INL/CON-25-83190-Revision-0

Trustworthiness and Trust: Identifying Factors that Drive Successful Human-AI Interaction in Nuclear Power Plant Applications

June 2025

Yusuke Yamani, Austin Jackson, Jeffrey C Joe, Jeremy David Mohon, Casey R Kovesdi

Changing the World's Energy Future

INL is a U.S. Department of Energy National Laboratory operated by Battelle Energy Alliance, LLC

DISCLAIMER

This information was prepared as an account of work sponsored by an agency of the U.S. Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness, of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

INL/CON-25-83190-Revision-0

Trustworthiness and Trust: Identifying Factors that Drive Successful Human-AI Interaction in Nuclear Power Plant Applications

Yusuke Yamani, Austin Jackson, Jeffrey C Joe, Jeremy David Mohon, Casey R Kovesdi

June 2025

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

Trustworthiness and Trust: Identifying Factors that Drive Successful Human-AI Interaction in Nuclear Power Plant Applications

Yusuke Yamani[1], Austin Jackson[1], Casey Kovesdi[2*], Jeffrey Joe[2], Jeremy Mohon[2]

[1]Department of Psychology, Old Dominion University, Norfolk, VA
[2]Human Factors and Reliability Department, Idaho National Laboratory, Idaho Falls, ID

ABSTRACT

Emerging technologies such as artificial intelligence (AI) and machine learning are rapidly evolving and promising tools for efficient and continued safe operations of the U.S. nuclear power plants (NPPs). Emerging AI techniques like large language models (LLMs) are one such technology that may help personnel at existing NPPs perform work more efficiently. For example, operators may query the current operational status of a power plant via a chat interface, leveraging LLMs to access plant-related information in an interactive manner rather than manually collecting various sensor data for surveillance or work order tasks. This is a fundamental shift in the way operators perform their tasks today. The literature of human-automation interaction indicates that trust is a crucial factor that drives a successful interaction between a human operator and an automated system, like an AI-infused NPP application. This work presents the results of a literature review on key factors that relate to trust in AI/LLM technologies for NPP applications. These key factors include trustworthiness, performance characteristics, operator skill, and perceived risk. This preliminary literature review will guide model development and evaluation involving the key factors influencing trust in AI and will develop a framework for a human-centered design for an interface between humans and AI. By addressing trust, this work supports developing a technical basis for designing key characteristics of AI/LLM to support calibrated trust, which will ultimately support widescale adoption of AI/LLM technologies, as well as ensure their safe, effective, and reliable use.

Keywords: Nuclear Plant Modernization, Artificial Intelligence, Large Language Models, Trust, Trustworthiness

1. INTRODUCTION

Nuclear power in the United States represents the highest capacity of other electricity generating sources [1], necessary for economic viability and environmental sustainability. Unfortunately, many existing nuclear power plants (NPPs) in the United States face challenges that present economic risks in continuing operation beyond their existing licensing period, such as managing an aging workforce, managing the obsolescence of existing instrumentation and control equipment, and identifying ways these plants can reduce their operating and maintenance costs. These challenges necessitate modernizing existing reactors with emerging technologies. Notably, advanced technologies involving artificial intelligence (AI) and machine learning may provide a promising solution to address these challenges. Emerging AI techniques like large language models (LLMs) are one such technology that may help personnel at existing NPPs perform work more efficiently. For instance, Visualization for PrEdictive maintenance Recommendation (VIPER) provides an interactive, user-centered visualization that allows for an analysis of heterogenous data, work orders, and accident and condition reports and an assessment and diagnostic of the health of the nuclear system for predictive maintenance [2], see Fig. 1.

* Casey.Kovesdi@inl.gov

Figure 1. A screenshot of the VIPER interface [2].

One way to extend VIPER's capabilities is using a chat interface leveraging LLMs to support the user in interpreting the information provided by VIPER [2], accessing real-time information about the current operational status of the power plant. For instance, operators who may be unfamiliar with machine learning can access plant-related information and explanations in an interactive manner through the VIPER chat interface by querying text-based questions or providing graphical content (e.g., plant schematics). The level of questioning by the user to the chatbot can be of different levels of granularity. The responses for the chatbot may provide relevant or irrelevant information to the user's question. For instance, a key challenge with LLMs is that they can hallucinate, providing erroneous information. If the chatbot presents irrelevant or inaccurate information, the user may make erroneous conclusions if these inaccuracies are not recognized. Moreover, trust in the system for future use may also be diminished if these inaccuracies are recognized.

Thus, trust is a critical area to ensuring the safe and reliable use of AI, as well as to accelerating the rate of adoption. The successful integration of advanced technologies like AI into existing NPPs requires an interdisciplinary approach to identify and control factors related to operator attitude, behavior, and performance for their continued safe and reliable operation. With the ever-advancing capabilities of AI, a human-centered approach to the design interface between human operators and AI-powered information systems is increasingly crucial for ensuring successful human-systems integration in the nuclear power space. This paper provides a preliminary literature review of factors that influence trust in LLMs and offers directions for future human factors research to characterize and optimize interactions between human operators and AI-enabled information support for safe and cost-effective NPP operations.

2. LITEREATURE REVIEW

2.1. Human-Automation Trust

Insights from the literature on human-automation trust and interaction may set a useful foundation for research on trust in AI/LLM. Automation trust is commonly defined as "an attitude that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" [3]. Previous work indicates that trust influences the performance of human operators interacting with technologies equipped with varying levels of automation [4], response behaviors [5], visual scanning patterns [6], attentional distribution [7], and behavioral intention to accept the technology [8][9]. While trust is known to influence many behaviors of operators interacting with the technology, a psychological construct critical for the safety of high-consequence domains such as NPP operation, there are a number of factors that can influence trust in automation.

Perhaps the most comprehensive model of automation trust is the three-layered model of trust offered by Hoff and Bashir [10]. Briefly, their model distinguishes dispositional trust, situational trust, and learned trust to capture both the static and dynamic nature of trust evolution and maintenance during human-automation interaction. First, dispositional trust refers to an individual's general tendency to trust automation, independent of context or specific characteristics of the system. Culture, age, gender, and personality traits of the operator are factors that fall under this category.

Second, situational trust captures the development of trust in automation that highly depends on variability in each situation (e.g. operators having high or low workload levels to complete work tasking). Factors influencing situational trust are classified into either external or internal variability. External variability refers to performance characteristics of the system such as transparency [11], complexity [12], and difficulty of the task [13] that determine the strengths and weaknesses of the system. Internal variability refers to characteristics of the operator that are more transient than dispositional, such as self-confidence, subject matter expertise, automation complacency, and attentional capacity [14][15].

In the realm of NPP operations fused with AI/LLM, subject matter expertise is likely to play a big role in accounting for performance variabilities in the success of task operations. Finally, learned trust, perhaps the hallmark of Hoff and Bashir's [10] model, contrasts initial and dynamic learned trust that develops prior to and during human-automation interaction, respectively. Preexisting knowledge, such as expected system characteristics, its reputation, and the operator's existing mental model of the system, sets their initial learned trust, followed by trust calibration based on their perception and experiences on system performance, such as reliability [16][17], predictability and dependability [18][19][20], and type of errors [5][21][22].

The development of learned trust is modeled as an evolving psychological state that adapts to dynamic experiences with the automation. Interestingly, Lee and See [3] already conceptualized this aspect of learned trust in their model using three attributional abstracts of trust—performance, process, and purpose. Namely, performance-based trust represents trust that develops solely based on observed behaviors of automation such as reliability, types of errors, and response speed. Process-based trust denotes trust that develops based on the underlying algorithm of automated processes while purpose-based trust refers to trust defined by the ultimate goal of why the automation was developed. Both Hoff and Bashir's [10] and Lee and See's [3] models identify trust as a psychological construct is at times dynamic, develops during human-automation interaction, can be influenced by several factors, and characterizes human performance in complex task environments.

