Copyright 2023 Carnegie Mellon University.
This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.
The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.
NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.
[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.
Internal use:* Permission to reproduce this material and to prepare derivative works from this material for internal use is granted, provided the copyright and "No Warranty" statements are included with all reproductions and derivative works.
External use:* This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other external and/or commercial use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.
* These restrictions do not apply to U.S. government entities.
Carnegie Mellon® is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.
DM23-1032

Overview

Today we are going to look at ways to Operationalize Responsible AI so that we are building trustworthy systems that support mission needs.

Imagine using an AI-enabled virtual reality headset for the first time at work. This headset is able to ask you questions and integrate your answers into the experience. You might expect to be asked basic questions about the purpose for your use, etc. Since this is a work device, consider what might feel too personal to share with the device. Next imagine the second time you put it on – what would you expect that experience to be like? Would you expect the system to remember your preferences?

To make a Responsible AI system, prior to designing those questions, someone needs to make decisions about what data are collected, how they are stored, and who has access to it. If someone accidentally provides access to their social security card or protected customer information, that information would need to be removed from the system.

Additionally, the system would need to work for the different people in the organization – they likely have different head sizes and heights, and would vary in their eye movements and eye colors. For example, someone with very light blue eyes is more sensitive to bright lights than someone with darker eyes. If the team testing the system only have darker eyes, the system may be designed in a way that is uncomfortable for people with lighter eyes.

Responsible AI is designing systems to work with and for people. There are three aspects of this work. First doing research to understand the complexity of the context of use. The second aspect is

understanding what is needed to enable human-machine teaming. The third is once the system is in use, teams need to engage in critical human oversight. The goal is to have end users choose to engage with the system and use it for appropriate scenarios – in short, we want to make a system that's trustworthy.

I. Understand Complexity of Content

First, we need to understand the complexity of the context the systems are working in. This includes the environmental and the human complexities, situational awareness, as well as the AI system's capabilities.

As teams evaluate these different pieces, they can begin to form an idea of what that collaboration is going to be like, and what types of interactions are expected to happen. What are the initial catalysts? What initiates the interaction, and how does it get to closure? We need to start thinking about how the system will work within existing situations and processes that the people are in.

Picture an autonomous vehicle moving down a road. If this vehicle encounters an obstacle in the road, it needs to determine what to do. One of those decisions may be to give control to a human. But how that occurs, when that occurs, and what the human operator's responsibilities are, are the key decisions that the team making that system need to consider when they're doing the development work. By conducting user experience research and other human-centered research activities from the start, teams can begin to identify those needs early on so they aren't surprises when the system is in use.

These are all aspects of understanding the complexity of context: looking at the primary user needs, their use cases and context, and then translating that information into design and functionality.

II. Designing for Human-Machine Teaming

The second piece of this work is designing the system for human-machine teams. This means taking the research finding and translating them into a system that people are willing to use - that they find to be trustworthy. Trustworthy systems augment human abilities. They partner with humans to attain a goal. They provide evidence of their capabilities to the people teaming with them so that people can build confidence in the system and understand when and how to use it.

Designing systems for human-machine teams needs to be based in the team's shared understanding of what they are making and for whom, as well as what the capabilities of the system will be. Additionally, the team needs to identify the inherent bias in data sources, and other inputs to the system, as well as consider how algorithm selection and other choices made throughout development will affect the system. We can use technical ethics to prompt those conversations. For example, the Department of Defense has a set of ethical principles for AI that include : responsible, equitable, traceable, reliable, and governable. Teams can leverage these principles to understand the goals of the system, reveal risks, and support inspection and mitigation planning. By having conversations about biases inherent in the system and that are introduced by the situation that teams are developing this AI system for, they can begin to make sure that they are building a system that is trustworthy.

The AI system will need to provide evidence, which means designing the outputs and the overall system people will interact with to be interpretable, understandable, and verifiable by the people who will interact with it and those affected by it

Teams also need to provide transparency regarding the limits of the system , and consider what boundaries and unfamiliar scenarios may be problematic.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution.

Ultimately - humans need to be in control of the systems, as they will be responsible for all final decisions the systems make. We need to ensure understanding of what the responsibilities are between the system and the human, and explicitly define them.

When authority is given to a system, any significant or final decisions it makes should be able to be overridden, be appealable, and reversible. For example, if an autonomous system is used to determine ticketing at traffic lights, individuals who receive those tickets should be able to appeal the ticket, and have a human make the final determination. There are situations where we may want to give the system ultimate control - particularly in instances that would protect human life (such as automatic brakes), but those should be rare.

Teams developing AI systems need to coalesce on a set of technical ethics so they have a shared understanding to base design decisions on, and to make AI systems that provide sufficient evidence to be trustworthy and enable strong human-machine teams.

III. Engage in Critical Oversight

The third area is engaging in critical oversight.

Critical oversight is only possible when we realize that these systems are dynamic. Because of that, they require proactive identification of risks, and continuous human oversight. Teams need to take deliberate steps to minimize unintended harm, and to make sure the system they're building is one they're willing to be responsible for.

AI systems, and all systems, have some form of bias. That's because they are created by humans for a reason. AI systems are based on data. Data is collected and curated by humans for a purpose, so complete objectivity is misleading. The purpose of the data collection imbues the initial bias, and the dataset can also include historical biases that aren't apparent, and curation of data can further add bias. . Bias can have purpose and it can be helpful, and bias can be harmful. The people responsible for the system need to make sure that they are aware of the bias, that they prevent risks where possible, and monitor and control accepted risks..

Harmful biases that are derived from data and the system design can result in unwanted or unintended consequences. Teams need to understand the data, as data is the primary source of bias for AI systems. They need to understand the creator's motivation, the inherent bias and variance within that data - as well as its recommended uses. Awareness of these factors will confirm that they have the right data, and it is a match with the context of use that they're designing the system for. By doing the work to identify risks and hazards early, they can ensure that they are building a system that is responsible.

Many types of harm can be identified through speculative work - thinking not only about harmful or malicious use, but also the good and beneficial use, can help prevent unintended harms. This work encourages curiosity, and speculative activities that most people enjoy once they begin doing it.

For example, starting conversations with your team to discuss: "What do we value? Who could be hurt? What lines won't our AI cross? And how are we shifting power?" can be difficult but valuable to gain

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution.

understanding of what's at stake. Framing these with a set of ethical principles, such as the DoD ethical principles for AI or other similar sets of technical ethics, can drive constructive conversations. Continuous human oversight requires monitoring, evaluation, and auditing. This work helps to make sure that the system outcomes are what is intended, and that the AI is still an effective collaborator with the end users. Teams need to do this consistently because these systems are dynamic. They're not stable, as compared to typical software. And this work is time-consuming and can be expensive, so it's important to plan for it.

To engage in critical oversight, we need to understand the bias inherent in the system, speculate about the misuse and abuse that could occur, prevent or plan to mitigate those risks, and then communicate awareness of bias in the system and the affects it may have.

Operationalizing Responsible AI is really about building trustworthy AI, understanding the complexity of context, designing for human-machine teaming, and engaging in critical oversight. Responsible AI systems are designed to work with and for people.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution.