aeon.co

What are the values that drive decision making by AI? – Paula Boddington | Aeon Essays

Paula Boddington

19-23 minutes

A five-year-old boy is helping his grandmother cook by cutting out biscuits from the dough she's made, and he's doing it rather badly. He instructs the family robot to take over and, even though the robot's never done this before, it quickly learns what to do, and cuts out the biscuits perfectly. The grandmother is rather disappointed, remembering fondly the lopsided biscuits, complete with grubby fingerprints, that her son had charmingly baked for her at that age. Her grandson continues to use the robot for such tasks, and will grow up with pretty poor manual dexterity.

When the boy's parents come home, he says: 'Look, I've made these biscuits for you.' One parent says: 'Oh how lovely, may I have one?' The other thinks silently: 'No you didn't make these yourself, you little cheat.'

Artificial intelligence (AI) might have the potential to change how we approach tasks, and what we value. If we are using AI to do our thinking for us, employing AI might atrophy our thinking skills.

The AI we have at the moment is narrow AI – it can perform only selected, specific tasks. And even when an AI can perform as well

as, or better than, humans at certain tasks, it does not necessarily achieve these results in the same way that humans do. One thing that AI is very good at is sifting through masses of data at great speed. Using machine learning, an AI that's been trained with thousands of images can develop the capacity to recognise a photograph of a cat (an important achievement, given the predominance of pictures of cats on the internet). But humans do this very differently. A small child can often recognise a cat after just one example.

Because AI might 'think' differently to how humans think, and because of the general tendency to get swept up in its allure, its use could well change how we approach tasks and make decisions. The seductive allure that tends to surround AI in fact represents one of its dangers. Those working in the field despair that almost every article about AI hypes its powers, and even those about banal uses of AI are illustrated with killer robots.

The impact of technology on shaping our values is well-established. At a recent roundtable discussion on the ethics of AI, the group I was in spent most of our time discussing the well-known example of the washing machine, which did not simply 'take over' the laundry, but which has had a major impact on attitudes to cleanliness and housework, and on the manufacture of clothing. Because AI is designed to contribute not merely to the laundry, but to how we think and make decisions over an indeterminate number of tasks, we need to consider seriously how it might change our own thought and behaviour.

It's important to remember that AI can take many forms, and be applied in many different ways, so none of this is to argue that using AI will be 'good' or 'bad'. In some cases, AI might nudge us to

improve our approach. But in others, it could reduce or atrophy our approach to important issues. It might even skew how we think about values.

We can get used to technology very swiftly. Change-blindness and fast adaptation to technology can mean we're not fully aware of such cultural and value shifts. For example, attitudes to privacy have changed considerably along with the vast technological shifts in how we communicate and how data is shared and processed. One of the very things driving progress in AI is indeed the vast amounts of data now available, much of it about us, collected as we go about our daily lives. Many people are extremely wary of the organisations that have control of our data, while nonetheless continuing to post large amounts of very personal information that even a few years ago would have been considered private. Research shows that people's concerns about data privacy are inconsistent from one situation to the next. This is not to say that technology 'alone' has done this, since there are always other social changes operating at the same time.

And perhaps we are especially blind to the effects of some technology because it does so much to shape how we see the world. The challenge of AI is that it might operate in ways we aren't fully aware of. It helps to mould how we communicate with each other, how we think, how we find the world. This is not completely new: writing technology, the printing press and the telephone have already altered how we perceive and interact with our world, and even changed our brains. But AI might be even more powerful. Algorithms embedded into the technology through which we access so much information could be shaping what information we receive, how we receive it, even how we react to it. And AI might be shaping

our behaviour, not as an unintended consequence of its use, but by design. Technology, often aided by AI, is exploiting human psychology to shape how we behave. Phones and social media platforms are designed by drawing upon psychological research about how to produce addictive responses to their use.

So let's explore a few examples of the use or potential use of AI, focusing on how machines and humans use and analyse data.

First, let's be clear that there can be great advantages in using AI over human decision making. The fast sharing and robust data-analysis that AI performs can be extremely advantageous. For example, the information engineer Paul Newman of the Oxford Mobile Robotics Group points out that learning from accidents in vehicles driven by humans is a slow and complex process. Other humans can't learn directly from each individual case, and even the human involved might learn little or nothing. But whenever an autonomous car has an accident, all the data can immediately be shared among all other autonomous vehicles, and used to reduce the chances of a future accident.

This aspect of AI – the ability to share information like a hive mind and to analyse data rapidly and rigorously – might then constitute a real improvement in how we solve problems. Sharing pooled data is something AI is extremely good at. Analysing data quickly is another. That's how training AI on thousands of pictures of cats works. In fact, it's access to vast pools of data, together with the ability to analyse this data at speed, that's helping to drive the

current boom in AI.

