Tech Talk | Artificial Intelligence | Machine Learning

14 Sep 2017 | 20:33 GMT

Interview: Max Tegmark on Superintelligent AI, Cosmic Apocalypse, and Life 3.0

In his new book, Max Tegmark says the outcome of AI research—and of the universe—is in our hands

By **Eliza Strickland**

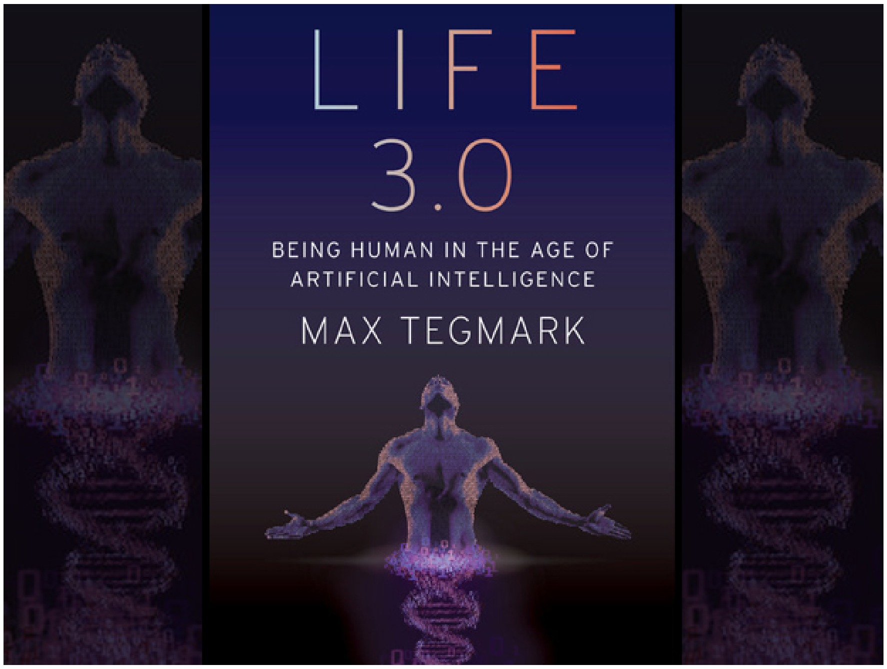

Image: Penguin Random House

Ask Max Tegmark why people should read his new book and get involved in the discussion about artificial intelligence, and you get a weighty answer. Forget about book sales, this is about cosmic destiny: The fate of the universe may well be determined by the decisions made "here on our little planet during our lifetime," he says.

In his book, *Life 3.0: Being Human in the Age of Artificial Intelligence*, Tegmark first explains how today's AI research will likely lead to the creation of a superintelligent AI, then goes further to explore the possible futures that could result from this creation. It's not all doom and gloom. But in his worst case scenario, humanity goes extinct and is replaced with AI that has plenty of intelligence, but no consciousness. If all the wonders of the cosmos carry on without a conscious mind to appreciate them, the universe will be rendered a meaningless "waste of space," Tegmark argues.

Tegmark, an MIT physics professor, has emerged as a leading advocate for research on AI safety. His thoughtful book builds on the work of Nick Bostrom, who famously freaked out Elon Musk with his book *Superintelligence*, which described in meticulous detail how a supercharged AI could lead to humanity's destruction.

Max Tegmark on . . .

Why He Disagrees With Yann LeCun

Why He Disagrees With Yann LeCun

***IEEE Spectrum*: Last Friday you had a discussion about AI with Yann LeCun, one of the most important computer scientists working on AI. LeCun said that since we don’t know what form a superintelligent AI would take, it’s premature to start researching safety mechanisms to control it. Why do you disagree?**

Max Tegmark: Just because we don’t know quite what will go wrong doesn’t mean we shouldn’t think about it. That’s the basic idea of safety engineering: You think hard about what might go wrong to prevent it from happening. Lots of people conflate safety engineering with alarmism. But when the leaders of the Apollo program carefully thought through everything that could go wrong when you sent a rocket with astronauts to the moon, they weren’t being alarmist. They were doing precisely what ultimately led to the success of the mission.

This is how we should think about AI. I like the rocket metaphor a lot. Until now, the main focus of AI has been trying to make the technology more powerful, without worrying about steering it. Like in early rocket research before they had good rocket motors—once they had the motors, they started working on how to steer the rockets too.

We need to think about how to steer AI. How do we transform today’s buggy and hackable AI systems into systems we can really trust, and make sure they really do what we intend them to do? How do we teach machines to understand our goals, adopt our goals, and retain our goals?

BACK TO TOP↑

“I Really Don’t Like It When People Ask What Will Happen in the Future”

***Spectrum:* Your book describes 12 possible futures that could result from humanity's attempt to build superintelligent AI, ranging from utopias to extinction events. Online, you've asked people to vote on which scenario they'd prefer. Given that we're still so far from creating superintelligent AI, why do you think it's important for people to contemplate these scenarios now?**

Tegmark: What I found very interesting about these scenarios: No matter how hard I tried to make every scenario sound as positive as possible, there was no scenario that I didn't have some misgivings about. It's much harder to imagine heaven than to imagine hell, and you tend to see that in world religions.

