https://www.wsj.com/tech/ai/how-worried-should-we-be-about-ais-threat-to-humanity-even-tech-leaders-cant-agree-46c664b6

How Worried Should We Be About AI's Threat to Humanity? Even Tech Leaders Can't Agree.

Artificial-intelligence experts debate whether to focus on averting an AI apocalypse or problems such as bias, disinformation

By Sam Schechner Follow *and Deepa Seetharaman* Follow

Sept. 4, 2023 8:00 am ET

Artificial-intelligence pioneers are fighting over which of the technology's dangers is the scariest.

One camp, which includes some of the top executives building advanced AI systems, argues that its creations could lead to catastrophe. In the other camp are scientists who say concern should focus primarily on how AI is being implemented right now and how it could cause harm in our daily lives.

Dario Amodei, leader of AI developer Anthropic, is in the group warning about existential danger. He testified before Congress this summer that AI could pose such a risk to humankind. Sam Altman, head of ChatGPT maker OpenAI, toured the world this spring saying, among other things, that AI could one day cause serious harm or worse. And Elon Musk said at a Wall Street Journal event in May that "AI has a nonzero chance of annihilating humanity"—shortly before launching his own AI company.

Altman, Musk and other top AI executives next week are expected to attend the first in a series of closed-door meetings about AI convened by U.S. Senate Majority Leader Chuck Schumer (D., N.Y.) to consider topics including "doomsday scenarios."

The other camp of AI scientists calls those warnings a science-fiction-fueled distraction—or even a perverse marketing ploy. They say AI companies and regulators should focus their limited resources on the technology's existing and imminent threats, such as tools that help produce potent misinformation about elections or systems that amplify the impact of human biases.

The dispute is intensifying as companies and governments worldwide are trying to decide where to focus resources and attention in ways that maximize the benefits and minimize the downsides of a technology widely seen as potentially world-changing.

"It's a very real and growing dichotomy," said Nicolas Miailhe, co-founder of the Future Society, a think tank that works on AI governance and is working to bridge the divide. "It's the end of the month versus the end of the world."

For all the attention it has been getting, serious public discussion of AI's existential risk—or "x-risk" as those most worried about it like to call it—had until recently remained confined to a fringe of philosophers and AI researchers.

That changed after OpenAI's release of ChatGPT late last year and subsequent improvements that have delivered humanlike responses, igniting warnings that such systems could gain superhuman intelligence. Prominent researchers including Geoffrey Hinton, considered one of the godfathers of AI, have contended it contains a glimmer of humanlike reasoning. Hinton left his role at Alphabet's Google this year to more freely discuss AI's risks.

With existential risk warnings, "there's been a taboo that you'll be mocked and treated like a crazy person and it will affect your job prospects," said David Krueger, a machine learning professor at the University of Cambridge. Krueger helped organize a statement in May saying that extinction risk from AI was on par with the dangers of pandemics and nuclear war. It was signed by hundreds of AI experts, including top officials and researchers at Google, OpenAI and Anthropic.

"I wanted researchers to know that they're in good company," Krueger said.

Some in the field argue that there is a paradoxical upside for AI companies to emphasize the x-risk of the systems because it conveys a sense that their technology is extraordinarily sophisticated.

"It's obvious that these guys benefit from the hype still being fueled," said Daniel Schoenberger, a former Google lawyer who worked on its 2018 list of AI principles and now is at the Web3 Foundation. He said policy makers should focus more on near-term risks, such as AI making it cheaper to mount campaigns to disseminate false and misleading information, or concentrating more power in Silicon Valley.

"There is the risk of dominance, of Big Tech becoming Big AI," Schoenberger said.

AI leaders worried about existential risks say their concerns are genuine, not a ploy. "To say, 'Oh the governments are hopeless, so the call for regulation is some sort of 4D chess move'—it's just not how we think. This is an existential risk," OpenAI's Altman said in June.

So-called doomers don't say that AI will necessarily rise like Skynet in the Terminator movies to destroy humans. Some worry that AI systems trained to seek rewards could end up with hidden power-seeking urges, inadvertently harm humans while carrying out our wishes or simply outcompete humans and take control of our destiny. Research in this community focuses largely on what is called alignment—how to make sure tomorrow's computer minds have goals intrinsically in sync with ours.

Specialists in AI ethics and fairness, by contrast, are concerned about how the tools are, accidentally or intentionally, exploiting workers and deepening inequality for millions of people. They want tech companies and regulators to implement training standards and techniques to reduce that threat.

Diversity is a flashpoint. AI ethicists have shown how AI systems trained on historical data can bake past discrimination into future high-stakes decisions such as housing, hiring or criminal sentencing. Research also has shown that generative AI systems can produce biased images. They also argue that a lack of diversity among AI researchers can blind them to the impact AI could have on people of color and women.

The debate can get spirited. "What is your plan to make sure it doesn't have an existential risk?" Max Tegmark, president of the Future of Life Institute, demanded of Melanie Mitchell, a prominent AI researcher and professor at the Santa Fe Institute, during a public forum on x-risk in June. "You're not answering my question."

"I don't think that there is an existential risk," Mitchell shot back, adding that people are working hard "on mitigating the more immediate, real-world risks," while Tegmark widened his eyes.

Mitchell said in an interview that the discussion over existential risk is "all based on speculation, there's really no science."

Tegmark, a professor at the Massachusetts Institute of Technology whose nonprofit aims to prevent technology from creating extreme, large-scale risks, said that he thinks companies have an interest in stoking a divide between people focused on fairness issues and existential risk to avoid regulation.

"People on both sides of this are doing themselves a disservice if they don't agree with the other side," he said in an interview.

The conflict has caused sparks for years. In 2015, some academics and scientists gathered to discuss AI's risks on the sidelines of a conference hosted on Google's campus. One side confronted existential-risk proponents, arguing that the focus should be on present-day harms, including bias.

The x-riskers retorted that with humanity's future in the balance, no one should worry about AI's causing a quarter-point difference on a mortgage, recalled Steven Weber, a professor at the University of California, Berkeley who was present.

"I almost thought it was going to be a fistfight at an academic meeting," Weber said.

Those concerned about an apocalypse are far from unified, with some doomers arguing that even executives at big companies who say they are worried aren't doing enough to avoid it.

"We're seeing a ridiculous death race to godlike AIs from all the major players," said Connor Leahy, chief executive of Conjecture, an AI company working on solutions to the alignment problem. He said that he takes larger tech companies' professions of concern about existential risk with a grain of salt. "Watch the hands, not the mouth," Leahy says.

There are efforts to bridge the divide, too. Some ethics researchers say they don't entirely discount existential risk, but just think it should be tackled as part of more well-defined problems that exist today. Some doomers say that the path to catastrophe could well come from concerns highlighted by the ethics community, such as industrialized disinformation toppling governments or starting wars. Both sides are interested in being able to pierce the black-box of how AI thinks, called the interpretability problem.

"Some people are really trying to bridge these two spaces," said Atoosa Kasirzadeh, an assistant professor of AI ethics at University of Edinburgh who previously worked for Google DeepMind. "Hopefully those communities can be convinced they are all concerned about the same sort of things deep down."

Write to Sam Schechner at Sam.Schechner@wsj.com and Deepa Seetharaman at deepa.seetharaman@wsj.com

Appeared in the September 5, 2023, print edition as 'Tech Leaders Are Divided On AI's Threat to Humanity'.