Living with artificial intelligence—how do we get it right?

February 28 2018

Credit: GIC on Stocksy

Powerful AI needs to be reliably aligned with human values. Does this mean that AI will eventually have to police those values? Cambridge philosophers Huw Price and Karina Vold consider the trade-off between safety and autonomy in the era of superintelligence.

This has been the decade of AI, with one astonishing feat after another. A chess-playing AI that can defeat not only all human chess players, but also all previous human-programmed chess machines, after learning the game in just four hours? That's yesterday's news, what's next?

True, these prodigious accomplishments are all in so-called narrow AI, where machines perform highly specialised tasks. But many experts believe this restriction is very temporary. By mid-century, we may have artificial general intelligence (AGI) – machines that are capable of human-level performance on the full range of tasks that we ourselves can tackle.

If so, then there's little reason to think that it will stop there. Machines will be free of many of the physical constraints on human intelligence. Our brains run at slow biochemical processing speeds on the power of a light bulb, and need to fit through a human birth canal. It is remarkable what they accomplish, given these handicaps. But they may be as far from the physical limits of thought as our eyes are from the Webb Space Telescope.

Once machines are better than us at designing even smarter machines, progress towards these limits could accelerate. What would this mean for us? Could we ensure a safe and worthwhile coexistence with such machines?

On the plus side, AI is already useful and profitable for many things, and super AI might be expected to be super useful, and super profitable. But the more powerful AI becomes, the more we ask it to do for us, the more important it will be to specify its goals with great care. Folklore is full of tales of people who ask for the wrong thing, with disastrous consequences – King Midas, for example, who didn't really want his breakfast to turn to gold as he put it to his lips.

So we need to make sure that powerful AI machines are 'human-friendly' – that they have goals reliably aligned with our own values. One thing that makes this task difficult is that by the standards we want the machines to aim for, we ourselves do rather poorly. Humans are far from reliably human-friendly. We do many terrible things to each other and to many other sentient creatures with whom we share the planet. If superintelligent machines don't do a lot better than us, we'll be in deep trouble. We'll have powerful new intelligence amplifying the dark sides of our own fallible natures.

For safety's sake, then, we want the machines to be ethically as well as cognitively superhuman. We want them to aim for the moral high ground, not for the troughs in which many of us spend some of our time. Luckily they'll have the smarts for the job. If there are routes to the uplands, they'll be better than us at finding them, and steering us in the right direction. They might be our guides to a much better world.

However, there are two big problems with this utopian vision. One is how we get the machines started on the journey, the other is what it would mean to reach this destination. The 'getting started' problem is that we need to tell the machines what they're looking for with sufficient clarity and precision that we can be confident that they will find it – whatever 'it' actually turns out to be. This is a daunting challenge, given that we are confused and conflicted about the ideals ourselves, and different communities might have different views.

The 'destination' problem is that, in putting ourselves in the hands of these moral guides and gatekeepers, we might be sacrificing our own autonomy – an important part of what makes us human.

Just to focus on one aspect of these difficulties, we are deeply tribal creatures. We find it very easy to ignore the suffering of strangers, and even to contribute to it, at least indirectly. For our own sakes, we should hope that AI will do better. It is not just that we might find ourselves at the mercy of some other tribe's AI, but that we could not trust our own, if we had taught it that not all suffering matters. This means that as tribal and morally fallible creatures, we need to point the machines in the direction of something better. How do we do that? That's the getting started problem.

As for the destination problem, suppose that we succeed. Machines who are better than us at sticking to the moral high ground may be expected to discourage some of the lapses we presently take for granted. We might lose our freedom to discriminate in favour of our own tribes, for example.

Loss of freedom to behave badly isn't always a bad thing, of course: denying ourselves the freedom to keep slaves, or to put children to work in factories, or to smoke in restaurants are signs of progress. But are we ready for ethical overlords – sanctimonious silicon curtailing our options? They might be so good at doing it that we don't notice the fences; but is this the future we want, a life in a well-curated moral zoo?

These issues might seem far-fetched, but they are already on our doorsteps. Imagine we want an AI to handle resource allocation decisions in our health system, for example. It might do so much more fairly and efficiently than humans can manage, with benefits for patients and taxpayers. But we'd need to specify its goals correctly (e.g. to avoid discriminatory practices), and we'd be depriving some humans (e.g. senior doctors) of some of the discretion they presently enjoy. So we already face the getting started and destination problems. And they are only going to get harder.

This isn't the first time that a powerful new technology has had moral implications. Speaking about the dangers of thermonuclear weapons in 1954, Bertrand Russell argued that to avoid wiping ourselves out "we have to learn to think in a new way". He urged his listener to set aside tribal allegiances and "consider yourself only as a member of a biological species... whose disappearance none of us can desire."

We have survived the nuclear risk so far, but now we have a new powerful technology to deal with – itself, literally, a new way of thinking. For our own safety, we need to point these new thinkers in the right direction, and get them to act well for us. It is not yet clear whether this is possible, but if so it will require the same cooperative spirit, the same willingness to set aside tribalism, that Russell had in mind.

But that's where the parallel stops. Avoiding nuclear war means business as usual. Getting the long-term future of life with AI right means a very different world. Both general intelligence and moral reasoning are often thought to be uniquely human capacities. But safety seems to require that we think of them as a package: if we are to give general intelligence to machines, we'll need to give them moral authority, too. That means a radical end to human exceptionalism. All the more reason to think about the destination now, and to be careful about what we wish for.

Provided by University of Cambridge

Citation: Living with artificial intelligence—how do we get it right? (2018, February 28) retrieved 6 May 2024 from https://techxplore.com/news/2018-02-artificial-intelligencehow.html