Trustworthy by Design

Carol J. Smith
cjsmith@andrew.cmu.edu
Trust Lab, AI Division, Software Engineering Institute, Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

The relatively recent public release of generative artificial intelligence (AI) systems has ignited a significant leap in awareness of the capabilities of AI. In parallel, there has been a recognition of AI system limitations and the bias inherent in systems created by humans. Expectations are rising for more trustworthy, human-centered, and responsible software connecting humans to powerful systems that augment their abilities. There are decades of practice designing systems that work with, and for humans, that we can build upon to face the new challenges and opportunities brought by dynamic AI systems.

CCS CONCEPTS

• Human-centered computing • Computing methodologies → Artificial intelligence

KEYWORDS

Keynote, ethics, trust, emerging technology, AI

1. Introduction

Trustworthy systems are designed to work with, and for, people, and are developed using a combination of AI engineering and human-computer interaction (HCI) practices. Trustworthy AI systems are carefully crafted so that their capabilities are understood, and that continuous monitoring and oversight are a priority when they are in use. There are decades of practice designing software and systems that work with, and for humans. We can build upon that work to face the new challenges and opportunities brought by dynamic AI and autonomous systems.

ICSE '24, April 14–20, 2024, Lisbon, Portugal

ACM ISBN 979-8-4007-0217-4/24/04.
https://doi.org/10.1145/3597503.3649400

2. AI Appropriate

For a system to be trustworthy, it should be built for a specific context of use and closely fit with user needs and their tasks, utilize appropriate data, and be reliable (robust and secure). However, the inherent ambiguity for development teams of the context of use, the end users' needs, and the availability, appropriateness, and usefulness of data can lead to poor development decisions such as [1] [2] [3].

Ambiguity can be clarified by conducting research to determine what the users' need are and what challenges they currently have, starting with whether the situation can be improved with AI. AI is not an appropriate solution for every problem. Selecting the right tool to solve the problem in the context of use is key to success.

For example, if the topic is too broadly scoped, AI systems that are currently available will not be able to be successful. Large language models (LLMs) are extremely broad and can appear to respond to a wide variety of topics and context, but due to their generative nature, they cannot be relied on for consistent results or factual information without significant verification. If data are limited in amount or availability, are divergent, include harmful bias, and/or are of low quality, it will be extremely difficult to build a good AI system. If humans are faster or more accurate than the proposed system, it is likely not worth the effort. The total cost of AI system ownership can be quite high as AI systems need consistent monitoring and maintenance by specialized teams over the entire course of usage.

We should always be looking for ways to capitalize on human strengths such as exposing bias, identifying downstream impacts, judgment, recognizing bias, responding to change, socio-political nuance, and taking context into consideration [4]. AI systems are very good at computation, repetition, replication, scale, short and long-term memory, simultaneous operation, and velocity. Problems that humans are still challenged by such as social problems are not able to be solved by AI systems. Systems that reliably identify people's emotions, intent, or cognition, have not been successful, and any application of them must be carefully reviewed to avoid privacy violations and general ethical considerations. Humans are extremely bad at vigilance tasks and teams need to be careful not to rely on their attention during boring or repetitive activities particularly when safety is a factor such as with vehicles and robotics.

Trust is a common discussion point for AI systems. Trust is a psychological state based on expectations of the system's behavior—the confidence that the system will fulfill its promise. [5] Trust is complex, transient, and personal, and these factors make the human experience of trust hard to measure. [5] Therefore, we need to make AI systems that are trustworthy, and not focus on measuring human trust. Our goal is to have people gain a level of calibrated trust of the system - appropriate for the context of use and capabilities of the system. [6] [7] This paper discusses methods for making trustworthy AI.

3. Data Are Biased

As all data are created and curated by humans for a specific reason. This reason is bias, and since data is always historical in nature, it will also contain patterns of behavior that are both good and bad. Every decision made while creating a system creates and affects bias such as what data to use during training, what model to apply, etc. Bias can have purpose and can be helpful, but it cannot be removed from a system [8]. Removing location, gender, and other bias indicators also removes the ability to track any bias in the system. Negative bias inherent in the system needs to be communicated to the people using the system and those affected by the system.

Knowing and communicating the provenance of proposed data is exceptionally important. Knowing the researcher's motivation, the collection process, what data was included and excluded and why, and what the recommended uses of the data are.

It is important to be mindful of the data that are being proposed for use and to dig deeply into the data. Include subject matter experts in the review, and consider the composition, variance, and suitability for the expected context of use. If the data includes information about people, the system will have higher risk than one that does not. The context of use will introduce more complexity including the environment it will be used in, the capabilities of the people who will use the system, the society and government.

