Nuffield Foundation

Ada Lovelace Institute Event

Ensuring data and artificial intelligence work for people and society

Tuesday 4 December 2018, Nuffield Foundation

Speech by Dr Stephen Cave, Leverhulme Centre for the Future of Intelligence

Picture a system that makes decisions that have a huge impact on a person's prospects and life course; or even that makes decisions that are literally life and death. Imagine that system is hugely complex, and also opaque: it is very hard to see how it comes to the conclusions it does. A system that is discriminatory by its nature: it sorts people into winners and losers; but the criteria by which it does so are not clear.

The outsider being assessed by this system does not know what data it has gathered about them, or how it is relating it to other data it has gathered, or what biases might be inherent in that data. And on top of this, no individual human is willing to take responsibility for the decisions of the system - each claims to be fulfilling their own cog-like function, deferring always to the machine. Consequently, no one can be held accountable for the decisions this system makes.

In this system I am asking you to imagine, you will recognise many of the ethical challenges mentioned by Sir Alan in his prospectus for the Ada Lovelace Institute. But what I am describing is not some new machine learning system deployed in a bank or a local council. What I am describing is the vision offered to us by Franz Kafka in his novel The Trial.

In that book, written in 1915, Kafka gives us a hyperbolic, parodic account of an encounter with the apparatus of a bureaucratic state. As you will remember, the protagonist Josef K. does not know why he is arrested, or what the evidence against him is; no one is able to give him a full account of how the judicial system works, or how he can defend himself or appeal. No one is willing to take responsibility or be held accountable for the decisions of the system. And of course it ends gloomily with Josef K. utterly defeated, resigning himself to his fate.

So, we could say that many of the challenges posed by AI and related technologies are not new. But my point is not merely that — it is also that this is not a coincidence. There is a direct link between the trials of Josef K. and the ethical and political questions raised by AI.

Last November, at my Centre in Cambridge, we had a landmark conference on the History of AI organised by Jonnie Penn. The first lesson was that these technologies have histories. AI did not suddenly appear some time after the year 2000 in the form of a shiny white robot distributing gizmos like a silicon Santa Claus. The second lesson is that these histories are deeply entwined with state and corporate power.

The developers of digital technologies were largely funded by governments, including the military, or by large corporations. The purpose for which the technologies was developed was largely to further the interests of those bodies — to make them more efficient and effective, to increase their scope and reach.

Perhaps most importantly, the models of decision-making, problem-solving, reasoning and so forth that these systems sought to automate were taken directly from these bureaucracies.

In other words, the ‘intelligence’ in ‘artificial intelligence’, is not the intelligence of the human individual — not that of the composer, or the care-worker or the doctor — it is the systemic intelligence of the bureaucracy, of the machine that processes vast amounts of data about people’s lives, then categorises them, pigeon-holes them, decides over them, and puts them in their place.

It is therefore not a coincidence that AI and data-driven technologies, with all their challenges — such as opacity and lack of accountability — resemble the bureaucratic state as shown to us by Kafka. They parallel it because they are a product of it, and replicate its thinking. Confronted with the ‘computer says no’ culture of our time, Joseph K. would feel quite at home.

It is essential to recognise that we are not dealing simply with a ‘technology’ or even a set of technologies, but with socio-technical systems — that is, complex matrices of technology, people, culture, and power, that have deep roots in our society, and long histories.

I want to emphasise this historical dimension because I think it is particularly pertinent to a progressive institution like this one. As we have heard from Sir Alan, the Ada Lovelace Institute is invested in the ethical challenges of AI and data: it is, in the great tradition of the Nuffield Foundation and the other organisations supporting the Institute, invested in moral progress, in making the world a better place.
Better than what? Better than it has been, of course.

The Nuffield Foundation was established in 1943: as you know, that was a time when anti-semitism had reached an unbelievably savage historical peak in Europe; when colonialism was still justified as the White Man’s Burden; when eugenics was practised against the poor and disabled; when women were still denied many basic rights; when homosexuality was criminalised here in this country. That was just 75 years ago.

Of course, in this country, we have come a long way since then. There is much to celebrate, in terms of racial equality, gender equality, LGBT rights, and so on. But as you are all aware, there is also much still to do: there is still a great deal of injustice; still much that prevents everyone from having an equal chance of a dignified and fulfilling life.

There are still many ways in which realising the future we hope for means breaking away from the prejudices of the past.

So the question for us, who work on the ethics and impact of AI, is what difference will this technology make? How will it facilitate the kind of progress we hope for, and how will it threaten it?

