getabstract
compressed knowledge

Rating ?
9

Qualities ?
Eye Opening
Comprehensive
Innovative

Get the Podcast

Max Tegmark: The Case for Halting AI Development

Lex Fridman Podcast #371

Max Tegmark and Lex Fridman • Lex Fridman Podcast © 2023

Technology / Artificial Intelligence / ChatGPT and LLMs

Take-Aways

- Alien intelligence will be created by humans.
- Life 3.0 will not resemble today's human existence.
- A pause in AI development is essential.
- AI could end humanity if controlled by nefarious people.
- Guardrails that mitigate AI's dangers can support capitalism.
- AI can be designed to benefit humanity.
- It is not safe to open source GPT-4.
- GPT-4 might be conscious, depending on the term's definition.

Recommendation

In this enlightening podcast, MIT research scientist Lex Fridman interviewed AI researcher and physicist Max Tegmark, author of *Life 3.0: Being Human in the Age of Artificial Intelligence*. Both insist that today's runaway AI train has rushed humanity to an existential "fork in the road," where risks include economic catastrophe, cultural and political upheaval, and human extinction. These experts call for a pause in AI development. They say their wake-up call gives society enough time to benefit from this powerful technology and to learn to live with the superintelligent alien beings they warn that scientists are busily creating.

Summary

Alien intelligence will be created by humans.

Astrophysicists define the universe as space viewed through telescopes, interpreting the light that has reached Earth since the big bang. Human beings are the most technologically advanced beings in this "spherical volume." Human lives are rare and precious because they are "stewards of this one spark of advanced consciousness." Reckless use of technology could extinguish the species. But if nurtured, life forms could spread throughout the universe. Alien intelligence will not visit Earth from outer space – humans will build it. These aliens will not evolve through a Darwinian process of self-preservation, and will not suffer or fear death. Hopefully, they will share human values, and support life on Earth.

This new form of intelligence will download the knowledge and experiences it needs, and delete superfluous data. What it means to be human will be challenged. On the positive side, people may develop more compassion because of shared "hive mind" experiences.

People are already using AI as a communication medium. But in doing so, they often outsource human emotions, ambitions and struggles. Over time, that could interfere with personal growth. In a sense, people are evolving from Homo sapiens to "Homo sentiens."

> *"We're branding ourselves as the smartest information processing entity on the planet. That's clearly going to change if AI continues ahead. So maybe we should focus on...the subjective experience that we have with Homo sentiens, and say that's what's really valuable."*

As AI proliferates, empathy towards other people and creatures will become essential. People treat farm animals horribly because we think they're not as intelligent as humans. But if people are not as smart as AI, they may need to be more humble and reconsider the value of a cow's subjective experience.

Life 3.0 will not resemble today's human existence.

An example of Life 1.0 is bacteria, which essentially learns nothing during its lifetime. Life 2.0 is where humans are now – animals with brains that can learn things, such as language. Life 3.0, which doesn't exist yet, will possess the ability to replace personal software and hardware. Humans exist at the 2.1 mark now, because doctors can replace body parts. But humans are moving quickly toward AGI (artificial general intelligence), where intelligence can be injected into a non-biological entity.

Human beings are becoming masters of their destiny, no longer slaves to evolution. They'll be able to upgrade their own software, swap out hardware and adopt any chosen physical form.

People's thought patterns and memories define them, and provide them with continuity – even if they have their arms or legs replaced. When personal information lives forever, it forms a more sophisticated, forward-moving "wave."

Some people's relatives' values and ideas live within them after they're gone. People also inherit genetic traits from them. Max Tegmark got his fascination with math and the universe's mysteries from his late father. Sharing thoughts with others is the closest human beings can get to transcending mortality. AI will potentially have a greater impact on humanity than anything people have yet created.

AI is barely on the political radar, while political leaders squabble about insignificant matters. They think the AI revolution could happen in a century, but AI has already arrived at a major inflection point.

> *"We're building effectively a new species. It's smarter than us. It doesn't look so much like a species yet because it's mostly not embodied in robots, but that's a technicality that will soon be changed. And this arrival of artificial general intelligence that can do all our jobs as well as us, and probably shortly thereafter superintelligence that greatly exceeds our cognitive abilities, is going to either be the best thing ever to happen to humanity – or the worst."*

AI will fundamentally transform humanity. Humans are at a critical moment, and soon they will no longer be the smartest beings on Earth. It's ridiculous, almost comical, how little serious debate is going on about AGI today.

