DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States 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. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Reference herein to any social initiative (including but not limited to Diversity, Equity, and Inclusion (DEI); Community Benefits Plans (CBP); Justice 40; etc.) is made by the Author independent of any current requirement by the United States Government and does not constitute or imply endorsement, recommendation, or support by the United States Government or any agency thereof.

Pacific Northwest
NATIONAL LABORATORY

PNNL-37971

Yes, No, Maybe So

Human Factors Considerations for Fostering Calibrated Trust in Foundation Models Under Uncertainty

July 2025

Brandon D Dreslin
Jessica A Baweja

U.S. DEPARTMENT of ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-37971

Yes, No, Maybe So

Human Factors Considerations for Fostering Calibrated Trust in Foundation Models Under Uncertainty

July 2025

Brandon D Dreslin
Jessica A Baweja

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

High-stakes analytical environments require analysts to evaluate evidence and generate conclusions to inform critical decisions often under conditions of uncertainty. Researchers are developing expert systems built on foundation models (FMs) to support analysts' decision-making processes, in part by quantifying and expressing uncertainty information. To ensure effective human-artificial intelligence (AI) teaming, it is imperative to address analysts' needs when interpreting and using uncertainty information. However, it remains unclear how analysts engage with FM-generated uncertainty information and the extent to which these interactions influence trust in, and reliance on, expert systems. We will review the state of the science and propose our research design and methodology of an exploratory, qualitative study currently under review to (a) understand how properly communicated uncertainty fosters calibrated trust and appropriate reliance, and (b) identify strategies for conveying FM-generated uncertainty information during analytical work. Through semi-structured interviews, analysts will share their current experiences with job-related uncertainty and assess FM outputs that communicate uncertainty. The results will help us to understand how analysts currently interpret and use uncertainty information. Our findings may inform human factors recommendations for effectively conveying uncertainty information to foster calibrated trust in, and appropriate reliance on, expert systems. Practitioners can use this knowledge to enhance human-AI teaming and promote responsible FM-based expert system deployment.

Acknowledgments

This work was supported by the NNSA Office of Defense Nuclear Nonproliferation Research and Development, U.S. Department of Energy, and Pacific Northwest National Laboratory, which is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract DE-AC0576RLO1830. This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program.

Acronyms and Abbreviations

AI	Artificial intelligence
CLIP	Contrastive Language-Image Pre-training
DOE	U.S. Department of Energy
FM	Foundation model
GPT-3	Generative Pre-trained Transformer 3
IRB	Institutional Review Board
ML	Machine learning

Contents

1.0 Introduction

On July 3, 1988, the crew aboard the U.S.S. *Vincennes* faced a critical decision in the Persian Gulf. Engaged in a skirmish with Iranian gunboats, tactical personnel detected an approaching aircraft on radar. Officers reported that the aircraft was attacking because it was descending toward the ship, but its transponder indicated that it was a commercial flight. Multiple data sources provided conflicting signals about the aircraft's identity, altitude, and intentions. The crew's interpretation of this ambiguity led to a devastating decision: they fired two missiles at what they believed was an attacking Iranian F-14 Tomcat. However, the target was Iran Air Flight 655, a civilian Airbus A300 carrying 290 passengers and crew, all of whom were killed (U.S. Department of Defense, 1988). This tragedy demonstrates how the interpretation and communication of uncertain information can have catastrophic consequences. It represents a "just-so" story, a plausible but untestable narrative where the conclusions of "yes," "no," or "maybe so" appear equally valid.

Such "just-so" stories are characteristic of high-stakes analytical environments, where analysts systematically analyze information from multiple sources to assess threats and inform critical decisions (Drumhiller et al., 2024). This work inherently involves uncertainty because analysts must often draw conclusions from incomplete, conflicting, or ambiguous evidence under demanding circumstances (Amiram et al., 2018). For example, an analyst might need to assess the credibility of a threat based on one data source with unknown reliability, fragmented communications intercepts, or satellite imagery that shows unusual but not definitively suspicious activity. These judgments require both analytical skill and explicit communication of uncertainty information.