2.2. Factors that Drive Human-Automation Trust

In AI-infused NPP operations, which of the various factors in Section 2.1 influence trust in LLMs? We conducted a systematic review of the human factors literature surrounding autonomous systems such as AI and LLM that identified various factors related to trust. In total, we identified 41 papers published in the human factors literature by searching using the keywords "automation," "AI," "LLM," and "human factors." Within the 41 papers, we found 21 papers discussing *trustworthiness,* nine on *performance characteristics,* five on *operator skill,* and six on *perceived risk.*

2.2.1. Trustworthiness of LLMs

While trustworthiness and trust are sometimes used interchangeably in the literature, trustworthiness of a system can refer to an intrinsic property of the system (the trustee) manifested to the automation user (the trustor) [23]. Attributes like automation reliability and explainability [24] are used as estimates of automation trustworthiness, but a comprehensive list of attributes of trustworthiness await a more detailed literature review and analysis [24]. More broadly, according to Fukuyama [25], trustworthiness of a system comes from guidelines, certification agencies, and organizations that regulate the system (e.g., U.S. Nuclear Regulatory Commission), calling for a broader definition of trustworthiness of technologies related to nuclear energy.

Currently, a diverse set of approaches exist for defining trustworthiness in LLMs. Yet, Huang and his team [26] proposed probably the most comprehensive benchmark of trustworthiness in LLMs available to date across six attributes, trustfulness, safety, fairness, robustness, privacy, and machine ethics, in a comprehensive study called TrustLLM evaluating 16 mainstream LLMs using 30 datasets. These attributes are likely to additively contribute to characterizing overall trustworthiness in LLM, thus influencing operator trust. Stanton and Jensen [27] proposed trust (T) as a function of user trust potential (UTP) and perceived system trustworthiness (PST) as the following equation:

$$T(u, s, a) = f\,(UTP(u), PST(u, s, a)) \tag{1}$$

where a user u interacts with an LLM with characteristics s within a context a. Here, incorporating the attributes identified in Huang et al. [26], the parameter s may be defined as:

$$s = \sum attribute_i \tag{2}$$

where $attribute_i$ denotes i^{th} attribute in TrustLLM. Further research is necessary for investigating the function between UTP and PST that form user trust in LLMs and the weights among the six attributes in TrustLLM that drive optimal trust in LLM.

2.2.2. Performance of LLM

Operators may develop their performance-based trust in an automated system especially when they have access to perceive system behaviors [6]. A meta-analysis study that examined factors predicting trust in AI showed that both the performance and reliability of AI significantly predict trust in AI [28], aligning with the common findings in human-automation interaction literature. Interestingly, perceived AI attributes such as personality, anthropomorphism, reputation, and transparency, in addition to system performance, significantly predict trust in AI. This suggests that AI performance should be clearly communicated to human users to gain their trust. These findings may also extend to trust in LLMs as responses from an LLM may contain statements with various levels of accuracy and precision along with nuanced delivery depending on the context. Further research should focus on how performance and its presentation to users affect trust in LLMs.

2.2.3. Operator skill

In the same meta-analysis study [28], ability-based factors related to operator skills such as operators' competency and understanding of the AI system as well as their expertise in their task were significant predictors of trust in AI. This suggests that operators with more experience with AI systems are more likely to trust an AI system. One consideration in the modern nuclear industry is that operators with extensive experience in reactor operations are also aging, potentially facing issues of cognitive aging [29][30], which calls for designing emerging technologies that consider the characteristics of an aging workforce. However, the results of the meta-analysis show no significant relationship between operators' ages and their trust in the AI system. More research should examine the effect of expertise on trust in AI independent of aging, because the results may reflect the cohort effect. Relatedly, their findings suggest individual differences—culture, gender, and personality traits—significantly predict trust in AI, suggesting that there is a considerable variability in how operator skills interact with individual characteristics to shape their trust in AI. For example, individuals with more innovative mindsets tend to trust in the AI system more, implying that future LLMs may be perceived as more trustworthy to those who hold more positive attitudes toward novel technologies.

2.2.4. Perceived risk

Decisions are often characterized by an element of risk, in which risk is defined as "the extent to which there is uncertainty about whether potentially significant and/or disappointing outcomes of decisions will be realized" [31]. Perceived risk is an essential component of trust development, where trust in an agent is influenced by how risky a situation is. Depending on the context, a high-risk situation can cause operators to rely on automation more than human teammates [33], elevate trust in high-workload environments [32], and increase the accuracy of participants at the cost of spending more time verifying the automation responses [34]. Yet, in other contexts, increased risk can cause a decrease in use behavior and trust in automation [35]. For example, if users are concerned that the chatbot may provide incorrect information or are concerned that the technology will threaten their job security [2], then they may be less likely to use the technology. System aspects, such as automation transparency, have been shown to mitigate the negative effects of risk [34]. However, a further investigation of risk within the context of human-automation interaction is necessary. Moreover, the unique risks of NPP operations have yet to be studied within the context of trust in LLMs.