Although autonomous vehicles can also make errors, this example demonstrates the human faults that AI can overcome. There are all sorts of ways in which humans fail to take in or analyse the data needed to make good decisions and to act on them. An autonomous vehicle will never be ashamed to admit fault, never be too vain to wear driving glasses, never insist on driving when tired, never refuse to go on an advanced driver course. Overcoming bias, partiality and irrationality is one way of improving human decision making – especially where issues of value are concerned. Some of these biases and irrationalities involve the rejection of, or failure to process, relevant information. So this model of using AI to pool data seems to be an advantage we can apply to decision making.

But such a conclusion might be hasty. Not all our problems can be solved by a purely data-led approach. It is pretty clear that avoiding car accidents is good. It's a safety issue where what we're doing is mostly applying technological fixes, and it's pretty easy to measure success. The vehicle either crashes or it doesn't, and deaths and injuries can be determined. It's also pretty easy to measure near-misses. But for problems that are less purely technical, it's not so clear that a data-driven, 'hive mind' approach is always good.

There's a danger that AI might prematurely shut off options or lead us down particular treatment routes

Take medicine, for example, one of the most promising areas of AI. Medicine is both a science, and an art: it combines science and technology with the pursuit of values: the value of health, the value of good patient relations, the value of person-centred care, the value of patient autonomy, and others. In medicine, we are not just

looking for a technological fix.

The use of AI in diagnosis is very promising, for example, in assisting with the interpretation of medical images by trawling through vast amounts of data. The evidence seems to be that AI can detect minute differences between images that the human eye doesn't notice. But it can also make blatant errors that a human would never make. So, currently, combining AI with human skills seems the best option for improving diagnosis. So far, this is excellent news.

But a piece in *The New England Journal of Medicine* in 2018 by Danton Char, Nigam Shah and David Magnus of the Stanford University School of Medicine in California raises serious questions about using AI in diagnosis and treatment decisions. Think about medicine as a science. If AI forms the 'repository for the collective medical mind', we'd have to be extremely careful before using it in a way that moves towards uniformity of professional thinking, which might foreclose independent thought and individual clinical experience. At the moment, it's recognised that there might be different bodies of medical opinion about diagnosis and treatment. If we could be utterly confident that AI was only improving accuracy, then greater uniformity of medical thinking would be good. But there's a danger that AI might prematurely shut off options or lead us down particular treatment routes. Moreover, the authors warn that such machine learning might even be used to nudge treatment towards hitting targets or achieving profits for vested interests, rather than what's best for patients. The data might drive the medicine, rather than the other way around.

Think about medicine as an art. It involves relating to patients as real individuals living their own lives. Although AI might help us

better achieve the goal of health, treatments with a lower chance of success might be the better option for some patients, all things considered. A data-driven approach alone cannot tell us this. And we need to be cautious we're not carried away by the power of technology. For we already know that free and informed consent is extremely hard to achieve in practice, and that the medical establishment influences patients' consent. But with the added gravitas of technology, and of blanket professional agreement, the danger is that wedding the existing power of the medical profession to the added power of AI, 'Computer says take the drugs' might become a reality.

The relationship between physician and patient is at the heart of medicine, and of our understanding of medical ethics. But the use of AI could subtly, even radically, alter this. Precisely how we implement the morally laudable aim of using AI to improve patient care needs careful consideration.

AI's ability to manipulate and process vast amounts of data might push us into giving undue or sole prominence to data-driven approaches to identifying and solving problems. This might lead to uniformity of thinking, even in cases where there are reasons to aspire to variety of thought and approach. It might also eclipse other factors and, in doing so, distort not just our thinking, but our values.

How a decision is made, and by whom; how an action is performed, and by whom – these are critical issues in many circumstances. It's especially the case where values are concerned. The parent who was skeptical that the boy had really made the cookies *himself* has a point. Perhaps if he'd first designed and built the robot, his claim would have had more validity. The significance of these factors will

vary from case to case, as will the potential significance of replacing or supplementing human intelligence with machine intelligence.

Take the use of juries. We all know that juries are flawed: they sometimes get the wrong answer. Algorithms are already helping US courts come to some decisions regarding sentencing and parole, drawing on data such as information about recidivism rates, to considerable controversy. There are fears that this can help entrench existing biases against certain groups. But imagine that we have reached the point at which feeding all the available evidence into a computer has led to more accurate verdicts than those reached by juries. In such a case, the computer would be able to pool and analyse all data with speed and (in this imagined example) accuracy and efficiency. Compare how actual juries work, where individuals might have made differing notes about the case, recall different things and, even after hours of deliberations, still have different views of the evidence. The power of AI to gather and analyse data might go a long way to address these shortcomings.

But this example readily demonstrates that we care about more than simply getting things right. Even if, by using a machine, we get a more accurate answer, there can still be some reason to value the distinctive contribution of having humans serving on juries. Consider the history of how legal reforms and individual rights have been fought for, and the value of the common person that's enshrined in the notion of being tried by 'a jury of one's peers'. Perhaps we want to hand that over to an AI – but perhaps not.