The challenge of effectively communicating uncertainty will become increasingly complex once artificial intelligence (AI) systems enter analytical workflows. Foundation models (FMs) are being developed to augment human analytical capabilities by processing vast amounts of data, identifying patterns, and generating insights more quickly than human analysts alone (Vaccaro et al., 2024). However, the future integration of these AI systems introduces new questions about how uncertainty should be communicated from machines to humans. While FMs can quantify and express their uncertainty in various ways, it remains unclear how analysts interpret and respond to this uncertainty information. Additionally, it is unknown whether uncertainty expression approaches can foster calibrated trust in, and appropriate reliance on, AI-assisted decision-making. To address these gaps, this paper proposes an exploratory, qualitative study design to better understand how analysts engage with uncertainty information both in their current work and from FM outputs. The study has not yet been conducted; this paper presents our research design and methodology.

2.0 Literature Review

2.1 Foundation Models (FMs) in Decision Support

FMs are large-scale AI models characterized by unprecedented scale, generality, and adaptability across diverse tasks through extensive pretraining on massive datasets (Bommasani et al., 2021). These models serve as general-purpose platforms that can be adapted for specialized applications, functioning through five major decision-making paradigms: optimization, prediction, planning, recommendation, and control (Schneider, 2022; Zhang et al., 2023). FMs encompass multiple types: natural language models like Generative Pre-trained Transformer 3 (GPT-3) demonstrate sophisticated text generation capabilities, while vision-language models such as Contrastive Language-Image Pre-training (CLIP) enable cross-modal understanding and image classification for improved zero-shot performance and flexible transferability (Awais et al., 2025; Zhou et al., 2024). However, evaluation and benchmarking remain challenging. Because significant gaps exist between pretraining tasks and real-world performance, there are ongoing concerns regarding data quality and alignment (Zhou et al., 2024).

End-user interactions with FMs present significant challenges for effective decision support. Users often develop inadequate mental models of these complex systems, leading to misaligned expectations about their capabilities and limitations (Passi & Vorvoreanu, 2022). The inherently unexplainable nature of AI systems (particularly very large models) limits transparency and user understanding (Bommasani et al., 2021; von Eschenbach, 2021). This “black box” problem becomes more pronounced as FMs exhibit emergent behaviors that cannot be directly understood or predicted. Thus, risks of under- or over-reliance based on perceived competence arise that may not be justified in real-world scenarios (Bommasani et al., 2021). A lack of transparency may challenge users who must make critical decisions based on system outputs without fully understanding the system’s underlying reasoning processes (Schneider, 2022). Given these challenges, it is important to understand how users develop and maintain trust in AI systems.

2.2 Trust in Artificial Intelligence (AI) Systems

Trust can be defined as the “reliance by an agent that actions prejudicial to their well-being will not be undertaken by influential others” (Hancock et al., 2011, p. 24). A meta-analysis of trust in AI reveals that appropriate trust formation depends on antecedents in three categories: characteristics of the trustor, trustee, and their shared situation (Kaplan et al., 2023). For example, trust may depend on an individual’s level of technology knowledge, the accuracy of the technology itself, or the difficulty of the task they are performing together. Glikson and Woolley (2020) explain that calibrating trust with system performance becomes difficult with increased model complexity because outputs cannot be interpreted as simply correct or incorrect. This requires users to consider the various ways in which outputs may be accurate or flawed (Glikson & Woolley, 2020). If trust remains uncalibrated, users may inappropriately mistrust or distrust the system, which can lead to actions of over-trust and under-trust (Jacovi et al., 2021). These actions may result in users relying on a system when it is incorrect or failing to rely on it when it is correct (Glikson & Woolley, 2020). This underscores the need for calibrated trust rather than simply high or low trust levels. Several mechanisms can be employed to achieve calibrated trust. From interviews with 12 expert decision-makers in various domains, Bedué and Fritzche (2022) report that access to knowledge, transparency, explainability, certification, and self-imposed standards and guidelines are most important for user trust

calibration. While calibration mechanisms guide the maintenance of appropriate trust levels, trust must first be established.

