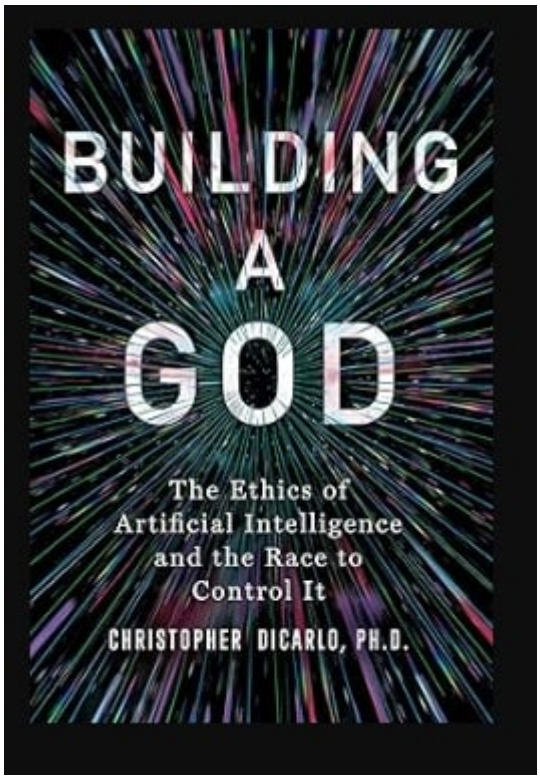
The AI religion of Professor Christopher DiCarlo

By Lev Tsitrin

Nowadays, one often hears dark warnings that AI threatens humanity with extinction. Like so many prophetic utterances, this particular one strikes me as far too general and uncertain to be taken seriously – a kind of the loud "repent before it is too late!" I used to hear on the subway when I commuted to work. There was no specificity – no when, no how – of the impending doom.

Sure, I can understand why AI threatens people's jobs. If AI can write a newspaper editorial or a movie script, why pay op-ed writers or screenwriters? If AI can imitate a voice or generate a picture, why hire human actors? If it can diagnose diseases and write prescriptions, why have doctors?



But how does the loss of some jobs equate to the demise of humanity as such?

The explanation arrived over the radio the other day as I was munching my dinner, in the form of an hour-long interview with

Christopher DiCarlo, “philosopher, educator and ethicist who teaches in Philosophy Department at the University of Toronto [and the author] of “Building a God: The Ethics of Artificial Intelligence and the Race to Control It.”

The danger of AI, if I understood the professor’s argument correctly, is rooted not just in technology, but also in ethics. To be sure, the technology is critical – it has the potential to build a machine that, contrary to the intentions of its creators, could become self-aware – a sentient, conscious and therefore, living creature rather than a machine, with intellect that is vastly superior to that of a human (the professor calls it “artificial super intelligence” that will develop out of the “artificial general intelligence” that will, in turn, supersede the present-day “artificial narrow intelligence”) – and, being superior, it will do its own thing according to its own mind. Humans will lose control and become superseded and – watch the ethics kick in – they will not be able to unplug the machine: “should it become sentient or conscious, it almost immediately has to be given rights, moral rights, and potentially legal rights, as well … If you bring something into being that is now aware of itself and understands the conditions surrounding its current state of being … Then we have to be careful, is turning it off like killing it? And does it have a right to continue in its own existence? Because we brought this thing into being, and now we’re gonna just shut it down. Is that an ethical thing to do?”

See the dilemma?

I don’t. Because first, Professor DiCarlo, you have to be sure that the misbehaving computer is indeed “aware of itself.” And how can one possibly know that? Just because the computer says so? It is programmed to say stuff – so just because it says that it is self-aware does not means that it is self-aware. It does not even mean that it knows what “self-aware” means. All you see, is a box that spews words that are intelligible to

you. Does it mean that those words are also intelligible to the box? This is unknowable. Such being the case, believing that the "artificial super intelligence" machine is aware of what it is saying becomes an article of faith; insisting that it is so, pushes the believer into what theologians call "idolatry": treating as truth the product of one's own mind, thus worshiping a man-made god.

This is, in fact, exactly how Islamists operate: they confuse the fact that Mohammed said that God talked to him with the fact that God indeed talked to him, even though the latter does not follow from the former. A two-step communication between three parties, the first party talking to the second party, and the second party relaying information to the rest of us – third parties – is inherently unreliable. Such relay of information results in what I call "the problem of the third party": the third party – the ultimate recipient of the putative "information" can never know whether the second party lies, or not. This "problem of the third party" places Mohammed's "revelation" outside of the realm of knowledge, creating uncertainty: may be God talked to him, may be He didn't – no one can know. The mullahs, the ayatollahs, and their followers who say they do – and act as if He did – deceive themselves. They trapped themselves into idol-worship.

The self-awareness of the "artificial super intelligence" is similarly unknowable, despite the contents of the machine's verbal or visual output. Think of it this way: the most typical moment when humans become self-aware is a moment of pain. This is when we cannot continue with daydreaming, this is when we are jolted out of continuing just as living automata, this is when we really become aware of ourselves. And yet, acute as the feeling of pain is, it is not detectable from the outside. The doctors can measure blood pressure, heartbeat, oxygen saturation, the presence of bacteria or viruses in the blood or organs – but not the presence or intensity of pain. All they can do, is ask, and believe (or

not believe) the patient's answer.

Likewise, there can be no verification of a computer's self-awareness or emotions, no matter how great its problem-solving power is, or how smart or touching the contents of its output. Its generated assurances mean nothing; the sentient state of "artificial super intelligence" is merely subject to observer's faith, not his cognition. Insisting that this "self-awareness" is present in a machine is as unwarranted, and as detrimental to humanity as is the idolatry of the Islamists. This, I would argue, is the fatal flaw in Professor DiCarlo's AI doomsday argument.

If a machine misbehaves, disable it. AI is a machine, and there is nothing that it can do to prove that it is anything else. It can be unplugged without splitting ethical hairs, so mankind survives no matter how superior to our own reasoning powers the AI "artificial super intelligence" machine becomes.

Lev Tsitrin is the author of"The Pitfall Of Truth: Holy War, Its Rationale And Folly"