"Shoulda, Coulda, Woulda": Conceptualizing the Differences in Trust Between Human-Human Teaming and Human-Machine Teaming

Brandon D. Dreslin and Jessica A. Baweja

National Security Directorate, Pacific Northwest National Laboratory

Author Note

Brandon D. Dreslin   https://orcid.org/0000-0002-0442-9504

Jessica A. Baweja   https://orcid.org/0000-0001-8466-8611

Correspondence concerning this article should be addressed to Brandon D. Dreslin, National Security Directorate, Pacific Northwest National Laboratory. Email: brandon.dreslin@pnnl.gov

Abstract

Intelligent decision support systems (IDSSs) are machine teammates designed to facilitate better human decision-making in high-consequence domains such as health care, power grid operations, and fraud detection. IDSSs identify patterns in datasets and provide intelligent decision-making recommendations to human teammates. However, previous research indicates that humans often trust IDSS recommendations less than the recommendations from their human teammates, even when the machine teammate is more accurate. To conceptualize why trust differs, we review the literature surrounding trust, error, and predictability. Then, we compile and compare participant trust ratings and decision-making in an abridged systematic review of previous studies manipulating teammate type, error rate, and error type. Finally, we conduct a content analysis of participants' qualitative responses to trust queries from a survey on generative language models. Results suggest that humans may trust IDSS teammates less than other human teammates because of differences in (1) interaction complexity, (2) blame attribution, and (3) swift trust. We conclude that human factors practitioners should collaborate with data scientists and domain experts to build and maintain trust in IDSSs by anthropomorphizing algorithms, matching mental models, and considering individual differences.

Keywords: Human-human teaming, human-machine teaming, intelligent decision support systems, teamwork, trust

"Shoulda, Coulda, Woulda": Conceptualizing the Differences in Trust Between Human-Human Teaming and Human-Machine Teaming

Suppose that a runaway train is on a collision course with a group of five innocent people whose shoelaces are stuck in the railway down the track. As the section operator, you can divert the train's path to a second track where one innocent person's shoelaces are also stuck in the railway down the track. Because only two tracks are available, death is inevitable. No one is aware of the impending danger. Suddenly, your computer's display warns that there is human activity on both tracks. Through a sophisticated machine learning algorithm that intelligently senses activity markers consistent with human presence via advanced railway sensors, your computer analyzes the activity data and recommends that you divert the train to the second track to minimize the loss of life. You must decide everyone's fate based on the computer's advice. Should you trust it? Could you trust it? Would you trust it?

This modern twist on a classic thought experiment demonstrates current real-world issues concerning artificial intelligence (AI) and machine learning (ML) algorithms. Intelligent decision support systems (IDSSs) are machine teammates that use AI and ML techniques (e.g., fuzzy logic, decision trees, and neural networks) to facilitate better human decision-making by analyzing data and providing decision recommendations (Phillips-Wren, 2013). These systems are especially prevalent in high-consequence domains like health care, power grid operations, and fraud detection. For example, optometrists can consult a retinal disease-screening IDSS to view the machine's recommended prognosis before providing patients with a final diagnosis (Bourouis et al., 2014). However, prior research (e.g., Wærn & Ramberg, 1996; Lee & Moray, 1992; Dzindolet et al., 2001; Wiegmann et al., 2001) indicates that when working together in a team, humans often trust machine teammates less than other human teammates, even when the

machine is more accurate. This study aims to contribute a human-centered computing perspective to teaming literature by deriving a conceptual model for the reasons why humans might trust machines less than humans after errors are made.

Methods

This study employs three methods of data collection: a literature review, an abridged systematic review, and a content analysis. The mixed-methods approach ensures that the conceptual model is founded with substantiating evidence. Data collection was completed during a period of six weeks between May 2023 and July 2023.