But it's when we envision positive outcomes that we foster collaboration, make people put their petty interests aside, and focus on this great goal they're envisioning. If you focus on risk, it causes fragmentation in society, which we have too much of already. Hollywood is not helping. The pictures of the future in movies tend to almost always be dystopias. When I ask my friends what they would like society to be like in 30 or 50 years, they usually haven't thought about it much, and just give me a laundry list of things they want to avoid.

Think about how many engineers got excited by reading Jules Verne's descriptions of interplanetary travel and telecommunication systems. This is why I wanted to write a more hopeful book. I think it's so important for people to think about what kind of future they want. If they don't think about it, they're less likely to get it.

***Spectrum:* Is there any one of these 12 scenarios that seems most plausible to you?**

Tegmark: I really don't like when people ask what will happen in the future, as if the future is just going to happen to us. I'd rather they ask, what can we do today to make the future good?

"With things like nuclear weapons and superintelligent AI, we don't want to learn from mistakes."

—Max Tegmark

We can create a great future with technology as long as we win the race between the growing power of technology and the wisdom with which we manage it. At the moment we're still using our old outdated strategy of learning from mistakes. We invented fire, screwed up a bunch of times, and invented the fire extinguisher. We invented the car, screwed up a bunch of times, and invented the safety belt. But with things like nuclear weapons and superintelligent AI, we don't want to learn from mistakes. We need to get it right the first time, because that might be the only time we have.

That means working on AI safety, which we're being very flippant about now. The vast majority of the billions being spent on AI research is going to make the tech more powerful. I'm not advocating slowing down that research, I'm advocating for also accelerating research on AI safety.

BACK TO TOP↑

What Types of AI Safety Research We Should Fund Now

***Spectrum:* You've been instrumental in getting that research started. You established the Future of Life Institute, which hosted the 2015 conference in Puerto Rico that brought together leading AI researchers and resulted in Elon Musk giving $10 million for research grants on AI safety. And other institutions and research groups have followed suit.**

Tegmark: Yes, but we wanted to do more than that. We wanted to seed the idea, with the hope that large funding agencies around the world would step up and support this growing field… which hasn't really happened so far. I very much hope that we can get any government that funds computer science research to view AI safety research as part of that, so the safety research gets a little slice of the pie. You'd never fund nuclear power research without also doing research on nuclear safety.

***Spectrum:* At the conference in Puerto Rico established research priorities for AI safety. If a superintelligent AI is still far in the future, what practical computer science research can be now? And are you also funding philosophers and ethicists?**

Tegmark: There is an emerging field of AI safety research. Before our Asilomar conference (in January 2017), we had a two-day workshop in which all 37 teams funded by our initial grants presented their research. Moshe Vardi said that you know that the field is a success when the talks start getting boring: These presentations were full of equations and matrices. You can see a list of all the research topics and publications on the Future of Life website.

Most of the funding went to hard core computer science researchers to work on things like AI transparency, robustness, and value alignment. But we're also funding people working on more ethical issues like lethal autonomous weapons and the impact of AI on work. They're looking at questions like, if you produce all this wealth with AI, how do you distribute it? Ensuring that AI has a beneficial impact on our civilization can't be done with equations alone.

***Spectrum:* Will this research give us techniques that are up to the task of controlling a superintelligent AI?**

Tegmark: It's important not to conflate the short-term and long-term questions relating to AI. Maybe decades from now we'll have the question of how to deal with a superintelligent AI. What's convenient is that for most of the long-term technology challenges you have, there are corresponding questions in the short term.

Suppose someone tells you that they've come up with a fantastic system for a superintelligent AI that will do everything you want… but it gets hacked. If we can't even solve cyber security 101, all that advanced work is useless. So verification, validation, and security work is a stepping stone in that direction. We can also work now on transparency: If you're going to trust a very advanced system, you want to know what its goals are, and make sure they're aligned with yours. There was an investigation by ProPublica last year about courts using machine learning to recommend who gets probation and who doesn't. It turned out the system was racially biased. Because the system wasn't transparent enough, it was doing things we didn't want it to do. That's why my research group at MIT is working on what we call intelligible intelligence.

BACK TO TOP↑

The Question of Consciousness

***Spectrum:* Why do you think that AI research should include the study of consciousness?**

Tegmark: First we need to define the word "consciousness." When I talk about consciousness I simply mean subjective experience: It feels like something to be me.

I find it remarkable how many very intelligent people in science can fail to notice that we haven't understood something yet, because it feels so familiar. People spent hundreds of years being aware that the moon doesn't fall down without asking the reason for it, until Newton found the reason.

We know that light of 700 nanometers is experienced one way, and light of 400 nanometers is experienced another way: We call one red, the other violet. People are so used to it, they forget there's something from physics we don't understand there: Why don't we experience them the other way around? We frankly have no clue. It's that kind of subjective experience that goes beyond wavelengths of light.

