

Trust is Not a Virtue: Why We Should Not Trust Trust

By Matthew L. Bolton 

FEATURE AT A GLANCE:

There is currently significant research and industry interest in engineering machines and algorithms that humans will trust. This is justified as a means for facilitating the adoption of developing technology. However, there are many problems with trust that directly relate to its epistemological validity, usefulness, ethical implications, and potential for human disempowerment. This article explores trust from this perspective in the hopes of encouraging the human factors engineering community to de-emphasize trust as an end goal and replace it with more objective measures and good human factors engineering practices.

KEYWORDS:

trust, human-machine interaction, human-automation interaction, psychometrics, validity, artificial intelligence, machine learning

At the time of writing, there is intense interest in human-machine trust, especially for systems where critical decisions and behaviors could be assumed by highly-automated, artificial intelligence (AI), machine learning (ML), or autonomous systems. There appear to be two primary reasons for this. The first is the human factors perspective (Bainbridge, 1983; Lee & See, 2004; Parasuraman & Riley, 1997): that over and under trust can cause people to, respectively, not use automation when it would be beneficial or use it when it is inappropriate. Thus, the human factors perspective advocates for calibrated trust: trust that encourages people to use a system only when it is appropriate. The other motivation appears to come from a more administrative and technology advancement perspective. From this viewpoint, human trust is seen as a major barrier to the acceptance and adoption of new technologies (see evidence of this mindset in Table 1). As a result of both perspectives, significant effort is attempting to determine what trust is, how it can be measured, how it can be modeled, and how it can be predicted (both individually and in aggregate) for different human-machine contexts (see Meyer & Lee, 2013 for a historical perspective).

Trust is a psychological concept. Thus, any measurement of trust puts it into the realm of psychometrics. This inherently makes trust measurement and modeling controversial. There is an ongoing debate about the validity of psychometric measures: whether they are real or merely artificial/folk constructs (Annett, 2002; Dekker & Hollnagel, 2004; Dekker & Nyce, 2015); whether they converge with objective measures (Dekker & Nyce, 2015; Matthews et al., 2020); and whether they satisfy the requirements for cardinal levels

of measurement (Bolton et al. n.d.-a, n.d.-b; Wei et al., 2019, 2020).

Despite this controversy, subjective measures of trust appear to be reliable and (at least across a population, not for all individuals) produce interval-level data (Wei et al., 2019, 2020). This means that as an engineering measure, trust can be used with some consistency (enough to allow for modeling; e.g., Lee & Moray, 1992) and (in most situations) produce numbers compatible with parametric statistics (Wei et al., 2020). While the “folkiness” of the models does not invalidate trust (Parasuraman et al., 2008), there are deeper issues with it that belie not only its usefulness, but its ethic. In this paper, I examine trust from several perspectives to, hopefully, provide a convincing case that trust is not to be trusted.

TRUST IS NOT USEFUL

Trust is Difficult to Define and Highly Contextual

Any discussion around trust usually produces debates about what trust is and how to define it. To this end, trust can be defined as anything ranging from a belief, to an expectation, to an attitude, to an intention, to a behavior, and to a personality trait (see Cho et al., 2015; James Jr., 2002; Lee & See, 2004; Rousseau et al., 1998). Further complicating the matter is that the dominant definitions vary based on who or what is being trusted. This means there are separate definitions for human-machine trust, interpersonal trust, institutional trust, social trust, economic trust, epistemic trust, and organizational trust (there are likely more). It is fatuous to say that these concepts are distinct from one another. For example, human-machine trust will inherently be impacted by trust the

Table 1. Examples of Funding Agencies, Government Organizations, and Private Interests Discussing the Importance of Trust in Technology Adoption.