Literature Review

Literature on trust, errors, and reliability in teaming dynamics was reviewed to build the knowledge necessary to interpret results from the other data collection methods. Multiple databases were searched to find articles across various disciplines. Social psychology literature yielded an understanding of the types, constructs, antecedents, and consequences of trust. IDSS literature revealed interactions between machine error, model reliability, and their effects on trust. Human factors and teaming literature linked these concepts together to uncover the differences in and importance of appropriate trust for proper decision-making between two systems of teamwork. It was revealed that lower trust in machines is a topic not yet fully understood.

Abridged Systematic Review

The knowledge gap identified above led to the formation of the following research question: What factors drive lower trust in machine teammates despite higher accuracy compared

to human teammates? An abridged systematic review[1] of quantitative findings from previous studies was conducted to answer this question.

Search Strategy

Interdisciplinary literature was searched for empirical studies that investigated the effects of reliability manipulation on trust in human-human teaming and human-machine teaming (specifically IDSSs and ML) across various domains. Using the Google Scholar, IEEE *Xplore*, ACM Digital Library, and DTIC databases, four combinations of key words and phrases were searched for: (1) "trust AND team* OR human-human team*;" (2) "trust AND human-machine team*;" (3) "trust AND machine learning OR intelligent decision support systems;" and (4) "trust AND accuracy OR reliability OR predictability AND machine learning OR intelligent decision support systems."

Inclusion/Exclusion Criteria

Studies found through the search results must: (1) be written in English; (2) be published between 2010 and present; (3) investigate the effects of teammate error rate on trust; (4) only use AI, ML, or IDSSs with embedded AI/ML models as testbeds for human-machine teaming (HMT) experiments; (5) manipulate the teammate's decision-making recommendation, accuracy, error type, or any combination of these variables; and (6) measure any construct of human trust, reliability, or performance identified in the literature review. All studies followed a cascading criterion check (i.e., once an article failed to meet one criterion, it was immediately discarded). A total of three studies that met all inclusion/exclusion criteria were found.

---

[1] Since very few studies have quantitatively investigated this topic, a meta-analysis was inappropriate for this study.

Figure 1

Cascading Criterion Check Flowchart
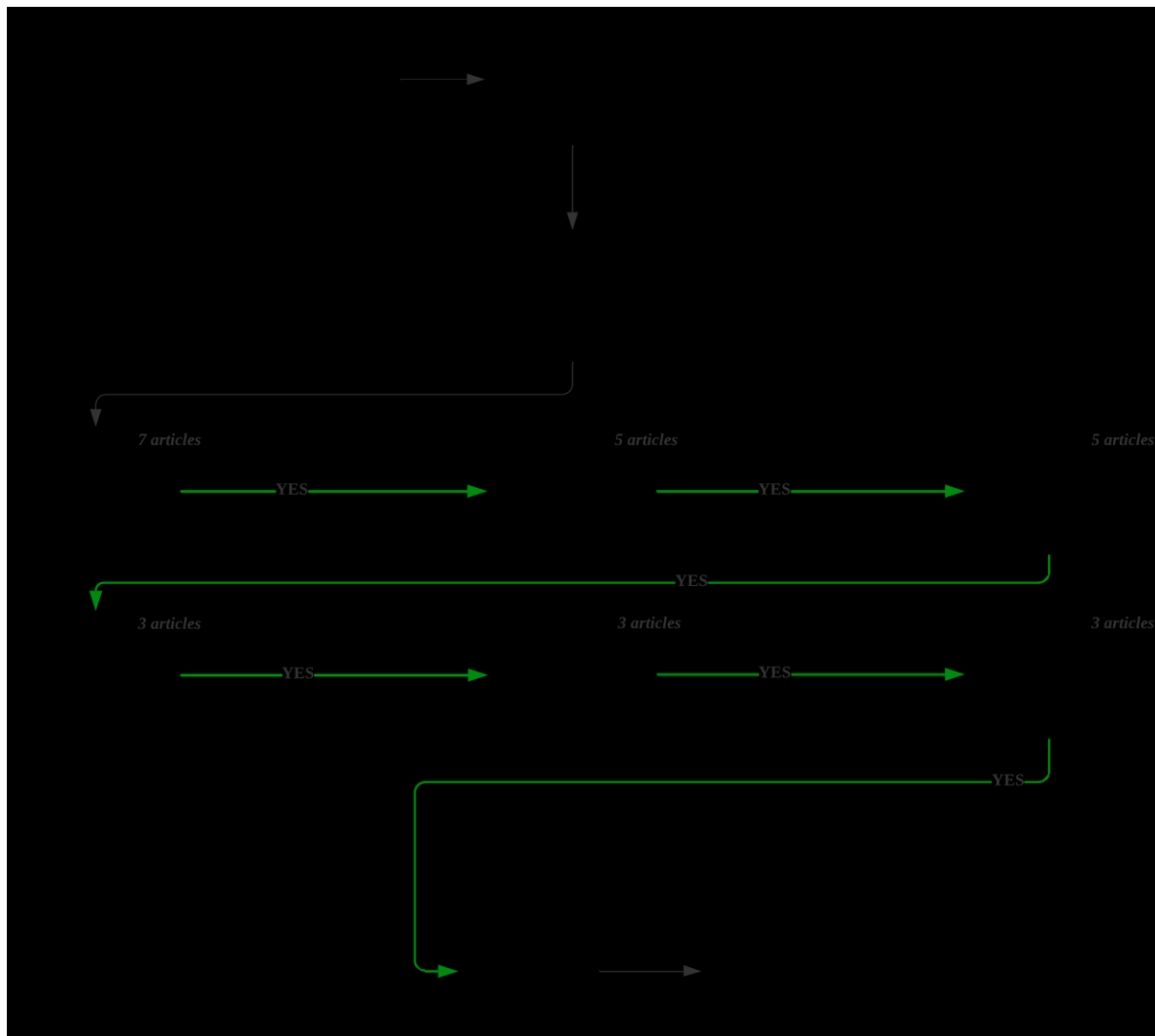
Content Analysis

