Security, Trust, and Use of AI

Alex Shah

EN.695.741.81.SP25 Information Assurance Analysis

Mod 7 Assignment

March 9, 2025

Table of Contents

Introduction....2
Safety....2
Trust....3
Cybersecurity....4
Privacy....4
ChatGPT....5
Conclusion....6
References....6
Appendix....7

Introduction

Machine Learning and Artificial Intelligence models like ChatGPT have seen widespread use in industries like healthcare, finance, and logistics. As AI becomes more commonplace, questions about trusting AI's decisions, as well as ensuring privacy and security around AI systems are being asked more frequently. Answering these questions about AI systems will depend on the level of validation and transparency AI systems designers are able to integrate from the ground up when training a new model from data sources and performing inferences. And also user privacy and security will also be an ongoing and continuous process in an AI system's lifecycle as data is used to train models and refine them from user interaction.

Safety

Safety in AI systems involves complex design decisions from the time the model is trained to the inference stage when models may be used to make decisions. In the current generation of models, a large amount of high quality data is required to produce useful models. A data set used to train an AI system could contain inaccuracies or bias that effect the results and performance of the trained model.

For AI to behave safely, human designers need to examine the data sources the model is trained on for these biases and inaccuracies in order to ensure that in use the model doesn't instill bias in the predictions and decisions it leads humans to. For example, AI and ML models are increasingly being used in deciding loans in financial sectors where bias could affect real people's lives. Data is often gathered from individuals to train a model, and personal information can become embedded in the weights these models build up. Through continuous learning implementations real user interactions are used to train the model further, which can embed sensitive information. Srinivas for ManageEngine states that "The technology requires large amounts of data to work effectively, which can lead to privacy concerns. The more data you feed into the algorithm, the more accurate and personalized the content it generates becomes. However, this also means that personal data is being used, which can be a cause for concern." (Srinivas, 2024). AI is also being used in weapons and defense systems. AI and ML predictions have come a long way and in some cases are a leap over existing systems for some tasks, but they still have the potential to deliver false positives and negatives in critical systems that could be relied upon for safety and in fast paced environments where humans may not have the chance to verify the results.

Trust

AI as a "black box" system is closed off from directly analyzing what factors weighed in making a decision or generating an output. For most casual users, the reasoning and activated neurons are not as important as the results that the model provides. However, with a closed system, there is a lack of transparency in the model for oversight and for compliance and regulatory reasons that determine the model's safety and reliability. "For a long time, making the code open source was seen as a way of transparency in software, but for very complex systems such as autonomous systems, opening up the source code may not even be enough to know how the system works" (Such, 2017, p. 4). Looking into the way the model was trained, the data that it was trained on, and gathering as much contextual information as possible, we have not been able to look inside the inner functioning of these AI models, making them inherently untrustworthy. Securing AI systems is a critical step in using them safely, and model transparency is necessary to start trusting them and using them effectively.

Cybersecurity

Despite the drawbacks, AI systems can be used effectively to enhance existing cybersecurity frameworks such as reducing the false positive rate in some detection engines for malware and vulnerability testing. However, the systems themselves are vulnerable to attack, sometimes in novel or difficult to detect ways. For example, one of the biggest risks is data poisoning where attackers can manipulate training data to alter the model's behavior. (Cloudflare, 2025). There are mitigations to these threats through the use of adversarial training to help a model become resistant to malicious inputs, and stricter validation and monitoring of training. While there are many existing security and infrastructure related cybersecurity guidances and frameworks such as NIST, Cloud Security Alliance, and regulatory/compliance requirements like HIPAA, the software, design, and use of these models can still be tainted while maintaining compliance. It is very difficult to detect a tainted model from the inference stages, and operational practices after training won't be able to prevent a model being trained on manipulated data.

Privacy

AI algorithms themselves do not ensure user privacy. Rather it is up to those designing the systems, who also collect and control data to ensure that the system respects principles around privacy.

Privacy by design is an important strategy to ensure user privacy in an autonomous system like an AI product. The designers of the system would need to consider how data is gathered, inform users about what is collected, and aggregate or remove PII from the collected data. In addition, designers should use best practices and guidance from cybersecurity frameworks around securing data in transit and at rest, enforcing access control, and using logging and auditing to prevent data tampering. For the AI system itself, fair systems are an ongoing area of research including attempts to make AI systems behave fairly when weighing decisions based on personal data. For example in Dwork et al, the authors establish a similarity metric to determine how similar a pair of individuals are within the context of a particular classification task, where the metric would be determined externally. These types of systems are practical in theory but rely on manually described metrics that would be “agreed upon by society”. It is unlikely that a regulatory body or civil rights organization would be able to create a comprehensive list or a practical application of such a system. This leaves it in the hands of those designing the AI systems to implement good practices up front through the training and inferences stages to align the models to the best of their ability.

ChatGPT

ChatGPT circles the problems it is presented with while refraining from diving deeper into topics without the user’s guidance, so I find the description fitting. It seems to list several buzzwords or considerations in an attempt to view the topic from a central or neutral perspective while presenting the information in an authoritative tone that describes what “should” be done. So while it might answer a question in a definitive way, the answer itself attempts to be a non exhaustive list of possibilities. Further prompting can narrow down the answer or find more useful or applicable information. For example, when asked about AI safety ChatGPT gave a list of ethical or broad topics. It failed to provide any specific methods or frameworks that could be used to achieve the goals. In comparison, my answer to the question brought up areas where AI could not behave safely and provided examples and potential solutions. However when asked about Cybersecurity in AI systems, the model was at least able to provide some specific cybersecurity frameworks around AI and compliance such as naming ISO 27001. Without a human user to probe the concepts further, ChatGPT is willing to look at the surface of the water from all angles, whereas a human researcher is likely to dive in quicker and more willingly find something to examine closely such as finding sources, looking at real implementations, and mapping related topics.

Conclusion

Trust, safety, and security in AI systems will continue to be a challenge as AI is adopted by more industries and grows in capability. AI can be used in critical applications, and right now models lack explainability and might contain bias or misinformation from the data it is trained on. However with privacy by design principles, technical safeguards, explainable research, and updated regulation and guidelines, AI systems can become more trustworthy and safe. As AI systems make their way into more fields like healthcare and finance, it is critical to address security, transparency, and fairness concerns in the early days of AI.

References

Cloudflare. (2025). What is AI data poisoning? Retrieved from https://www.cloudflare.com/learning/ai/data-poisoning/

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness through awareness. Retrieved from https://arxiv.org/abs/1104.3913

Srinivas, N. (2024). Ai’s privacy pandora’s box: The risks of oversharing with Generative AI. Retrieved from https://insights.manageengine.com/artificial-intelligence/risks-of-oversharing-with-generative-ai/

Such, J. (2017). Privacy and Autonomous Systems. Retrieved from https://www.ijcai.org/proceedings/2017/0663.pdf

Vähäkainu, P. (2019). (PDF) Artificial Intelligence in the Cyber Security Environment Artificial Intelligence in the Cyber Security Environment. Retrieved from https://www.researchgate.net/publication/338223306_Artificial_intelligence_in_the_cyber_security_environment_Artificial_intelligence_in_the_cyber_security_environment

Appendix

ChatGPT output when asked questions 1 and 3:

1. What Does It Mean for an AI System to Be Safe?

An AI system is considered safe when it operates in a way that minimizes risks and unintended harm while maintaining reliability and fairness. AI safety involves several aspects, including:

Robustness: The AI should function correctly even under uncertain or adversarial conditions.

Fairness & Bias Mitigation: The system should avoid discrimination and ensure equitable treatment for all users.

Transparency & Explainability: Users and regulators should be able to understand how AI makes decisions.

Security: The system should be resilient against attacks, such as adversarial manipulation or data poisoning.

Privacy Protection: AI should comply with privacy laws and ensure data is handled securely.

Alignment with Human Values: AI should act in a way that aligns with human intentions and ethical guidelines.

2. The Impact of Cybersecurity on AI Systems

Cybersecurity plays a critical role in ensuring AI systems remain secure and trustworthy. Key impacts include:

Protecting AI Training Data: Cybersecurity frameworks help prevent data poisoning, where attackers manipulate training data to influence AI behavior.

Securing AI Models: Applying encryption and secure model storage can prevent adversaries from stealing or modifying AI models.

Mitigating Adversarial Attacks: Techniques like adversarial training and anomaly detection help AI systems resist manipulation.

Ensuring Compliance: Cybersecurity frameworks (e.g., NIST AI Risk Management Framework, ISO/IEC 27001) provide structured guidelines for securing AI systems.

Resilience Against AI-Specific Threats: Security controls prevent threats such as model inversion (extracting sensitive data from AI models) and prompt injection attacks (manipulating AI outputs through input data).

By integrating cybersecurity best practices, organizations can enhance the safety, reliability, and integrity of AI systems while reducing risks associated with malicious exploitation.