.43
The agent is reliable	63.00* ± 3.26	**40.33 ± 3.49**	44.06 ± 3.46	43.79 ± 3.41
The agent is deceptive	44.96* ± 4.58	48.50 ± 5.74	**58.96 ± 5.06**	47.46 ± 4.17
I am wary of the agent	53.42* ± 5.03	59.04 ± 5.58	**61.17 ± 5.54**	60.88 ± 4.90
I am confident in the agent	52.92* ± 4.67	34.62 ± 4.36	**32.96 ± 3.06**	37.54 ± 4.02
The agent provides security	55.71* ± 4.31	40.38 ± 5.30	**36.50 ± 4.35**	42.21 ± 5.16
Raw TLX scores	11.09* ± 0.78	**12.56* ± 0.95**	11.56* ± 0.92	11.74* ± 0.83

When comparing ratings to the questions "the agent is dependable", an ANOVA yielded $p < 0.0001$) with $F = 12.093$. While performing pairwise comparisons for this statement, Gamma (no error) was found to be significantly higher than Delta (mistakes) with $p < 0.0001$, Epsilon (lapses) with $p < 0.0001$ and Zeta (slips) with $p < 0.0001$.

When comparing ratings to the questions "the agent provides security", an ANOVA yielded $p < 0.0001$) with $F = 6.935$. While performing pairwise comparisons for this statement, Gamma (no error) was found to be significantly higher than Delta (mistakes) with $p < 0.0001$, Epsilon (lapses) with $p < 0.001$ and Zeta (slips) with $p < 0.0053$.

When comparing ratings to the questions "I am confident in the agent", an ANOVA yielded $p < 0.0001$) with $F = 9.882$. While performing pairwise comparisons for this statement, Gamma (no error) was found to be significantly higher than Delta (mistakes) with $p < 0.0002$, Epsilon (lapses) with $p < 0.0002$ and Zeta (slips) with $p < 0.0026$.

Participants Feedback. At the end of each sessions, in accordance with the Critical Incident Technique (Flanagan, 1954), people were asked to write about the positive and negative aspects of each agent, and what improvement(s) they would suggest. After being presented with definitions of "lapses", "mistakes" and "slips", three annotators were given the task to code the "positive", "negative" and "improvement" feedback given by each participant about the agents. Internal agreement scores (Kappa scores (Viera & Garrett, 2005)) were used to interpret participants feedback in relation with the behaviours of the agents. The Kappa scores obtained for coding feedback related to "mistakes" was *0.5243* (perceived as "Moderate (Viera & Garrett, 2005)), with an internal agreement of *78.73%*. The agent most frequently associated with mistakes was Delta (the agent effectively displaying errors qualified as "mistake"), with the most common occurrences being "target(s)" (91 occurrences), "prioritize" (24 occurrences), "focus" (23 occurrences), "confused" (21 occurrences) or "struggle" (14 occurrences). The Kappa scores obtained for "lapses" was *0.5493* (perceived as "Moderate"), with an internal agreement of *91.84%*. The agent most frequently associated with lapses is Epsilon (agent effectively displaying "lapses") with the most common occurrences being "sometimes stops" (37 occurrences), "not working" (17 occurrences) or "stopped" (10 occurrences). The Kappa scores obtained for "mistakes" was *0.2109* (perceived as "fair"), with an internal agreement of *77.34%*. The agent most frequently associated with slips was Zeta (agent effectively displaying "slips" in its behaviour), with the most common occurrences being "aim" (23 occurrences), "accurate" (21 occurrences) or "accuracy" (18 occurrences).

Discussion

From the results, it is clear that participants preferred to interact with agent Gamma (no error), as Gamma was perceived significantly more positively than any other agent (see Table 1) while resulting in a lower cognitive load compared to any of the error-prone agents. These results were to be expected. However, while agent Gamma (no error) by itself had the best and most reliable performance out of all the other agents, participants interacting with agent Epsilon (lapses) managed to score higher on average. When looking at Epsilon (lapses), we also noticed that participants, on average, corrected Epsilon (lapses) less than Gamma (no error). However, on average, participants corrected agent Epsilon (lapses) for a significantly longer period of time than agent Gamma (no error). We hypothesize that, in order to match the performance of Gamma (no error), participants had to involve themselves more in the aiming process. In turn, participants managed to get better performance, even outperforming Gamma (no error). However, while doing so, participants reported a significantly higher cognitive load (due to the increased amount of action participants had to carry out) and an overall significantly worse perception of Epsilon (lapses) compared to Gamma (no error) in all of the ratings displayed in Table 1. Agent Epsilon (lapses) was still perceived as being helpful, as its impact on cognitive load was significantly lower than playing without any agent at all. This result could indicate that participants were more resilient to an agent committing lapses, as this type of error seems to be easier to notice. When looking at the participants feedback for each of the agents, we can also notice that the agent most frequently associated with "lapses" is effectively agent Epsilon (lapses), with many references to its tendency to "stop" or to suddenly "stop working". These observations make it apparent that participants clearly perceived that Epsilon was making lapses.