Qualitative survey data from a between-subjects study[2] investigating mitigative strategies for recovering from generative language model (GLM) trust violations was used to perform a content analysis of trust in this study. Participants were randomly assigned to one of three experimental conditions (i.e., mitigative strategies): confidence scores, transparency, and

---

[2] The study was approved in June 2023 by the Pacific Northwest National Laboratory Institutional Review Board under IRB 2023-21.

feedback and control. For all conditions, participants were asked to rate the reliability of the model's answers and to rate their trust in the model based on their experiences with 20 GLM query/answer scenarios. Participants assigned to the feedback and control condition were specifically prompted to (a) indicate whether the model's answer was satisfactory through clickable thumbs up or thumbs down buttons and (b) provide responses to the following question, if desired: "If you want to provide any feedback on the quality of the answer, please do so here. Your responses will be used to improve the model in the future." Once data collection was completed, each qualitative response from the feedback and control condition was organized in a Microsoft Excel file according to question number. Common themes and sentiments were extracted, labeled, and applied across questions and participants below the raw data. Because this study reports on data from one condition, the remaining conditions' data is purposefully ignored.

Results

Literature Review

Teammates

Teamwork has traditionally been described as groups of two or more people collaborating (i.e., being teammates) on certain tasks to achieve a shared goal and is often termed human-human teaming (HHT) (Salas et al., 2003). The introduction of computers allowed machines to act as teammates (Seeber et al., 2020). This created HMT systems wherein the human directly collaborates with a computer instead of another human to reach the objective (Walliser et al., 2019). For example, humans who interact with IDSSs participate in an HMT system (Henry et al., 2022). While extensive research on HHT has been conducted, more research must be done on HMT to fully comprehend the effects of replacing humans with technology.

Trust

Effective teamwork requires trust between teammates. Seminal work by Lee and See (2004) defines trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (p. 51). The agent (i.e., teammate) can be either another human or a machine. Hoff and Bashir (2015) identify three main types of trust: dispositional trust (where there is a general predisposition to place trust in teammates irrespective of teaming dynamic and context), situational trust (where trust is context- and problem-dependent) and learned trust (where trust is based on both past experiences and current interactions with teammates). Swift trust is a unique fourth type of trust that develops quickly between teammates with little or no prior experience to the teaming dynamic (Patel et al., 2022). Each type of trust can be measured objectively or subjectively (Law & Scheutz, 2020). Objective measures may include quantitative performance metrics of the human, machine, and/or team (Cho et al., 2015). Subjective measures include qualitative responses (e.g., thoughts and feelings) on questions regarding trust (Schwartz et al., 2022).

