Thinking Responsibly About AI Systems

Eliza Wells

Embedded EthiCS

11/29/21

When we talk about being a responsible computer scientist, there are at least two things we might mean. We might be talking about **negative responsibility:** who should be blamed when things go wrong? Determining negative responsibility can be difficult, especially in the context of AI systems. Many people are involved in designing those systems, and there are many elements that can contribute to eventual problems. We often see people trying to deflect blame: "I didn't make that decision, the AI did!"

A more helpful way to think about responsibility is to focus on **positive responsibility** and ask, "How can I be aware of the impacts of my decisions?" This is the model of responsibility that the Association for Computing Machinery emphasizes (see https://www.acm.org/code-of-ethics). In this lecture, we cultivate positive responsibility by introducing tools for ethical decision-making.

Whether you are starting out on a project or evaluating a system that has already been put into place, ethical decision-making relies upon your awareness of the different impacts your decisions can have. Impacts aren't just abstract concepts; they *happen* to *stakeholders.* Anyone or anything that can be impacted by your project, from individuals to the environment, is a stakeholder who should be considered.

The first step to thinking in a positively responsible way is thus to identify the stakeholders. Ask: "Who will be impacted by this system?" It's important to think big – almost anyone can be affected by technologies – and to consider small subsets. Philosophers have developed the concept of *intersectionality* to explore how different features of one's identity can overlap in distinct ways: while women and Black people both experience oppression in America, for example, White and Black women experience oppression very differently. Take time to consider different intersectional groups who might be affected by your project.

The second step is to ask how these stakeholders might be impacted. There are three important ethical lenses we should consider.

One is **benefits and harms.** Exploring this involves asking, "What are the potential consequences of this system for each stakeholder?"

Another lens is **respect.** We should ask, "How does this system show respect for each stakeholder's autonomy?" When thinking about autonomy, it's important to consider how the system incorporates elements like transparency, consent, and user control. A state-mandated system that used AI to predict which medications I needed and then immediately injected me with them might benefit me in that (if accurate) it would improve my health, but it wouldn't show respect for my autonomy.

A final lens to consider is **justice.** This asks whether the system treats each stakeholder fairly *and* whether it leads to fair outcomes. A system that gave loans to people based on calculated likelihood of defaulting might treat each individual fairly, but might not lead to fair outcomes if there are background conditions of oppression. On the other hand, a hiring system that aimed to hire equal numbers of different groups when members of one group are far more likely to apply might not treat each individual fairly, but could lead to fair outcomes on a broader scale.

These lenses overlap: an injustice can also lead to harms, a lack of respect can be an injustice, etc. Additionally, it is not always possible to secure everything for every stakeholder. For example, making the hypothetical medical system above more respectful might mean decreasing the benefits it can offer.

Once you have considered the different impacts your system can have on different stakeholders, you can then turn to the technical side to ask how best to serve as many stakeholders as possible. When designing AI systems, there are three important areas of technical choices that can influence how your system impacts stakeholders.

The first is your **data**: which data sources are you using to train your model? Is it appropriately representative of the population? Does it contain the appropriate variables?

The second is your **design:** how are you defining your objectives? How did you build your model?

Finally, you should consider **deployment.** Even if your design and data are perfect, the impacts of your system will depend on user interaction.