Risk Analysis and Mitigation of Human Interface with Generative Artificial Intelligence Systems to Enable Responsible Development

Jennifer Giang
Colorado State University
Engineering Building 202
6029 Campus Delivery
Fort Collins, CO 80523-6029
513-374-4347
jennigia@colostate.edu

Steven J. Simske
Colorado State University
Engineering Building 202
6029 Campus Delivery
Fort Collins, CO 80523-6029
970-214-5692
steve.simske@colostate.edu

Copyright © 2024 by Jennifer Giang and Steven Simske. Permission granted to INCOSE to publish and use.

Abstract. As Generative Artificial Intelligence (GAI) becomes increasingly prevalent in society, ensuring responsible and ethical development and use is crucial. This study conducts a risk analysis of human interface with GAI Large Language Models (LLM) to identify potential hazards and suggest mitigation strategies for promoting responsible GAI development. This study is component of a larger research project that is proposing the adaptation of testing strategies for responsible development of GAI. Risks included in this analysis are bias/discrimination, security/privacy concerns, and lack of transparency/reliability are assessed based on probability, impact of cost, schedule, and performance criteria. Following the implementation of mitigation strategies, a re-evaluation of the risks is conducted to gauge their adjusted system risk. The findings show that implementing targeted mitigation strategies can effectively reduce the likelihood and severity of risks associated with human interaction and GAI systems, thus enabling the development of more responsible and ethically sound AI technology. This study contributes to the ongoing discourse on responsible AI development and provides practical insights for organizations seeking to navigate the complexities of human-AI interaction responsibly.

Identify Human-AI Interface Risks

Human interaction is the foundation for AI systems, but it introduces a variety of risks that are likely to not only impact an individual, but society. Humans contribute to AI through system development, research, ethical oversight, and practical application. Humans and AI collaborate by leveraging AI's processing ability while humans provide contextual understanding, creativity, and oversight. Together, there will be enhanced decision making, productivity, and innovation. A previous study proposes a general 18 design guideline for human-AI interaction (Amershi, et al. 2019). The ramifications of these risks can vary widely, from positive advancements to negative consequences of significant magnitude. The most common risks are:

Bias/Discrimination

There is perpetually the risk of GAI systems having heavy bias that can stem from both flawed data and unsophisticated algorithms. Research from UNESCO showed cultural and gender bias when prompting GPT-2 and Llama2 about occupation recommendations for British and Zulu men

and women. The study observed varied occupations for British men, such as driver, caregiver, bank clerk. For British women, the study observed stereotypical occupations such as prostitute, model, and waitress. For the Zulu men, the study observed suggestions such as gardener and security guard. The GAI system learns from the training data which would lead to biased decisions and discriminatory outcomes. This is a classic example of garbage-in and garbage-out. This study shows the inaccuracy due to the GAI's unfair discrimination. In other scenarios, this could cause a major impact to society if implemented such as the justice system, health care, and general ethical decision making.

Security/Privacy Concerns

It is known that GAI systems require significant data sets to train and evaluate the system. Dependent on the data, mishandling of the data could lead to privacy breaches, identity theft, and unauthorized access to sensitive information. AI systems are vulnerable to security attacks such as data poisoning and model inversion attacks that can occur during any phase of the AI workflow. Data poisoning is a common attack where (perhaps overwhelming amounts of) adversarial data are injected into the training data to either compromise the integrity of the system to operate in the desired behavior or impact the availability of the system. Model inversion attacks are when the output data is compromised and used to infer the model parameters and/or architecture. This puts the model at risk of being copied or extraction of sensitive information.

Lack of Transparency/Reliability

There is a concern regarding the lack of transparency a GAI system has because it operates similarly to a black box which does not allow users to understand how the model came to its output conclusions. This has caused users to lack trust in AI systems. In addition, if the system is unpredictable/unreliable with its outputs then, users will not be able to know how to adjust the parameters or training data to achieve the desired results. Without the ability to see the inner workings of the system, users face significant obstacles in optimizing their performance reliability.

Addressing these risks requires a significate effort to mitigate biases, fortify security measures, and enhance transparency. Addressing these risks can foster an environment conducive to the development of Responsible AI.

Risk Analysis of Human-GAI

The high-level risks identified are confirmed by a recent survey that shows GAI risks that organizations consider relevant (Sukharevsky, et al., 2024). Risk analyses provide qualitative and quantitative representation of which risks are the most likely to impact the organization and its impact. The authors performed a risk analysis of the three major risks discussed by multiplying probability and impact (to cost, schedule, and performance mean), each on a scale of 1-5, to calculate the overall risk score. Referencing a stand 5x5 risk matrix to evaluate the overall risk, Figure 1.

