# Is AI Dangerous? That Depends…

Somewhere in the long list of topics that are relevant to astrobiology is the question of ‘intelligence’. Is human-like, technological intelligence likely to be common across the universe?

Are we merely an evolutionary blip, our intelligence consigning us to a dead-end in the fossil record? Or is intelligence something that the entropy-driven, complexity-producing, universe is inevitably going to converge on?

All good questions. An equally good question is whether we can replicate our own intelligence, or something similar, and whether or not that's actually a good idea.

In recent months, once again, this topic has made it to the mass media. First there was Stephen Hawking, then Elon Musk, and most recently Bill Gates. All of these smart people have suggested that artificial intelligence (AI) is something to be watched carefully, lest it develops to a point of existential threat.

Except it's a little hard to find any details of what exactly that existential threat is perceived to be. Hawking has suggested that it might be the capacity of a strong AI to 'evolve' much, much faster than biological systems - ultimately gobbling up resources without a care for the likes of us. I think this is a fair conjecture. AI's threat is not that it will be a sadistic megalomaniac (unless we deliberately, or carelessly make it that way) but that it will follow its own evolutionary imperative.

It's tempting to suggest that a safeguard would be to build empathy into an AI. But I think that fails in two ways. First, most humans have the capacity for empathy, yet we continue to be nasty, brutish, and brutal to ourselves and to pretty much every other living thing on the planet. The second failure point is that it's not clear to me that true, strong, AI is something that we can engineer in a pure step-by-step way, we may need to allow it to come into being on its own.

What does that mean? Current efforts in areas such as computational 'deep-learning' involve algorithms constructing their own probabilistic landscapes for sifting through vast amounts of information. The software is not necessarily hard-wired to 'know' the rules ahead of time, but rather to find the rules or to be amenable to being guided to the rules - for example in natural language processing. It's incredible stuff, but it's not clear that it is a path to AI that has equivalency to the way humans, or any sentient organisms, think. This has been hotly debated by the likes of Noam Chomsky (on the side of skepticism) and Peter Norvig (on the side of enthusiasm). At a deep level it is a face-off between science focused on underlying simplicity, and science that says nature may not swing that way at all.

An alternative route to AI is one that I'll propose here (and it's not original). Perhaps the general conditions can be created from which intelligence can emerge. On the face of it this seems fairly ludicrous, like throwing a bunch of spare parts in a box and hoping for a new bicycle to appear. It's certainly not a way to treat AI as a scientific study. But if intelligence is the emergent - evolutionary - property of the right sort of very, very complex systems, could it happen? Perhaps.

One engineering challenge is that it may take a system of the complexity of a human brain to sustain intelligence, but of course our brains co-evolved with our intelligence. So it's a bit silly to imagine that you could sit down and design the perfect circumstances for a new type of intelligence to appear, because we don't know exactly what those circumstances should be.

Except perhaps we are indeed setting up these conditions right now. Machine learning may be a just piece of the behavioral puzzle of AI, but what happens when it lives among the sprawl of the internet? The troves of big and small data, the apps, the algorithms that control data packet transport, the sensors - from GPS to thermostats and traffic monitors - the myriad pieces that talk to each other directly or indirectly.

This is an enormous construction site. Estimates suggest that in 2014 some 7.4 billion mobile devices were online. In terms of anything that can be online - the internet of 'things' (from toilets to factories) - the present estimate is that there are about 15 billion active internet connections today (via a lovely service by Cisco). By 2020 there could be 50 billion.

If this were a disorganized mush of stuff, like the spare parts in a box, I think one would have little hope for anything interesting to happen. But it's not a mush. It's increasingly populated by algorithms whose very purpose is to find structures and correlations in this ocean - by employing tricks that are in part inspired by biological intelligence, or at least our impression of it. Code talks to code, data packets zip around seeking optimal routes, software talks to hardware, hardware talks to hardware. Superimposed on this ecosystem are human minds, human decision processes nursing and nurturing the ebb and flow of information. And increasingly, our interactions are themselves driving deep changes in the data ocean as analytics seek to 'understand' what we might look for next, as individuals or as a population.

Could something akin to a strong AI emerge from all of this? I don't know, and neither does anyone else. But it is a situation that has not existed before in 4 billion years of life on this planet, which brings us back to the question of an AI threat.

If this is how a strong AI occurs, the most immediate danger will simply be that a vast swathe of humanity now relies on the ecosystem of the internet. It's not just how we communicate or find information, it's how our food supplies are organized, how our pharmacists track our medicines, how our planes, trains, trucks, and cargo ships are scheduled, how our financial systems work. A strong AI emerging here could wreak havoc in the way that a small child can rearrange your sock drawer or chew on the cat's tail.

As Hawking suggests, the 'evolution' of an AI could be rapid. In fact, it could emerge, evolve, and swamp the internet ecosystem in fractions of a second. That in turn raises an interesting possibility - would an emergent AI be so rapidly limited that it effectively stalls, unable to build the virtual connections and structures it needs for long term survival? While that might limit AI, it would be cold comfort for us.