4. Risk Identification

The addition of AI technology will increase the risk of any system or process. The team needs to identify the level of risk that is already present in the system or process, so that the expected risk with AI can be assessed. Some systems will have a higher potential for negative impacts and harms such as complex systems and processes and systems used to make crucial and emergent decisions.

AI system decisions and AI system outputs used for decisions that will affect people in ways that are irreversible such as their safety, health, life, quality of life, reputation, and autonomy are particularly high risk. Humans must retain responsibility for final decisions so that the integrity of the system is appropriately considered. Humans will make mistakes and then face consequences which machines cannot. Additionally, humans must be able to monitor and control risk, and have ultimate control of the system, being able to shut it down or revert to a previous version when it fails.

For any system, the proactive identification of potential harms and risks is extremely important. There are existing templates and tools for this purpose such as [9] [10] [11] [12] and more are being developed. Teams should be encouraged to activate their curiosity and to use simple speculative activities such as [11] to identify potential unintended and unwanted consequences, and negative impacts. Including people from historically and commonly marginalized groups can help to identify and prevent additional harms.

5. Good Design Mitigates Risk

HCI research informs trustworthy AI system development, and good design mitigates the risks of AI systems. Understanding the context of use by doing initial research to identify what is being made and for whom, and understanding the benefits and risks, will inform the design and support the team in building the right system and building it the right way.

Consideration for the user needs and goals such as quality of responses, need for provenance and the importance of anomalies also affects design. Interactions that are short and hectic require very different types of information and interaction techniques than longer, cyclical, and iterative tasks.

Ensure that the current state of the system is clear for users and that actions to get into or maintain a safe state are easy to do, while actions that can lead to an unsafe state (hazard) should be hard to do. [13] [14]

Understanding the constraints that the system will be used in can be extremely helpful in designing a good system. For example, knowing about:

- physical limitations due to weather, temperature, and clothing (gloves reduce ability for touch).
- Screen use, accessibility of information and concerns for screen visibility and brightness.
- Audio access and accessibility with considerations loud background sounds).

Identifying the expected improvements to the existing context is another important aspect of then being able to measure success. Different systems will be build depending on what the team thinks the system is expected to improve such as the quality of work, the speed of work, and/or a reduction in repetition.

Initial designs for interactions that are new, or complex should be prototyped using low-fidelity methods. The team should pursue feedback on the prototype from the intended end users (or close proxies). For example, users should be able to understand when the system is experiencing model drift (using language appropriate to the user), and outputs of the system should be appropriate for their needs. For example, when driving a car, we expect to be able to tell if there is an issue with the engine, even if the car is unfamiliar to

us. Similarly, users do not need to understand exactly how the system works (algorithms, etc.), but they should be able to diagnose the system's state and understand its capabilities in their context.

The team should adopt technology ethics and be encouraged to have regular conversations to reach shared understanding on difficult topics such as their shared values, what the system will and will not do, who could be hurt by the system, and how the implementation of the system will shift power. [15]

Future systems are discussed as being a human-machine team where people will exchange information and take turns with the AI system. This concept is already in use to some extent with regard to robots and autonomy and will improve as systems become more capable. In the meantime, it is worth taking the time to think about how responsibilities will be defined and conveyed to the user. They will need to know whose turn it is and if the system is making significant decisions autonomously, they need to be explained, able to be overridden by the responsible person, and appealable by the individuals affected by the decision.

6. Conclusion

Using qualitative measures, tools such as Datasheets for Datasets [16] and Model Cards for Models [17], and frameworks such as [18] [19] [9], paired with technical ethics will bridge the gap between potentially vague language and the reality of building an AI-enabled software system.

The activities described in this paper will help your team reduce risk in the systems you build and support the team in mitigation planning for when systems fail.

Trustworthy, human-centered, and responsible AI requires good design that meets the needs of their intended users in context. Users and those affected by the system need to understand the systems' capabilities and see that the system was built intentionally to be trustworthy by design.

ACKNOWLEGEMENTS

This paper was supported by Carnegie Mellon University Portugal, the ACM Distinguished Speakers Program, and the Carnegie Mellon University, Software Engineering Institute.

REFERENCES

[1] J. R. Young, "What happened after this college student's paper was falsely flagged for AI use after using Grammarly.," 04 04 2024. [Online]. Available: https://www.fastcompany.com/91074029/can-using-grammarly-set-off-ai-detection-software. [Accessed 25 04 2024].

[2] C. Lecher, "NYC's AI Chatbot Tells Businesses to Break the Law," 29 03 2024. [Online]. Available: https://themarkup.org/news/2024/03/29/nycs-ai-chatbot-tells-businesses-to-break-the-law. [Accessed 25 04 2024].

[3] K. Armstrong, "ChatGPT: US lawyer admits using AI for case research," 27 05 2023. [Online]. Available: https://www.bbc.com/news/world-us-canada-65735769 . [Accessed 25 04 2024].

[4] A. Muller and C. J. Smith, "Perceptions of Function Allocation between Humans and AI-Enabled Systems," in *UXPA 2022*, San Diego, 2022.

[5] C. Gardner, K.-M. Robinson, C. J. Smith and A. Steiner, "Contextualizing End-User Needs: How to Measure the Trustworthiness of an AI System," Carnegie Mellon University, Software Engineering Institute, 17 07 2023. [Online]. Available: https://insights.sei.cmu.edu/blog/contextualizing-end-user-needs-how-to-measure-the-trustworthiness-of-an-ai-system/. [Accessed 07 01 2024].

[6] J. D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *International Journal of Human-Computer Studies,* vol. 40, no. 1, p. 153–184, 1994.

[7] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Hum Factors,* vol. 46, no. 1, p. 50–80, 2004.

[8] K. Bode, "Why You Can't Model Away Bias," *Modern Language Quarterly,* vol. 81, no. 1, p. 95–124, 1 March 2020.

[9] L. Dance, "3Q-Do No Harm Framework - Categories of Harm," ServiceEase, 2019.

[10] M. Hoffmann and H. Frase, "Adding Structure to AI Harm: An Introduction to CSET's AI Harm Framework," 01 July 2023. [Online]. Available: https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/. [Accessed 13 02 2024].

[11] N. Martelaro and W. Ju, "What Could Go Wrong? Exploring the Downsides of Autonomous Vehicles," in *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '20)*, New York, 2020.

[12] Microsoft, "Types of harm - Harms Modeling Documentation," 24 01 2023. [Online]. Available: https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/type-of-harm. [Accessed 18 05 2023].

[13] N. Leveson, Safeware: System Safety and Computers, Addison Wesley, 1995.

[14] N. G. Leveson, "The Therac-25: 30 Years Later," *Computer,* vol. 50, no. 11, pp. 8-11, November 2017.

[15] P. Kalluri, "Don't ask if artificial intelligence is good or fair, ask how it shifts power," *Nature,* vol. 583, no. 7815, p. 169–169, July 2020.

[16] T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, H. Daumé III and K. Crawford,

"Datasheets for Datasets," *Communications of the ACM,* vol. 64, no. 12, pp. 86-92, 1 December 2021.

[17] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji and T. Gebru, "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, New York, NY, USA, 2019.

[18] C. J. Smith, "Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development," in *AAAI Symposium FSS-19: Artificial Intelligence in Government and Public Sector*, Arlington, VA, USA, 2019.

[19] J. Dunnmon, B. Goodman, P. Kirechu, C. Smith and A. Van Deusen, "DIU Responsible AI Guidelines in Practice: Lessons Learned from the DIU AI Portfolio," Defense Innovation Unit, 2021.

[20] University of North Carolina at Chapel Hill , "Generative AI Usage Guidance," [Online]. Available: https://provost.unc.edu/generative-ai-usage-guidance-for-the-research-community/.

[21] S. Ren, "How much water does AI consume? The public deserves to know," Organisation for Economic Co-operation and Development (OECD), OECD.AI Policy Observatory, Academia, 30 Nov 2023. [Online]. Available: https://oecd.ai/en/wonk/how-much-water-does-ai-consume. [Accessed 15 01 2024].

[22] Carnegie Mellon University, Software Engineering Institute, "Building AI Better: SEI Introduces Three Pillars of AI Engineering Press Release," 30 June 2021. [Online]. Available: https://insights.sei.cmu.edu/news/building-ai-better-sei-introduces-three-pillars-of-ai-engineering/. [Accessed 15 01 2024].

[23] M. Chapman, "What Exactly Does a Data Scientist Do?," Towards Data Science, 22 June 2023. [Online]. Available: https://towardsdatascience.com/what-exactly-does-a-data-scientist-do-42c53db57df5. [Accessed 15 01 2024].

[24] J. Bhalla and N. J. Robinson, "'Techno-Optimism' is Not Something You Should Believe In," Current Affairs, 20 October 2023. [Online]. Available: https://www.currentaffairs.org/2023/10/techno-optimism-is-not-something-you-should-believe-in. [Accessed 15 01 2024].

[25] H. Barmer, R. Dzombak, M. Gaston, J. Palat, F. Redner, C. J. Smith and T. Smith, *Human-Centered AI, White Paper,* Carnegie Mellon University Software Engineering Institute, 2021.