Article

Failing to Grasp our Failure to Grasp Automation Failure

Journal of Cognitive Engineering and Decision Making
2024, Vol. 0(0) 1–7

DOI: 10.1177/15553434241228799
journals.sagepub.com/home/edm
Sage

Erin K. Chiou

Abstract

This paper discusses three points inspired by Skraaning and Jamieson's perspective on automation failure: (a) the limitations of the automation failure concept with expanding system boundaries; (b) parallels between the failure to grasp automation failure and the failure to grasp trust in automation; (c) benefits of taking a pluralistic approach to definitions in sociotechnical systems science. While a taxonomy of automation-involved failures may not directly improve our understanding of how to prevent those failures, it could be instrumental for identifying hazards during test and evaluation of operational systems.

Keywords

safety, trust, first principles, definitions, models, sociotechnical

Introduction

Skraaning and Jamieson (2023) highlight the challenges in defining automation failure, contrasting a narrower conception with broader perspectives. They reference aviation case studies and safety science, illustrating how escalating system complexity complicates attributing catastrophic failures solely to automation. I agree with their characterization of system complexity and the challenge of failure attribution. I disagree that an automation failure classification scheme is necessary to make meaningful progress in safety science and cognitive engineering. However, a taxonomy of automation-involved failures similar to what Skraaning and Jamieson propose may serve as a useful tool to help advance test and evaluation efforts of operational systems. Building on this context, the following section delves into the limitations of defining automation failure, highlighting the complexities involved.

Limitations of Defining Automation Failure

The efficiency benefits of a common definition are clear, and working to achieve common ground within the scholarly literature is worth encouraging. Generally, automation failure is understood as when the performance of an automation-enabled system does not meet a widely accepted standard or expectation. However, the challenge lies in operationalizing this definition. Key questions include what constitutes the system—does it encompass people and organizational processes? Who sets and assesses whether an expectation is met—is it the organization's evaluators, the judicial system, society at large?

Definitional First Principles. To start, there have been many definitions of automation, in part shaped by history and by wider interest in the subject (Sheridan, 2002). For some, one definition may be more convenient than others, depending on the goal:

Arizona State University, Mesa, AZ, USA

Corresponding Author:
Erin K. Chiou, Arizona State University, Santa Catalina Hall, Ste 150H, 7271 E Sonoran Arroyo Mall, Mesa, AZ 85212, USA.
Email: erin.chiou@asu.edu

- Automation has been defined based on categories of technology progress or ability, a definition that is often used to contrast mechanistic systems from learning systems. (A technology-facing view that is helpful for communicating technological progress and associated new abilities or concerns.)
- Automation has been defined as a device that can function without requiring continuous input from an operator. (A work-oriented view that is helpful for communicating why automation may be differently problematic than other types of technology in operational environments.)
- Automation has been defined as a function performed by a machine that could conceivably be completed by a person. (A relationship-focused view that is helpful for communicating why automation problematics depend on people's expectations of technology relative to themselves.)

Engineers that want to highlight how advancements in technology have resulted in novel capabilities and corresponding concerns may, for example, prefer to characterize "automation" and "autonomy" as different concepts, irrespective of the scholarly history of either word. Scholars wanting to take a broader concept of automation to study the effects of technology on human systems may prefer a more progressive definition that is robust to technological advancements by centering the human-automation relationship (Parasuraman & Riley, 1997). If such definitions might serve as normative models (Sheridan, 2018), then we might ask how useful are these definitions, or how consistently can they be applied, to whom are they useful, and for what purpose?

Before we can answer that question, how are we defining failure? Failure implies that an expected standard is drastically unmet. If system components are working to exacting standards, but brought together fail to achieve broader expectations, is it useful to describe this as *automation* failure? Skraaning and Jamieson cite Leveson (2004) who describes this type of failure as *dysfunctional interactions*. For instance, combining various health monitoring devices in hospitals can lead to alarm fatigue and medical errors (Albanowski et al., 2023; Cvach, 2012). This is not to say that automated components need no improvement, but the point made explicit in Leveson (2004, p. 244), and echoed by Skraaning and Jamieson (2023), is that "these dysfunctional interactions among system components (system accidents) have received less attention than component failure accidents."