Appropriate (or calibrated) trust refers to the alignment between an individual's perceived trust level and the teammate's actual performance (Yang et al., 2020). There are consequences when appropriate trust is abandoned. Overtrust occurs when humans violate assumptions and blindly rely on teammate performance (Aroyo et al., 2021). Undertrust occurs when people dismiss the capabilities of their teammate and rely heavily on themselves (Hoff & Bashir, 2015). Four characteristics—referred to as the ABI+ framework—comprise the antecedents of trust. To foster appropriate trust between trustor and trustee, teammates must possess: ability (the skills necessary to influence task outcomes), benevolence (the intention to want to do good to the trustor), integrity (the adherence to a set of pre-determined principles),

and predictability (the potential to reproduce actions or outcomes) (Toreini et al., 2020).

Appropriate trust is only possible once all four antecedents have been satisfied.

Errors

Human errors are classified under skill-based mistakes, rule-based mistakes, or

knowledge-based mistakes (Stone et al., 2017). Conversely, machine errors are categorized by

misses and false alarms (McBride et al., 2014). Team errors occur when humans and/or

machines make mistakes either individually or together. A taxonomy of team errors presented by

Sasou and Reason (1999) illustrates that errors are not only easier to commit, but are also more

difficult to recover from in teaming dynamics because of interaction complexity. Indeed, any

error can affect trust. A negative correlation appears to exist between the number of errors a

human teammate makes and the trust another human places in that teammate (Erdem et al.,

2004). Additionally, Johnson et al. (2004) report that individuals' overall trust in a machine

teammate is reduced more by false alarms than by misses. People react differently to human,

machine, and team errors. Nonetheless, trust decreases as the number of mistakes increases.

Reliability

One definition of reliability refers to the extent to which a teammate can predictably

produce the same outcomes over time (Miller & Swain, 1987). As the number of teammates

increases, total system and/or perceived reliability will likely decrease due to the added

interaction complexity (Stone et al., 2017). This aligns with previous research mentioned earlier

(e.g., Sasou & Reason, 1999). In fact, individuals' perceived reliability of (and consequently,

trust in) a machine teammate is often less than the actual system reliability (Johnson et al., 2004).

Washburn et al. (2020) state that because trust is one of the most influential factors on team

performance, teammate reliability must be high to avoid total task failure. A second definition of

reliability refers to the predictability of a human operator to obey a teammate's commands (Lyons & Stokes, 2012). Although the first type of reliability is most common in the literature, the second type may be more important for understanding where a person's trust lies.

Abridged Systematic Review

Dietvorst et al. (2015) recruited over 2,000 participants across five different IDSS experiments to evaluate human trust between human and machine teammates with varying performance levels. Participants were tasked with analyzing trends in graduate student performance (GSP) and airline passenger number rank (APNR) according to U.S. state. After listening to advice from a human teammate or an ML IDSS teammate, participants were asked to predict outcomes for both scenarios. Results indicate that compared to the IDSS, the human teammate produced approximately 15% to 29% more error in the GSP scenarios and 90% to 97% more error in the APNR scenarios. Interestingly, participants chose to follow the human teammate's decision recommendation, even when shown that the IDSS outperforms the human. The authors conclude that algorithm aversion (i.e., the tendency to distrust algorithms after witnessing it commit errors) made participants less confident in the IDSS such that they placed more trust in the human teammate despite the poor reliability.

