Admin

* Who has already had their subgroup discussion?
* Remember to submit before your tutorials next week.
* Hopefully everyone has had a look at the essay, there have definitely been questions on Piazza, but I encourage you to look sooner rather than later

Intro

* We’ve covered some important background concepts in talking about Responsibility and Power.
* 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
* First is 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. Hopefully it was touched on by FDS, but it’s important enough to be worth some overlap.
* Followed by a topic that is talked about less often, which is the value that comes from having people involved in systems, and what you might lose when you automate them out.
* These were originally two lectures, but I think they’re compactable into one slot to compensate for Monday and not lose too much momentum.

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 “failing 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, subconsciously 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, or assuming a particular name format, 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.
  + Other examples?
* 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.
  + Other examples?
* 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.
* In the former case, the bias could be said to be the bias of the labeler being input to the model, but in the latter it’s a lack of understanding that we might be less inclined to call bias.
* It isn't just the person doing the 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? Last I checked, the Dulux paint website offers 41 different *popular* shades of *white.* Good luck getting different labelers to reliably delineate those.
* Side note, I checked this stat today and they’re down to 41. That is possibly the funniest change they could have made, so hat’s off to them.
* Bias can also be introduced by deployment or usage of a system, for example by deciding that you’re going to use it on one group of people and not another. This could be face recognition used on crowds at the Nottinghill carnival, or a medical screening tool only being offered to customers at a private clinic.
  + Other examples?

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 have similar outcomes, 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 or just noise. 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.

<questions and breather>

Humans

Intro

* The topic we’re looking at is one that, with a few exceptions, **gets a lot less press** I think than bias because it’s a lot harder to quantify (and as established in the week 1 readings, computer scientists love quantitative data)
* That topic is **human beings**, and specifically what we might unintentionally affect when we **remove people from a system**, or at least change their involvement in it.

What do we do when we automate?

* So why, beyond the obvious, are people relevant to us?
* Well one of the main things technology does is the automation of existing systems, or tasks within those systems.
* That means that often what we’re doing when we design a technological artifact and put it out into the world is essentially saying “here is a different way to do that task than the way people currently do it”
* This can be just implicitly, by creating the automated version as an option, or in a lot of cases explicitly, in the process of trying to sell it
* OpenAI, for example, is putting a lot of effort trying to embed ChatGPT into other systems
* This obviously isn’t always a bad thing. There’s plenty of tasks we’re probably better off not doing ourselves, not least because they’re dangerous (like mining for coal) or boring (like quality control on a conveyor belt)
* But we should be mindful that often what we’re doing by creating these automated versions of systems or processes, what we’re doing is *competing* with the version where people are more involved.

Automation’s unfair advantage

* A lot of the time, that competition is rigged.
* This comes back again to the dangers of focusing too much on quantitative methods for things; in this case, quantitative measures for success.
* The kinds of improvements provided by automation are often things that are easy to measure. Things like how fast a task is completed, how many tasks can be done simultaneously, how accurate something is.
* Some of this is because these are the things computers are naturally good at (precision and repetition), and
* some is because when we’re designing things we *want* easily measured metrics so we can know if we’re making progress.
* This is especially true now, in the age of machine learning, because having quantifiable criteria is often a necessity as a part of the feedback loop that lets a system learn.
* These are unsurprisingly also the kinds of metrics used by the people who are selling a system (and I mean that in a broad sense, not just at the point of sale but right through from the conception of the idea) because they highlight the advantages of that system.
* The benefits brought about by continuing to include people in these processes are harder to quantify with these performance measures, and so if we let the conversation be dominated by only the easily quantified factors then in the end people will lose out.

The ways in which human systems might be better

* So what are some of these reasons for keeping people involved in our processes?
* For me, three important categories could be: humans as robustness, humans as value, and humans as benefactors.
* I’ll elaborate a bit on these in the rest of the video.