Motivation for research and standards: "Increasing trust in AI technologies is a key element in accelerating their adoption for economic growth and future innovations that can benefit society." (National Science Foundation, National Institute of Food and Agriculture, Department of Homeland Security, Science & Technology Directorate, U.S. Department of Transportation, Federal Highway Administration, & U.S. Department of Veterans Affairs 2020; National Institute of Standards and Technology, 2019, p. 8)
Concerning examples of fundable projects: "Some examples of AI-driven solutions are provided below...Decision tools for markets and value chains that explicitly build trust through producer and consumer-oriented methods such as modeling of cooperative market-making and peer-to-peer input and feedback." (National Science Foundation, Department of Homeland Security, Science & Technology Directorate, et al., 2020)
Concerning a new artificial intelligence initiative for the Department of Defense (DOD): "DOD's operators must come to trust the outputs of AI systems" (Cronk, June 22, 2021)
The former CEO of Google discussing the use of machine learning in the DOD: "DoD does not have an innovation problem; it has an innovation adoption problem." (Schmidt, April 2018, p.1)
Results of the National Telecommunications and information Administration's July 2015 current population Survey's computer and Internet use Supplement concludes that "lack of trust in internet privacy and security may deter economic and other online activities" (Goldberg, May 13, 2016)
On establishing guidelines for AI in aviation safety: "trust is considered to be essential and critical to the general acceptability of the AI-based systems." (European Union Aviation Safety Agency, 2021, p. 50)

person has for the institutions that built the machines and potential economic impacts associated with use (and vice-versa). Thus, you cannot say (for example) that institutional trust is an intention (Rousseau et al., 1998) while human-machine trust is an attitude (Lee & See, 2004) and that any associated economic trust is an expectation (James Jr., 2002).

This surplus of definitions, contexts, and inter-dependencies fails to add clarity. The effect of this can be seen in the large number of recent review articles that attempt to make sense of trust (Braga et al., 2018; Cho et al., 2015; Gebru et al., 2022; Hancock et al., 2011; Hoff & Bashir, 2015; Israelsen & Ahmed, 2019; Meyer & Lee, 2013; Schaefer et al., 2016; Shahrdar et al., 2018; Zhou et al., 2022; this is not an exhaustive list). Reducing the study of trust to one definition may allow for some consistency between studies and communities. However, doing so is a quintessential example of what Dekker and Nyce (2015) call "ontological alchemy:" defining something based on how it is measured or modeled instead of what it actually is (which is not real, at least materially).

How real or well-defined trust is will be less of an issue if its measures and models are useful. As the following sections show, trust's usefulness is highly questionable.

Trust is Not Selective or Diagnostic

If a measure or concept is to be useful in engineering, it should (among other things) be selective and diagnostic (Eignor, 2013).

Selectivity refers to the ability of a measure to be sensitive to the quality being measured, but not other qualities (Eignor, 2013). As the abundance of definitions for trust suggests, it is not clear what a measure of trust is capturing. Trust is also

highly related to other similar concepts; particularly perceived risk (human interpretation of the probability and consequence of actions), confidence (the assessed capabilities and competencies of the entity being engaged with), and cooperation (how well something works with you to accomplish tasks) (Chancey, et al., 2017; T. Earle & Siegrist, 2008; T. C. Earle, 2010; Hoff & Bashir, 2015; Lyons & Stokes, 2012; Siegrist, 2021). The literature is filled with contradictory models, interpretations, and definitions that attempt to determine whether these concepts are distinct and if/how they influence each other. For example, in human factors, perceived risk is regarded as being upstream, and important to, trust formation (e.g., Chancey et al., 2017; Hoff & Bashir, 2015; Lee & See, 2004; Lyons & Stokes, 2012; Mayer et al., 1995). Conversely, the financial and risk management communities (who have also done extensive analyses on the subject) regard trust as being a predictor of (input to) perceived risk (see reviews in Earle, 2010; Siegrist, 2021) or a wholly independent phenomenon (Delbufalo, 2015). This situation inherently means trust measurement lacks selectivity: if the subtleties between these concepts are difficult for researchers to pull apart, they will be confounded when measuring or modeling normal people who will bring their own definitions to bear.