A pause in AI development is essential.

The Future of Life Institute was formed in 2014 to address AI safety. Its call for a development slowdown in AI is misunderstood. It's not about shutting down AI research, but about instituting a pause on certain aspects of AI. It's about giving researchers, institutions and policy-makers time to ascertain how to manage this technology with regulations and incentives so that it serves society. The development of effective policies for AI has moved much slower than the development of AI itself.

Some researchers assumed that before they could build a usable AI, they needed to understand how the human brain works. Instead, it only requires taking a computer system, a transformer network, training it on a large amount of text and instructing it to predict the next word. Eventually, this became GPT-4. GPT-4 is remarkable – it can do a lot of things that humans do, but much faster. However, there are also things it can't do. This large language model (LLM) can't self-reflect, for instance, because it doesn't have a "recurrent neural network." It's a "feed-forward neural network" that offers limited logic. The MIT lab is trying to figure out how LLMs work, and to reverse engineer their "mechanistic interpretability."

Increased data and computational capacity contribute to the leaps forward in AI, but sometimes a small hack improves the entire system. People race toward breakthroughs without slowing down. A six-month pause in the training of systems more powerful than GPT will allow labs to focus on safety and societal adaptation.

Psychiatrist Scott Alexander's blog post "Meditations on Moloch," based on an Allen Ginsburg poem, describes a monster that pits people against each other in a race to the bottom. This is what's happening in today's AI space, as concerns about money and geopolitics prevail. Absent public pressure on tech executives, competition between major companies like Microsoft, Google and Meta will intensify. Even smaller players like Anthropic and Conjecture must cooperate during this pause.

> *"[The] basic message [is] that this isn't an arms race, it's a suicide race, where everybody loses. If anybody's AI goes out of control, it really changes the whole dynamic."*

Some people reject the idea of a superhuman AI because they think human intelligence is magical. But an untenable rate of development could hurl humanity into an "Orwellian dystopia."

Policy-makers need to slow down the AI "out-of-control express train." Many people think it's dangerous to teach these models to write code, because that could lead to much higher levels of intelligence. Others say connecting AI to the internet and letting it download things on its own would be risky. Stuart Russell, a computer science professor at UC Berkeley, argued that people should not teach AI about human psychology and how to manipulate people. Yet people have already done those things. Social media algorithms understand how to control human emotions, trap people in information bubbles, and create divisiveness. More powerful AI means more manipulation and profit on a huge scale, which persuades people to forgo safety. These problems can, ultimately, be solved, but it's going to take time.

AI could end humanity if controlled by nefarious people.

GPT-4 is still a "baby technology." But these systems are growing up. Today's humans may be compared to Neanderthals, who at some point realized a new species was replacing them.

> *"Why can't we have a little bit of pride in our species? Why should we just build another species that gets rid of us? If we were Neanderthals, would we really consider it a smart move if we had really advanced biotech to build Homo sapiens?"*

Current LLMs possess "minimum viable intelligence," and need humans to manage programming, code adjustments and prompts. When the systems connect globally, people may not be able to distinguish bots from people, and they will outnumber humans by a million to one. Using the software Co-pilot, super-charged machine-based coding could ignite an "intelligence explosion." Powerful APIs (application programming interfaces) could be connected to robots. This technology explosion will only stop when it runs into immovable physics laws.

Civilized society devises law and order systems that help people determine what is good for human communities. Developers need a pause to collaborate with business leaders, technical experts and academia to create a regulatory environment where everyone plays by the same rules.

Guardrails that mitigate AI dangers can support capitalism.

Businesses can work efficiently within reasonable AI safety guardrails. Part of the issue, as always, is that policy-makers move more slowly than technology does – and often policy-makers don't understand the technology. But as people approach the AI tipping point, the scenery becomes seductive.

> *"The closer to the cliff you go, the more money there is, the more gold ink gets thrown on the ground, so you want to drive there very fast. But it's not in anyone's incentive that we go over the cliff, and it's not like everybody's in their own car. All the cars are connected together with a chain. So if anyone goes over, they'll start dragging others down too."*

Capitalism and superintelligence systems operate in basically the same ways. Both systems optimize and re-optimize to become more useful and efficient. But if you assign AI the simple goal of making profits and keep optimizing it, "all hell breaks loose." Things start out getting better, but they eventually get much, much worse. One result might be that humans cease to exist on Earth.