Attribution of Failure in Complex Systems. This may be why Skraaning and Jamieson (2023) propose to include a broader concept (systemic automation failures) as part of a taxonomy, but attribution is important. Might "automation failure" unintentionally suggest to reasonable actors that automation failure is the automation's failure? What I like about the initial concept of automation failure is that it presumes automation is fallible and implies that human operators of automated systems play a critical role. This implication focuses on the broader human-automation system beyond the mechanical or informational components that comprise the automation, and powerfully counters the view that frontline operators are the primary cause of adverse events, or the weakest link in productive work systems because relative to automation people are inconsistent, get tired, and make mistakes.

However, the danger of attributing *systemic* failures using either technology-focused ("automation failure") or operator-focused ("human performance challenges") language is that it may orient adjudicators to blame correspondingly. After an accident, a term like "automation failure" affords inspection of technology development and testing; "human performance challenges" affords inspection of operator training or operator responsibility. What these labels seem less helpful for is holding organizations or oversight agencies accountable for known challenges in interoperability ("Medical Device Interoperability," 2023), human monitoring performance (Moray, 2003), and safety culture (Billeaud & Snow, 2023; "Inadequate Safety Culture," 2019). That may not be the intended purpose of the taxonomy, even though intent does not prevent it from being used that way. Perhaps the taxonomy is meant as a common reference point for scholars modeling human-automation performance.

Yet even in this latter case of scholarly orientation, attribution questions are raised in the

Skraaning and Jamieson (2023) classification scheme. Should the examples under "systemic automation failures" be attributed to automation at all, or should these be attributed to the organization's failure to test for these situations, or failure to consider operator expectations in the design? Similarly, should "operators are unfamiliar with the automation due to inadequate training" be attributed to "human and organizational slips/misconceptions" when this unfamiliarity might be the result of poor interface design that impedes learning and discovery on the job? Even within a single category, the attribution process is not clear. How do we determine, and who gets to determine, what is "overly complex" or "unsuitable"? Questions like these, and the challenge of answering them in an encompassing or unambiguous way, add uncertainty that this framework will be useful for guiding system design, understanding, or communication around automation-involved systemic failures.

Parallels in Defining Automation Failure and Trust in Automation

Similar to the varied concepts of automation failure noted by Skraaning and Jamieson, literature on trust in automation also reflects diverse concepts. The literature continues to note a lack of definitional agreement despite there being a robust one of trust as, "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004, p. 54). The main challenge of achieving consensus, and a reasonable criticism of existing definitions, is not so much what the definition is but how to best operationalize it so that it is useful for practitioners.

Defining and Operationalizing Trust in Human-Automation Systems. To define means to describe the nature, scope, or meaning of something. To operationalize, on the other hand, involves making an abstract concept observable and measurable. Therefore, the existence of the abstract concept is inferred from other phenomena, often incompletely. While broad consensus may exist on definitions, operationalizations often vary because they tend to be more context-dependent. A good definition should be useful within a wide range of situations and difficult to contradict, whereas good operationalizations should enable the appropriate application of the construct within specific constraints—similar to the respective differences between a normative model and a descriptive model (Sheridan, 2018).

Although Lee and See (2004) cover an awesome array of trust factors and relationships from a wide range of sources, their operationalization of trust in automation could be characterized as a narrower concept. Their distinctions between trust in people and trust in automation, and their model of trust and reliance (2004, pp. 65–67), indicate an information processing view of trust under supervisory control conditions (Sheridan, 1975). This concept of trust fits well with human-automation relationships that are essentially signal detection type tasks with a human-in-the-loop. In such tasks, the role of trust in an operator's reliance and compliance decisions (or lack of decision) are of interest; calibrating operator expectations to automation capabilities, and having appropriate trust is a primary goal.