Research on trust formation shows several patterns and challenges in understanding human-AI relationships. Glikson and Woolley (2020) distinguish between cognitive trust (competence-based) and affective trust (emotional-based). Different system representations elicit different types of trust: whereas anthropomorphic agents (e.g., physical robots) generate more affective trust, feedback (e.g., transparency cues) strengthens cognitive trust. Multiple factors influence how trust is formed and the types of trust that manifest. In a review of 23 empirical studies, Bach et al. (2024) conclude that user characteristics (e.g., gender, education, AI experience/interactions) are the most influential to trust formation. However, the authors also suggest that socio-ethical considerations (e.g., continuous feedback, legal boundaries) and technical and design features (e.g., anthropomorphism, social presence) are paramount, especially when considering the diverse contexts and environments in which AI systems can be used. Moreover, other factors influence trust formation such as task attributes (e.g., complexity, time pressure) and system properties (e.g., robustness, reliability, transparency; Bach et al., 2024). These influential factors may interact with each other. For instance, Bedué and Fritzsche (2022) illustrate how user characteristics interact with system properties through their finding that AI certifications increase novice user trust but may lower expert trust when viewed as superficial marketing. These findings underscore that trust in AI systems is a complex result of the user, the technology, and the situation in which they are operating. Central to trust formation is the clear communication of expected system performance to users. With FMs, this involves appropriately conveying how certain or uncertain the model is about its results.

2.3 Uncertainty Information

Uncertainty in AI systems stems from multiple sources and manifests in different forms. Jalaian et al. (2019) identify various sources of uncertainty in AI, including model selection, data noise, and extrapolation beyond training contexts. Two types of uncertainty exist: (1) aleatoric uncertainty, which represents irreducible randomness inherent in data, and (2) epistemic uncertainty, which reflects limitations in model knowledge that could potentially be addressed with additional information (Abdar et al., 2021; Bhatt et al., 2021). Senge et al. (2014) clarify that distinguishing between these uncertainty types is crucial in domains like medical diagnosis for determining whether additional data collection is warranted. Beyond these traditional categories, Wenskovitch et al. (2024) introduce the concept of interaction uncertainty in human-machine teaming, which arises from mismatches in behavior, communication, or goal understanding between humans and AI systems.

The quantification and expression of uncertainty information significantly impact how users interpret and respond to FM outputs. Emerging uncertainty quantification methods for FMs include Bayesian methods, ensemble approaches, conformal prediction, and entropy-based metrics (Abdar et al., 2021). Bhatt et al. (2021) propose uncertainty quantification as a complement to explainability for achieving algorithmic transparency, advocating for visual and probabilistic formats to communicate uncertainty to different stakeholder groups. However, quantifying uncertainty is only part of the challenge; the format and framing of expressing uncertainty also matter. Dhami et al. (2025) compared verbal probability expressions with visual encodings for uncertainty communication in intelligence analysis, finding that analysts were slightly more sensitive to verbal probability cues but showed poor consistency across all formats. Further, the linguistic framing of uncertainty communications influences user perceptions and decision-making: first-person expressions like “I am certain” are perceived differently depending on speaker expertise compared to general perspective statements like “It

is uncertain" (Juanchich et al., 2017). This suggests that how FMs quantify and express uncertainty can affect both judgment and decision-making.

Users can employ various strategies to reduce uncertainty when interacting with AI systems. In their seminal work on uncertainty reduction during interpersonal communication, Berger and Calabrese (1975) present a three stage-based theory with six axioms and 21 theorems on the psychosocial behavior people exhibit to respond to and reduce uncertainty when uncertain information is presented. As an example, as the amount of verbal communication between strangers increases during the entry stage, the level of uncertainty each person perceives will decrease. Kramer (1999) reconceptualizes this theory for organizational contexts, proposing a "Motivation to Reduce Uncertainty" model influenced by context, goals, and uncertainty tolerance. Reduction strategies can be categorized as passive (observing entity behavior), active (seeking information from external sources), and interactive (directly engaging with the entity). Wenskovitch et al. (2024) highlight that uncertainty in human-AI collaboration is bidirectional because machines may also experience uncertainty about human behavior (e.g., in understanding the task that the user is trying to complete). The bidirectional nature of human-AI uncertainty presents unique challenges for ensuring that uncertainty can be effectively reduced not only for the human decision-maker, but also for the AI assistant. Building on these theoretical foundations, research has begun to investigate how humans actually perform when provided with AI-generated uncertainty information.

2.4 Human Performance Under AI-Generated Uncertainty

When properly communicated, uncertainty information can significantly improve human decision-making performance. Schaekermann et al. (2020) compared conventional AI assistants with ambiguity-aware systems that highlighted cases likely to lead to expert disagreement and presented arguments for conflicting classification choices. Their ambiguity-aware AI altered expert workflows by significantly increasing the proportion of contentious cases reviewed, with the relevance of AI-provided arguments affecting experts' accuracy at revising AI-suggested labels. Users benefit most when uncertainty information is paired with clear explanations of underlying reasoning, particularly regarding why the system is uncertain rather than simply being told that uncertainty exists. This approach enables rapid trust calibration and helps users adjust their reliance levels appropriately, emphasizing the importance of interpretable uncertainty that clarifies the source, scope, and implications of uncertain recommendations (Tomsett et al., 2020).

The communication of uncertainty information critically influences user responses, particularly in how users recover from AI errors. Siegling (2020) demonstrates that when AI systems disclose their uncertainty, users experience less severe trust degradation following system failures and develop better awareness of system limitations. Participants preferred AI systems that communicated uncertainty, perceiving them as more trustworthy and valuable compared to systems that did not disclose uncertainty. However, the design must balance clarity with information visualization, as overly complex visualizations can reduce utility despite their transparency benefits. For example, Reyes et al. (2025) reveal that the size and visual prominence of uncertainty displays emerge as key factors affecting both trust and decision confidence, with continuous uncertainty visualization significantly enhancing trust for 58% of participants who initially held negative attitudes toward AI.

Task characteristics and individual differences can moderate how uncertainty affects human-AI collaboration. Salimzadeh et al. (2024) found that complex and uncertain tasks lead users to rely more heavily on AI systems while paradoxically demonstrating lower appropriate reliance

compared to simpler tasks. Interestingly, their research reveals that trust in AI systems appears less sensitive to task characteristics than reliance behavior, suggesting these constructs respond differently to uncertainty. User expertise also plays a crucial role in uncertainty interpretation. For example, novice users reported slowing down and thinking more analytically about their decisions when uncertainty about machine learning (ML) predictions were properly communicated, demonstrating greater vigilance and reducing overreliance (Prabhudesai et al., 2023). Beyond individual responses, users actively develop strategies to manage uncertainty when interacting with AI systems. Chang et al. (2025) found that interactive approaches such as asking follow-up questions were more effective than passive observation strategies, with their study of 566 users revealing that consulting peer feedback was the most effective strategy for reducing uncertainty while transparency concerns, information accuracy issues, and privacy worries serve as key sources of uncertainty that users must navigate. Given these findings about the complex relationship between uncertainty communication and human performance, it becomes imperative to understand how analysts engage with uncertainty information to make decisions during their analytical work.

3.0 Methods

The following study is under review with the Central Department of Energy (DOE) Institutional Review Board (IRB). We propose a qualitative study to understand how analysts working in a specific high-stakes analytical environment conceptualize, produce, and use uncertainty information in their analytical work, with the goal of informing FM-based AI system design for effective uncertainty communication. We aim to explore how analysts currently understand and interpret uncertainty, gather feedback on different approaches to uncertainty, and identify their preferences for uncertainty representation in FM-generated outputs.

