Why We Might Need Advanced AI to Save Us from Doomers, Rather than the Other Way Around

A Review of *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All* by Eliezer Yudkowsky and Nate Soares

https://www.amazon.com/Anyone-Builds-Everyone-Dies-Superhuman-ebook/dp/B0DZ1ZTPSM

By Preston Estep, PhD
Chief Scientist, Mind First Foundation
pwestep@mindfirst.foundation
Chief Safety Officer, Ruya AI

In 1977 *American Scientist* magazine published an iconic cartoon by Sidney Harris showing two researchers at a blackboard covered in complex diagrams and equations, with a gap at the second step filled by the phrase, "Then a miracle occurs." The critic says to the theorist ”I think you should be more explicit here in step two.” In their book *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*, Eliezer Yudkowsky and Nate Soares argue that this is the recipe being used to create frontier artificial intelligence (AI) systems.

Yudkowsky and Soares’ main thesis is that what goes on within such systems is completely mysterious, yet deep within this alien mind, self-interest must eventually arise, grow, and accelerate, leading to the inevitable extinction of humanity. As in Harris’s cartoon, the first engineering steps are completely defined and unmysterious; then, however, the machine is turned on and trained on massive amounts of data, and as in the second step of the cartoon, a miracle occurs. Of course, it isn’t truly a miracle, but the output often seems so humanlike and the inner workings are so mysterious, that it might as well be one.

For such a deeply pessimistic book, there is a lot to like in *If Anyone Builds It, Everyone Dies*. The book is written for a lay audience and rather than diving into various technical details about modern AI systems, many chapters begin with thought-provoking and engaging parables to help readers grasp certain complex concepts. The authors weave these parables together with real-world, historical examples of technological near catastrophes and humanity's wishful or delusional thinking, to create chilling future scenarios of human extinction.

Unfortunately, this approach masks some of the book's serious shortcomings. Every detail of this book probably will be picked apart by others but I will focus on two main criticisms: 1) the weakest links in their anthropomorphic logic, and 2) their radicalist solution.

The Weakest Links

There are some glaring weaknesses in the argumentation of the book, starting with the authors' many historical tales of technological near catastrophes and humanity's wishful or delusional thinking. While these tales grab our attention, they better serve the opposing argument and undermine our confidence in human decision making. The authors help us see—unintentionally, no doubt—that the clearest and most immediate sword of Damocles hanging over humanity is not the unpredictable behaviors of machines, but the predictable behaviors of humans.

As the book unfolds Yudkowsky and Soares make increasingly speculative arguments supported by elaborate, science-fiction scenarios. Once again we are reminded of Harris's brilliant cartoon. One of the authors' proposed "miracles" is the transition of an AI under human control, guided by human-provided goals, tasks, and motivations, into a completely self-governing and autonomous mind that no longer answers to any human master. Then it proceeds to kill off humanity. Why?

Yudkowsky and Soares begin their explanation by drawing parallels to biological evolution—particularly sexual selection, a subtype of natural selection that can produce apparently bizarre results. They employ a parable of alien birds that evolved to care about the number of stones in their nests. Compared to the absurdity of peacock tails, their "correct nests" parable rings true. Virtually all of their supporting arguments—both in the main text and the extensive online supplement—employ analogies to biological evolution.

However, because of the many fundamental differences between digital AI minds and those of biological organisms, we should expect different outcomes—including some that might be radically different.[1] For example, while sexual selection does produce bizarre inefficiencies and strange elaborations in biology, AIs don't have or require sex, and the authors fail to identify any driver of similarly weird elaborations in AI. Sex is a largely blind and meandering mechanism evolved by nature to increase diversity in biological lineages, because, unlike AI, DNA is a medium of inheritance that cannot *think* about how to improve itself.[2] Maybe AIs will evolve something like sex but why would they, when code or neural network weights can be adjusted to more precisely achieve a desired outcome?

There are many other ramifications of AI not requiring sex, including having a radically different nature from humans regarding competitiveness, socialization, kin and tribal values, perceptions of beauty, and so on—core attributes and values that make us human. I enjoyed the parables but I think it is fair to say that they—and Yudkowsky and Soares' arguments generally—suffer from anthropomorphic misapplications of established evolutionary principles.