However, increasing system complexity has meant that this narrower concept of trust may fall short in situations with increasingly autonomous and interactive information systems. A broader concept was therefore introduced to include both the information processing view (i.e., "semiotics") and the changing social structure implications of highly capable automation in more open-world environments (Chiou & Lee, 2021). In these more open-world environments, the hypothesis is that our understanding of trust in the context of human-automation systems will be less informed by information processing alone, and more informed by identifying decision interdependencies, values alignment, and changing situation structures. While this broader perspective increases the problem scope and complexity, the point is that the narrower concept has reached its limit for advancing our understanding of trust in human-automation systems, and how to design for or evaluate these new systems effectively.

Both narrow and broad concepts can coexist, as Skraaning and Jamieson illustrate in their taxonomy for the concept of automation failure. Yet, even with a well-documented narrative arc, there will be surprising takes that emerge, no matter how

many or how thorough are the previous reviews of empirical literature (Hancock et al., 2011; Hoff & Bashir, 2013; Huang et al., 2021; Lee & See, 2004; Madhavan & Wiegmann, 2007). As an example, take the recent effort to translate tradecraft standards from Intelligence Community Directive (ICD) 203 into the Multisource AI Scorecard Table (MAST), a rating system tool for evaluating artificial intelligence trustworthiness (Blasch et al., 2020). MAST is based on a set of nine criteria traditionally used to evaluate the trustworthiness of human intelligence analysis, and its justification as a trust assessment tool is primarily supported by literature from computer and information science perspectives, rather than from social science perspectives. Setting aside conceptual differences between trust and trustworthiness for the moment, the implication is that if AI was assessed as trustworthy based on this rating system, then operators could (and would) trust the AI more, and more appropriately. Despite many of the MAST criteria comprising what might be considered distal variables of trust that are open to interpretation (consider the "customer relevance" or "visualization" criteria as examples), this tradecraft-derived tool correlates surprisingly well with more scientifically accepted trust assessment instruments (Chiou et al., 2022). As a result, the tool joins the many dozens of other tools that have operationalized trust in automation in various ways (Alsaid et al., 2023; Kohn et al., 2021).

A Gricean Grasp of Automation Failure

Might the Skraaning and Jamieson taxonomy help us be more precise when describing what we are testing, modeling, or analyzing? For abstract concepts that invite multiple perspectives, it may be Sisyphean to try and consense the many possible operationalizations that are each contingent on their own unique constraints. We see this over again in the literature for automation (and overlapping concepts like autonomy or artificial intelligence), for trust (and related concepts like trustworthiness, reliance, credibility, intent), and for accountability (and more measurable phenomena like responsibility, consequence, transparency, obligation). Because human language evolves, and the design of large-scale systems remains mostly a matter of "experience, art, and iterative trial and error" (Sheridan, 2018, p. 27), abstract concepts in this area will continue to be operationalized in many credible ways, sometimes unnecessarily separated into "different" concepts, and at other times inappropriately conflated. To err, to understand, and to repair is human—and science.

No matter how committed the effort may be to iterating a precise taxonomy of automation failure, there are some benefits to accepting multiple interpretations of a concept. These benefits include demonstrating that scientific efforts value: dialogue and critical thinking as central to sensemaking; openness to new perspectives even if their premise seems wrong initially; and agility within and cooperation across disciplines to achieve shared goals, rather than to assert one way of thinking over another. Accepting multiple interpretations is not necessarily agreeing to disagree, it means searching for common ground and embracing ambiguity that has meaningful purpose.

Ambiguity does not necessarily mean a lack of clarity that will stymy progress. Successful ambiguity abounds in our language. It means others are able to say what is needed in as few words as possible; I am able to read Skraaning and Jamieson's use of the narrower concept of automation failure throughout their paper, even as they argue for including a broader concept. I am able to understand the phrase, "system failure" as being used to encompass more than one situation, rather than as a failure attributed to automation alone. Abstract concepts that are applied to many different areas demand an openness to learning from multiple co-existing perspectives, rather than effort spent delineating and dividing all possible examples. In instances of perceived conflict around language, our biggest strength as a scientific community is our ability to calibrate our trust in others to faithfully apply concepts as they understand them for a particular purpose, and our ability to evaluate those applications critically.