Diagnosticity relates to the ability of a measure to explain changes in the measured phenomena. Here too, trust falls flat. This is due to the many dimensions that are critical to trust. Table 2 lists 84 factors researchers have systematically identified. That workload is included, a phenomena with at least seven dimensions (Hart & Staveland, 1988), means that at least seven more items could be added. Culling may be able to distill these items into their critical components. However,

various researchers have attempted this and discovered different lists (e.g., Hancock et al., 2011; Hoff & Bashir, 2015; Jian et al., 2000). Even active trust researchers (e.g. Yang et al., 2021) acknowledge that there are more than two dozen distinct factors. All of this means that there are too many factors influencing trust for it to be usefully diagnostic.

Trust Is Not Predictive

Another way that a concept can be useful is if it is predictive of human behavior. Unless you define trust as a behavior (the quintessential manifestation of ontological alchemy), here too, trust is deficient. An increase in trust does not indicate whether people will use something. Rather, it shows that they **may** be more likely to do so in certain situations. This is common knowledge in the trust community with even Lee and See (2004, p. 76) saying that “trust influences reliance on automation; however, it does not determine reliance.” There are many situations where people may trust something and choose to act divergently: they do not think the task the trusted agent would do needs to be done; they think they can do better than the (good/trusted) agent; they are bored and want to do the task manually; or they enjoy doing the task. There are also many contexts where people might rely on a system or agent that they do not trust: they are compelled to by an authority or social pressures; the people use the agent but monitor it extremely closely to detect any potential failures; the situation requires use of the agent due to task demands or environmental constraints (e.g., “I do not trust the autopilot, but I have to land this aircraft and I can’t see the terrain”); the people are lazy (do not want to perform the tasks manually) or have become apathetic. The point is, knowing whether and/or how much somebody trusts something does not inherently provide useful information.

This is in stark contrast with compliance (whether the human performs actions recommend by the automation) and reliance (stopping action performance when the automation indicates it is unnecessary; Meyer, 2004). Both concepts describe behaviors that can be examined objectively (Vashitz et al., 2009) without the de-precision inherent to trust.

If we return to the big picture: the reason for focusing on trust is to get people to use technology or (from the human factors perspective) to get them to use it only when it is appropriate. **Use of technology is compliance and reliance behavior.** Since trust is a difficult-to-define, subjective concept with selectivity and diagnosticity problems, there is no reason to examine it over objective compliance and reliance: the precise things we are trying to understand and design for. In fact, working with reliance and compliance can be more informative than looking at trust. This is because they account for situations (like those enumerated previously) where people’s behavior can be influenced by factors beyond trust.

THE ETHIC OF TRUST

Trust Is Not Inherently Humanistic

Unlike other subjective concepts in the human factors toolbox, trust is not inherently humanistic. To illustrate this, consider workload, usability, and situation awareness. Both workload and usability relate to work or interaction qualities that have direct implications for human experience: interactions that are pleasing to people and help them achieve their goals without unpleasantness and overload. Situation awareness relates to the human’s ability to maintain accurate knowledge and make predictions. Engineering around these concepts will help people do their jobs, facilitate system and workplace safety, improve human working conditions, and grow an individual’s or team’s expertise. Trust, as a concept, has no such claims to humanistic perspectives. If somebody trusts something, this *may* have some impact on whether or how the thing is used, but the use of the thing is not good or bad. If the thing people are trusting should not have been trusted in a work-, safety-, or financial-critical situation, this could have profound negative implications for human well-being. As such, trust can be quite problematic. Subsequent sections develop this idea further.

Trust is Not a Virtue

As indicated previously, the human factors community knows well that there are distinct dangers with both over and under trust (Bainbridge, 1983). As such, we emphasize the need for properly calibrated trust. This means that trust is not a virtue, but something that only has value when it corresponds to the system’s trustworthiness. But this perspective makes the science of trust kind of pointless. If we genuinely want people to trust systems for the right reasons, we need to provide them with honest, compelling evidence for why they should do so. This means engineering the system to be reliable (a.k.a “trustworthy”); it means making the automation and/or autonomy transparent about its behavior and limitations; it means making the system interfaces ecologically valid; it means ensuring that designs and tasks are consistent with human cognitive and physical constraints and capabilities; and it means ensuring that the automation integrates well into the humans’ tasks and workflows... This is just good human factors engineering. Actual analysis of trust or explicit consideration of it in design need not factor in.