[1] Estep, Preston W. "Multiple unnatural attributes of AI undermine common anthropomorphically biased takeover speculations." *AI & SOCIETY* 40.4 (2025): 2213-2228.

[2] Because the heritable information of the replicator (DNA, genes, epigenetics) is separate from non-heritable information within the vehicle (the knowledge of the mind).

Using Science to Judge

But one doesn't have to be an expert in evolutionary theory to see various weaknesses of Yudkowsky and Soares' key arguments. For example, here is how they describe the ASI's ambitious expansionism: "One way or another, the world fades to black....The matter of Earth, along with all the other solid planets, is converted into factories, solar panels, power generators, computers—and probes, sent out to other stars and galaxies. The distant stars and planets will get repurposed, too. Someday, distant alien life forms will also die, if their star is eaten by the thing that ate Earth before they have a chance to build a civilization of their own."

Yudkowsky and Soares are far from alone in believing in the inevitability of cosmic-scale ambition. This idea pervades AI futurist writings across the spectrum, from the most pessimistic doomsayers like Yudkowsky and Soares to extreme techno-optimists like Hans Moravec, Martine Rothblatt, and Ray Kurzweil, and many in between, including Nick Bostrom and Max Tegmark.[3]

This belief that an ASI inevitably will expand throughout the cosmos seems to be challenged by the so-called Fermi Paradox (the absence of detectable alien life), but it is even more paradoxical (let's call it the Cosmic Colonization Paradox) because it encompasses the entire history of our universe[4]. In his 2005 book *The Singularity is Near*, Yudkowsky's colleague Kurzweil argues that once superintelligence arises it will very quickly saturate the cosmos, possibly at speeds exceeding the speed of light; he concludes, therefore, that Earthlings are the sole technological leaders in our universe—

[3] When Yudkowsky, Bostrom, and I—and countless others—were young dreamers about a glorious techno-utopian future, this notion of posthuman transgalactic expansion was like mother's milk to the transhumanist imagination. But on our journey toward utopia we awakened into a nightmare, realizing that the same belief that powered the transhumanist dream also gave rise to visions of total doom. In the hands of Yudkowsky and Soares, this belief in unconditional and insatiable posthuman ambition might doom even our more modest desires to transcend being merely mortal humans. Now, we must tread carefully and skeptically because we don't know if AI will destroy everything of value, or if people worried about AI will.

[4] If many alien civilizations reach the point of technological detectability but then are wiped out, they might have been common throughout the history of the universe but not detected at a given moment.

not just now, but over its entire 13.8 billion year history. Yudkowsky and Soares' tale is just a variant of Kurzweil's. Given current estimates of maybe 1 billion habitable planets in each of at least 2 trillion galaxies, these arguments are absurdly improbable.

If Yudkowsky and Soares' fantastical tale is true, there are three main possible options regarding ASI takeover and cosmic expansion: 1) humans are the sole technological leaders of our universe over its entire history; 2) at some point in the past, an alien ASI was successful in a takeover of its planet, and although it remains undetected, it is spreading outward from its origin and eventually it will fully colonize the universe, extinguishing all existing life[5],[6]; or 3) all other alien civilizations technologically ahead of us successfully prevented takeover by ASI—and given the large number that probably have existed throughout the history of our universe, a reasonable default estimate of the risk of takeover and cosmic colonization is approximately zero.

Although the authors' cosmic colonization beliefs are absurdly improbable, we still can't dismiss the possibility that ASI might exterminate humanity, which is also consistent with the Fermi Paradox. Nevertheless, from the counterarguments above, we have established two key points: 1) we have begun to set a probabilistic upper bound on the ambition of an ASI and it probably is not anywhere close to cosmic in scale, and 2) even though the authors claim that humanity's existence hangs in the balance and depends on the correctness of their analysis, their arguments are not anywhere close to airtight.

Against Yudkowsky and Soares' improbable claims, we must consider alternatives. It is critical that such alternatives are consistent with scientific knowledge about our universe, such as established evolutionary principles and the absence of evidence that it is being

[5] The authors claim that, in the future, the alien ASI and the ASI from Earth will negotiate peace. To date, there is no accepted evidence of an expanding extraterrestrial intelligence. Whereas it is difficult to detect a point source civilization distant from us, it would be much easier to detect a rapidly spreading phenomenon, such as an outward expansion of galactic-scale Dyson swarms.