We've traditionally thought of both intelligence and consciousness as something mysterious and special to biological life. But from my experience as a physicist, I say they're just forms of information processing, with quarks and electrons moving around according to the laws of physics. I don't like this "carbon chauvinism" that says you can only have consciousness in something made out of carbon atoms. We're already seeing that you can have a lot of intelligence in something built out of silicon instead. I think it's plausible that you can also build something out of silicon that has consciousness.

***Spectrum:* If determining whether AI can or should be conscious is the long-term challenge, what are the short-term research questions that people can work on?**

Tegmark: There's no secret sauce to intelligence. I think there's no secret sauce to consciousness either, but there are certainly some equations we haven't discovered yet. We need to define the properties that information processing has to satisfy for there to be consciousness there.

Giulio Tononi has put forward one theory, and maybe it's right and maybe it's wrong. But it's a theory we can start testing on ourselves by looking at conscious and unconscious brain processes. If we find a theory that correctly predicts which of the information processing that happens in our own heads is conscious, then we could take it seriously for others: for pets, robots, unresponsive patients in the emergency room.

Life itself used to be considered mysterious. The difference between a living bug and dead bug was that the former had some *élan vital*, some life essence in it. But an engineer today would think of them as two mechanisms, one of which is broken. Similarly, the difference between an intelligent and a non-intelligent blob of matter is how good they are at processing information in certain ways. And the difference between a conscious and unconscious information-processing system is some property that we haven't managed to pin down yet with mathematical equations.

BACK TO TOP↑

Cosmic Optimism vs Cosmic Pessimism

***Spectrum:* Even though you think consciousness can be explained by the movement of quarks and electrons, that doesn't seem to take away the wonder you feel when contemplating the human species.**

Tegmark: I wrote the opening of chapter one to convey how awestruck I am by the marvelous patterns that our electrons and quarks make in our universe. When people say, "I can't believe I'm just a blob of quarks," I object to their use of the word "just." It's not the particles, but the pattern that matters. The pattern the particles are arranged into to form our brains is the most complex pattern the universe has ever seen.

***Spectrum:* You say in the book that all the glories of the universe are meaningless if there aren't conscious beings around to perceive them. So if humans create a superintelligent but mindless AI that wipes out our species, we may cause a cosmic apocalypse. The stakes are pretty high, huh?**

Tegmark: In the book, I contrasted the sentiments of two of my favorite scientists and friends: Steven Weinberg and Freeman Dyson. Steven Weinberg said that the more we understand our universe, the more pointless it seems. Whereas Freeman Dyson pointed out that on the scale of cosmic history, life has been a tiny perturbation in an otherwise lifeless universe—yet in the future it has the potential to completely transform our universe.

Look at an aerial photo of Manhattan today and compare it with an image of the island from 1 million years ago. Back then life had made a slight perturbation, the island is green because of life. But now it's completely transformed. And just like life has transformed Manhattan, there's no reason why life can't transform our cosmos.

I'm the optimist, not the pessimist: I'm with Freeman Dyson here. It's not our universe giving meaning to us, it's us giving meaning to our universe.

BACK TO TOP↑

AI as the "Child of All Humanity"

***Spectrum:* One of the possible futures you describe is the "descendant" scenario, in which humans create conscious AI that puts an end to our species, but then goes on to spread its form of intelligent life throughout the galaxy. Does this qualify as a good outcome?**

Tegmark: If we raise children who go on to fulfill the dreams that we couldn't fulfill ourselves, then we can be very proud of them, even if we don't live to see it all. But if we instead raise the next Adolf Hitler who destroys everything that we care about, we'll be less enthusiastic. That's why we put so much effort into raising our children: To make sure they adopt our values. That's what we as a society have to do to if we create a superintelligent AI, teach it to adopt our values. Which is easier said that done. It's a great responsibility to give birth to a child. And if it's the child of all humanity, the responsibility is even bigger.

***Spectrum:* How do you personally feel about this scenario? Are you okay with humanity being replaced?**

Tegmark: I'm not sure. The descendent scenario is my wife's favorite, and there's an active debate about it on our website. I want to think more about it, and I encourage everyone to think more about it.

The debate shouldn't be left to a bunch of computer geeks like myself. I think it's wonderful that the AI community has started to own this conversation on the technology side, that they're talking about how to build machines that are robust and can be trusted, that they're studying goal alignment. But there's also the question of what goals we want in the first place, and for that we need input from broader society.

That's partly why I wrote this book. This is a conversation that everyone needs to join. But for them to join it constructively, we need to educate them about what the challenges and opportunities actually are. Otherwise it degenerates into the scaremongering that the British tabloids do. Ultimately, this is a very exciting opportunity. Everything I love about civilization is the product of intelligence. If we can create a beneficial superintelligence, we can help humanity flourish better than ever before.

BACK TO TOP↑

About the Tech Talk blog

IEEE Spectrum's general technology blog, featuring news, analysis, and opinions about engineering, consumer electronics, and technology and society, from the editorial staff and freelance contributors.

Follow @IEEESpectrum Subscribe to RSS Feed