I can't resist positing a connection to another hoary old problem - the Fermi Paradox. Perhaps the creation of AI is part of the Great Filter that kills off civilizations, but it also self-terminates, which is why even AI has apparently failed to spread across the galaxy during the past 13 billion years...

**AI Is Not out to Get Us**

Elon Musk’s new plan to go all-in on self-driving vehicles puts a lot of faith in the artificial intelligence needed to ensure his Teslas can read and react to different driving situations in real time. AI is doing some impressive things—last week, for example, makers of the AlphaGo computer program reported that their software has learned to navigate the intricate London subway system like a native. Even the White House has jumped on the bandwagon, releasing a report days ago to help prepare the U.S. for a future when machines can think like humans.

But AI has a long way to go before people can or should worry about turning the world over to machines, says Oren Etzioni, a computer scientist who has spent the past few decades studying and trying to solve fundamental problems in AI. Etzioni is currently the chief executive officer of the Allen Institute for Artificial Intelligence (AI2), an organization that Microsoft co-founder Paul Allen formed in 2014 to focus on AI’s potential benefits—and to counter messages perpetuated by Hollywood and even other researchers that AI could menace the human race.

AI2’s own projects may not be very flashy—they include an AI-based search engine for academic research called Semantic Scholar, for example—but they do address AI areas such as reasoning, which will move the technology beyond developing what Etzioni calls “narrow savants that can do one thing super well.”

Scientific American spoke with Etzioni at a recent AI conference in New York City, where he voiced his concerns about companies overselling the technology’s current capabilities, in particular a machine-learning technique known as deep learning. This process runs large data sets through networks set up to mimic the human brain’s neural network in order to teach computers to solve specific problems on their own, such as recognizing patterns or identifying a particular object in a photograph. Etzioni also offered his thoughts on why a 10-year-old is smarter than Google DeepMind’s AlphaGo program and on the need to eventually develop artificially intelligent “guardian” programs that can keep other AI programs from becoming dangerous.

**Is there a rift among AI researchers over the best way to develop the technology?**

Some people have gotten a little bit ahead of themselves. We’ve had some real progress in areas like speech recognition, self-driving cars (or at least the limited form of that) and of course AlphaGo. All these are very real technical achievements. But how do we interpret them? Deep learning is clearly a valuable technology, but we have many other problems to solve in creating artificial intelligence, including reasoning (meaning a machine can understand and not just calculate that 2 + 2 = 4), and attaining background knowledge that machines can use to create context. Natural language understanding is another example. Even though we have AlphaGo, we don’t have a program that can read and fully understand a paragraph or even a simple sentence.

**It’s been said that deep learning is “the best we have” in terms of AI. Is that a knock against deep learning?**

When you have a large amount of data that is labeled so a computer knows what it means, and you have a large amount of computing power and you’re trying to find patterns in that data, we’ve found that deep learning is unbeatable. Again, in the case of AlphaGo, the system processed 30 million positions that taught the AI program the right moves in different situations. There are other situations like that—for example, radiology images—where images are labeled as having a tumor or not having a tumor, and a deep-learning program can be tuned to then determine whether an image it has not previously seen shows a tumor. There is a ton of work to do with deep learning, and, yes, this is cutting-edge technology.

**So what’s the problem?**

The problem is there’s much more to intelligence than just a situation where you have that much data that can be used to train a program. Think about the data available to a student studying for a standardized test like the SAT or the New York Regents [college entrance] exams. It’s not like they can look at 30 million previous exams that are labeled “successful” or “not successful” in order to get a high score. It’s a more sophisticated, interactive learning process. Intelligence also involves learning from advice or in the context of a conversation or by reading a book. But again, despite all of these remarkable advances in deep learning, we don’t have a program that can do something a 10-year-old can do, which is pick up a book, read a chapter and answer questions about what they read.

**How would an AI’s ability to pass a standardized be a significant advance in the technology?**

We’ve actually started working on that as a research program at the Allen Institute of AI. Last year we announced a $50,000 prize [for] anyone who could build AI software that could take a standard eighth-grade science test. More than 780 teams from around the world worked for several months to do that but nobody was able to score above 60 percent—and even that’s just on the multiple-choice questions in an eighth-grade exam. That showed us a realistic and quantitative assessment of where we are today.

**How were the top-performing AI systems able to answer questions correctly?**

Often there are cues in the language. The most successful systems used carefully curated information from science texts and other public resources, which was then searched over using carefully tuned information retrieval techniques to locate the best candidate answer for each multiple-choice question. For example, what’s the best conductor of electricity: a plastic spoon, a wooden fork or an iron bat? Programs are very good at formulas and could detect that electricity and iron or conductivity and iron co-occur in large collections of documents a lot more often than, say, plastic and conductivity. So sometimes a program can take a shortcut and kind of figure it out. Almost the way kids make educated guesses. With no system scoring above 60 percent, I would say those programs were using statistics to take educated guesses as opposed to carefully reasoning through the problem.

