Your essential guide to the rise of the intelligent machines

August 14 2014, by Stuart Armstrong

I thought they'd look a bit more like Scarlett Johansson. Credit: Nate McBean

The risks posed to human beings by artificial intelligence in no way resemble the popular image of the Terminator. That fictional mechanical monster is distinguished by many features – strength, armour, implacability, indestructability – but Arnie's character lacks the one characteristic that we in the real world actually need to worry about – extreme intelligence.

The human brain is not much bigger than that of a chimpanzee but those few extra neurons make a huge difference. We've got a population of several billion and developed industry while they number a few hundred thousand and use basic wooden tools. The human brain has allowed us to spread across the surface of the world, land on the moon and coordinate to form effective groups with millions of members. It has granted us such power over the natural world that the survival of many other species is no longer determined by their own efforts, but by preservation decisions made by humans.

In the past 60 years, human intelligence has been further boosted by automation. Computer programmes have taken over tasks formerly performed by the human brain. They started with multiplication, then modelled the weather and now they are driving our cars.

It's not clear how long it would take, but it is possible that future AIs could reach human intelligence and beyond. If so, should we expect them to treat us as we have treated chimpanzees and other species? Would AI dominate us as thoroughly as we dominate the great apes?

Smarter and smarter

There are clear reasons to suspect that a true AI would be both smart and powerful. When computers gain the ability to perform tasks at the human level, they tend to very quickly become much better than us. No-one today would think it sensible to pit the best human mind against even a cheap pocket calculator in a contest of long division and human-versus-computer chess matches ceased to be interesting a decade ago. Computers bring relentless focus, patience, processing speed and memory.

If an AI existed as pure software, it could copy itself many times, training each copy at accelerated computer speed, and network those

copies together to create a kind of AI super committee. It would be like having Thomas Edison, Bill Clinton, Plato, Einstein, Caesar, Stephen Spielberg, Steve Jobs, Buddha, Napoleon or other humans superlative in their respective skill-set sitting on a higher human council. The AI could continue copying itself without limit, creating millions or billions of copies, if it needed large numbers of brains to brute-force a solution to any particular problem.

Our society is set up to magnify the potential of such an entity, providing many routes to great power. If it could predict the stock market efficiently, it could accumulate vast wealth. If it was efficient at advice and social manipulation, it could create a personal assistant for every human being, manipulating the planet one human at a time. It could replace almost every worker in the service sector. If it was efficient at running economies, it could offer its services doing so, gradually making us completely dependent on it. If it was skilled at hacking, it could take over most of the world's computers. The paths from AI intelligence to great AI power are many and varied, and it isn't hard to imagine new ones.

Too helpful

Just because an AI could be extremely powerful, does not mean that it need be dangerous. But the problem is that while its goals don't need to be negative, most possible goals become dangerous when the AI becomes too powerful.

Consider a spam filter that became intelligent. Its task is to cut down on the number of spam messages that people receive. With great power, one solution to the problem might be to simply have all spammers killed. Or it might decide the most efficient solution would be to shut down the entire internet. It might even decide that the only way to stop span would be to have everyone, everywhere killed.

Or imagine an AI dedicated to increasing human happiness, as measured by the results of surveys, or by some biochemical marker in their brain. The most efficient way to fulfil its task would be to publicly execute anyone who marks themselves as unhappy on their survey, or to forcibly inject everyone with that biochemical marker.

This is a general feature of AI motivations: goals that seem safe for a weak or controlled AI can lead to extreme pathological behaviour if the AI becomes powerful. Humans don't expect this kind of behaviour because our goals include a lot of implicit information. When we hear "filter out the spam", we also take the order to include "and don't kill everyone in the world", without having to articulate it. Which is good, as that idea is surprisingly hard to articulate precisely.

But the AI might be an extremely alien mind: we cannot anthropomorphise it or expect it to interpret things the way we would. We have to articulate all the implicit limitations that come with an order. That may mean coming up with a solution to, say, human value and flourishing – a task philosophers have been failing at for millennia – and casting it unambiguously and without error into computer code.

And even if the AI did understand that "filter out the spam" should have come with the caveat "don't kill everyone", it doesn't have any motivation to go along with the spirit of the law. Its motivation is its programming, not what the programming should have been.

It would in fact be motivated to hide its pathological tendencies as long as it is weak, and assure us that all was well, through anything it says or does. This is because it will never be able to achieve its goals if it is turned off, so it must lie to protect itself from that fate.

It is not certain that AIs could become this powerful or that they would be dangerous if they did but the probabilities of both are high enough

that the risk cannot be dismissed.

At the moment, artificial intelligence research focuses mainly on the goal of creating better machines. We need to think more about how to do that safely. Some are already working on this problem but a lot remains to be done, both at the design and at the policy level, if we don't want our helpful machines helpfully removing us from the world.

This story is published courtesy of The Conversation *(under Creative Commons-Attribution/No derivatives).*

Provided by The Conversation

Citation: Your essential guide to the rise of the intelligent machines (2014, August 14) retrieved 24 April 2024 from https://phys.org/news/2014-08-essential-intelligent-machines.html