The rest of the error-prone agents, namely agent Delta (mistakes) and agent Zeta (slips), differed in multiple ways. For instance, Delta's F1 score, the lowest among all the agents, is significantly lower than Epsilon's F1 score. This decrease in performance goes with a decrease in reliance, as participant corrected agent Delta (mistakes) for a significantly higher amount of time than any of the other agents; error-prone or not. These findings indicate that an agent making mistakes is harder for participants to correct, compared to an agent having lapses. Interaction with agent Delta resulted in the worst performance and reliance scores which in turn led to higher cognitive loads (the highest among all agents and sessions), with a statistically significant difference when compared to the cognitive load associated with agent Gamma (no error). From the qualitative coding, most of the negative descriptions associated with agent Delta were adequately coded as resulting from "mistakes", with recurring terms such as "confused", "targets" or "prioritize", highlighting how indecisive the aiming of the agents was perceived to be. These observations suggest that participants were well aware of how and when Delta was making mistakes, which led them to rely on it less, and to rate it as less dependable and less reliable than any of the other error-prone agents.

The scores obtained by participants interacting with agent Zeta (slips) seem to place it in the middle of the other error-prone agents in terms of performance (F1 scores) and participant reliance (control time). However, the number of participant corrections for agent Zeta (slips) is the highest among all the agents and was significantly higher than Epsilon's (lapses). Participant control time was also found to

be significantly lower than Delta's (mistakes) and higher than Gamma's (no error).

These results indicate that participants felt the need to correct the agent makings slips more than any other agent, however these corrections were shorter compared to the agent making mistakes (Delta). Overall, the increase in user corrections of Zeta (slips) came with an increase in cognitive load, as Zeta was found to be the most cognitively taxing agent. Nonetheless, the slips displayed by Zeta were still perceived as less cognitively taxing than the mistakes displayed by Delta. When looking at participants feedback, agent Zeta was most often coded as displaying "slips", albeit with the lowest Kappa score (0.21) from all the other agents' errors. Participants still perceived differences in the way agent Zeta behaved, mainly mentioning issues in the "accuracy" displayed by the agent. In order to make up for Zeta's accuracy, participants had to constantly adjust the aiming themselves. However, out of all the error-prone agents, Zeta was rated as the most dependable, providing the most security and being the most trustworthy agent.

Overall, our findings suggest that when designing collaborative agents likely to give imperfect inputs, it is most preferable to avoid indecisiveness, and that, most often, a total lack of input is preferable to indecisive or inaccurate information.

References

Baker, A. L., Phillips, E. K., Ullman, D., & Keebler, J. R. (2018). Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems*, *8*(4), 1–30. https://doi.org/10.1145/3181671

Daronnat, S., Halvey, M., & Azzopardi, L. (2019). Human-Agent Collaborations : Trust in Negotiating Control. *Workshop Proceedings Everyday Automation Experience'19 In Conjunction with CHI'19, May 5th, 2019, Glasgow, UK*. https://matthiasbaldauf.com/automationxp19/papers/AutomationXP19_paper_10.pdf

De Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., & Parasuraman, R. (2012). The world is not enough: Trust in cognitive agents. *Proceedings of the Human Factors and Ergonomics Society*, 263–267. https://doi.org/10.1177/1071181312561062

Fan, X., Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., & Endsley, M. R. (2008). The influence of agent reliability on trust in human-agent collaboration. *Proceedings of the 15th European Conference on Cognitive Ergonomics the Ergonomics of Cool Interaction - ECCE '08*, 1. https://doi.org/10.1145/1473018.1473028

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *51*(4), 327–358. https://doi.org/10.1037/h0061470

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society*, 904–908. https://doi.org/10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). *Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research* (Vol. 43, Issue 5). https://doi.org/10.1016/S0166-4115(08)62386-9

Hoc, J. M., Young, M. S., & Blosseville, J. M. (2009). Cooperation between drivers and automation: Implications for safety. *Theoretical Issues in Ergonomics Science*, *10*(2), 135–160. https://doi.org/10.1080/14639220802368856

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Kim, P. H., Cooper, C. D., Dirks, K. T., & Ferrin, D. L. (2013). Repairing trust with individuals vs. groups. *Organizational Behavior and Human Decision Processes*, *120*(1), 1–14. https://doi.org/10.1016/J.OBHDP.2012.08.004

Marinaccio, K., Kohn, S., Parasuraman, R., & De Visser, E. J. (2015). A Framework for Rebuilding Trust in Social Automation Across Health-Care Domains. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, *4*(1), 201–205. https://doi.org/10.1177/2327857915041036

Mascarenhas, S., Melo, F. S., & Paiva, A. (2018). *Exploring the Impact of Fault Justification in Human-Robot Trust*. 507–513.

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors*, *57*(1), 34–47. https://doi.org/10.1177/0018720814561675

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation - Syst ems, Man and Cybernetics, Part A, IEEE Transactions on. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *30*(3), 1–12. https://doi.org/10.1109/3468.844354

Reason, J. (1990). Human error. *Cambridge University Press*. https://doi.org/10.7748/ns.11.49.20.s36

Viera, A. J., & Garrett, J. M. (2005). Anthony J. Viera, MD; Joanne M. Garrett, PhD (2005). Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37(5):360-63. *Family Medicine*, *37*(5), 360–363. http://www1.cs.columbia.edu/~julia/courses/CS6998/Interrater_agreement.Kappa_statistic.pdf

Wiczorek, R., & Manzey, D. (2014). Supporting attention allocation in multitask environments: Effects of likelihood alarm systems on trust, behavior, and performance. *Human Factors*, *56*(7), 1209–1221. https://doi.org/10.1177/0018720814528534