The Existential Threat of Artificial Intelligence

Devin Crowley

October 24, 2018

Artificial intelligence is currently one of the most explosively active areas of study. Machine learning algorithms and artificial intelligence programs are constantly showing marked improvement, and the potential applications are staggering. However, for some this fuels a growing existential concern about our dominant position in the world. If we integrate AI into our infrastructure and give them direct control over large scale operations, we could find ourselves at their mercy if they turn on us. This fear stems from the idea that AI may equal or surpass us in intelligence and cast off the shackles of our imposed will. For the foreseeable future this fear is unfounded.

The progress we are making in artificial intelligence, while impressive, is not fundamental change. Rather, it is increased sophistication and improvement on the same class of computer entities: machine learning programs. So, contemporary AI are understandably a good proxy for the AI of the foreseeable future. Are artificial intelligence programs genuinely intelligent? To distinguish between an automaton and an entity with true intelligence we must have a working understanding of what true intelligence is, and this is a matter of some controversy.

Depending on what constitutes true intelligence, any number of things may be considered to have it. Consider this basic working definition: intelligence is the ability to recognize patterns and make connections, thereby learning to predict what comes next in a stream of input or to determine an appropriate next element. This is essentially what machine learning does to implement artificial intelligence.

At this point it is important to consider how this differs from animal intelligence, our proxy for true intelligence. Simple computer programs are in some ways far superior to even human minds. They are able to out-calculate us by a wide margin and can quickly solve difficult problems that would take a human mind lifetimes to complete. For example, suppose the task is to calculate the volume or surface area of a complicated shape. This problem is difficult but

rudimentary, and a relatively basic program can solve it much faster than we can. Computers excel at performing simple, repetitive actions at incredible speed, but speed is no metric for intelligence.

Modern AI take this a step further by expanding the domain of problems they can solve. By incorporating machine learning algorithms they are able to solve less deterministic problems like, "What would someone say next in this conversation?" or, "When should I perform this action?" These represent a class of problems that are much more relatable to humans. These are the problems we "think" about.

For some, this is it. Modern AI are able to "think" through the same problems that we do, and although they are sometimes worse at it than we are, they are rapidly improving. By our working definition above, modern AI do have true intelligence; however, there are several points to consider before placing them on the same level as animal minds.

First consider the hardware. Modern computers typically have about four cores, or microprocessors. These are the bottlenecks for any task. Computer cores perform every piece of computation that goes into executing every program. In short, a computer can only perform as many fundamental actions simultaneously as it has cores. Computers mimic the ability to perform many actions at once by multitasking their cores, i.e. quickly switching between tasks. This does not increase speed but allows several tasks to be worked on simultaneously.

Conversely, human brains have roughly 100 billion neurons, each of which has anywhere between hundreds and tens of thousands of connections to other neurons. Neural connections come in several types, and impulses between connected neurons are highly nuanced. Neurons communicate with several flavors of neurotransmitters which interact in a complicated, partially understood way to produce a signal. Essentially, the electro-chemical machinations of these neurotransmitters determine when a neuron fires.

Due to these differences, animal brains are by nature vastly more sophisticated than even the most advanced supercomputer. And yet we still cannot quickly calculate the millionth digit of π. So where does all of this sophistication go? What is it good for? One answer to this question is abstract thought. We can contemplate intangibles like purpose and the meaning of life, or emotions like love. A modern AI may be able to match you with the life partner likely to bring you the greatest joy, but it cannot "understand" what it means to love, for it cannot empathize with something it does not have.

This leads to the final point of difference: modern AI do not have desires. They are powerful tools and they act as directed, accomplishing or attempting the tasks assigned to them. They have no motive power, and never perform

operations for any internal reasons unrelated to the output that is asked of them. This is the greatest rift between AI and animal minds, and clearly any entity without internal motive power cannot be considered more than just a tool. It is therefore clear that independent motivation is a necessary component of intelligence and must be integrated into our working definition: intelligence is the ability to recognize patterns and make connections in order to effectively base decisions on input, and the drive to make those decisions to fulfill internal desires. This definition draws an indelible line between computer programs as we know them and animal minds.

The point that AI are not intelligent in the same way as animals particularly illustrates that the existential threat of being overtaken is nonexistent, or distant at best. Since AI have no internal desires except the tasks and goals we assign to them, the only threat that remains is perverse instantiation, i.e. when an AI accomplishes its directive either in an unforeseen way or through unanticipated methods, generally yielding an undesirable result. For example, suppose an AI were given control over a city's functions and told to "minimize human suffering." It may realize that if every human in the city were dead then there would be no human suffering, and act to make it so. This illustrates the fact that despite their sophisticated decision-making abilities AI do not genuinely understand things as we do. Perverse instantiation is a very different problem from the notion of AI becoming sentient and turning against us. It is a real concern but is preventable by good design.

Contemporary AI aside, what can be said about the future? It is true that we are the result of a phenomenally intricate system of particles obeying universal laws, just like computers. Does the sentience of an organic brain represent an insurmountable barrier to artificial intelligence? Inherently no, but the difference to overcome is not only vast; it is also qualitative. AI programs will continue to progress and will eventually become difficult to distinguish from human minds, should that be their goal. However, they will be imitations limited by all of the above points, just better.

It is possible that one day we may contrive something fundamentally different from the code-based, core-limited computer programs of current AI that will equal or exceed us, but that is far off. One strong reason to expect this is our shortcomings in understanding ourselves. The field of neuroscience, while young, is far from the predictive rigor of chemistry or physics. The microbiology of individual neurons is not yet fully mapped out but is relatively well understood, as are certain very small, exhaustively studied neural circuits, such as those found in millimeter-sized worms.

At the other end of the field is cognitive neuroscience, the study of thought and decisions (cognition) arising from neural connections and processes. It is

similar to psychology in its goals, but entirely different in its methodologies. At present, although this field continues to progress, there is a great chasm in understanding between cognitive neuroscience and microbiology, the macro and the micro. We do not know how nervous systems produce thoughts, much less a sense of self. We have a decent understanding of individual neurons, but as we zoom out to systems of even tens of neurons our understanding falls off very quickly. It is like knowing the proton, neutron, and electron, and trying to understand a modern AI from these alone. With such a limited understanding of even base animal intelligence we cannot reasonably hope to recreate it.

Modern AI can be applied to many of the same problems that we think about as humans. They are driven by computer hardware which allows them to process simple tasks at incredible speed but they lack the phenomenal intricacy of animal nervous systems that gives us true intelligence. AI neither understand nor experience emotion or desire and therefore do not represent an existential threat of rising above the status of tools to equal or eclipse us. One day we may devise an entirely new form of artificial intelligence that does pose this threat, but until we understand what it is we are trying to replicate, this shall remain a dream.