When you have well-validated and/or objective measures and methods for evaluating and designing for compliance, reliance, transparency, user-centeredness, ecological-validity, and human performance, you do not need concepts as abstruse and messy as trust.

As Table 2 shows, there are dimensions of trust not captured by traditional human factors and reliability engineering. These relate to emotional or affective responses that people may have (e.g. attachment, faith, wariness) or

Table 2. List of Factors that Impact Trust (Cho et al., 2015; Hoff & Bashir, 2015; Jian et al., 2000; Lee & See, 2004; Madsen & Gregor, 2000).

Ability	Interpersonal Competence
Accessibility	Judgment (of truster)
Adoption (has the trustee been adopted)	Level of control
Age (of truster)	Loyalty
Altruism (does the trustee display it)	Mood (of truster)
Appearance	Motives
Attachment (of truster to trustee)	Openness
Attentional capacity (of truster)	Organizational setting
Attitudes/expectations (of truster)	Persistence
Availability	Personality traits (of truster)
Benefits offered	Power (had by the trustee)
Benevolence	Predictability
Business sense	Reciprocation (does the trustee show it)
Communication style	Relational capital
Competence	Reliability (both general and context-specific)
Concern (does the trustee show it)	Reputation of system and/or brand
Confidence	Responsibility (does the trustee take it)
Confidentiality (does the trustee maintain it)	Risk (associated with the operating environment)
Congeniality	Security (how secure is the trustee)
Consistency	Self-confidence (of truster)
Contract (is the trustee under one)	Sincerity
Controllability	Social capital
Cooperativeness	Subject matter expertise (of truster)
Credit (how the trustee shares it)	Suspicion (of truster)
Culture (of truster)	System complexity (the complexity of the trustee)
Deception (does the trustee deceive)	Tact (does the trustee display it)
Delegation	Task difficulty (for tasks done by the truster with the trustee)
Dependability	Timeliness
Difficulty of error (for the trustee)	Timing of error (made by trustee)
Discreetness	Transparency/feedback (does the trustee provide it)
Ease-of-use	Type of error (made by the trustee)
Experience (of truster)	Type of system (what the trustee is)
Expertise (does the trustee have it)	Underhandedness (does the trustee display it)

(Continued)

Table 2. (Continued)

Faith (of truster in trustee)	Understandability
Familiarity (of trustee to truster)	Understanding (of truster)
Fear (of truster towards trustee or situation)	Usefulness
Feeling (of truster)	Valence
Framing of task	Validity
Gender (of truster)	Value congruence (between truster and trustee)
Harmful or injurious outcomes (are they possible)	Wariness (of truster)
Integrity	Willingness to reduce uncertainty
Intentions	Workload (of truster when interacting with trustee)

Unless otherwise specified in parentheses, all the above relate to the object of trust (the trustee). Cases representing properties of the agent doing the trusting (the truster) and trustee-truster interactions are noted. Truster and trustee are used instead of human and machine (respectively) because factors can describe inter-agent trust relationship beyond just human-machine ones.

machine displays of unconventional affective and/or ethereal qualities (e.g. concern, congeniality, loyalty, sincerity). Thus, it may be possible to engineer trust without or beyond what is achievable with traditional human factors. But this is the essence of trust’s ethical problem. If technology cannot overcome adoption limitations through objective performance and honest communication with human users, then we move into the realm of manipulation and coercion. That is, the technology (or the designer or owner of the technology) is asserting that it knows better than the person what they should do. This is why trust is humanistically deficient compared to other cognitive engineering concepts (i.e., workload, usability, and situation awareness): it is a vector for removing human autonomy, not enhancing it. There is also a contradiction here. Presumably, humans are part of a system because they bring experience, expertise, instincts, and creativity that can be beneficial. If engineers manipulate people into behaving the way they (or others) want, why include the people at all?

