Artificial Intelligence Ethics: 2nd lesson – Human-centered Design for AI

Before selecting data and training models, it is important to carefully consider the human needs an AI system should serve - and if it should be built at all. Human-centered design (HCD) is an approach to designing systems that serve people’s needs.

In this tutorial, you will learn how to apply HCD to AI systems. Then, you will test your knowledge in an exercise, by applying HCD to design issues in interesting real-world scenarios.

Approach:

HCD involves people in every step of the design process. Your team should adopt an HCD approach to AI as early as possible - ideally, from when you begin to entertain the possibility of building an AI system. The following six steps are intended to help you get started with applying HCD to the design of AI systems. That said, what HCD means for you will depend on your industry, your resources, your organization and the people you seek to serve.

* ***Understand people’s needs to define the problem***

Working with people to understand the pain points in their current journeys can help find unaddressed needs. This can be done by observing people as they navigate existing tools, conducting interviews, assembling focus groups, reading user feedback and other methods. Your entire team – including data scientists and engineers – should be involved in this step, so that every team member gains an understanding of the people they hope to serve. Your team should include and involve people with diverse perspectives and backgrounds, along race, gender, and other characteristics. Sharpen your problem definition and brainstorm creative and inclusive solutions together.

* ***Ask if AI adds value to any potential solution***

Once you are clear about which need you are addressing and how, consider whether AI adds value:

* Would people generally agree that what are you trying to achieve is a good outcome?
* Would non-AI systems such as rule-based solutions, which are easier to create, and audit and maintain, be significantly less effective than AI system?
* Is the task that you are using AI for one that people would find boring, repetitive or otherwise difficult to concentrate on?
* Have AI solutions proven to be better than other solutions for similar use cases in the past?

If you answered no to any of these questions, an AI solution may not be necessary or appropriate.

* ***Consider the potential harms that the AI system could cause***

Weigh the benefits of using AI against the potential harms, throughout the design pipeline: from collecting and labeling data, to training a model, to deploying the AI system. Consider the impact on users and on society. Your privacy team can help uncover hidden privacy issues and determine whether privacy-preserving techniques like *differential privacy* or *federated learning* may be appropriate. Take steps to reduce harms, including by embedding people - and therefore human judgment - more effectively in data selection, in model training and in the operation of the system. If you estimate that the harms are likely to outweigh the benefits, do not build the system.

* ***Prototype, starting with non-AI solutions***
* Develop a non-AI prototype of your AI system quickly to see how people interact with it. This makes prototyping easier, faster and less expensive. It also gives you early information about what users expect from your system and how to make their interactions more rewarding and meaningful.
* Design your prototype’s user interface to make it easy for people to learn how your system works, to toggle settings and to provide feedback. The people giving feedback should have diverse backgrounds – including along race, gender, expertise and other characteristics. They should also understand and consent to what they are helping with and how.
* ***Provide ways for people to challenge the system***

People who use your AI system once it is live should be able to challenge its recommendations or easily opt out of using it. Put systems and tools in place to accept, monitor and address challenges. Talk to users and think from the perspective of a user: if you are curious or dissatisfied with the system’s recommendations, would you want to challenge it by:

* Requesting an explanation of how it arrived at its recommendation?
* Requesting a change in the information based on your input?
* Disabling certain features?
* Reaching out to the product team on social media?
* Taking some other action?
* ***Build in safety measures***

Safety measures protect users against harm. They seek to limit unintended behavior and accidents, by ensuring that a system reliably delivers high-quality outcomes. This can only be achieved through extensive and continuous evaluation and testing. Design processes around your AI system to continuously monitor performance, delivery of intended benefits, reduction of harms, fairness metrics and any changes in how people are actually using it.

The kind of safety measures your system needs depends on its purpose and on the types of harms it could cause. Start by reviewing the list of safety measures built into similar non-AI products or services. Then, review your earlier analysis of the potential harms of using AI in your system. Human oversight of your AI system is crucial:

* Create a human ‘red team’ to play the role of a person trying to manipulate your system into unintended behavior. Then, strengthen your system against any such manipulation.
* Determine how people in your organization can best monitor the system’s safety once it is live.
* Explore ways for your AI system to quickly alert a human when it is faced with a challenging case.
* Create ways for users and others to flag potential safety issues.