3.1 Participants

We will recruit analysts working at four national laboratories. Participants must have experience conducting analytical work in high-consequence decision environments. They will have varying levels of familiarity with AI/ML systems. We will collect demographic information related to their analytical background, technical knowledge, and experience with AI/ML systems.

3.2 Measures

Semi-structured interviews will explore participants' current experiences with uncertainty information in their analytical work and their interpretations of AI-generated uncertainty outputs. The interview protocol will include questions about how analysts currently understand, use, and communicate uncertainty information in their professional practice. Additionally, participants will discuss their thoughts, beliefs, and reactions regarding the presentation of several FM outputs containing different representations of uncertainty information. The interviews may also capture participants’ impressions of how uncertainty information might influence their trust in, and reliance on, FM-based AI systems.

3.3 Procedure

Following informed consent, we will conduct individual semi-structured interviews with each participant. Interviews will begin with questions about participants' current experiences with uncertainty in their analytical work, including how they interpret, use, and communicate uncertain information. Then, participants will view examples of FM outputs that display uncertainty information in different formats and provide feedback on their interpretations and preferences. Finally, we will ask participants to share demographic information related to their experience and expertise. Throughout the interview, participants can share their thoughts about how uncertainty information might affect their trust in AI-assisted analysis.

3.4 Data Analysis

We will employ thematic analysis to identify patterns and themes in participants' responses regarding their experiences with and interpretations of uncertainty information. Following data cleaning and preparation, we will develop a coding framework to systematically analyze interview notes and transcripts. We will leverage the resulting themes to create an affinity diagram, which may allow us to organize and visualize relationships between different themes and concepts that emerge from the data. This analytical approach will help us identify common patterns in how analysts conceptualize uncertainty, their preferences for uncertainty representation, and factors that influence their trust and reliance on FM-generated uncertainty information.

4.0 Anticipated Outcomes

4.1 Research Objectives

This study may reveal several key insights about how analysts work with uncertainty information. For example, analysts might have domain-specific conceptualizations of uncertainty that differ from technical definitions used in AI research. We will seek to identify current practices for communicating uncertainty in reporting and discover gaps between these practices and what analysts need from AI systems. The study also aims to reveal how different types of uncertainty information are valued depending on the analytical situation. As such, the interviews will explore whether analysts' preferences for uncertainty representation vary based on their specific tasks, expertise levels, and analytical contexts. Additionally, findings will inform whether analysts' trust in and reliance on AI-generated information is influenced by how uncertainty is communicated, with some representation formats potentially fostering more appropriate trust calibration than others.

4.2 Limitations

Several limitations may affect the generalizability and interpretation of our findings. First, while a small sample size is appropriate for qualitative research, it may limit the breadth of perspectives captured during this study. Additionally, the cross-sectional nature of the study means we will capture analysts' perspectives at a single point in time rather than observing how their views change with experience. Further, the study focuses on analysts from national laboratories, which may not represent the full range of analytical environments. Finally, findings may be specific to the high-consequence domain examined here and may not generalize to other types of analytical work or high-consequence domains.

4.3 Contributions to Knowledge

Our research seeks to contribute to multiple areas of knowledge. For instance, the findings should enhance understanding of how domain experts conceptualize and use uncertainty information in high-consequence decision environments. The goal is to provide practical insights for AI system designers about how to effectively communicate uncertainty information to decision makers. Thus, our findings aim to inform the development of more effective uncertainty visualization and communication strategies for FMs used in analytical workflows. Overall, this study attempts to bridge the gap between technical uncertainty quantification methods and user needs of uncertainty representation in real-world analytical contexts.

