Guillaume von der Weid, Philosopher

Ethics and the limits of AI

AI systems have the potential for great decision-making power but have no consciousness or moral conscience. The challenge now is to make sure that AI reflects and respects the ethical dimension of human society.

Something has gone badly wrong. A self-driving car controlled by AI is hurtling towards a barrier at high speed. If it crashes, the lone passenger will be killed instantly. If the AI system swerves the car to the right, the vehicle will run into a group of five pedestrians and kill them all. What decision should it make? Should the car be programmed to protect its passengers or should it be prepared to sacrifice them for the greater good?

What would be the right answer if there were two people in the car and just one pedestrian? If there was one passenger in the car and one person in the road? Or if the pedestrian was crossing illegally on a red light? Or if the sole occupant of the car was a pregnant woman and the potential victims in the street all elderly and infirm? These so-called “trolley problems,” familiar to a generation of philosophy students as thought experiments involving runaway trains, have assumed a new relevance in the age of AI.

Now that machines have the capacity to make decisions, dilemmas such as these have moved to the frontline of an urgent debate about moral decision-making, accountability and responsibility in the age of AI.

Most people accept that ethical criteria will have to be programmed into AI systems so that they make the right choices. But in many instances the utilitarian calculus is not sufficient for this. Surveys by carmakers have comprehensively concluded that hardly anyone would buy a self-driving car which, in the scenarios above, would choose to kill them rather than a pedestrian.

The logic of utilitarianism looks hard to defeat on its own terms, but the principle of self-preservation seems to outweigh them all. Having said that, people would not buy a car exercising moral choices, even if they agreed with the moral principles embedded in its programming. The ultimate meaning of our life can’t be encoded into a machine. For now, it seems that we want to be able to justify our choices in terms of principles, and conversely, to be able to hold a person accountable for their actions. Even if they can pilot themselves, we still want someone in front of the controls of our planes.

Putting humans first

It is the confrontation between, on the one hand, the utilitarian calculations that determine maximum well-being and on the other the ethical duties that we need to resolve, if human life is to have any value at all.

A celebrated example of the limits of simple utilitarianism concerns hospital surgery. Imagine a doctor with five severely ill patients who need respectively a new liver, heart, kidney, pancreas and lung if they are to survive. Should the doctor consider using the organs of a young and healthy patient who is asleep in the waiting room in order to save the lives of those five people, killing the healthy patient in the process? The calculation looks the same as in our initial example, but in the hospital case no-one would agree with killing one person in order to save five people’s lives. There is little room for doubt here that the ethical principle – the doctor’s duty to life – overrides the calculation of maximizing well-being.

Whatever science fiction may tell us, the real risk of AI is not that machines become conscious and turn against us. The real risk is that AI reduces human beings to mere numbers and units, with no respect for human values, principles and aspirations.

It is our responsibility now to set the right boundaries for AI. We must ensure that this most promising of technologies obeys ethical criteria and puts human beings at its heart. Our future depends on making sure that the AI systems of tomorrow respect the principles and values that make us human.

> **Machines do not, in terms of classical autonomy, comprehend the moral or legal rules they follow. They move according to their programming, following rules that are designed by humans to be moral.**
>
> The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems[1]

[1] https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e_classical_ethics.pdf