We Should be Suspicious of Trust

Human factors engineers are not the primary drivers of interest in trust. Technology companies, government agencies, and researchers excited by the potential of autonomous and AI systems are. Those working in these areas are rarely human factors engineers and they are likely encountering the need to account for human-automation interaction for the first time. Critically, technologies they are advocating for (like those using artificial neural nets), have significant human factors problems due to their complexity, non-linearity, and explainability limitations (Biran & Cotton, 2017; Wang et al., 2020). Most of the associated researchers are not familiar with the intricacies of human-machine interaction nor are they particularly interested in engineering things from

a humanistic perspective. They are concerned with advancing their technology whether it has good human factors or not. You can observe this sentiment when these interests emphasize the need for understanding machine trust in humans (e.g., [Air Force Research Laboratory, State University of New York, IBM, NYSTEC, and National Security Innovation Network, 2021](#); [Llinas, 2022](#); [Summit on trusted autonomy research and technology: Agenda, 2022](#)): a silly anthropomorphism of machines given that trust is inherently subjective and thus impossible for a machine; adding trust to machine behavior could only serve to make it less predictable and less reliable; it also suggests an authoritarian relationship of machines over people, one where the human must earn the trust of a machine to gain its permission to do something. For this set, I am not surprised they are concerned with trust: they may not be able to engineer their systems to be reliable and have good human factors (it may be impossible), but if they can make people trust them anyway, then the technology will still succeed. This naive emphasis on trust is wrongheaded, but it is not nefarious. The same cannot be said for the next consideration.

When con artists are trying to defraud somebody, they will frequently do two things (Braucher & Orbach, 2015; Levine, 2014; Orbach & Huang, 2018). First, they will create situations that are complex. This makes it difficult for targets to keep track of what is going on and thus disguises the true nature of the scam. It also makes investigation and legal enforcement difficult. Second, the con artists do everything they can to get the target to trust them. Most AI and autonomous systems are complex and (literally) beyond the limits of explainability. The business models of some of the biggest companies that are interested in advancing human trust in AI are based on extracting human data resource from the population, automating away skilled human labor (to save costs), and transferring decision making away from individuals and

democratic processes and into automated systems. These efforts serve to shift power from people and into the hands of those that develop the agents, systems and algorithms they want people to trust (Helbing et al., 2019). I think it is good for people not to trust these companies or their technologies because they are acting like con artists. Trust should be earned and continually re-justified, not simply designed into a system.

CONCLUSIONS

The discussion above shows that there is very little justification for the human factors or larger scientific community to support research on trust. Trust is an imprecise folk concept that is difficult to define. Trust is confounded with many other similar concepts and formed from a surfeit of factors, this means its representation (either as a measure or a model) lacks both selectivity and diagnosticity. Trust is not predictive of human behavior nor does its explicit consideration inherently advance a humanistic perspective in engineered technology. Finally, there is a dark side to trust. The focus on trust is actually facilitating a dodge around good human factors engineering and enabling human disenfranchising by giving more control and power to non-democratic organizations.

It is important to note that I did not write this article to criticize the research work that the human factors community has done on trust. I think our intentions have generally been good and that we have put in an honest effort to determine how real and to what extent trust can be measured, modeled, and accounted for in engineering. That said, research on human-machine trust has been going since the 70s (Halpin et al., 1973; Sheridan & Ferrell, 1974). That there are still so many fundamental problems should make all of us seriously question it. We can only spend so much effort trying to define the composition of invisible fibers, their interactions in garment formation, or their dynamics when the garments are worn in different environments, and fail to get cohesive results, before we acknowledge that the emperor has no clothes.