**The DeepMind team behind AlphaGo now has an AI program that goes beyond deep learning by using an external memory system. What impact might their work have on creating more humanlike AI?**

DeepMind continues to be a leader in moving deep neural networks (AI designed to mimic the human brain) forward. This particular contribution is an important but small step towards reasoning over facts connected in a graph structure—for example a subway map. Existing symbolic programs can easily perform this task but the achievement here—which merited a Nature paper—is for a neural network to learn how to perform the task from examples. Overall, a big step for DeepMind but a small one for humankind.

**How might someone use a combination of approaches—deep learning, machine vision and memory, for example—to develop a more complete AI?**

That’s a very appealing notion and actually a lot of my research back when I was a professor at the University of Washington is based on the idea of using the Internet as a database for an AI system. We built a technique called open-information extraction and it indexed five billion Web pages, extracted the sentences from them and tried to map them into actionable knowledge for the machine. The machine had the supernatural ability to suck down Web pages and get all of the sentences. The problem is that the sentences are in text or pictures. We as humans with our brains have the remarkable ability—that we [computer scientists] haven’t yet cracked—to map that action to reasoning and so on. What makes this idea of a universal database and an artificially intelligent interface science fiction is that we haven’t figured out how to map from text and images into something that the machine can work with the same way that people can.

**You’ve mentioned that human-level AI is at least 25 years away. What do you mean by human-level AI, and why that time frame?**

The true understanding of natural language, the breadth and generality of human intelligence, our ability to both play Go and cross the street and make a decent omelet—that variety is the hallmark of human intelligence and all we’ve done today is develop narrow savants that can do one little thing super well. To get that time frame I asked the fellows of the Association for the Advancement of AI when we will achieve a computer system that’s as smart as people are in the broad sense. Nobody said this was happening in the next 10 years, 67 percent said the next 25 years and beyond, and 25 percent said “never.” Could they be wrong? Yes. But who are you going to trust, the people with their hands on the pulse or Hollywood?

**Why do so many well-respected scientists and engineers warn that AI is out to get us?**

It’s hard for me to speculate about what motivates somebody like Stephen Hawking or Elon Musk to talk so extensively about AI. I’d have to guess that talking about black holes gets boring after awhile—it’s a slowly developing topic. The one thing that I would say is that when they and Bill Gates—someone I respect enormously—talk about AI turning evil or potential cataclysmic consequences, they always insert a qualifier that says “eventually” or this “could” happen. And I agree with that. If we talk about a thousand-year horizon or the indefinite future, is it possible that AI could spell out doom for the human race? Absolutely it’s possible, but I don’t think this long-term discussion should distract us from the real issues like AI and jobs and AI and weapons systems. And that qualifier about “eventually” or “conceptually” is what gets lost in translation.

**Given AI’s shortcomings, should people be concerned about carmakers’ growing interest in self-driving vehicles?**

I'm not a big fan of self-driving cars where there’s no steering wheel or brake pedal. Knowing what I know about computer vision and AI, I’d be pretty uncomfortable with that. But I am a fan of a combined system—one that can brake for you if you fall asleep at the wheel, for example. A human driver and the automated system together can be safer than either one alone. This isn’t simple. Taking new technology and incorporating into how people work and live is not easy. But I'm not sure the solution is to have the car do all of the work.

**Google, Facebook and other prominent tech companies recently launched the** [**Partnership on Artificial Intelligence to Benefit People and Society**](https://www.partnershiponai.org/) **to set ethical and societal best practices for AI research. Is the technology advanced enough for them to have meaningful conversations?**

The leading tech corporations in the world coming together to think about these things is a very good idea. I think they did this in response to concerns about whether AI is going to take over the world. A lot of those fears are completely overblown. Even though we have self-driving cars, it’s not like 100 of them are going to get together and say, “Let’s go take the White House.” The existential risks that people like Elon Musk talk about are decades away if not centuries away. However, there are very real issues: automation, digital technology and AI in general do affect the jobs picture, whether it’s robotics or other situations, and that’s a very real concern. Self-driving cars and trucks will substantially improve safety but they will also have an impact on the large number of workers in our economy who rely on driving to earn a living. Another thing for the new group to talk about is the potential for discrimination. If AI techniques are used to process loans or credit card applications, are they doing that in a legal and ethical way?

**How do you ensure that an AI program will behave legally and ethically?**

If you’re a bank and you have a software program that’s processing loans, for example, you can’t hide behind it. Saying that my computer did it is not an excuse. A computer program could be engaged in discriminatory behavior even if it doesn’t use race or gender as an explicit variable. Because a program has access to a lot of variables and a lot of statistics it may find correlations between zip codes and other variables that come to constitute a surrogate race or gender variable. If it’s using the surrogate variable to affect decisions, that’s really problematic and would be very, very hard for a person to detect or track. So the approach that we suggest is this idea of AI guardians—AI systems that monitor and analyze the behavior of, say, an AI-based loan-processing program to make sure that it’s obeying the law and to make sure it’s being ethical as it evolves over time.

**Do AI guardians exist today?**

We issued a call to the community to start researching and building these things. I think there might be some trivial ones out there but this is very much a vision at this point. We want the idea of AI guardians out there to counter the pervasive image of AI—promulgated in Hollywood movies like The Terminator—that the technology is an evil and monolithic force.