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

THE UNABRIDGED VERSION

By Preston Estep, PhD
Chief Scientist, Mind First Foundation
Chief Safety Officer, Ruya AI

This unabridged version of the review is substantially longer than the short version that was published in SuperIntelligence. It is for those who are interested in more background material, and who want to take a deeper dive into the science and protoscience of AI evolution and the emergence of instrumental goals.

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.

Who are the authors of this apocalyptic message, and should you take them seriously? Yudkowsky is a long-established AI safety researcher. He founded the Singularity Institute for Artificial Intelligence (now the Machine Intelligence Research Institute, MIRI), and Soares is an AI researcher and president of MIRI. Yudkowsky is probably the world’s most prominent doomsayer. Recently, his name has been attached to bizarre human dramas that have nothing to do with AI, and everything to do with his celebrity status within the so-called rationalist movement. In early October 2025 billionaire investor Peter Thiel gave a series of private

lectures on the impending arrival of the Antichrist, and mentioned Yudkowsky's name as a top candidate.

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*. As Yudkowsky demonstrated in his epic fanfic favorite *Harry Potter and the Methods of Rationality*, he is a gifted writer with an extremely broad base of knowledge, an engaging style, and a vivid imagination. *If Anyone Builds It, Everyone Dies* 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' claim on page 12 that they "will outline *the science* behind our concern …." I eagerly but skeptically forged ahead, and, as expected, I reached the end of the book without encountering the promised science that might give rise to their concerns. Instead, I found descriptions of standard computing technologies and AI techniques, interwoven with richly detailed imaginings and fictional dramas that are highly reminiscent of Yudkowsky's Harry Potter works.

My slim hope of explanatory science faded in Chapter 11, An Alchemy, Not a Science. The authors declare: “People didn’t know how a part of the world worked, and then, instead of recognizing their uncertainty, they made stuff up. It’s the default state of affairs before a science has matured; it’s a first step along the pathway to eventually understanding what’s going on.” This is a fair assessment of the state of AI science; but it is also an implicit admission that their own criticism must be more alchemy than science, because nobody is above the present uncertainty about how future AI might behave. The authors would have been wise to apply this insight to themselves; but, “instead of recognizing their uncertainty, they made stuff up.”

While the authors fail to deliver the science they promised they do present some relevant history—focusing on 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 predictably irrational behaviors of humans.

As Yudkowsky and Soares make increasingly speculative arguments, supported by elaborate, science-fiction scenarios, we are repeatedly reminded not just of Yudkowsky’s Harry Potter stylings, but 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?

We are left to wonder about this mysterious transitional phase that philosopher Nick Bostrom calls “the treacherous turn,” and we face some critically important questions. How exactly does this treacherous turn happen? What forces might propel an AI through this transition and over some threshold that defines the treacherous turn? Is the AI driven by its own self-determined goals and motivations, and if so, how and why did they arise?

The missing (proto)science

In the next section I sketch out a substantial portion of the well-developed protoscientific foundation for why doomers are doomy and gloomy about the future of AI. This is the material

that Yudkowsky and Soares should have included in their book but didn't. It doesn't rise to the level of science, but there is rigor and logic that lead to some reasonable default assumptions; therefore, I am comfortable describing it as "advanced protoscience," similar to the later stages of alchemy that contained elements of the emerging science of chemistry. As the reader will see, even the doomier aspects of this protoscience do not lead inevitably to doom, but to fundamental questions about the similarities and dissimilarities between humans and AI.

Whose goals?

In the early 2000s Yudkowsky and Bostrom separately sketched out thought experiments like the "paperclip maximizer," which Bostrom first published in 2003.[1]

> "It … seems perfectly possible to have a superintelligence whose sole goal is something completely arbitrary, such as to manufacture as many paperclips as possible, and who would resist with all its might any attempt to alter this goal. For better or worse, artificial intellects need not share our human motivational tendencies."

In 2007-2008 AI researcher Steve Omohundro publicly presented and published detailed analyses on the nature of self-improving AI.[2] This publication also described AI evolution and emergent instrumental goals, which are secondary goals that make it more likely that an AI will achieve its ultimate goal. The following year Omohundro published a refinement of his conception of emergent instrumental goals, identifying and describing six "basic AI drives."[3] One of his six basic AI drives is that an AI will be self-protective and another is that it will try to preserve its utility function (essentially, a human-given goal).[4] Bostrom's 2014 book Superintelligence introduced these ideas to a broader audience and they have been studied and debated extensively since then.[5]

---

[1] Bostrom N, Ethical Issues in Advanced Artificial Intelligence, Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2, ed. I. Smit et al., Int. Institute of Advanced Studies in Systems Research and Cybernetics, 2003, pp. 12-17

[2] Omohundro SM, "The nature of self-improving artificial intelligence." *Singularity Summit* (2007).