Despite how good our community's intentions may be, developing the science of trust beyond what is achievable with good reliability and human factors engineering will advance the science of manipulation and coercion. The ability of engineers to recognize and resist problematic developments is fundamental to engineering ethics (National Society of Professional Engineers, 2019). Thus, it is our moral and professional responsibility to resist the ethical dilemmas trust research enables. If human factors engineers insist on continuing to research trust, then a more ethical approach would investigate how to prevent its emotional dimensions from being a factor in technology adoption and use. This will help ensure that human-machine interactions cannot be subject to coercion and will push human factors engineering to the fore. On this front, there are many opportunities for improving system trustworthiness/reliability and human-

machine interaction that need not consider trust. By deemphasizing the importance of trust, we can focus more energy on how to precisely examine and honestly designing the relationships between humans and machines. We can also, hopefully, stop debating the drab, pseudophilosophical issues surrounding trust definition, measurement, and modeling; sidestep the ethically wrought implications of trust; and move forward creating more-reliable, humanistic technologies and societies.


REFERENCES

- Air Force Research Laboratory, State University of New York, IBM, NYSTEC, & National Security Innovation Network. (2021). *Trusted AI challenge series*. Innovare Advancement Center. https://assets.website-files.com/5f47f05cf743023a854e9982/60885a1587a12f5393a33f67_TAITopic3RFPNSIN_NYSTECfinal26Apr.pdf
- Annett, J. (2002). Subjective rating scales: Science or art? *Ergonomics*, 45(14), 966–987. <https://doi.org/10.1080/00140130210166951>
- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775–780. [https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8)
- Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)* (8, pp. 8–13).
- Bolton, M. L., Biltkoff, E., & Humphrey, L. (n.d.-a). The level of measurement of subjective situation awareness and its dimension in the situation awareness rating technique (SART). *IEEE Transactions on Human-Machine Systems*. (In Press). <https://doi.org/10.1109/THMS.2021.3121960>
- Bolton, M. L., Biltkoff, E., & Humphrey, L. (n.d.-b). The level of measurement of the NASA task load index and its constituent dimensions. In *IEEE transactions on human-machine systems*, 10 pages. (Under Review).
- Braga, D. D. S., Niemann, M., Hellingrath, B., & Neto, F. B. D. L. (2018). Survey on computational trust and reputation models. *ACM Computing Surveys (CSUR)*, 51(5), 1–40. <https://doi.org/10.1145/3236008>
- Braucher, J., & Orbach, B. (2015). Scamming: The misunderstood confidence man. *Yale Journal of Law and the Humanities*, 27(2015), 249. <https://doi.org/10.2139/ssrn.2314071>
- Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. (2017). Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors*, 59(3), 333–345. <https://doi.org/10.1177/0018720816682648>
- Cho, J.-H., Chan, K., & Adali, S. (2015). A survey on trust modeling. *ACM Computing Surveys (CSUR)*, 48(2), 1–40. <https://doi.org/10.1145/2815595>
- Cronk, T. M. (June 22, 2021). *Hicks announces new artificial intelligence initiative*. DOD News. <https://www.defense.gov/News/News-Stories/Article/Article/2667212/hicks-announces-new-artificial-intelligence-initiative/>
- Dekker, S., & Hollnagel, E. (2004). Human factors and folk models. *Cognition, Technology & Work*, 6(2), 79–86. <https://doi.org/10.1007/s10111-003-0136-9>
- Dekker, S., & Nyce, J. M. (2015). From figments to figures: Ontological alchemy in human factors research. *Cognition, Technology & Work*, 17(2), 185–187. <https://doi.org/10.1007/s10111-015-0321-7>
- Delbufalo, E. (2015). Subjective trust and perceived risk influences on exchange performance in supplier–manufacturer relationships. *Scandinavian Journal of Management*, 31(1), 84–101. <https://doi.org/10.1016/j.scaman.2014.06.002>
- Earle, T., & Siegrist, M. (2008). Trust, confidence and cooperation model: A framework for understanding the relation between trust and risk perception. *International Journal of Global Environmental Issues*, 8(1–2), 17–29. <https://doi.org/10.1504/IJGENVI.2008.017257>