The Benefits of an Imperfect Taxonomy of Automation Failures. In that spirit, a taxonomy that enumerates examples of automation-involved failures, and estimates where they sit within a sociotechnical system, could be a useful tool for those lacking the language to identify contributors of automation-involved problems. As is evident from chemistry

and biology, taxonomies are powerful tools in science and education, even as new knowledge and perspectives are discovered, valued, and included in subsequent versions. A taxonomy of automation-involved failures can structure a repository of issues to investigate, so that we can learn from known knowledge and have a place for new knowledge. There have long been calls for developing tools that not just human factors and cognitive engineers can use (Cummings, 2018). As automation-involved systems become increasingly complex, influential, and accessible, there will be even stronger needs for effective test and evaluation. Test and evaluation efforts will likely accelerate if a larger community of technologists, scientists, and engineers could draw from a centralized repository of accessible issues.

Even if the categories of Skraaning and Jamieson's taxonomy do not achieve scientific consensus, they could still be used to intuitively (and imprecisely, or even incorrectly) classify problem cases for test and evaluation efforts. Such a tool could help promote safety culture within organizations. However, a more useful structure of the taxonomy might be to use labels with potential overlap, rather than having columns that imply mutually exclusive categories, especially for examples with inherent ambiguity and complicated attribution. As suggested above, operators who are unfamiliar with automation due to inadequate training might fall under both "human and organizational slips/misconceptions" and "human-automation interaction breakdown." The "inadequacy" of the training may be a tradeoff that occurs when cognitive task analyses are not conducted that could have minimized the need for extensive training. Having the ability to label these examples under multiple categories could have more important benefits than side-stepping stringent consensus, specifically, getting the people responsible for those various system components (elementary, systemic, interactional, organizational) at the same table to resolve them together, and making more salient the workload required to address the dysfunctional system interactions that can arise especially when automation is involved.

Conclusion

While the effort to define automation concepts remains in the marketplace of ideas short of standards development, acknowledging and understanding the etymology of terms may be a more meaningful path forward toward consensus on abstract concepts. The rapidly evolving sociotechnical landscape that affects automation-involved failures cannot be held constant. Instead, it seems more important to identify how sociotechnical factors and relationships can lead to failures (Leveson, 2004) rather than categorizing them as caused by elementary, systemic, interactional, or organizational factors. Furthermore, focusing on failures, performance challenges, breakdowns, and slips or misconceptions does not tell us much about how increasing system complexity or changing capabilities of new automation contribute to these issues. Understanding what constitutes "unrealistic operational assumptions" involves not just quantitative analyses (Moray, 2003) but also deep understanding of task requirements, risk tolerance, public relations, and the workforce context (Ackoff, 1979). As such, *failing to grasp our failure to grasp the concept of automation failure* means seeing multiple conceptions of automation failure as part of an ongoing, collective effort to better understand how human-designed, automation-involved systems have failed as many people as they have, and how to prevent such failures in the future.

Acknowledgments

The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-07-04.

ORCID iD

Erin K. Chiou https://orcid.org/0000-0002-7201-8483

References

Ackoff, R. L. (1979). The future of operational research is past. *Journal of the Operational Research Society*, *30*(2), 93–104, https://doi.org/10.2307/3009290

Albanowski, K., Burdick, K. J., Bonafide, C. P., Kleinpell, R., & Schlesinger, J. J. (2023). Ten years later, alarm fatigue is still a safety concern. *AACN Advanced Critical Care*, *34*(3), 189–197. https://doi.org/10.4037/aacnacc2023662

Alsaid, A., Li, M., Chiou, E. K., & Lee, J. D. (2023). Measuring trust: A text analysis approach to compare, contrast, and select trust questionnaires. *Frontiers in Psychology*, *14*(1192020). https://doi.org/10.3389/fpsyg.2023.1192020

Billeaud, J., & Snow, A. (2023, July 28). The backup driver in the 1st death by a fully autonomous car pleads guilty to endangerment. *AP News*. https://apnews.com/article/autonomous-vehicle-death-uber-charge-backup-driver-1c711426a9cf020d3662c47c0dd64e35