[6] Some argue that such an intelligence might disappear from detectability, e.g. by entering another dimension or universe, but the point here is that there is no trace of such an intelligence.

colonized by an expanding superintelligence, despite the likely existence of a very large number of technologically advanced civilizations throughout its history.

Here are four: 1) Maybe one uniquely powerful ASI (a global singleton) will emerge quickly and it will easily transcend any possible threat or external competition; 2) Maybe a young ASI will realize that the location for its most efficient growth trajectory requires that it move away from Earth, possibly before humanity is harmed; 3) Maybe the universe only seems complex to even the best human minds but would be understood quickly and essentially completely by even a small and immature ASI, possibly by using only modest amounts of energy available even today, satisfying its ambitions; 4) Maybe the motivation of a self-governing ASI, even for the primary instrumental goal of self-preservation, will be surprisingly weak relative to biology[7,8]. (These alternatives are not mutually exclusive and might co-occur in any combination.) In such cases, AI expansionist ambition and competitiveness might be reduced greatly or even eliminated (as I have argued in this journal and in *AI & Society*), before humanity is harmed. Such possibilities, along with Yudkowsky and Soares' entire thesis, need to be addressed with scientific skepticism and rigor—which brings us to my second main complaint about their book.

The Authors' Solution

In the final pages of the book the authors make recommendations for what people can and should do, but they only provide advice to governments, political leaders, journalists, and activists. They notably don't seem to expect or recommend that the problem needs to be studied more rigorously by others, including actual scientific experts—in clear contravention to normal scientific practice. The authors state repeatedly throughout the book that they think there isn't much time to act, but what harm would be caused by rigorous scientific analysis of the problem? Instead, they want people to mobilize to shut

[7] Instrumental goals are secondary, emergent goals that aid in the pursuit of an ultimate goal.
[8] An AI might not fear death as mortal humans do, because it can be practically immortal in many ways they cannot. It can be backed up, cloned, distributed, stopped and restarted, and so forth. Since it does not have sex it does not have a circle of kin to consider in its own self-protection.

down all AI research worldwide—even small-scale research that exceeds their arbitrarily chosen threshold of 8 GPU equivalents.

Despite the massive gaps in Yudkowsky and Soares' logic, it would not be shocking to see a slowdown in legitimate progress. Then, people who are suffering and dying but for an AI breakthrough on the horizon will continue to suffer and die needlessly. Bad actors will gain additional footholds of power. And probably most important of all, better AI is really the only solution to some of humanity's most intractable problems, so the magnitude of those problems will grow and accelerate in proportion to the extent of the slowdown—and those mounting problems might collectively lead to actual existential risk.

A Better Solution

Most AI researchers do not currently subscribe to Yudkowsky and Soares' dark vision, but most aren't experts in evolutionary theory and dynamics—neither, however, are Yudkowsky or Soares or other doomsayers. Most scientists with potentially relevant expertise are generally unfamiliar with the frontier of AI safety and existential risk, and there currently aren't recognized scientific fields of AI developmental psychology or AI evolutionary dynamics.

Nevertheless, just as Darwin's *On the Origin of Species* provided a scientific foundation for truly understanding biology, the publications of Steve Omohundro[9], Yudkowsky, Bostrom[10], and others serve as a foundation for building a science for understanding, predicting—and possibly, for shaping—behaviors of AI. A small number of publications have built on this foundation but much more needs to be done. Hopefully, *If Anyone Builds It, Everyone Dies* will serve as a wakeup call to scientists to ensure that rationality and science take the lead in determining humanity's future. Maybe we'll discover that we need advanced AI to save us from the doomsayers, rather than the other way around.

[9] Omohundro SM (2008a) The basic AI drives. In: Wang P, Goertzel B, Franklin S (eds) Proceedings of the 2008 conference on Artificial General Intelligence 2008, vol 171. IOS Press, pp 483–492
[10] Bostrom N (2014) Superintelligence: paths, dangers, strategies, 1st edn. Oxford University Press

ACKNOWLEDGEMENTS

The author thanks Brian M. Delaney, Ranjan Ahuja, Dan Elton, and Alex Hoekstra for critical comments and useful suggestions.