- Earle, T. C. (2010). Trust in risk management: A model-based review of empirical research. *Risk Analysis: An International Journal*, 30(4), 541–574. <https://doi.org/10.1111/j.1539-6924.2010.01398.x>
- Eignor, D. R. (2013). *The standards for educational and psychological testing*. American Psychological Association.
- European Union Aviation Safety Agency (2021). *EASA concept paper: First usable guidance for level 1 machine learning applications: A deliverable of the EASA AI roadmap (tech. Rep. No. Proposed issue 01)*. European Union Aviation Safety Agency.
- Gebru, B., Zeleke, L., Blankson, D., Nabil, M., Nateghi, S., Homaifar, A., & Tunstel, E. (2022). A review on human-machine trust evaluation: Human-centric and machine-centric perspectives. *IEEE Transactions on Human-Machine Systems*, 52(5), 952–962. <https://doi.org/10.1109/THMS.2022.3144956>
- Goldberg, R. (May 13, 2016). *Lack of trust in internet privacy and security may deter economic and other online activities*. National Telecommunications and Information Administration, United States Department of Commerce. <https://www.ntia.doc.gov/blog/2016/lack-trust-internet-privacy-and-security-may-deter-economic-and-other-online-activities>
- Halpin, S. M., Johnson, E. M., & Thornberry, J. A. (1973). Cognitive reliability in manned systems. *IEEE Transactions on Reliability*, 22(3), 165–170. <https://doi.org/10.1109/TR.1973.5215932>
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53(5), 517–527. <https://doi.org/10.1177/0018720811417254>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology*, 52, 139–183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R. V., & Zwitter, A. (2019). Will democracy survive big data and artificial intelligence? In *Towards digital enlightenment* (pp. 73–98). Springer.
- 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>
- Israelsen, B. W., & Ahmed, N. R. (2019). dave... I can assure you... that it's going to be all right..." A definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. *ACM Computing Surveys (CSUR)*, 51(6), 1–37. <https://doi.org/10.1145/3267338>
- James, H. S., Jr (2002). The trust paradox: A survey of economic inquiries into the nature of trust and trustworthiness. *Journal of Economic Behavior & Organization*, 47(3), 291–307. [https://doi.org/10.1016/S0167-2681\(01\)00214-1](https://doi.org/10.1016/S0167-2681(01)00214-1)
- 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
- 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
- Levine, T. R. (2014). *Encyclopedia of deception Vol. (2)*. Sage Publications.
- Llinas, J. (2022). Putting AI in the critical loop: Assured trust and autonomy in human-machine teams. In *Association for the advancement of artificial intelligence*. <https://aaai.org/Symposia/Spring/sss22symposia.php>
- 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>
- Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *11th australasian conference on information systems* (Vol. 53, pp. 6–8). Springer.
- Matthews, G., De Winter, J., & Hancock, P. A. (2020). What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. *Theoretical Issues in Ergonomics Science*, 21(4), 369–396. <https://doi.org/10.1080/1463922X.2018.1547459>
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709–734. <https://doi.org/10.2307/258792>
- Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, 46(2), 196–204. <https://doi.org/10.1518/hfes.46.2.196.37335>
- Meyer, J., & Lee, J. D. (2013). Trust, reliance, and compliance. In J. D. Lee & A. Kirlik (Eds.), *The oxford handbook of cognitive engineering*. Oxford University Press.
- National Science Foundation, Department of Homeland Security, Science & Technology Directorate, Institute of Education Sciences, U.S. Department of Education, National Institute of Food and Agriculture, National Institute of Standards and Technology, Department of Defense, & IBM Corp. (2020). National artificial intelligence (AI) research institutes: Accelerating research, transforming society, and growing the American workforce: Program solicitation (tech. Rep. No. NSF 22-502). National Science Foundation.
- National Institute of Standards and Technology. (2019). *US leadership in AI: A plan for federal engagement in developing technical standards and related tools. Prepared in response to executive order 13859 (tech. Rep.)*. US Department of Commerce Washington, DC. https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf
- National Science Foundation, National Institute of Food and Agriculture, Department of Homeland Security, Science & Technology Directorate, U.S. Department of Transportation, Federal Highway Administration, & U.S. Department of Veterans Affairs. (2020). National artificial intelligence (AI) research institutes: Accelerating research, transforming society, and growing the American workforce: Program solicitation (tech. Rep. No. NSF 20-503). National Science Foundation.
- National Society of Professional Engineers. (2019). *NSPE code of ethics for engineers*. National Society of Professional Engineers. <https://www.nspe.org/sites/default/files/resources/pdfs/Ethics/CodeofEthics/NSPECodeofEthicsforEngineers.pdf>
- Orbach, B., & Huang, L. (2018). Con men and their enablers: The anatomy of confidence games. *Social Research: An International Quarterly*, 85(4), 795–822. <https://doi.org/10.1353/sor.2018.0050>
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230. <https://doi.org/10.1518/001872097778543886>
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making*, 2(2), 140–160. <https://doi.org/10.1518/155534308X284417>
- Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, 23(3), 393–404. <https://doi.org/10.5465/amr.1998.926617>
- Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, 58(3), 377–400. <https://doi.org/10.1177/0018720816634228>
- Schmidt, E. (April 2018). *Statement of Dr. Eric Schmidt: House armed services committee (Vol. 17)*. House Armed Services Committee. <https://docs.house.gov/meetings/AS/AS00/20180417/108132/HHRG-115-AS00-Wstate-SchmidtE-20180417.pdf>
- Shahrdar, S., Menezes, L., & Nojournian, M. (2018). A survey on trust in autonomous systems. In *Science and information conference* (pp. 368–386).
- Sheridan, T. B., & Ferrell, W. R. (1974). *Man-machine systems; information, control, and decision models of human performance*. the MIT press.
- Siegrist, M. (2021). Trust and risk perception: A critical review of the literature. *Risk analysis*, 41(3), 480–490. <https://doi.org/10.1111/risa.13325>

