‘What should the limits be?’ The father of ChatGPT on whether AI will save humanity – or destroy it

Publication Date: 2023-06-07

Author: Alex Hern

Section: Technology

Tags: Artificial intelligence (AI), ChatGPT, Computing, Consciousness, features

Article URL: https://www.theguardian.com/technology/2023/jun/07/what-should-the-limits-be-the-father-of-chatgpt-on-whether-ai-will-save-humanity-or-destroy-it

When I meet Sam Altman, the chief executive of AI research laboratory OpenAI, he is in the middle of a world tour. He is preaching that the very AI systems he and his competitors are building could pose an existential risk to the future of humanity – unless governments work together now to establish guide rails, ensuring responsible development over the coming decade. In the subsequent days, he and hundreds of tech leaders, including scientists and “godfathers of AI”, Geoffrey Hinton and Yoshua Bengio, as well as Google’s DeepMind CEO, Demis Hassabis, put out a statement saying that “mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war”. It is an all-out effort to convince world leaders that they are serious when they say that “AI risk” needs concerted international effort. It must be an interesting position to be in – Altman, 38, is the daddy of AI chatbot ChatGPT, after all, and is leading the charge to create “artificial general intelligence”, or AGI, an AI system capable of tackling any task a human can achieve. Where “AI” is bandied about to describe anything more complex than a Tamagotchi, AGI is the real thing: the human-level intelligence of stories such as Her, Star Trek, Terminator, 2001: A Space Odyssey and Battlestar Galactica. On his world tour, further complicating his position, Altman is also preaching something besides the dangers of unchecked AI development. He is arguing that the benefits of developing “superintelligence” – an AGI turned up to 11, capable of solving problems humanity has been unable to crack – are so great that we should risk the destruction of everything to try to do it anyway. It is a gruelling few weeks. On the day I meet him, he woke up in Paris having met with Emmanuel Macron the night before. A Eurostar trip to London and a quick hop to Oxford later, he is giving a talk to the Oxford Guild, a business-focused student society, before a few more meetings, then off to Number 10 for a sit down with Rishi Sunak. Later he boards a flight to Warsaw before heading to Munich the following morning. His PR team is rotating in and out, but Altman’s in it for a five-week stint. “I love San Francisco and the Bay Area,” he says on stage at the Oxford event. “But it is a strange place, and it’s quite an echo chamber. We set up this trip to start to answer this question, with leaders in different places, about, like, what should the limits of these systems be, to decide how should the benefits be shared. And there are very different perspectives between most of the world and San Francisco.” To the exasperation of his team, hearing as many perspectives as possible clearly takes priority over plans for the day. After an event at UCL, he wanders down into the audience – a casual conversation that leads to headlines in Time and the FT. Just as he is about to sit down and start talking to me, he goes outside to speak to a small collection of protesters holding signs exhorting OpenAI to “stop the AGI suicide race”. “Stop trying to build an AGI and start trying to make sure that AI systems can be safe,” says one of the protesters, an Oxford University student called Gideon Futerman. “If we, and I think you, think that AGI systems can be significantly dangerous, I don’t understand why we should be taking the risk.” Altman, a classic dropout founder in the Mark Zuckerberg mould – he quit Stanford university in his third year to launch a social network called Loopt – seems in full politician mode as he tries to find middle ground. “I think a race towards AGI is a bad thing,” Altman says, “and I think not making safety progress is a bad thing.” But, he tells the protester, the only way to get safety is with “capability progress” – building stronger AI systems, the better to prod them and understand how they work. Altman leaves Futerman unconvinced, but as we head back down, he’s sanguine about the confrontation. “It’s good to have these conversations,” he says. “One thing I’ve been talking a lot about on this trip is what a global regulatory framework for superintelligence looks like.” The day before we meet, Altman and his colleagues published a note outlining their vision for that regulation: an international body modelled on the International Atomic Energy Agency, to coordinate between research labs, impose safety standards, track computing power devoted to training systems and eventually even restrict certain approaches altogether. He was surprised by the response. “There’s a ton of interest in knowing more; more than I was expecting, from very senior politicians and regulators, about what that might look like. I’m sure we’ll talk about much near-term stuff, too.” But that distinction, between the near and the long-term, has earned Altman no shortage of criticism on his tour. It’s in OpenAI’s interest, after all, to focus regulatory attention on the existential risk if it distracts governments from addressing the potential harm the company’s products are already capable of causing. The company has already clashed with Italy over ChatGPT’s data protection, while Altman started his trip with a visit to Washington DC to spend several hours being harangued by US senators over everything from