Admin

Intro

* This week we continue the theme of looking at general kinds of harm technology can cause if we’re not careful with it.
* 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, we’ll be moving on from this middle section of material, which was about some specific harm categories
* The remaining weeks will look at different approaches to try and make development more ethical
* 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!