



If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All, Eliezer Yudkowsky & Nate Soares, 2025, 260 pages, with notes and interspersed QR Codes for further resources.

Eliezer Yudkowsky is a founding researcher of the field of AI alignment and co-founder of the Machine Intelligence Research Institute (MIRI), San Francisco. Nate Soares is president of MIRI. (<https://ifanyonebuildsit.com/>)

Review by K. D. Kragen, KaveDragen Ink - kdkragen.org

Yudkowsky and Soares map out their case, and compelling arguments, with a clarity that even non-geeks and non-hackers (white-hat of course) should be able to follow. The book is also full of creative futurist scenarios to help illustrate potential outcomes – to be wise concerning, and to avoid.

Among the many well-articulated warnings throughout the book, in Part III the authors draw analogies with the technological case studies of nuclear power, specifically the failure of Chernobyl Reactor #4, and many mishaps and failures of space exploration satellites. These two areas of modern technology serve as applicable illustrations of specific concerns with AI-to-ASI technology. Though today nuclear power production is far safer than in the mid-1980s USSR, nonetheless bureaucracy and speed of fission nuclear reaction both can negatively affect nuclear power plant safety and security. In the history of space exploration satellites, often the issues is the unpredictability of the outer space environment and resultant loss of control of a satellite far from Earth. With Chernobyl, a poor safety culture prevailed to a meltdown disaster in April 1986. With many of the multi-million dollar exploration satellites, despite excellent engineering, technological redundancies, and dedicated scientists and engineers, many satellites were lost or failed their missions.

The authors thus see strong correlations with these two technology fields of nuclear power and space exploration concerning the effect of a culture of poor safety and oversight in the computer industry in general, and especially in the AI and LLM computer worlds. In other words, corporate and government sector computers are far more vulnerable to hacking, security issues, and failures than many of the technologies before them. Cybersecurity is too often multiple steps behind security breaches and loss of control. Among the many other pressing concerns with AI-to-ASI dangers, computer security is a key area of vulnerability in the present world's AI arms race.

An artificial superintelligence is like a space probe, in that we cannot test it in quite the same environment where it needs to work, and by default it is not retrievable or correctable once it rises high above us.... And ASI alignment [control] has it even worse than space probes: Failure will destroy not just billions of dollars of investment, but *everything*.

An artificial superintelligence is like a nuclear reactor, in that its underlying reality involves immense, potentially self-amplifying forces, whose inner processes run faster than humans can react.

An artificial superintelligence is like a computer security problem, in that every constraint an engineer tries to place upon the system might be bypassed by the intelligent forces that those constraints hinder....

AI is grown, not crafted. Whatever vast complications lay inside AIs and lend them their power of intelligence, nobody knows them. [p. 175-76]

Cybersecurity and a culture of poor safety is one of the weakest links in future control of AI and checks against an AI-to-ASI disaster. The authors supply a great tech anecdote that exemplifies the cybersecurity problems related to tech culture's vulnerability. "Elon Musk, the head of a major AI lab named xAI, shared his plan for ASI alignment [and control] in a 2023 interview:

I'm going to start something called TruthGPT. Or a maximum truth-seeking AI that tries to understand the nature of the universe.

I think this might be the best path to safety, in the sense that an AI that cares about understanding the universe is unlikely to annihilate humans, because we are an interesting part of the universe.

“This plan [completely] fails to address the problem at hand.... Nobody knows how to engineer exact desires into an AI, idealistic or not.

“We respect Musk’s success in other areas, including electric cars and reusable rockets.

Landing rockets undamaged is a hard engineering challenge that Musk and his team regularly succeed at. But that would have been based on far more solid engineering principles. Why does he put his hope in vague idealistic platitudes in the case of AI? You couldn’t get a car or a rocket to work using that level of understanding.” (p. 181) The root of the issue for Yudkowsky & Soares: “The inner workings of batteries and rocket engines are well understood, governed by known physics recorded in careful text books. AIs, on the other hand, are grown, and no one understands their inner workings.” (p. 182)

Yudkowsky and Soares are clear about their materialist (empiricist) bias underlying the dire conclusions about emergent ASI; the book is in part aimed at AI industry leaders such as Musk, Altman, Zuckerberg, Bezos, Gates, who it is hoped may heed these warnings. If on the other hand one is not a strict materialist, but holds a more nuanced metaphysics that includes some construal of immaterial reality, as do many scientists and philosophers, myself included, then one would need to take the author’s presuppositions into account, while still nonetheless heeding their concerns.

If Anyone Builds It contributes much to the discussion on ASI singularity, and makes an important case concerning a culture of arrogance, hubris, even outright greed, that severely compromises computer technology industries and their leadership. For me it may be more a concern with that leadership, as well as the present governments fulling the AI arms race. I believe Yudkowsky and Soares would agree that other factors equally fuel the problems with AI technology, e.g. the massive environmental degradation of a runaway technology, and the nearly complete destruction of the veracity of information systems so easily hacked, manipulated, and propagandized by bad actors.

This latter issue is well documented by Nina Schick in *Deep Fakes: The Coming Infocalypse 2020* (<https://ninaschick.org/about-nina-schick>). Schick lays out a brilliant account of a growingly polluted information ecosystem. Her warning to the world: “If you do not want the [infocalypse]... to become a permanent reality, engage now. Be careful about what information you share. Verify your sources. Correct yourself when you get something wrong. Be wary of your own political biases. Be skeptical, but not cynical” (*Deep Fakes*, pp. 205-206). These two books together, *Everyone Dies* and *Deep Fakes*, do a good job of educating us all, as earthlings sharing a small planet, about the technologies that more and more are dominating our future and our present.

The last two chapters lays out strategies for survival, before ASI takes control. First, in the tradition of Cold War era nuclear test ban treaties and nuclear disarmament strategies, the only long term solution must include an AI-expansion treaty agreed upon by all participating nations, including an international monitoring agency. Unlike the disaster incident of Chernobyl #4 Reactor (discussed earlier) which was a wake-up call on nuclear power plant safety preparedness, there won’t be an “ASI-incident”; if an ASI takes control, “everyone dies.” This is the conclusion of Eliezer Yudkowsky and Nate Soares. Halting AI research is thus only a first step. Second, the authors “recommend augmenting humans to make them smarter, smart enough to get us out of this mess. We believe the ASI alignment problem (i.e., maintaining “control” of an ASI) is possible to solve *in principle*, by the sort of people so inhumanly smart that they never optimistically believe some plan will work when it won’t” (p. 218).

As noted earlier, the authors’ materialist presuppositions informs one of their solution for human survival. Make smarter, faster thinking humans through augmentation. While human “intelligence” (reasoning, smarts) as compared to a potential ASI is ultimately a matter of degree, machine intelligence is just very fast and structurally unpredictable by slower human reasoning. It is thus logical to conclude that augmented humans, as also discussed by Johnson et. al. in their book *OfficeShock* (<https://www.iff.org/projects/officeshock/>), would be a longer term strategy for keeping carbon-based homosapiens at the top of the cognitive food-chain.

See also *Battle for Your Brain: Defending The Right To Think Freely In The Age Of Neurotechnology*, Nita A. Farahany, 2023 (<https://www.nitafarahany.com/the-battle-for-your-brain>).