Humans as robustness

* The first benefit to including people in systems is perhaps the one most linked to the mindset of performance measures; humans as robustness.
* No matter how well an automated system performs at the tasks it has been designed for, it is still only designed for those tasks. When the unexpected happens, even if it’s something mundane like the wrong data type in a field, these systems can fail.
* Of course, if they’re well designed they will fail as gracefully as possible, but we certainly are not anywhere near having systems with the general intelligence of humans, or the corresponding ability to reflect on what they are meant to be doing and what the new information means for that task. Humans adapt on the fly.
* This goes double for errors that can would otherwise go uncaught. Automated processes can easily miss that something has changed they were unprepared for, and continue processing the same data in the same way without ever getting suspicious, or trying to run some extra checks, or getting another opinion.
* Linked with this is the notion of trust. One of the main roadblocks (pun intended) to the uptake of self driving cars is that a lot of people rightly realise that no matter how well a robot performs on driving performance metrics, it still can’t be *trusted* the way a human driver could be to deal with unusual situations.
* Is there a technical solution for this? Is there a certain number of certain kinds of tests that an autonomous car could pass at which point we would finally go “Oh OK I trust it”, or is trust just a different kind of thing than test performance?
* Trust in autonomous systems is a big field at the moment, that doesn’t necessarily have an answer to this yet. What a lot of research focuses on is designing systems that explain their decision making, in the hopes that we can use this as a different way to interrogate performance, but whether we would trust this remains to be seen.

Humans as value

* The next advantage to including humans in processes are the extra benefits that this brings for other users of that system; humans as value.
* This benefit is covered pretty explicitly in the readings about care bots.
* There are lots of situations where the ability to interact with another person is a core part of the value of that process, not just care. Look at all the people who still intentionally use checkouts staffed by people over self-checkouts. Look at the difference between ringing a helpline and getting a real person vs and automated voice.
* Humans are social animals, and many of us, at least, still derive happiness from interacting with each other at whatever opportunity.
* This isn’t just a binary, either. As a lot of people have found out during the pandemic, it matters *how* you interact with other people. However much we might be told that all of these video conferencing technologies can provide us with the same core functionality as we had in person, the lack of human interaction is still felt, even though we are technically still interacting with humans.
* It’s worth mentioning that social media is a form of this too, where interactions with other people are enabled to the extent that they’re numerous and, importantly, countable (clicks, likes, posts). But there’s a risk that these end up competing with and crowding out the deeper more meaningful “real world” interactions.

Humans as benefactors

* This all leads into the final benefit of including people in systems, that those people themselves can benefit from being included.
* Partly this can be for human interactions reasons, as above. It is likely not just the care patient or the shopper who benefits from interacting with a human system over a machine one, but the human nurse or the person on the checkout too.
* They also benefit from having work more generally. Not just in the sense that they have a source of income, but that actually getting to do work that you find meaningful or interesting is of such great value to human flourishing. Jobs that keep you active, too, can be incredibly valuable to people.
* Even if we believe the rhetoric that automation won’t take away jobs, but instead change what kind of jobs people do, we have to be wary of what those new jobs are. If we shift people away from meaningful interactions and towards more monitoring and maintenance of the automated systems, we risk giving everyone work, but unfulfilling work. All in the name of a more efficient system, but a system that as meant to exist for people in the first place.

Going forward

* Looking forward to the next few weeks, availability of guest lecturers and my birthday mean we’ll go back to the original one lecture a week format (on Fridays). Next week Jacquie Rowe, who works in NLP for underrepresented languages, will talk about data harms
* The following week we’ll have Meenakshi Mani, who studies engineer’s experiences
* The remaining weeks will look at different approaches to try and make development more ethical tech
* We’ll view these in order scaling up, so starting with individual approaches next week, through to design methodologies that teams might use, then to measures organisations or whole fields might put in place.
* Next week’s main lecture is from Professor Shannon Vallor, who is an expert in virtue ethics, but shes’s super busy so rather than have her in person this year we have a recording.
* In the lecture slot, I will talk a little bit about final year projects and particularly the kinds of things you might be interested in doing if you want to follow on from any of the topics we’ve covered in this course.
* The following two weeks, we’ll have guest lectures, one on approaches to design and one on collective organising as a tool for sector change.
* In the final week of the semester we’ll have a guest lecture from the System Design Project, which will talk about what to expect from the course and announce the groups.
* Look out for an email asking you to fill out a questionnaire for that!