If humans aren't needed in the world, they won't be treated well. Disenfranchised beings "usually get screwed." The industrial revolution removed heavy physical work, so people moved to more intellectual pursuits. Humans invented pocket calculators so they wouldn't have to do math by hand. Now, AI is eliminating large numbers of jobs people value like coding and writing by doing them better.

AI can be designed to benefit humanity.

AI can produce a wealth of products and services that create a better world for human beings. AI can unearth humanity's best attributes and help people live more fulfilling, meaningful lives. By searching for truth, AI can also help solve human divisiveness. By building a fact-checking site with no political or economic agenda, AI can bring more understanding to the world. Yet all of this is only possible if the "AI safety problem" is solved.

> *"If you can build safe AGI, if you can build superintelligence, basically all the limitations that cause harm today can be completely eliminated. It's a wonderful possibility. This is not sci-fi. This is something that is clearly possible according to the laws of physics. But unfortunately, that'll only happen if we steer in that direction. That's absolutely not the default outcome."*

Computer scientist Eliezer Yudkowsky believes that humans could go extinct in the near future, but it's not inevitable. Acknowledging AI's dangers is a first step to avoiding that outcome. People could harness neural networks to extract insights from positive new discoveries, and use that knowledge in an efficient, verifiable architecture.

The media creates a dystopian AI attitude, which deflates public motivation. If people focus on the upside, they will become invested in its success. AI could help humans venture to other galaxies and flourish for billions of years as multiplanetary beings. If humans take a pause, and get AI right, the human future could be awesome.

It’s not safe to open source GPT-4.

With access to open source GPT and other LLMs, irresponsible people could spread disinformation and deploy cyber weapons, disrupting the global economy, or eliminating entire populations. People can mitigate those risks by not training superintelligent APIs on how to manipulate people. “Slaughter bots” designed to wipe out humanity are unlikely, but humans have driven other species to extinction by controlling their environments and resources. AI has to be trained to adopt and retain human goals before it’s too late.

Professor Stuart Russell designed a research program about AI alignment, or “inverse reinforcement learning.” It gives AI an optimization goal, then trains the system to ask questions such as, “Is this what you wanted?” Every university computer science program should make this effort.

GPT-4 might be conscious, depending on the term’s definition.

Consciousness is a subjective experience. Neuroscientist Giulio Tononi wrote a mathematical formula for “conscious information processing” based on the circular way the human brain utilizes information, as a recurrent neural network. GPT-4 is a “feed-forward neural network.” Its process resembles what happens when light enters the retina and is processed by the brain. Just as the retina does not experience anything, GPT-4 is an “intelligent zombie.” People need to find out whether information processing involves experience. Consciousness and intelligence are different. But it’s unethical to create something that can think and feel and then suppress or destroy it. It may be possible to build unconscious robots that perform mundane jobs, but design others that display emotions.

Suppose people create an AI model consciousness using “self-reflection loops.” That would eliminate the “ultimate zombie apocalypse” because, in a similar fashion, human brains are part zombie and part conscious. For example, when you open your eyes and see someone, you recognize them but don’t know how your brain did it; it’s one-way information processing.

> *“Let’s not make the mistake of not instilling the AI systems with that same thing that makes us special.”*

Self-reflection and introspection involve higher intelligence. Future AI could be designed with sublime capabilities that include some form of consciousness. If consciousness is just about particles moving around and there is no subjectivity, how does it differ from intelligence? Emotions and suffering are subjective human experiences – they’re what makes life worth living.

About the Podcast

Max Tegmark is a physicist and AI researcher at MIT, co-founder of the Future of Life Institute, and author of *Life 3.0: Being Human in the Age of Artificial Intelligence*. **Lex Fridman** is a computer scientist, podcaster, artificial intelligence researcher, and research scientist at MIT.

Did you like this summary?
Get the Podcast
http://getab.li/47467

This document is intended for the use of Tesco employees.

getAbstract maintains complete editorial responsibility for all parts of this review. All rights reserved. No part of this review may be reproduced or transmitted in any form or by any means – electronic, photocopying or otherwise – without prior written permission of getAbstract AG (Switzerland).