misinformation to copyright violations. “It’s funny,” Altman says, “the same people will accuse us of not caring about the short-term stuff, and also of trying to go for regulatory capture” – the idea that, if onerous regulations are put in place, only OpenAI and a few other market leaders will have the resources to comply. “I think it’s all important. There’s different timescales, but we’ve got to address each of these challenges.” He reels off a few concerns: “There’s a very serious one coming about, I think, sophisticated disinformation; another one a little bit after that, maybe about cybersecurity. These are very important, but our particular mission is about AGI. And so I think it’s very reasonable that we talk about that more, even though we also work on the other stuff.” He bristles slightly when I suggest that the company’s motivations might be driven by profit. “We don’t need to win everything. We’re an unusual company: we want to push this revolution into the world, figure out how to make it safe and wildly beneficial. But I don’t think about things in the same way I think you do on these topics.” OpenAI is indeed unusual. The organisation was founded in 2015 as a non-profit with a $1bn endowment from backers including Elon Musk, PayPal co-founder Peter Thiel and LinkedIn co-founder Reid Hoffman. Altman initially acted as co-chair alongside Musk, with a goal “to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return”. But that changed in 2019, when the organisation reshaped itself around a “capped profit” model. Altman became CEO, and the organisation began taking external investment, with the proviso that no investor could make more than 100 times their initial input. The rationale was simple: working on the cutting edge of AI research was a lot more expensive than it had first seemed. “There is no way of staying at the cutting edge of AI research, let alone building AGI, without us massively increasing our compute investment,” OpenAI chief scientist Ilya Sutskever said at the time. Altman, already independently wealthy – he made his first fortune with Loopt, and his second as the president of startup accelerator Y Combinator – didn’t take any equity in the new company. If AI does end up reshaping the world, he won’t benefit any more than the rest of us. That’s important, he says, because while Altman is convinced that the arc bends towards the reshaping being broadly positive, where he’s less certain is who wins. “I don’t want to say I’m sure. I’m sure it will lift up the standard of living for everybody, and, honestly, if the choice is lift up the standard of living for everybody but keep inequality, I would still take that. And I think we can probably agree that if [safe AGI] is built, it can do that. But it may be a very equalising force. Some technologies are and some aren’t, and some do both in different ways. But I think you can see a bunch of ways, where, if everybody on the Earth got a way better education, way better healthcare, a life that’s just not possible because of the current price of cognitive labour – that is an equalising force in a way that can be powerful.” On that, he’s hedging his bets, though. Altman has also become a vocal proponent of a variety of forms of universal basic income, arguing that it will be increasingly important to work out how to equitably share the gains of AI progress through a period when short-term disruption could be severe. That’s what his side-project, a crypto startup called Worldcoin, is focused on solving – it has set out to scan the iris of every person on Earth, in order to build a cryptocurrency-based universal basic income. But it’s not his only approach. “Maybe it’s possible that the most important component of wealth in the future is access to these systems – in which case, you can think about redistributing that itself.” Ultimately, it all comes back to the goal of creating a world where superintelligence works for us, rather than against us. Once, Altman says, his vision of the future was what we’d recognise from science fiction. “The way that I used to think about heading towards superintelligence is we were going to build this one extremely capable system. There were a bunch of safety challenges with that, and it was a world that was going to feel quite unstable.” If OpenAI turns on its latest version of ChatGPT and finds it’s smarter than all of humanity combined, then it’s easy to start charting a fairly nihilistic set of outcomes: whoever manages to seize control of the system could use it to seize control of the world, and would be hard to unseat by anyone but the system itself. Now, though, Altman is seeing a more stable course present itself: “We now see a path where we build these tools that get more and more powerful. And, there’s billions, or trillions, of copies being used in the world, helping individual people be way more effective, capable of doing way more. The amount of output that one person can have can dramatically increase, and where the superintelligence emerges is not just the capability of our biggest single neural network, but all of the new science we’re discovering, all of the new things we’re creating. “It’s not that it’s not stoppable,” he says. If governments around the world decided to act in concert to limit AI development, as they have in other fields, such as human cloning or bioweapon research, they may be able to. But that would be to give up all that is possible. “I think this will be the most tremendous leap forward in quality of life for people that we’ve had, and I think that somehow gets lost from the discussion.”