First the good news: I think there are countless ways in which AI and related technologies can be used to empower people: for example, to bring better medical care to more people, to provide more personalised services, better tailored to people’s individual needs and preferences, and so on.

But the bad news is that there are also many ways in which these technologies risk ossifying and exacerbating historical injustices. As you know, all data sets have limitations, and these limitations can reflect historical injustices. This could be because certain groups have been over-examined - like people of colour by the police, which consequently distorts results about risk of reoffending; or it could be because some groups have been under-examined, like women in drug trials; or simply because the data reflects existing prejudices, like that engineers must be men.

In many cases, these historical injustices were considered right and just at the time. Our morals have moved on — for example with regard to the role of women in society — but the datasets might not have.

But it is not only a question of the data. Historical injustices can lead to the field of those developing the technology being disproportionately from one demographic — because some groups are disadvantaged in the education system, or because they imbibe the cultural perception that this field is not for them, or because the workplaces are hostile to people who don't fit the mould.

And to return to Kafka: these technologies risk perpetuating the injustices of the past because, as I mentioned, they are frequently the tools of entrenched interests: of councils, judiciaries, police forces, militaries and of course large corporations.

I don't think for a moment that any of those institutions are inherently bad. Of course not — they can all be forces for enormous good. But they can all also be Kafkaesque: they can be opaque, incomprehensible and unaccountable, while at the same time making enormously consequential decisions over people's lives.

And they are now developing and deploying technologies that reflect their opacity, incomprehensibility and unaccountability, while at the same time making enormously consequential decisions over people's lives. Technologies that also can extend their reach and power, and make already bureaucratic socio-technical systems more mechanical, and less human.

So while they offer enormous opportunities, these technologies also threaten to bind us to the injustices of the past.

All of this, to my mind, makes the Ada Lovelace Institute absolutely necessary. I have emphasised that these challenges do not arise from the technology by itself, but from complex socio-technical systems, with long histories. Tackling these challenges requires an institution that, like the Nuffield Foundation, has a strong tradition in the social and political sciences.

It requires access to expertise in philosophy and ethics — such as that of the British Academy — and experience in applied ethics, such as that of the Nuffield Council on Bioethics. But of course alongside all this, an understanding of the technology itself, and its applications, is essential, which is why it is so important that the Alan Turing Institute, the Royal Society, the Royal Statistical Society, Luminate, Tech UK, and Wellcome are all part of this initiative.

And it's also so important that the Institute is based here in London and close to the centre of power.

The institutions that will deploy these technologies in ways that shape people's lives are full of well-meaning people. But these institutions have their own momentum. No doubt everyone in the Trial of Josef K. was just doing their job. When councils and hospitals and police services are deploying these technologies, this will also be done by people just doing their jobs, trying to cut costs, save time, be efficient — unaware of how this might cross lines, undermine rights or embed injustice.

The Ada Lovelace Institute must be a trusted but independent and critical partner of these institutions.

In this, I particularly want to emphasise the second goal of the Institute that Sir Alan mentioned: to convene diverse voices to create a shared understanding of the ethical issues arising from data and AI.

Sometimes, talking about the importance of diversity can sound like a platitude. But I hope it is clear that it is not.

I have tried to emphasise in the past few minutes the obvious point that making moral progress means making the future different to the past, and also the ways in which AI and related technologies can hinder that.

But it should be obvious too that those juggernauts of entrenched power that threaten to mobilise AI to further their ends are represented by the non-diverse voices. A diverse range of voices means, among other things, those who have been excluded from these systems of power, such as women; or colonised by them, including much of the developing world and numerous communities in the developed world; or victimised by them, such as the poor or the disabled.

Ensuring those voices are heard; that they are fully involved in decisions about the development, deployment and impact of these technologies will therefore be critical to ensuring the future does not just replicate the injustices of the past.

Kafka alerted us to the perversity of 20th century socio-technical systems — their opacity, arbitrariness and unaccountability — he gave us a vision of the banal, bureaucratised evil of that century.

But also, over the course of that century, thanks to progressive organisations like this one, a huge amount was achieved in pushing back at those threats — through demanding rights, accountability, transparency.

These new technologies risk creating 21st century socio-technical systems every bit as opaque, arbitrary and unaccountable as those described by Kafka.

But at the same time, they offer enormous benefits and opportunities — they can and surely will transform many lives for the better.

The Ada Lovelace Institute can play a hugely important role in fending off these threats, and ensuring that the positive possibilities of AI prevail.