Here’s Why Robots Will Never Achieve Consciousness

You know the doomsday movie scenarios: An army of robots we’ve made to serve us decides to enslave or even replace the inefficient, refractory human race, and to that end wages a pitiless war of extermination on us.

But is all that mere sensationalism?

It would be flippant to dismiss the possibility. As science writer Bobby Azarian, who holds a PhD in neuroscience, notes:

> *“Among the fearful are intellectual heavyweights like Stephen Hawking, Elon Musk, and Bill Gates, who all believe that advances in the field of machine learning will soon yield self-aware A.I.s that seek to destroy us—or perhaps just dispose of us, much like scum getting obliterated by a windshield wiper. In fact, Dr. Hawking told the BBC, ‘The development of full artificial intelligence could spell the end of the human race.’”*

And our own Devin Foley, having mused on Aldous Huxley’s classic book *Brave New World,* concluded one post thus:

> *“Is it easier to pour our energy into building replacements for humans than to actually figure out how to completely condition a human? Maybe.”*

But Azarian also argues that, to become intelligent enough to replace us, robots would have to acquire the same sort of “consciousness” we have. I think he’s right about that. To be sure, a mere simulation of human consciousness could be a fearsome weapon for some people to use against others. Yet, for reasons Azarian explains, he thinks it unlikely at best that robots will ever achieve genuine *intentionality and* *subjectivity*. And achieving that is what it would take for robots to become an existential threat to humanity as such.

In that view, Azarian has got quite respectable company. For example, Berkeley philosopher John Searle has been arguing for 35 years that it is impossible for computers to achieve genuine “consciousness”. This TED talk sums up his reasoning briefly:

Still, having studied philosophy of mind as a graduate student, I suspect that authors like Azarian and Searle get this issue only half right.

They are right to argue that “…a strict symbol-processing machine can never be a symbol-understanding machine.” Computers are just symbol-processing machines; if a robot’s brain were only a computer, it would only be a processor—not a cognizant, living thing. Symbols don’t *mean* anything to entities that only manipulate symbols according to the rules given them. An entity that cannot grasp meaning cannot generate it either. And if you can’t do those things, you’re not conscious in the relevant sense.

My doubts arise when Searle, Azarian, and others argue that it’s a scientifically open question whether we could fabricate brains physically similar enough to ours to be conscious: to understand and generate meaning like we do. To affirm that possibility, one has to assume that consciousness is *merely* a biological phenomenon, so that if you reconstruct the biology correctly, you get consciousness. But of course, there’s a long philosophical tradition of arguing that consciousness is not *merely* biological.

In that tradition, consciousness is no more reducible to the right biology than “semantics” (the meaning of linguistic symbols) is reducible to the right “syntax” (the symbols themselves, and the rules for processing them).

That view ought to be taken seriously. For if it’s correct, it’s even less likely that robots will achieve the consciousness they’d need to be true replacements for us.

And that would be reassuring indeed.