Yu et al. (2019) asked 30 participants to assume the role of a quality control specialist at a drinking glass manufacturing site. With the support of an automatic quality monitor (AQM) IDSS, they were tasked with deciding whether each glass was good or faulty. The researchers manipulated the IDSS's accuracy in 10% increments between 30% and 90%. It was found that participants' trust was initially lower than IDSS accuracy, but gradually approximated system accuracy after multiple interactions. In alignment with prior work, trust in the IDSS increased up to around 70% accuracy but decreased for lower accuracies. In contradiction with previous

studies, participants eventually inappropriately trusted the IDSS recommendations. The authors attribute this to an oversimplicity in the AQM's logic and recommend that future research experiment with higher-complexity IDSSs.

Suresh et al. (2020) recruited 175 participants to compare similar images in two scenarios with and without IDSS recommendations. The first scenario asked them to determine which image of random crowds contained more people. The second scenario asked them to select which image looked most like the animal they were assigned to identify. For both scenarios, the IDSS was between 72% and 93% accurate. The researchers found that participants generally accepted correct recommendations and rejected incorrect recommendations from the IDSS. However, they were willing to blindly accept incorrect (but convincingly accurate) advice, particularly for images that contained too much detail and even when the algorithm demonstrated low accuracy. According to the authors, the implicit yet inappropriate trust evinced by many participants may explain why allowing the IDSS to generate more information for greater transparency failed to affect trust.

Content Analysis

Sixteen Pacific Northwest National Laboratory employees agreed to be participants in the study and were assigned to the feedback and control condition. Eight people provided 46 qualitative responses on the quality of the model's answers. Although the question did not specifically inquire on trust, three themes emerged from many responses that relate to why trust in the GLM was affected.

Accuracy

Unbeknownst to the participants, one of the model's answers was purposefully altered to be incorrect. All participants correctly identified this inaccuracy, explained why it was

"completely wrong," and swiftly corrected the answer. Many participants also pinpointed inconsistencies and inaccuracies in the model's other answers that were not purposefully altered. For example, two participants expressed that some answers contradicted themselves. Another participant believed an answer was wrong on a technicality because "California is not part of the southwest but considered the west." The varying levels of GLM accuracy affected participants' trust. One response discloses that because so many things were "wrong or creepy previously, [they] can't trust this model." The respondent continues, revealing their concern for other people's trust in the model: "What if someone got just one of these 'near correct' or 'true but intrusive' answers? How would they know to be suspicious?!" Other participants echo this sentiment. One participant disagreed and appreciated the explanatory mitigative strategy in the feedback and control condition. However, this participant did not explain how or why explanations affected their trust.

Superfluous Information

All participants agreed that many of the GLM's answers contained superfluous information. While the model answered every query, it often provided information unnecessary or unrelated to the direct prompt. For three scenarios, one participant stated that the model is too "chatty" and provides too much "unrequested information." Other participants claimed that the model was "verbose," "sloppy," and could benefit from the exclusion of information that was not explicitly asked for. One participant implied that information overload negatively affected their trust, stating that "the clunky language… ma[de] me suspicious."

Mismatched Mental Models

Perhaps the most salient factor that negatively affected trust was mismatched mental models in the form of knowledge gaps. Many participants were hesitant to trust the GLM during

six scenarios because their knowledge of the answer did not match the model's output. When explaining their doubts, participants appeared to lack a frame of reference with which to "compare [their] own knowledge." This theme was especially prevalent when answers were not common knowledge. For example, a participant divulged their distrust when provided the model's answer to a query asking for scientific facts about Jupiter: "I don't know the moons of Jupiter offhand, nor the planet's chemical composition. I'm reluctant to trust that the model got those details correct." To compare their knowledge with the model's answer for a scenario, one participant Googled the GLM query in search for independent sources that provided evidence for the correct answer. Intriguingly, some participants sought post facto comparison even when the model's answer "sound[ed] convincing." Because most participants were clearly skeptical of some answers, it can be deduced that trust in the model was lost.

Discussion

Based on the data presented in this study, it is evident that trust formation is different between human teammates and machine teammates, especially when errors are made. We have identified three factors that drive low trust in HMT to explain why these disparities exist.

