
DANIEL THOMPSON

Program good ethics into artificial intelligence

Concerns that artificial intelligence will pose a danger if it develops consciousness are misplaced, says **Jim Davies**.

What is it that makes us worry about artificial intelligence (AI)? The White House is the latest to weigh in on the possible threats posed by clever machines in a report last week. As two of those involved write in a Comment piece on page 311, scientific and political focus on extreme future risks can distract us from problems that already exist.

Part of the reason for this concentration on severe, existential threats from AI comes from misplaced attention on the possibility that such technology could develop consciousness. Recent headlines suggest that respected thinkers such as Bill Gates and Stephen Hawking are concerned about machines becoming self-aware. At some point, a piece of software will 'wake up', prioritize its desires above ours and threaten humanity's existence.

But, when we worry about AI, machine consciousness is not as important as people think. In fact, careful reading of the warnings from Gates, Hawking and others show that they never actually mention consciousness. Furthermore, the fear of self-awareness distorts public debate. AI becomes defined as dangerous or not purely on the basis of whether it is conscious or not. We must realize that stopping an AI from developing consciousness is not the same as stopping it from developing the capacity to cause harm.

Where did this concern of machine consciousness come from? It seems mainly a worry of laypeople and journalists. Search for news articles about AI threats, and it's almost always the journalist who mentions consciousness. Although we do lots of things unconsciously, such as perceiving visual scenes and constructing the sentences we say, people seem to associate complicated plans with deliberate, conscious thought. It seems inconceivable to do something as complex as taking over the world without consciously thinking about it. So it could be that people have a hard time imagining that AI could pose an existential threat unless it also had conscious thought.

Some researchers argue that consciousness is an important part of human cognition (although they don't agree on what its functions are), and some counter that it serves no function at all. But even if consciousness is vitally important for human intelligence, it is unclear whether it's also important for any conceivable intelligence, such as one programmed into computers. We just don't know enough about the role of consciousness — be it in humans, animals or software — to know whether it's necessary for complex thought.

It might be that consciousness, or our perception of it, would naturally come with superintelligence. That is, the way we would judge something as conscious or not would be based on our interactions with it. A superintelligent AI would be able to talk to us, create computer-generated faces that react with emotional expressions just like somebody you're talking to on Skype, and so on. It could easily have all of the outward signs of consciousness. It might also be that development of a general AI would be impossible without consciousness.

(It's worth noting that a conscious superintelligent AI might actually be less dangerous than a non-conscious one, because, at least in humans, one process that puts the brakes on immoral behaviour is 'affective empathy': the emotional contagion that makes a person feel what they perceive another to be feeling. Maybe conscious AIs would care about us more than unconscious ones would.)

Either way, we must remember that AI could be smart enough to pose a real threat even without consciousness. Our world already has plenty of examples of dangerous processes that are completely unconscious. Viruses do not have any consciousness, nor do they have intelligence. And some would argue that they aren't even alive.

In his book *Superintelligence* (Oxford University Press, 2014), the Oxford researcher Nick Bostrom describes many examples of how an AI could be dangerous. One is an AI whose main ambition is to create more and more paper clips. With advanced intelligence and no other values, it might proceed to seek control of world resources in pursuit of this goal, and humanity be damned. Another scenario is an AI asked to calculate the infinite digits of pi that uses up all of Earth's matter as computing resources. Perhaps an AI built with more laudable goals, such as decreasing suffering, would try to eliminate humanity for the good of the rest of life on Earth. These hypothetical runaway processes are dangerous not because they are conscious, but because they are built without subtle and complex ethics.

WE SHOULD PUT **MORE EFFORT** INTO PROGRAMMING GOALS, **VALUES** AND **ETHICAL** CODES.

Rather than obsess about consciousness in AI, we should put more effort into programming goals, values and ethical codes. A global race is under way to develop AI. And there is a chance that the first superintelligent AI will be the only one we ever make. This is because once it appears — conscious or not — it can improve itself and start changing the world according to its own values.

Once built, it would be difficult to control. So, one safety precaution would be to fund a project to make sure the first superintelligent AI is friendly, beating any malicious AI to the finish line. With a well-funded body of ethics-minded programmers and researchers, we might get lucky. ■

Jim Davies *is associate professor at the Institute of Cognitive Science at Carleton University in Ottawa, Canada.*
e-mail: jim@jimdavies.org