- Summit on trusted autonomy research and technology: Agenda (2022). *West Lafayette: Purdue University*.
- Vashitz, G., Meyer, J., Parmet, Y., Peleg, R., Goldfarb, D., Porath, A., & Gilutz, H. (2009). Defining and measuring physicians' responses to clinical reminders. *Journal of Biomedical Informatics*, 42(2), 317–326. [10.1016/j.jbi.2008.10.001](https://doi.org/10.1016/j.jbi.2008.10.001)
- Wang, X., Bisantz, A. M., Bolton, M. L., Cavuoto, L., & Chandola, V. (2020). Explaining supervised learning models: A preliminary study on binary classifiers. *Ergonomics in Design*, 28(3), 20–26. [10.1177/1064804620901641](https://doi.org/10.1177/1064804620901641)
- Wei, J., Bolton, M. L., & Humphrey, L. (2019). Subjective measurement of trust: Is it on the level? In *Proceedings of the human factors and ergonomics society annual meeting* (pp. 212–216). Sage.
- Wei, J., Bolton, M. L., & Humphrey, L. (2020). The level of measurement of trust in automation. *Theoretical Issues in Ergonomics Science*, 22(3), 274–295. <https://doi.org/10.1080/1463922X.2020.1766596>
- Yang, X. J., Schemanske, C., & Searle, C. (2021). Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors*.
- Zhou, S., Sun, X., Liu, B., & Burnett, G. (2022). Factors affecting pedestrians' trust in automated vehicles: Literature review and theoretical model. *IEEE Transactions on Human-Machine Systems*, 52(3), 490–500. <https://doi.org/10.1109/THMS.2021.3112956>



Matthew L. Bolton  is an associate professor in the Department of Engineering Systems and Environment at the University of Virginia. Dr Bolton received the B.S. in computer science, M.S. in systems engineering, and PhD in systems engineering from the University of Virginia in 2004, 2006, and 2010, respectively. He was previously a senior researcher at the NASA Ames Research Center and has held academic appointments at the San Jose State University, the University of Illinois, Chicago, and the University at Buffalo. His research focuses on using human performance modeling and formal methods to engineer critical systems. ORCID iD: <https://orcid.org/0000-0002-7943-0497>



Copyright 2022 by Human Factors and Ergonomics Society. All rights reserved.
DOI: 10.1177/10648046221130171
Article reuse guidelines: sagepub.com/journals-permissions