3. CONCLUSIONS

Recent development and proliferation of advanced technologies such as AI and LLMs are likely to support significantly reducing costs by enabling human operators to rapidly access a large volume of plant data (e.g., including sensor data, incident reports, condition reports) to make more timely and strategic decisions. Specifically, advanced computing is expected to allow information extraction from unstructured data via natural language processing and visualization of the data to spark insights. One promising way for human operators to interact with AI is through a chatbot that leverages natural language processing and LLMs (e.g., OpenAI). In this context, our paper focused on identifying factors pertinent to driving trust in LLMs for work performed at NPPs.

In the vast literature of human-automation trust, our literature review has revealed four major factors—LLM trustworthiness, LLM performance, operator skill, and perceived risk—crucial for the success of AI-infused NPP operations (Table 1).

Table I. Identified factors that are likely to influence trust in AI-infused NPP operations.

Factors	Areas of Further Study

LLM Trustworthiness	Trustfulness, safety, fairness, robustness, privacy, and machine ethics
LLM Performance	Transparency, personality, anthropomorphism, reputation
Operator Skill	Competency and understanding of the AI system, task expertise, individual differences (culture, gender, and personality traits)
Perceived Risk	The effects of risk on reliance, trust, and use behavior

First, trustworthiness is an LLM characteristic that influences the growth of trust in the LLM through interactions between the human user and the LLM. Six attributes—trustfulness, safety, fairness, robustness, privacy, and machine ethics—are considered pertinent constituents of trustworthiness that shape trust in LLMs. Second, significant predictors of trust in LLMs include not only an LLM's behaviors and performance in different levels of transparency [11] but also perceived attributes including personality, anthropomorphism, and reputation of LLMs. Third, operator skills, specifically their competency and understanding of the AI system and expertise in the very tasks that they routinely perform in the NPP, also predict trust. Various individual differences such as culture, gender, and personality traits were also found significant predictors of trust, calling for a large study to model relationships among these variables known to shape human-automation trust. Lastly, risk in general plays an essential role for trust to emerge, and previous research showed that operators under use and trust the system more under higher levels of perceived risk. However, more research awaits to characterize the effect of risk on human-automation interaction and trust, especially trust in LLMs.

A series of human-subject experiments exploring the identified factors may give insights into psychological mechanisms that control trust development, maintenance, and recovery through dynamic interactions between human operators and LLMs in NPP operation tasks. In the context of LLMs, trustworthiness may be better characterized by various evaluation criteria being developed for LLMs such as relevance, hallucination, and question-answering accuracy, beyond the attributes identified in the TrustLLM study [26]. To improve LLM performance, retrieval-augmented generation (RAG) allows "crosschecking" answers an LLM is generating internally with an external knowledge base (e.g., accident and condition reports) to provide accuracy verification and avoid hallucinations and privacy violations. Controlled experiments on RAG's influence on operator behaviors and trust may test how operators explore and check the accuracy of LLM's responses for tuning their trust.

Additionally, with RAG, human operators may access levels of "confidence" for each output from an LLM. However, how to implement this additional information to the existing interface warrants extensive user-experience research. For example, various color-coding schema may influence their perception, comprehension, and integration of the LLM outputs for their decision-making. Our team is currently conducting human-subject experiments to examine the effectiveness of color-coding on operator behavior, trust, and eye movements (Fig. 2).

The adoption and actual use of rapidly evolving AI-inspired technologies such as LLMs likely require tapping into each of the four factors identified in this study. That is, RAG is just one mechanism that can improve reliability of the information internally generated by an LLM, influencing LLM trustworthiness and perhaps transparency in LLM performance (abundance of RAG-based information can bury important messages in it). For example, prompt engineering allows optimizing input prompts to guide the AI model toward more accurate, useful, or context-relevant responses, thus influencing the reliability of generated information. In the context of NPP applications, though, the experiences and risk perception of LLM users interactively and organically shape the development, maintenance, and calibration of trust in LLMs. For example, background knowledge in nuclear engineering and reactor operations may affect users' understanding of LLM-generated responses. Relatedly, users with greater knowledge of, or experience

with, NPPs may anticipate specific risks when LLM outputs contradict their expectations. Currently, our team is extending the human-subjects research to compare performance data between university students and employees at Idaho National Laboratory to elucidate individual differences affecting trust in LLMs.

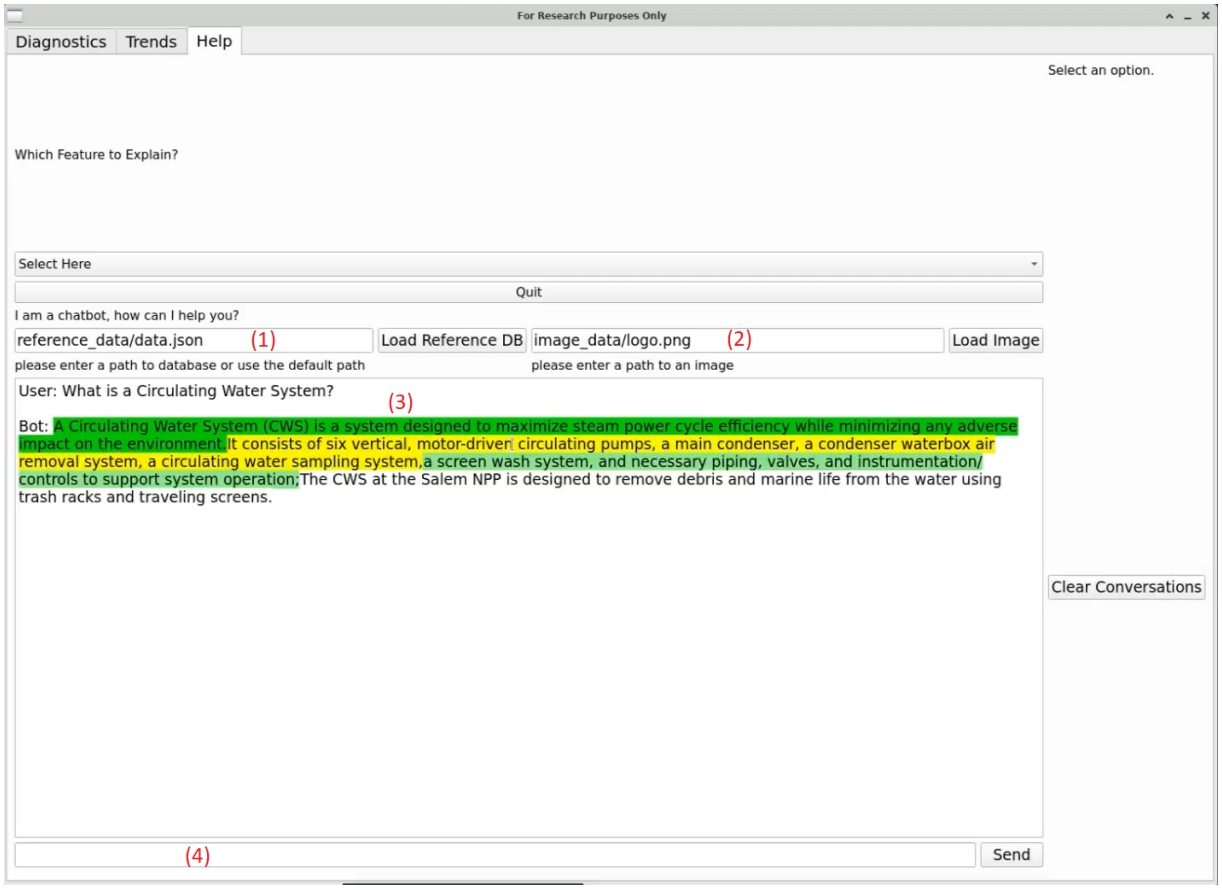

Figure 2. A simulated LLM chat interface in the NPP context [2].

In conclusion, this paper outlines the results of our initial literature review on factors crucial for operator trust in LLM as a novel technology expected to streamline and improve NPP operations powered by AI. Integrating AI and LLMs into existing NPP fleets is a promising and attractive solution for continued safe and more efficient operations. However, more experiments and efforts to model interactions between human operators and advanced technologies such as AI and LLMs, focusing on trustworthiness, performance, operator skills, and perceived risk, are necessary for successful human-systems integration that will ensure the accelerated adoption of these technologies and ensure their safe and reliable use in the modern era of nuclear electricity generation.

ACKNOWLEDGMENT

This work of authorship was prepared as an account of work sponsored by Idaho National Laboratory (under Contract DE-AC07-05ID14517), an agency of the U.S. Government. Neither the U.S. Government, nor any agency thereof, nor any of their employees makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes.

REFERENCES

1. C. R. Kovesdi, R. M. Spangler, J. D. Mohon, and P. Murray, *Development of Human and Technology Integration Guidance for Work Optimization and Effective Use of Information*, Idaho National Laboratory, Idaho Falls, USA (2024).
2. C. Walker, L. Linyu, V. Agarwal, N. Lybeck, A. Hall, R. Hill, and R. Boring, "Demonstration and Evaluation of Explainable and Trustworthy Predictive Technology for Condition-based Maintenance"*, INL/RPT--24-80727-Rev000, 2474859* (2024).
3. J. Lee and K. See, "Trust in Automation: Designing for appropriate reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **46**(1), pp. 50–80 (2004).
4. B. D. Adams, L. E. Bruyn, S. Houde, P. Angelopoulos, K. Iwasa-Madge, and C. McCann, "Trust in Automated Systems," Ministry of National Defense (2003).
5. E. T. Chancey, J. P. Bliss, Y. Yamani, and H. A. Handley, "Trust and the Compliance–Reliance Paradigm: The effects of risk, error bias, and reliability on trust and dependence," *Human Factors*, **59**(3), pp. 333–345 (2017).
6. T. Sato, S. Islam, J. D. Still, M. W. Scerbo, and Y. Yamani, "Task Priority Reduces an Adverse Effect of Task Load on Automation Trust in a Dynamic Multitasking Environment," *Cognition, Technology & Work*, **25**(1), pp. 1–13 (2023).
7. N. D. Karpinsky, E. T. Chancey, and Y. Yamani, "Modeling Relationships among Workload, Trust, and Visual Scanning in an Automated Flight Task," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, **60**, pp. 1550–1554 (2016).
8. K. Wu, Y. Zhao, Q. Zhu, X. Tan, and H. Zheng, "A Meta-Analysis of the Impact of Trust on Technology Acceptance Model: Investigation of moderating influence of subject and context type," *International Journal of Information Management*, **31**(6), pp. 572–581 (2011).
9. R. C. Mayer, J. H. Davis, and F. Schoorman, "An Integrative Model of Organizational Trust," *The Academy of Management Review*, **20**(3), pp. 709–734 (1995).
10. K. A. Hoff and M. Bashir, "Trust in Automation: Integrating empirical evidence on factors that influence trust," *Human Factors*, **57**(3), pp. 407–434 (2014).
11. G. A. Jamieson, G. Skraaning, and J. Joe, "The B737 MAX 8 Accidents As Operational Experiences with Automation Transparency," *IEEE Transactions on Human-Machine Systems*, **52**(4), pp. 794–797 (2022).
12. N. R. Bailey and M. W. Scerbo, "Automation-Induced Complacency for Monitoring Highly Reliable Systems: The role of task complexity, system experience, and operator trust," *Theoretical Issues in Ergonomics Science*, **8**(4), pp. 321–348 (2007).
13. P. Madhavan, D. A. Wiegmann, and F. C. Lacson, "Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **48**(2), pp. 241–256 (2006).

14. R. Parasuraman and D. H. Manzey, “Complacency and Bias in Human Use of Automation: An attentional integration,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **52**(3), pp. 381–410 (2010).
15. Y. Yamani, S. K. Long, T. Sato, A. L. Braitman, M. S. Politowicz, and E. T. Chancey, “Multilevel Confirmatory Factor Analysis Reveals Two Distinct Human–Automation Trust Constructs,” *Human Factors*, **67**(2), pp. 166–180 (2024).
16. M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, “The Role of Trust in Automation Reliance,” *International Journal of Human-Computer Studies*, **58**(6), pp. 697–718 (2003).
17. J. P. Bliss and S. A. Acton, “Alarm Mistrust in Automobiles: How collision alarm reliability affects driving,” *Applied Ergonomics*, **34**(6), pp. 499–509 (2003).
18. B. M. Muir, “Trust in Automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems,” *Ergonomics*, **37**(11), pp. 1905–1922 (1994).
19. B. M. Muir and N. Moray, “Trust in Automation: Part II. Experimental studies of trust and human intervention in a process control simulation,” *Ergonomics*, **39**(3), pp. 429–460 (1996).
20. J. Lee, Y. Yamani, S. K. Long, J. Unverricht, and M. Itoh, “Revisiting Human-Machine Trust: A replication study of Muir and Moray (1996) using a simulated pasteurizer plant task,” *Ergonomics*, **64**(9), pp. 1132–1145 (2021).
21. S. R. Dixon, C. D. Wickens, and J. S. McCarley, “On the Independence of Compliance and Reliance: Are automation false alarms worse than misses?” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **49**(4), pp. 564–572 (2007).
22. A. Jackson, T. Sato, J. Glassman, M. Politowicz, and E. T. Chancey, “Automation Error Bias, Trust, and Dependence Behaviors in a Simulated Drone Collision Avoidance Task,” Manuscript submitted for publication (2025).
23. T. B. Sheridan, “Individual Differences in Attributes of Trust in Automation: Measurement and application to system design,” *Frontiers in Psychology*, **10**, 1117 (2019).
24. M. H. Khalid, P. F. HB, A. Al Rashdan, and Z. Mohaghegh, “Automation Trustworthiness in Nuclear Power Plants: A literature review,” *Probabilistic Safety Assessment and Management (PSAM) International Topical Meeting on Artificial Intelligence (AI) and Risk Analysis* (2023).
25. F. Fukuyama, *Trust: The Social Virtues and the Creation of Prosperity*, Simon and Schuster, New York, USA (1995).
26. Y. Huang, L. Sun, H. Wang, S. Wu, Q. Zhang, Y. Li,... and Y. Zhao, “TrustLLM: Trustworthiness in large language models,” arXiv preprint arXiv:2401.05561 (2024).
27. B. Stanton and T. Jensen, “Trust and Artificial Intelligence,” *Preprint*, *NIST Interagency/Internal Report (NISTIR)*–8332 (2021).
28. A. D. Kaplan, T. T. Kessler, J. C. Brill, and P. A. Hancock, “Trust in Artificial Intelligence: Meta-analytic findings,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **65**(2), pp. 337–359 (2023).
29. A. Hall, R. L. Boring, and T. M. Miyake, “Cognitive Aging As a Human Factor: Effects of age on human performance,” *Nuclear Technology*, **209**(3), pp. 261–275 (2023).
30. Y. Yamani, A. Hall, T. M. Miyake, and J. Joe, “Cognitive Aging and Emerging Technologies in Nuclear Power Plants: The roles of mental models from a systems engineering perspective,” Manuscript submitted for publication (2025).
31. S. B. Sitkin and A. L. Pablo, “Reconceptualizing the Determinants of Risk Behavior,” *Academy of Management Review*, **17**(1), pp. 9–38 (1992).

32. T. Sato, Y. Yamani, M. Liechty, and E. T. Chancey, "Automation Trust Increases under High-Workload Multitasking Scenarios Involving Risk," *Cognition, Technology & Work*, **22**(2), pp. 399–407 (2020).
33. J. B. Lyons and C. K. Stokes, "Human–Human Reliance in the Context of Automation," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **54**(1), pp. 112–121 (2012).
34. S. Loft, A. Bhaskara, B. A. Lock, M. Skinner, J. Brooks, R. Li, and J. Bell, "The Impact of Transparency and Decision Risk on Human–Automation Teaming Outcomes," *Human Factors*, **65**(5), pp. 846–861 (2023).
35. L. Perkins, J. E. Miller, A. Hashemi, and G. Burns, "Designing for Human-Centered Systems: Situational risk as a factor of trust in automation," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, **54**, pp. 2130–2134 (2010).