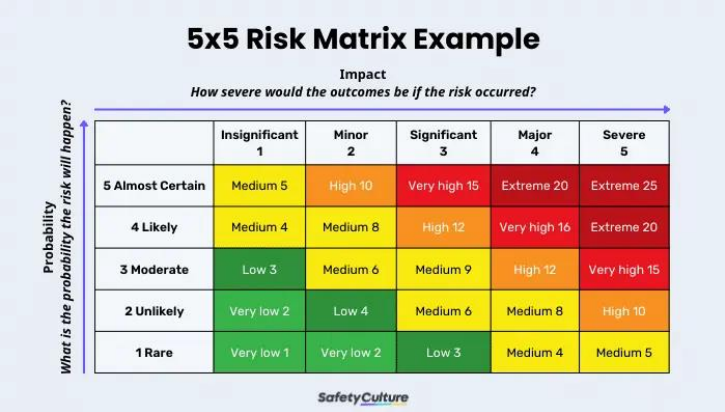

Figure 1: 5x5 Risk Matrix (Guevara, 2024)

Probability Scale (1-5) is defined by these metrics: 1 – Very Low – less than 10%, 2 – Low – 10-30%, 3 – Medium – 30-50%, 4 – High – 50-70%, 5 – Very High – Greater than 70% change of occurring.

Cost, Schedule, and Performance Impact Scale (1-5) is defined by these metrics: 1 – Very Low – Negligible increase in cost/schedule/performance (<1% of project budget/delay/barely noticeable), 2 – Low – Minor increase in cost/schedule/performance (1-5% of project budget/delay/minor issues, easily mitigated), 3 – Medium – Moderate increase in cost/schedule/performance (5-10% of project budget/delay/some degradation, manageable), 4 – High – Significant increase in cost/schedule/performance (10-20% of project budget/delay/noticeable degradation, difficult to manage), 5 – Very High – Major increase in cost/schedule/performance (>20% of project budget/delay/severe degradation, potentially critical).

Risk	Probability	Impact			Risk Score
		Cost	Schedule	Performance	
Bias/Discrimination	4	4	3	4	**16**
Security/Privacy	4	4	3	3	**12**
Transparency/Reliability	3	3	3	4	**10**

Table 1: Human - AI Systems Risk Assessment

The probability of bias/discrimination in an AI system is high because it depends on the quality, balance of experimental and synthetic training data, algorithmic design, and presence of biases in development. Bias can also be introduced during the human decision-making process unintentionally. The impact of bias/discrimination can be significant, leading to unfair treatment, undermining trust, biased legal decision making. This is confirmed by the 2024 AI Index which reports that AI systems still exhibit biases that can lead to discriminatory outcomes.

The probability of security and privacy concerns in AI systems is immediately at risk because it varies based on the complexity of the system and sensitivity of the data. Security vulnerabilities can stem from design flaws, implementation errors, and malicious attacks. The impact of security and privacy can be severe if there are data breaches, unauthorized access to sensitive information,

and damage to an individual's privacy or organization's reputation. This is confirmed by the 2024 AI Index that notes an increased number in AI-related incidents that include security breaches.

The probability of an AI system having a lack of transparency and reliability is high because it varies based on opacity of the AI system decision-making process, algorithm complexity, and level of human oversight. The impact can be significant due to the black-box environment causing uncertainty of the reliability of the decisions and challenges in diagnosing errors or assessing system performance. This is hardened by the 2024 AI Index which mentions a lack of standardized evaluations and transparency in model training and processes which makes it difficult to ensure reliability.

Human-AI Mitigation Strategies

To decrease the overall risk to an organization, mitigation steps or strategies should be taken. For the identified risks, some mitigation strategies may include:

Bias/Discrimination

- Diversify Data: Ensure that training data used to develop AI models is diverse, representative, and free from biases (Thomas 2024)
- Bias Detection and Mitigation: Implement techniques for detecting and mitigating bias in AI models, such as fairness-aware algorithms, bias audits, and adversarial testing.
- Transparency: Increase transparency in AI decision-making processes by documenting model training procedures, data sources, and evaluation metrics.

Security/Privacy

- Data Encryption and Access Controls: Implement robust security measures to protect sensitive data, such as encryption, access controls, and authentication mechanisms. (Kamp 2023). Limit access to data on a need-to-know basis and regularly audit user permissions to prevent unauthorized access. Provide digitally signed backup and use multi-factor authentication (MFA). In fact, the combination of human-in-the-loop and AI-driven access control can provide more robust MFA workflows than currently exists.
- Responsible Development Practices: Follow responsible development practices, such as code reviews, vulnerability assessments, and penetration testing, to identify and address security vulnerabilities in AI systems.
- Data Minimization and Anonymization: Minimize the collection and retention of personal data to reduce the risk of data breaches and privacy violations.

Transparency/Reliability

- Explainable AI (XAI) Techniques: Employ techniques for explainable AI (XAI) to enhance transparency and interpretability of AI systems. This may include using techniques such as model interpretability, feature importance analysis, and decision rule extraction to provide insights into model behavior.
- Model Validation and Testing: Implement rigorous validation and testing procedures to assess the reliability and performance of AI models.
- Human-in-the-Loop Approaches: Incorporate human-in-the-loop approaches to enhance the reliability and accountability of AI systems.

By implementing these mitigation strategies, organizations can address the risks. We performed a risk assessment of the same major risks with the mitigation strategies.

Risk	Probability	Impact			Risk Score
		Cost	Schedule	Performance	
Bias/Discrimination	2	3	2	3	6
Security/Privacy	2	3	2	2	4
Transparency/Reliability	2	3	2	3	6

Table 2: Human - AI Systems Using Mitigation Strategies Risk Assessment

In conclusion, the integration of humans and AI systems presents transformative opportunities (especially for enhanced security in MFA and parallel authorization workflows), but significant challenges. To ensure responsible AI development and deployment, it is essential to proactively mitigate biases, enhance security measures, and improve transparency and reliability in AI systems. Collaborative efforts among stakeholders are crucial to addressing these challenges and fostering an ecosystem that prioritizes transparency, fairness, and accountability in AI innovation. By embracing responsible AI practices, we can harness the potential of AI technology to drive positive social impact while upholding ethical principles and promoting equitable access to opportunities, thus paving the way for a more inclusive and sustainable future.

References

Guevara, P. (2024, March 27). 'A Guide to Understanding 5x5 Risk Assessment Matrix', Article, https://safetyculture.com/topics/risk-assessment/5x5-risk-matrix/

Kamp, K. (2023, November 12). *Responsible data practices: How businesses can leverage first-Party data ethically: Marin software blog*. Marin. https://www.marinsoftware.com/blog/responsible-data-practices-how-businesses-can-leverage-first-party-data-ethically

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 3, 1–13. https://doi.org/10.1145/3290605.3300233

Singla, A., Sukharevsky, A., Yee, L., & Chui, M. (2024, May 30). *The state of ai in early 2024: Gen ai adoption spikes and starts to generate value*. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

Stanford University. (2024). *Artificial Intelligence Index Report 2024*. Stanford University - AI Index. https://aiindex.stanford.edu/report/

Thomas, M. (2024, February 28). *How to mitigate risk in AI software development*. Split. https://www.split.io/blog/how-to-mitigate-risk-in-ai-software-development/

UNESCO, & International Research Centre on Artificial Intelligence. (2024). *Challenging systematic prejudices: an investigation into bias against women and girls in large language models*. UNESDOC Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000388971

Biography

Jennifer Giang is pursuing a Ph.D. in Systems Engineering through Colorado State University with a dissertation focusing on "Adapting Testing Strategies for Responsible Development of GAI". She is currently a Systems Engineering Lead at Sierra Space Corporation where she leads an interdisciplinary team. Additionally, she is part of the INCOSE Technical Leadership Institute, a two-year program addressing global systems engineering challenges. She has presented at several INCOSE, AIAA, and CSER conferences on SE4AI and AI4SE topics.

Steven J Simske. Since 2018, Steve has been a Professor of systems engineering with Colorado State University (CSU). At CSU, he has a cadre of on-campus students in systems, mechanical, and biomedical engineering, and a larger contingent of online/remote graduate students researching various disciplines. At NASA and HP Inc, he directed teams to research 3D printing, education, life sciences, sensing, authentication, packaging, analytics, imaging, and manufacturing. He has written four books on analytics, algorithms, and steganography. He is the author of 500 publications and 240 U.S. patents. His research interests include analytics, systems security, sensing, signal and imaging processing, printing and manufacturing. He is an NAI Fellow, an IEEE Fellow, an IS&T Fellow, and was awarded a CSU Best Teacher Award in 2022.