4.4 Future Work

Several research directions build on this work. A controlled experiment that presents participants with different uncertainty representations and measures how their trust, reliance, and decision-making accuracy differ between each representation would help to explore whether these representations influence those factors. Within high-stakes analytical work, a longitudinal study could examine how analysts' trust in and use of AI-generated uncertainty information changes over time with repeated exposure. Outside of the domain examined here, research across other high-consequence domains could investigate whether findings from this study generalize to other tasks such as financial analysis, medical diagnosis, or air traffic control. Additionally, it may be fruitful to explore how explanations of AI system conservativeness or system limitations affect trust calibration when uncertainty information is

presented. This may be especially prevalent if future work examines how individual differences (e.g., age, gender, risk tolerance) influence trust formation and calibration under uncertainty.

5.0 Conclusion

When FMs enter analytical workflows, the challenge of effectively communicating uncertainty between AI systems and human analysts will become critical for national security decision-making. The proposed research addresses a significant gap in understanding how analysts conceptualize, interpret, and respond to uncertainty information, particularly as generated by AI systems. Through qualitative interviews with analysts, this study aims to bridge the divide between technical uncertainty quantification methods and the practical needs of domain experts who must make consequential decisions based on uncertain information. The findings may inform the design of more effective human-AI collaboration systems and contribute to appropriate trust calibration in FM-assisted analysis. Ultimately, this research seeks to ensure that the integration of AI into workflows enriches rather than undermines analysts' ability to navigate the complexity of “yes,” “no,” or “maybe so” possibilities that characterize their indispensable work.

6.0 References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Lu., L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion, 76*, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008

Amiram, D., Landsman, W. R., Owens, E. L., & Stubben, S. R. (2018). How are analysts' forecasts affected by high uncertainty? *Journal of Business Finance and Accounting, 45*(3–4), 295–318. https://doi.org/10.1111/jbfa.12270

Awais, M., Naseer, M., Khan, S., Anwer, R. M., Cholakkal, H., Shah, M., Yang, M-H., & Khan, F. S. (2025). Foundation models defining a new era in vision: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 47*(4), 2245–2264. https://doi.org/10.1109/TPAMI.2024.3506283

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2024). A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human-Computer Interaction, 40*(5), 1251–1266. https://doi.org/10.1080/10447318.2022.2138826

Bedué, P., & Fritzsche, A. (2021). Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. *Journal of Enterprise Information Management, 35*(2), 530–549. https://doi.org/10.1108/JEIM-06-2020-0233

Berger, C. R., & Calabrese, R. J. (1975). Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human Communication Research, 1*(2), 99–112. https://doi.org/10.1111/j.1468-2958.1975.tb00258.x

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G. G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., & Xiang, A. (2021). Uncertainty as a form on transparency: Measuring, communicating, and using uncertainty. In *AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 401–413). Association for Computing Machinery. https://doi.org/10.1145/3461702.3462571

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C.,… Liang, P. (2022). *On the opportunities and risks of foundation models*. arXiv. https://doi.org/10.48550/arXiv.2108.07258

Chang, Y.-H., Silalahi, A. D. K., & Lee, K.-Y. (2025). From uncertainty to tenacity: Investigating user strategies and continuance intentions in AI-powered ChatGPT with uncertainty reduction theory. *International Journal of Human-Computer Interaction, 41*(11), 6570–6588. https://doi.org/10.1080/10447318.2024.2381930

Dhami, M. K., Witt, J. K., & De Werd, P. (2025). Visualizing versus verbalizing uncertainty in intelligence analysis. *Intelligence and National Security, 40*(2), 302–327. https://doi.org/10.1080/02684527.2025.2468049

Drumhiller, N. K., Burch, J., & Skvorc, C. (2024). Warning intelligence and high consequence environments: A comparative assessment to integrate human factors to support warning

analysis. *Journal of Policing, Intelligence and Counter Terrorism, 19*(4), 448–465. https://doi.org/10.1080/18335330.2024.2367500

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Hancock, P. A., Billings, D. R., & Schaefer, K. E. (2011). Can you trust your robot? *Ergonomics in Design, 19*(3), 24–29. https://doi.org/10.1177/1064804611415045

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 624–635). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445923

Jalaian, B., Lee, M., & Russell, S. (2019). Uncertain context: Uncertainty quantification in machine learning. *AI Magazine, 40*(4), 40–49. https://doi.org/10.1609/aimag.v40i4.4812

Juanchich, M., Gourdon-Kanhukamwe, A., & Sirota, M. (2017). "I am uncertain" vs. "It is uncertain": How linguistic markers of the uncertainty source affect uncertainty communication. *Judgment and Decision Making, 12*(5), 445–465. https://doi.org/10.1017/S1930297500006483

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2021). Trust in artificial intelligence: meta-analytic findings. *Human Factors, 65*(2), 337–359. https://doi.org/10.1177/00187208211013988

Kramer, M. W. (1999). Motivation to reduce uncertainty: A reconceptualization of uncertainty reduction theory. *Management Communication Quarterly, 13*(2), 305–316. https://doi.org/10.1177/0893318999132007

Passi, S., & Vorvoreanu, M. (2022). *Overreliance on AI: Literature review* (Report No. MSR-TR-2022-12). Microsoft Corporation. https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/

Prabhudesai, S., Yang, L., Asthana, S., Huan, X., Liao, Q. V., & Banovic, N. (2023). Understanding uncertainty: How lay decision-makers perceive and interpret uncertainty in human-AI decision-making. In *IUI '23: Proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 379–396). Association for Computing Machinery. https://doi.org/10.1145/3581641.3584033

Reyes, J., Batmaz, A. U., & Kersten-Oertel, M. (2025). Trusting AI: Does uncertainty visualization affect decision-making? *Frontiers in Computer Science, 7*, Article 1464348. https://doi.org/10.3389/fcomp.2025.1464348

Salimzadeh, S., He, G., & Gadiraju, U. (2024). Dealing with uncertainty: Understanding the impact of prognostic versus diagnostic tasks on trust and reliance in human-AI decision-making. In *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). Association for Computing Machinery. https://doi.org/10.1145/3613904.3641905

Schaekermann, M., Beaton, G., Sanoubari, E., Lim, A., Larson, K., & Law, E. (2020). Ambiguity-aware AI assistants for medical data analysis. In *CHI '20: Proceedings of the 2020 CHI*

Conference on Human Factors in Computing Systems (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376506

Schneider, J. (2022). *Foundation models in brief: A historical, socio-technical focus*. arXiv. https://doi.org/10.48550/arXiv.2212.08967

Senge, R., Bösner, S., Dembczyński, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., & Hüllermeier, E. (2014). Reliable classification: Learning classifiers that distinguish between aleatoric and epistemic uncertainty. *Information Sciences, 255*, 16–29. https://doi.org/10.1016/j.ins.2013.07.030

Siegling, L. (2020). *Uncertainty communication by AI assistants: The effects on user trust* [Master's thesis, Utrecht University]. Utrecht University Student Theses Repository. https://studenttheses.uu.nl/handle/20.500.12932/39347

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., & Kaplan, L. (2020). Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns, 1*(4), 1–9. https://doi.org/10.1016/j.patter.2020.100049

U.S. Department of Defense. (1988). *Formal investigation into the circumstances surrounding the downing of Iran Air Flight 655 on 3 July 1988* (DTIC Publication No. ADA203577). Defense Technical Information Center. https://apps.dtic.mil/sti/citations/ADA203577

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour, 8*(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology, 34*, 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Wenskovitch, J., Fallon, C., Miller, K., & Dasgupta, A. (2024). Characterizing interaction uncertainty in human-machine teams. In *IEEE ICHMS 2024: 2024 IEEE 4th International Conference on Human-Machine Systems* (pp. 1–6). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICHMS59971.2024.10555605

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P. S., & Sun, L. (2024). A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT. *International Journal of Machine Learning and Cybernetics*, 1–65. https://doi.org/10.1007/s13042-024-02443-6

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov