Intro

* We’ve covered some important background concepts in talking about Responsibility, Power and Data. The idea behind these has been to motivate some of why we should be interested in ethics in computer science, and why we are often going to be at least partially responsible for anticipating and mitigating harms that technical artifacts can bring about.
* For this middle portion of the course, we are looking at some broad categories of harm (some of the Data readings begun this process).
* This week was Bias and Fairness, which is a very popular topic of focus (for good reason) and the thing I think you’re most likely to have covered at least in passing in any ethics tutorials etc you’ve had in other courses. I know it was at least also covered by the ethics portion of FDS, but it’s important enough to be worth some overlap.
* Next week will talk about something that’s covered maybe less often, and that’s the impact of automating things that used to be done by people.

Definitions

* As usual, let’s start with some definitions
* **Bias** has more or less formal definitions. Colloquially it has a similar meaning to discrimination or prejudice, in that it is some kind of systematic advantage or disadvantage applied to an individual or group. More formally, we might want to define this in terms of its relation to the truth, as in “a system that fails to reflect true values more for some groups than others” but given how complex real-world data tends to be, we are not always going to be so lucky as to know everything about the “true” state.
* Similarly to in our discussion on power, the fact that bias is a preference built into a system means that it doesn’t have to be intentional. Some people or systems are almost definitely knowingly biased, but a lot more have biases that they are not necessarily aware of. In people we call this unconscious bias.
* If people are often not aware of biases in themselves, then it follows that they are also often going to end up encoding those biases into the technology they build, or that they might be unaware of bias working its way into the systems in any number of other ways.
* Also when I say the bias is built in to the system, that can manifest in a variety of ways. It could be in the actual design or code of the system, it could be present in the data used by that system and propagated through, or it could be a result of the way that system is being deployed into the world.
* Some kinds of bias are desired, in the sense that we are often designing systems such that they perform differently for some users than others. For example, we probably want a key card entry system to be biased against people whose key cards have expired. This kind of bias is often not included when people are talking about bias though, so it’s worth bearing in mind when you read about bias that authors might be using it as a shorthand for “unwanted” biases.
* When we are talking about **Fairness** in technical applications, then, we are usually talking about methods for avoiding or mitigating unwanted bias, such that we move towards a “fair” system in which they are minimised. Consequently, a lot of fairness discussions are also about how to measure fairness, because this makes the task of improving the fairness of a system, or choosing between multiple systems on grounds of fairness, easier.

Sources of Bias

* Hopefully most of you have covered this before, but where does Bias come from?
* Firstly we can have Bias in the design of Technical System. This is introduced by the design decisions consciously or accidentally made by the people who designed a given program, system, or technical artifact. This could be as straightforward as deciding to use race as one of the variables used to calculate whether someone should be approved for a mortgage, but it can also be much less obvious and come about from all kinds of assumptions about what is or isn’t important information for a task.
* Secondly, given how many modern systems use training data, Bias can be introduced in Data Collection. Maybe a dataset has far more examples for one group than another, or it was gathered in a way that misrepresents wider society, like how taking lots of your samples just outside a football stadium might well result in an algorithm that overestimates how much people in general like football. Or how using a dataset of CVs gathered from Amazon’s current employees led to them creating a screening algorithm that perpetuated the company’s historical bias against female applicants.
* Relatedly, Bias can be introduced in the way a dataset is labelled. Language is an inherently subjective and fluid thing, but labelling processes require people to apply definitive category labels. This means that the biases of the labelers can easily creep in. Perhaps they will more readily label a picture of an older person as a Doctor or Professor. Perhaps when labeling animals they are shown a koala, and they label it as a bear.
* It isn't just the person labeling either; someone has to decide what the set of labels they’re interested in are. Imagine you wanted to train a system to label the colour of an object. How would you decide what colour labels to use? Just primary colours? Or are we splitting it down as much as we can? The Dulux paint website currently offers 42 different *popular* shades of *white.* Good luck getting different labelers to reliably delineate those.

Individual vs Group Fairness

* When we’re talking about fairness, we also have to decide what kind of fairness we want.
* Individual fairness is interested in whether two people with similar traits are treated the same way, where group fairness is interested in whether whole groups of particular demographics seem to suffer more from mistakes. When we’re zoomed in at the individual level, it can be very hard to tell whether individual small differences between, say, evaluation of pairs of CVs, are significant. It might only be when you consider whole groups that you can see, say, that male CVs are consistently evaluated higher.
* But a major challenge in defining measures for group fairness is that it often requires us to specifically decide what groups we care about.
* This is an issue partly because it means we have to know what kind of bias we might be looking for, so maybe I check for fairness on grounds of gender and race, but I don’t think to check for something else like religion. There can always be these things we leave out.
* But perhaps more importantly, we have to decide what the delineation of those groups is. Maybe I want fairness on grounds of gender, but I do so only for “male” and “female”, and end up unaware of bias against non-binary people.
* Maybe I want fairness on grounds of religion, but I cluster all Judeo-christian religions into a single category, or I do the opposite and I split Christianity down into loads of denominations but don’t do the same for any other religion.
* Maybe I want fairness on grounds of race, but what even makes a race? Am I just delineating by skin tone? By country of origin? Both of those are certainly dreadfully insufficient.
* These are all design decisions just like those made in data collection and labelling and they all involve making decisions about how the world should be divided up that reflect the preferences of the designers.