Can machines make ethical choices?

Author: Barry Morisse

It seems uncontroversial to suggest that the sole reason the human race has survived everything thrown at it and got to where it is today is because we have the ability to augment our 'natural' selves with tools and technologies. The tools we create extend our capabilities, cover up for our weaknesses and allow us to harness the energy of the world around us, directing it towards our own goals. They allow us to transcend our own physical limitations.

More practically, if we look at the key turning points in our evolution, every single one is demarcated by a significant shift in our technology – which opened opportunities we simply couldn't imagine at the time. It's these catalysing moments that shift the way our species evolves, and I believe we are approaching another one of these key turning points.

The speed of development in the field of artificial intelligence (AI) is one that has captured the imagination of millions. As more and more financial investment and intellectual resources are being poured into the field, we are seeing real progress starting to accelerate. It's fascinating to watch and much of the success in the field can be attributed to one thing: data. In the information age, data has become the new oil – the resource that underlies the value in all the technology hubs around the world and in all the world's biggest organisations. Advances in storage, processing and computing power have provided the biggest companies in the world with tremendous amounts of data that were previously unusable. Machine learning techniques promise to take this data and use it to train our artificially intelligent machines – making them ever more powerful, on an exponential scale. The unfathomable scale of data processing transcends what we are capable of as humans and threatens to usher in a new era.

This new era will be categorised by our decision to outsource more and more decision-making to machines that can crunch big sets of data and deliver task-specific performance that outstrips their human counterparts. We see fascinating examples of this already, such as self-driving cars, automated medical diagnosis, algorithm-based creditworthiness decisions, economic forecasting and a variety of other real-world problems where machines are doing a much better job than us humans can. And this trend is not going to stop.

The danger of outsourcing these decisions is obvious. We need to be supremely confident that these systems will not only make the right decisions probabilistically, but also ethically. We want them to act in our best interests. Not only does the machine need to make the right decision when facing things it was programmed to deal with, but it will also have to make decisions in new, uncertain situations where it does not have prior experience – known in the industry as the problem of unsupervised learning. Can we trust it to make an ethical choice in those situations?

An example is often the best way to explain this intuition and the most tangible application of this thinking lies in the realm of self-driving cars. The infamous trolley experiment , a fundamental thought experiment in modern philosophy, is suddenly rendered a reality as a self-driving car will have to make decisions between two seemingly amoral outcomes – for example, whether to swerve to avoid a pedestrian but harm the driver, or to save the driver and harm the pedestrian.

These kinds of decisions are often written off as instinctual in a human setting, but when a machine makes them, most of us wouldn't be happy with writing off a mistake as 'instinctual'.

Whether machines can make ethical decisions or not depends on your point of view as to what an ethical decision is. Ethicists have spent millennia debating this because these decisions are never black and white. Every decision we make is inherently contextual and we choose different paths depending on a wide variety of factors.

So, when our programmers attempt to code machines that will perform to ethical standards, it is nigh-impossible to determine what those socially-agreeable standards are. It would be naïve to think that we could imagine every possible externality arising from a certain task and program unique decision-making capability for each one. Beyond the limits of our human imagination, it would simply require too much brute force computationally.

Instead, the only way that seems feasible is for the machine to learn patterns from our behaviour and use that to approximate our ethics/values. We want the machines to be able to make the complex, contextual decisions we are capable of. This is the foundation of deep learning – an artificial intelligence technique to deal with lots of unstructured information without clear rules. In essence, the machine would take in an infinite amount of data points about how we act and then translate that into a complex set of algorithms that approximate our ethical choices in situations it has never been in.

This concept is feasible technologically, but it is a difficult problem to solve. If we can continue to improve on our current algorithms, it could become incredibly powerful. However, what is often forgotten is the concern with us as humans. It is not clear-cut as to whether we are consistent enough with our own ethical choices to provide a machine with the type of information it needs for it to make the decisions we want it to make. If we were relying on pure brute-force programming, the machine would be making the decisions we would want our best selves to make. However, with deep learning, the machine acts in the way we do in reality. That's a marked difference.

I think that this debate will unearth some uneasy truths about how hypocritical and irrational we are as human beings. Once we start to see our ethics reflected back to us in the form of machine decision-making, are we going to like what we see? Where does the blame lie then?

We can't hide behind programming/algorithms – we are going to have to face the music. Yes, machines can make ethical decisions, but only if we make them first – at our best and at our worst.