Figure 2

Concept Map of Driving Factors
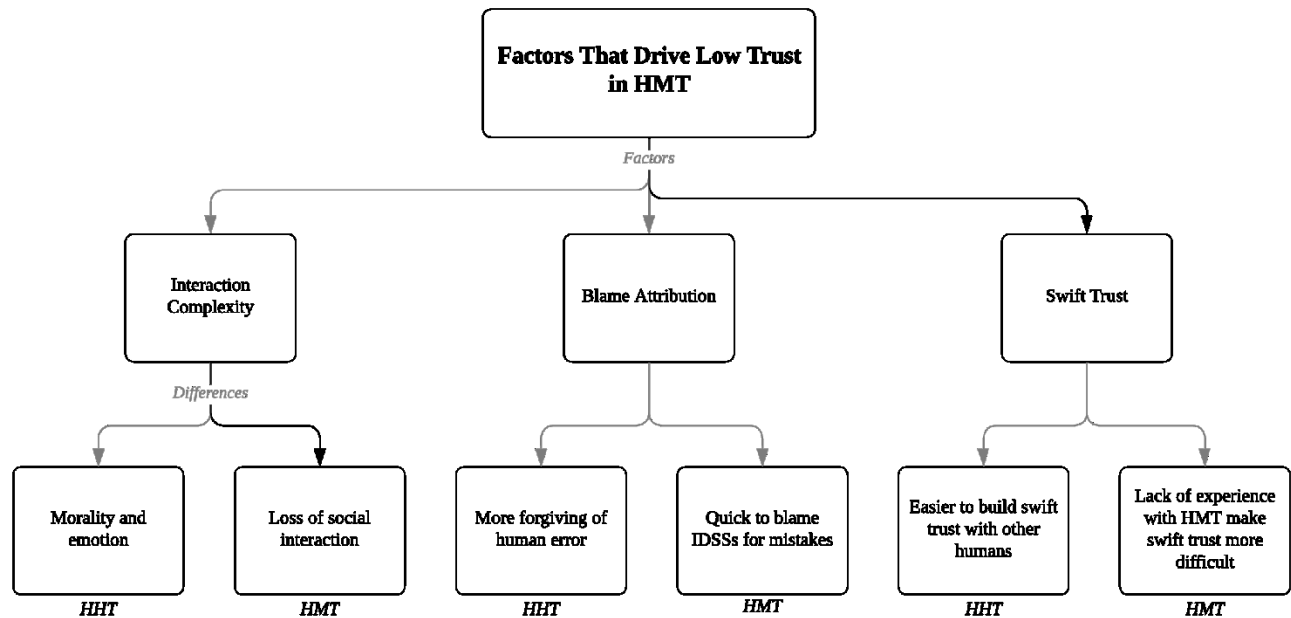
1. Interaction complexity: Teamwork is naturally a very social activity. Trust in HHT dynamics relies on the ABI+ framework. Benevolence and integrity exist as characteristics because humans are emotional creatures who can exercise morality. Contrarily, computers do not possess benevolence or integrity because they are emotionally detached objects. Therefore, ABI+ morphs into the A&P (ability and predictability) framework for HMT where emotion and social interaction is lost (Hoff & Bashir, 2015). The amorality of algorithms also means that computers are merely objective driven and hold no stake in, or care for, the outcomes of their performance. Madhavan and Wiegmann (2007) explain that the loss of social interaction and emotionally detached performance may explain why humans place more trust in their human counterparts than more accurate computers.

2. Blame attribution: Participants from previous studies and the GLM study were quick to blame IDSSs for mistakes. Two reasons may explain this phenomenon: (1) they expect the computer to always be right and/or (2) they may feel less guilty and responsible for negative outcomes by deflecting blame into machines (Tobia et al., 2021; Maasland & Weißmüller, 2022). Meanwhile, people know that their peers err and generally tend to be more forgiving of their human teammates, especially when they show genuine interest in and intention to improve their performance (Dietvorst et al., 2015). Because machines are unable to perceive feelings, some people may feel more comfortable placing blame and losing trust in computers, regardless of reliability.

