What an artificial intelligence researcher fears about AI

July 14 2017, by Arend Hintze

Will the robots come to control us? Credit: Peshkova

As an artificial intelligence researcher, I often come across the idea that many people are afraid of what AI might bring. It's perhaps unsurprising, given both history and the entertainment industry, that we might be afraid of a cybernetic takeover that forces us to live locked away, "Matrix"-like, as some sort of human battery.

And yet it is hard for me to look up from the evolutionary computer

models I use to develop AI, to think about how the innocent virtual creatures on my screen might become the monsters of the future. Might I become "the destroyer of worlds," as Oppenheimer lamented after spearheading the construction of the first nuclear bomb?

I would take the fame, I suppose, but perhaps the critics are right. Maybe I shouldn't avoid asking: As an AI expert, what do I fear about artificial intelligence?

Fear of the unforeseen

The HAL 9000 computer, dreamed up by science fiction author Arthur C. Clarke and brought to life by movie director Stanley Kubrick in "2001: A Space Odyssey," is a good example of a system that fails because of unintended consequences. In many complex systems – the RMS Titanic, NASA's space shuttle, the Chernobyl nuclear power plant – engineers layer many different components together. The designers may have known well how each element worked individually, but didn't know enough about how they all worked together.

That resulted in systems that could never be completely understood, and could fail in unpredictable ways. In each disaster – sinking a ship, blowing up two shuttles and spreading radioactive contamination across Europe and Asia – a set of relatively small failures combined together to create a catastrophe.

I can see how we could fall into the same trap in AI research. We look at the latest research from cognitive science, translate that into an algorithm and add it to an existing system. We try to engineer AI without understanding intelligence or cognition first.

Systems like IBM's Watson and Google's Alpha equip artificial neural networks with enormous computing power, and accomplish impressive

feats. But if these machines make mistakes, they lose on "Jeopardy!" or don't defeat a Go master. These are not world-changing consequences; indeed, the worst that might happen to a regular person as a result is losing some money betting on their success.

But as AI designs get even more complex and computer processors even faster, their skills will improve. That will lead us to give them more responsibility, even as the risk of unintended consequences rises. We know that "to err is human," so it is likely impossible for us to create a truly safe system.

Fear of misuse

I'm not very concerned about unintended consequences in the types of AI I am developing, using an approach called neuroevolution. I create virtual environments and evolve digital creatures and their brains to solve increasingly complex tasks. The creatures' performance is evaluated; those that perform the best are selected to reproduce, making the next generation. Over many generations these machine-creatures evolve cognitive abilities.

Right now we are taking baby steps to evolve machines that can do simple navigation tasks, make simple decisions, or remember a couple of bits. But soon we will evolve machines that can execute more complex tasks and have much better general intelligence. Ultimately we hope to create human-level intelligence.

Along the way, we will find and eliminate errors and problems through the process of evolution. With each generation, the machines get better at handling the errors that occurred in previous generations. That increases the chances that we'll find unintended consequences in simulation, which can be eliminated before they ever enter the real world.

Another possibility that's farther down the line is using evolution to influence the ethics of artificial intelligence systems. It's likely that human ethics and morals, such as trustworthiness and altruism, are a result of our evolution – and factor in its continuation. We could set up our virtual environments to give evolutionary advantages to machines that demonstrate kindness, honesty and empathy. This might be a way to ensure that we develop more obedient servants or trustworthy companions and fewer ruthless killer robots.

While neuroevolution might reduce the likelihood of unintended consequences, it doesn't prevent misuse. But that is a moral question, not a scientific one. As a scientist, I must follow my obligation to the truth, reporting what I find in my experiments, whether I like the results or not. My focus is not on determining whether I like or approve of something; it matters only that I can unveil it.

Fear of wrong social priorities

Being a scientist doesn't absolve me of my humanity, though. I must, at some level, reconnect with my hopes and fears. As a moral and political being, I have to consider the potential implications of my work and its potential effects on society.

As researchers, and as a society, we have not yet come up with a clear idea of what we want AI to do or become. In part, of course, this is because we don't yet know what it's capable of. But we do need to decide what the desired outcome of advanced AI is.

One big area people are paying attention to is employment. Robots are already doing physical work like welding car parts together. One day soon they may also do cognitive tasks we once thought were uniquely human. Self-driving cars could replace taxi drivers; self-flying planes could replace pilots.

Instead of getting medical aid in an emergency room staffed by potentially overtired doctors, patients could get an examination and diagnosis from an expert system with instant access to all medical knowledge ever collected – and get surgery performed by a tireless robot with a perfectly steady "hand." Legal advice could come from an all-knowing legal database; investment advice could come from a market-prediction system.

Perhaps one day, all human jobs will be done by machines. Even my own job could be done faster, by a large number of machines tirelessly researching how to make even smarter machines.

In our current society, automation pushes people out of jobs, making the people who own the machines richer and everyone else poorer. That is not a scientific issue; it is a political and socioeconomic problem that we as a society must solve. My research will not change that, though my political self – together with the rest of humanity – may be able to create circumstances in which AI becomes broadly beneficial instead of increasing the discrepancy between the one percent and the rest of us.

Fear of the nightmare scenario

There is one last fear, embodied by HAL 9000, the Terminator and any number of other fictional superintelligences: If AI keeps improving until it surpasses human intelligence, will a superintelligence system (or more than one of them) find it no longer needs humans? How will we justify our existence in the face of a superintelligence that can do things humans could never do? Can we avoid being wiped off the face of the Earth by machines we helped create?

The key question in this scenario is: Why should a superintelligence keep us around?

I would argue that I am a good person who might have even helped to bring about the superintelligence itself. I would appeal to the compassion and empathy that the superintelligence has to keep me, a compassionate and empathetic person, alive. I would also argue that diversity has a value all in itself, and that the universe is so ridiculously large that humankind's existence in it probably doesn't matter at all.

But I do not speak for all humankind, and I find it hard to make a compelling argument for all of us. When I take a sharp look at us all together, there is a lot wrong: We hate each other. We wage war on each other. We do not distribute food, knowledge or medical aid equally. We pollute the planet. There are many good things in the world, but all the bad weakens our argument for being allowed to exist.

Fortunately, we need not justify our existence quite yet. We have some time – somewhere between 50 and 250 years, depending on how fast AI develops. As a species we can come together and come up with a good answer for why a superintelligence shouldn't just wipe us out. But that will be hard: Saying we embrace diversity and actually doing it are two different things – as are saying we want to save the planet and successfully doing so.

We all, individually and as a society, need to prepare for that nightmare scenario, using the time we have left to demonstrate why our creations should let us continue to exist. Or we can decide to believe that it will never happen, and stop worrying altogether. But regardless of the physical threats superintelligences may present, they also pose a political and economic danger. If we don't find a way to distribute our wealth better, we will have fueled capitalism with artificial intelligence laborers serving only very few who possess all the means of production.

This article was originally published on The Conversation. Read the original article.

Provided by The Conversation