If a machine can do something quickly and efficiently, we might be more tempted to use it than is merited

The bias that humans can display, the tendency to be swayed by emotion, is of course a potential weakness in reaching a verdict. But it has also been the impetus for changes in the law. There are instances of 'jury nullification' where, swayed by those pesky human feelings of injustice, juries have simply failed to convict, even though the defendant is clearly guilty in terms of a strict application of the law. No matter how good at assessing evidence a machine might be, we're a long way off developing machines with a finely tuned sense of justice, an eye for the underdog, and the moral backbone to defy the apparatus of the legal system.

But the more general idea remains that juries perform the role of an independent source of judgment as a counter to the vested interests of the most powerful. As Lord Devlin said in the House of Lords in 2004: '[T]rial by jury is more than an instrument of justice and more than one wheel of the constitution: it is the lamp that shows that freedom lives.' And note this: the very feature of AI that is a strength in avoiding accidents for autonomous vehicles – the pooling of information, the melding of insights – in the context of the law directly undermines the basic moral principle of independence of juries. This independence is a counter to the ever-present possibility of powerful vested interests and gives a reason to keep trial by jury while being concerned with more than the mere processing of information presented in court.

A critic might say that we need this independence only because humans are so unreliable: the legal profession alone can't be left in charge of justice, but an accurate AI would solve the problem and maybe, with the passage of time, we'd get used to the idea, and hand over our justice system to the machines.

But it's utterly utopian to think that we'll ever escape the power

imbalances and vested interests that are the reason for having juries. There are alternatives to the miscarriage of justice problem than handing over justice to the machines, such as a fast, accessible appeals system. Perhaps, in the future, AI might *assist* judges and juries to come to decisions – but this is rather different to envisaging that AI might *replace* humans in legal decision making. Even here, we'd need to think carefully about the impact of AI, and whether it was nudging us towards a more technocratic approach. The law developed as a human political and social system through much struggle. But the use of AI within the law could, over time, help to change this. We need to consider this carefully, in full awareness of the many implications for justice and democracy.

One great attraction of using AI is simply the sheer speed at which it can analyse data. Efficiency is a virtue, but this virtue depends upon the ends to which it is being used. It's also by no means the only virtue. If a machine can accomplish something quickly and efficiently, we might be more tempted to use it than is really merited. And the speed with which it accomplishes tasks might make us overlook problems in how it achieves its ends. We might then end up placing too great a value on such efficiently generated results.

The Anti-Defamation League (ADL) in conjunction with D-Lab at the University of California, Berkeley, is developing an Online Hate Index (OHI) using machine learning in an attempt to detect hate speech online. This is an interesting example: it's a 'tech on tech' solution – the alleged proliferation of 'hate speech' online is (seemingly) a product of computerised technology. Yet hurling abuse at opponents is hardly new. Long before the World Wide

Web was even a twinkle in the eye of Tim Berners-Lee, the 17th-century French philosopher René Descartes described the work of rival mathematician Pierre de Fermat as ‘shit’. The long pedigree of insults is worth noting, given the way that hype around AI encourages us to think of issues as new, or uniquely dangerous. One way in which we can be steered in the direction of over-reliance on the narrow range of capacities that AI has is precisely by the combined assumptions that ‘tech cause = tech solution’, together with the idea that today’s tech is especially full of novel moral perils.

The concept of ‘hate speech’ is itself controversial. Some consider that attempting to eliminate certain uses of language, whether by law or by those managing online platforms, is necessary to achieve goals such as the elimination of discrimination. Others worry that this is a danger to free speech, and represents a form of censorship. There are concerns that what counts as hate speech to one person is merely banter to another, and that it’s extremely hard to classify something as ‘hate speech’ out of context. Additionally, there are concerns that, with the vast amount of material posted online, any policing of ‘hate speech’ will be patchy, and there are fears that some groups or individuals might be disproportionately targeted.

If we could automate the detection of hate speech online, it might help with classification and consistency. The OHI could address these issues by processing data faster than a human, and applying the policy correctly, without bias. (Of course, this depends on how it’s programmed. If it’s programmed to be biased, it will be biased with great efficiency.)

So here is the problem. Enthused with the idea that an AI can

detect, categorise and eliminate ‘hate speech’ with a speed and consistency that leaves humans standing, coupled with fears that the online world is turning so many of us into irresponsible hate-filled trolls, we could use this technology with such alacrity that the other problems of hate speech get somewhat overlooked. This then could help to drive the debate that could be roughly summarised as ‘hate speech versus free speech’ in one particular direction. In other words, it might help to mould our values. It might easily help to change the ways in which people communicate online, for fear that the hate-speech bot might oust them from the platform in question. Some aspects of this might be good, some not so good.

The virtues of AI include its particular ability to share data to reach a universal view of things; its capacity to help exclude human bias; the speed and efficiency with which it operates. It can transcend human capacity in all these things. But these virtues must all be measured up against our other values. Without doing so, we might be entranced by the power of AI into allowing it to take the lead in determining how we think about some of our most important values and activities.