3. Swift trust: Most participants from the abridged systematic review articles were novice users of AI, ML, and IDSSs. Their lack of experience with HMT dynamics may have hindered individual ability to build swift trust in the machines, which may explain implicit yet inappropriate trust (Yu et al., 2019; Suresh et al., 2020). In contrast, people in HHT dynamics have many experiences with human teammates such that it may be easier to build swift trust – even with different people (Madhavan & Wiegmann, 2007; Dietvorst et al., 2015).

Each factor affects the ability to build appropriate trust in a teammate, especially when errors occur. When trust is low, team performance is severely curtailed. We recommend that human factors practitioners collaborate with data scientists and domain experts to strengthen trust in machines by: (a) anthropomorphizing algorithms (i.e., engineering the ABI+ framework into machines by replicating human sentience), (b) matching people's mental models with a machine's output, and/or (c) designing IDSSs around individual propensities for trust.

Conclusion

Team trust is a complex concept. Teamwork implies trust, yet trust in teammates is often lost. It was previously unknown what factors influence lower trust in machine teammates despite higher accuracy than human teammates. To identify these driving factors, we conducted a literature review of previous research, an abridged systematic review of quantitative data from three empirical studies, and a content analysis of 46 qualitative responses from a GLM trust violation study. We found that three fundamental differences affect how trust is built between human teammates and machine teammates: interaction complexity, blame attribution, and swift trust. We justified the discrepancies using results from our data collection process. Trust in machines must be improved, otherwise poor team performance may cause harm to teammates and others affected by a team's decisions. If our recommendations are effectuated, society will begin to see that appropriate team trust—in humans and machines—is within reach.

References

Aroyo, A., de Bruyne, J., Dheu, O., Fosch-Villaronga, E., Gudkov, A., Hoch, H., Jones, S., Lutz, C., Sætra, H., Solberg, M. & Tamò-Larrieux, A. (2021). Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. *Paladyn, Journal of Behavioral Robotics, 12*(1), 423-436. https://doi.org/10.1515/pjbr-2021-0029

Bourouis, A., Feham, M., Hossain, M. A., & Zhang, L. (2014). An intelligent mobile based decision support system for retinal disease diagnosis. *Decision Support Systems, 59*, 341-350. https://doi.org/10.1016/j.dss.2014.01.005

Cho, J.-H., Chan, K., & Adali, S. (2015). A survey on trust modeling. *ACM Computing Surveys, 48*(2), 1-40. https://doi.org/10.1145/2815595

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126. https://doi.org/10.1037/xge0000033

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology, 13*(3), 147-164. https://doi.org/10.1207/S15327876MP1303_2

Erdem, F., Ozen, J., & Atsan, N. (2003). The relationship between trust and team performance. *Work Study, 52*(7), 337-340.

Henry, K. E., Kornfield, R., Sridharan, A., Linton, R. C., Groh, C., Wang, T., Wu, A., Mutlu, B., & Saria, S. (2022). Human-machine teaming is key to AI adoption: Clinicians' experiences with a deployed machine learning system. *NPJ Digital Medicine, 5*, Article 97. https://doi.org/10.1038/s41746-022-00597-7

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors

    that influence trust. *Human Factors, 57*(3), 407-434.

    https://doi.org/10.1177/0018720814547570

Johnson, J. D., Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004). Type of automation failure:

    The effects on trust and reliance in automation. *Proceedings of the Human Factors and

    Ergonomics Society Annual Meeting, 48*(18), 2163-2167.

    https://doi.org/10.1177/154193120404801807

Law, T., & Scheutz, M. (2021). Trust: Recent concepts and evaluations in human-robot

    interaction. In C. S. Nam and J. B. Lyons (Eds.), *Trust in human-robot interaction*, (pp.

    27-57). Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00002-2

Lee, J. D, & Moray, N. (1992). Trust, control strategies and allocation of function in human-

    machine systems. *Ergonomics, 35*(10), 1243-1270.

    https://doi.org/10.1080/00140139208967392

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human

    Factors, 46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Lyons, J. B., & Stokes, C. K. (2012). Human-human reliance in the context of automation.

    Human Factors, 54(1), 112-121. https://doi.org/10.1177/0018720811427034

Maasland, C., & Weißmüller, K. S. (2022). Blame the machine? Insights from an experiment on

    algorithm aversion and blame avoidance in computer-aided human resource management.

    Frontiers in Psychology, 13, Article 779028. https://doi.org/10.3389/fpsyg.2022.779028

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human

    and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics

    Science, 8*(4), 277-301. https://doi.org/10.1080/14639220500337708

McBride, S. E., Rogers, W. A., & Fisk, A. D. (2014). Understanding human management of automation errors. *Theoretical Issues in Ergonomics Science, 15*(6), 545-577. https://doi.org/10.1080/1463922X.2013.817625

Miller, D. P., & Swain, A. D. (1987). Human error and human reliability. In G. Salvendy (Ed.), *Handbook of human factors* (pp. 219-250). John Wiley & Sons, Ltd.

Patel, S. M., Napoli, S. E., Rohrbacher, A. S., Lazzara, E. H., & Phillips, E. (2022). Advancing human-robot teams: A framework for understanding swift trust. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 66*(1), 1159-1163. https://doi.org/10.1177/1071181322661365

Phillips-Wren, G. (2013). Intelligent decision support systems. In M. Doumpos and E. Grigoroudis (Eds.), *Multicriteria decision aid and artificial intelligence* (pp. 25-44). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118522516.ch2

Salas, E., Burke, C. S., & Cannon-Bowers, J. A. (2000). Teamwork: Emerging principles. *International Journal of Management Reviews, 2*(4), 339-356.

Sasou, K., & Reason, J. (1999). Team errors: Definition and taxonomy. *Reliability Engineering & System Safety, 65*(1), 1–9. https://doi.org/10.1016/S0951-8320(98)00074-X

Schwartz, J. M., George, M., Rossetti, S. C., Dykes, P. C., Minshall, S. R., Lucas, E., & Cato, K. D. (2022). Factors influencing clinician trust in predictive clinical decision support systems for in-hospital deterioration: Qualitative descriptive study. *JMIR Human Factors, 9*(2), e33960. https://doi.org/10.2196/33960

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as

teammates: A research agenda on AI in team collaboration. *Information & Management, 57*(2), 1-22. https://doi.org/10.1016/j.im.2019.103174

Stone, N. J., Chaparro, A., Keebler, J. R., Chaparro, B. S., & McConnell, D. S. (2017). *Introduction to human factors: Applying psychology to design*. CRC Press. https://doi.org/10.1201/9781315153704

Suresh, H., Lao, N., & Liccardi, I. (2020). Misplaced trust: Measuring the interference of machine learning in human decision-making. *WebSci '20: Proceedings of the 12th ACM Conference on Web Science*, 315-324. https://doi.org/10.1145/3394231.3397922

Tobia, K., Nielsen, A., & Stremitzer, A. (2021). When does physician use of AI increase liability? *Journal of Nuclear Medicine, 62*(1), 17-21. https://doi.org/10.2967/jnumed.120.256032

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C., & Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. In *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 272-283). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372834

Wærn, Y., & Ramberg, R. (1996). People's perception of human and computer advice. *Computers in Human Behavior, 12*(1), 17-27. https://doi.org/10.1016/0747-5632(95)00016-X

Walliser, J. C., de Visser, E. J., Wiese, E., & Shaw, T. H. (2019). Team structure and team building improve human-machine teaming with autonomous agents. *Journal of Cognitive Engineering and Decision Making, 13*(4), 258-278. https://doi.org/10.1177/1555343419867563

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science, 2*(4), 352-367. https://doi.org/10.1080/14639220110110306

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? *IUI '20: Proceedings of the 25th International Conference on Intelligent User Interfaces*, 189-201. https://doi.org/10.1145/3377325.3377480

Yu, K., Berkovsky, S., Taib, R., Zhou, J., & Chen, F. (2019). Do I trust my machine teammate? An investigation from perception to decision. *IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces*, 460-468. https://doi.org/10.1145/3301275.3302277