May 26, 2015

Suggested Title: Will AI Really Take Over the World?

Ashok K. Goel

Science fiction often fantasizes about Artificial Intelligence (AI) taking over the world and enslaving humans, as we've seen again this summer in the Hollywood blockbusters "Avengers: Age of Ultron" and "Terminator: Terminator Genisys". Even some leading thinkers, such as Stephen Hawking, Bill Gates, Steven Wozniak, and Elon Musk, have gone on the record to express concerns about future AI. However, the notion of rogue AI taking over the world is neither imminent nor plausible. Here's why.

To begin, human intelligence has evolved over 3.8 billion years, and is amazingly robust, flexible and dynamic. Cognitive science does not yet fully understand the core processes of human cognition such as visual thinking (thinking about images and thinking in images, like dreams) and analogical thinking (thinking about new experiences in comparison to similar or familiar situations). AI research thus far has had only modest success in reproducing these processes in robots. Realizing general, human-level intelligence in robots would require many major advances across several disciplines dealing with the bodies, brains, minds, behaviors, and interactions of both humans and machines. While producing general, human-level AI may not take billions of years, the development likely will unfold over centuries and millennia, not years or decades.

Next, human intelligence is fundamentally social. Many human accomplishments are the result of cooperation among humans. We organize ourselves into families and teams, tribes and countries, and we learn by imitating others. If and when robots acquire human-level intelligence, to challenge humans, they too would need to organize into societies. Anthropology suggests that human-level social organization may have started with *Homo habilis*, a biological ancestor that lived in East Africa about 2 million years back. Development of robot societies equal to the current level of human social organization also likely will take centuries and millennia, not mere years or decades.

Third, while it is easy to talk about a future AI that is self-aware and capable of self-modification, in fact cognitive science understands little about the metacognition that forms the basis of self-awareness and self-modification. We do know that metacognition in humans is quite limited. You and I are not really aware of much of what happens in our bodies, brains and minds, and cannot easily modify them. AI so far has paid little attention to metacognition, but we do know it has a significant computational cost (which may explain why it is so limited in humans). Thus, the prospect of an AI in the near future that is fully self-aware and fully capable of self-modification is implausible.

Perhaps most importantly, the notion that futuristic AI will be simultaneously superhuman and evil separates ethics from intelligence. Many philosophers and scientists assert that ethics are a part of intelligence. Human intelligence and ethics have gradually co-evolved into higher ethical values such as the Gandhian notion of non-violence. Thoughtful science fiction writers such as Arthur C. Clarke have made similar observations about extra-terrestrial intelligence: if we ever

do encounter superhuman aliens in the years and decades ahead, Clarke has argued, they will have superhuman ethics as well and likely will be benevolent towards humans. The same arguments apply to superhuman AI as well: if robots ever do evolve into superhuman intelligence, they likely would have superhuman ethics, too. Like most humans, they too will learn from experience that it is useful to be ethical. They too will have emotional needs to be accepted and admired, and they too will feel the pressure of social rules and cultural norms. Intelligence, ethics and values, emotions and feelings, society and culture, all go together: there is little prospect for human-level or superhuman intelligence in a society or a species without correspondingly high-level ethics, emotions, and culture – a point that many critics of future AI miss.

Future AI agents and societies cannot become sentient, self-aware and superhuman on one hand, and rogue, volatile and sinister on the other. While it is important to address legitimate concerns about AI, extreme fears of AI taking over the world or destroying humankind seem more a reflection of human fears of the unknown and tribal fears of the "other" than a reality of AI research. If left unaddressed, these irrational fears may become a hindrance to AI research and prevent us from enjoying the enormous potential benefits that robots co-working with humans could bring.

Ashok K. Goel is a Professor in the School of Interactive Computing at Georgia Institute of Technology. He conducts research into artificial intelligence, cognitive science, and human-centered computing to explore the fundamental processes of computational creativity.