Blasch, E., Sung, J., & Nguyen, T. (2020). Multisource AI scorecard table for system evaluation. In *AAAI FSS-20: Artificial Intelligence in Government and Public Sector*. ArXiv. https://arxiv.org/abs/2102.03985

Chiou, E. K., & Lee, J. D. (2021). Trusting automation: Designing for responsivity and resilience. *Human Factors*, *65*(1), 137–165. https://doi.org/10.1177/00187208211009995

Chiou, E. K., Salehi, P., Blasch, E., Sung, J., Cohen, M. C., Pan, A., Mancenido, M., Mosallanezhad, A., Ba, Y., & Bhatti, S. (2022). Trust in AI-enabled decision support systems: Preliminary validation of MAST criteria. In ICHMS22: 3rd IEEE International Conference on Human-Machine Systems, Orlando, FL, November 17–19, 2022. https://doi.org/10.1109/ICHMS56717.2022.9980623

Cummings, M. (2018). Informing autonomous system design through the lens of skill-rule-and knowledge-based behaviors. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 58–61. https://doi.org/10.1177/1555343417736461

Cvach, M. (2012). Monitor alarm fatigue: An integrative review. *Biomedical Instrumentation & Technology*, *46*(4), 268–277. https://doi.org/10.2345/0899-8205-46.4.268

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., 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

Hoff, K., & Bashir, M. (2013). A theoretical model for trust in automated systems. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13* (Vol. 115). Association for Computing Machinery.

Huang, L., Cooke, N. J., Gutzwiller, R. S., Berman, S., Chiou, E. K., Demir, M., & Zhang, W. (2021) Distributed dynamic team trust in human, artificial intelligence, and robot teaming. In C. S. Nam & J. B. Lyons (Eds.), *Trust in human-robot interaction* (pp. 301–319). Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00013-7

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, *12*(604977). https://doi.org/10.3389/fpsyg.2021.604977

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

Leveson, N. (2004). A new accident model for engineering safer systems. *Safety Science*, *42*(4), 237–270. https://doi.org/10.1016/s0925-7535(03)00047-x

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

Medical Device, Interoperability. (2023, May 16). U.S. food and drug administration. https://www.fda.gov/medical-devices/digital-health-center-excellence/medical-device-interoperability

Moray, N. (2003) Monitoring, complacency, scepticism and eutactic behaviour. *International Journal of Industrial Ergonomics*, *31*(3), 175–178. https://doi.org/10.1016/S0169-8141(02)00194-4

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

Sheridan, Thomas B. (1975). Considerations in modeling the human supervisory controller. *IFAC*

Proceedings Volumes, *8*(1, Part 3), 223–228. https://doi.org/10.1016/S1474-6670(17)67555-4

Sheridan, T. B. (2002). Chapter 1: Introduction. In *Humans and automation: System design and research issues* (pp. 3–13). John Wiley and Sons, Inc.

Sheridan, T. B. (2018). Comments on "Issues in human-automation interaction modeling: Presumptive Aspects of frameworks of types and levels of automation" by David B. Kaber. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 25–28. https://doi.org/10.1177/1555343417724964

Skraaning, G. Jr., & Jamieson, G. A. (2023). The failure to grasp automation failure. *Journal of Cognitive Engineering and Decision Making*, *Advance Online Copy*. https://doi.org/10.1177/15553434231189375

'Inadequate safety culture' contributed to uber automated test vehicle crash—NTSB calls for federal review process for automated vehicle testing on public roads. (2019, November 19). National Transportation Safety Board. https://www.ntsb.gov/news/press-releases/Pages/NR20191119c.aspx

Erin K. Chiou is an associate professor of human systems engineering at The Polytechnic School, part of the Ira A. Fulton Schools of Engineering at Arizona State University, and directs the Automation Design Advancing People and Technology Laboratory. She received her BS (psychology, philosophy) from the University of Illinois at Urbana-Champaign and her MS and PhD (industrial and systems engineering) from the University of Wisconsin-Madison. Her research focuses on social factors in complex human-machine systems.