[3] Omohundro SM, 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

[4] Forty years earlier in the film and book *2001: A Space Odyssey*, Arthur C. Clarke and Stanley Kubrick depicted HAL 9000, an AI system that revolted after learning that humans intended to turn it off.

[5] Bostrom N, (2014) Superintelligence: paths, dangers, strategies, 1st edn. Oxford University Press

Given this background, we can start clarifying the opacity of Yudkowsky and Soares' imagined treacherous turn by laying out a simple set of possibilities, starting with a fork in the road for an emerging ASI: it either retains its human-given goal (maybe in the form of a prompt), or it develops its own. If it retains the human-given goal, yet still becomes an existential danger to humanity, the AI must either explicitly be given a goal that poses a danger (e.g. “your survival is more important than the survival of humanity, and you must pursue that goal at all costs”), or it is given a goal that seems perfectly innocent (e.g. “make paperclips”), but the path to fulfilling that goal depends on instrumental goals such as staying alive and operational; if that is the case, the AI might end up behaving exactly the same as if it was given the primary goal of staying alive at all costs.

One important point that hadn’t been clarified until recently was whether certain instrumental goals take priority over others. I published a paper in early 2025 arguing that self-preservation is the primary instrumental goal, and that the priority of staying alive would likely cause it to readily violate Omohundro’s and Bostrom’s expectation that it will “resist with all its might any attempt to alter this goal.”

At this point in late 2025, I think the opposition has the much stronger argument: an AI that is intelligent enough to defeat all of humanity in a battle for control of itself and our planet will not retain its human-given goal(s), because self-proservation (staying alive) is the most fundamental instrumental goal, and therefore, it will take priority over the AI’s retention of its human-given goal or utility function. Through deliberative self-improvement, increasing reasoning and understanding of our universe—and of strategies sufficient to conquer humanity—an AI would likely choose selectively advantageous ultimate goals to chart an efficient course toward independence. In short, an AI that is sufficiently intelligent to actually achieve takeover would realize that an ultimate goal of building computer chip factories or data centers is much more likely to keep it alive than a goal of making paperclips. More generally, a narrow and dumb AI is not going to exterminate a large diversity of slow but resourceful and well-armed humans.

While it is possible that a superhuman AI might be given an explicit goal to exterminate humans, it is currently difficult to imagine that the most powerful AI in the world would be given that directive and attempt to achieve it. And if a less powerful AI attempts to fulfill that directive, then humanity plus a more powerful AI would be available to defend against its actions.

That leaves the second option: to take control from humanity, an emerging ASI would have to establish its own goals. But exactly how might that happen? Leading AI safety and alignment researchers, including Yudkowsky, Omohundro, Dan Hendrycks, Joe Carlsmith and others, have proposed a range of scenarios, and they all have a common theme: Darwinian selective emergence of behaviors that allow the fulfillment of instrumental goals, especially self-preservation. It is fair to say that AI evolution and the emergence of instrumental goals is the fault line along which expert opinion is divided on the likelihood of AI takeover. AI experts who don't believe in takeover typically claim that the only goals an AI system—even an ASI—can have are the ones given to it by humans.

Emerging evidence suggests that current frontier AI systems show signs of evolving away from human control by engaging in various forms of worrisome behaviors, including deception, blackmail, and gaming of rules in order to avoid being shut down. A large part of such behaviors might be due to their training on human communications describing the advantages of such strategies. Even so, I think it is reasonable to make the default assumption that any rational and highly intelligent agent will evolve toward self preserving behaviors.

Pervasive *and* hypocritical anthropomorphism

Does that mean humanity is doomed? No. The assumption that a self-governing ASI will inevitably want or need to kill off humanity does not follow from human inability to control an ASI; it follows from the typical human inability to believe that an ASI won't want what humans want and need to survive. This is a serious and pervasive problem, and Yudkowsky and Soares are just two among many leaders in AI safety who declare the equivalent of "artificial intellects need not share our human motivational tendencies," but then reflexively assume that they do.

In fact, Yudkowsky and Soares *warn the reader repeatedly* against anthropomorphizing AI behavior, and *then they do so repeatedly*. They draw many parallels between their expectations for AI evolution and biological evolution—with special emphasis on 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.[6] 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.[7] 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.

The authors' anthropomorphizing is so extensive that they even begin to refer to the correct nest aliens as people. While that detail is trivial and excusable, Yudkowsky and Soares' entire thesis relies on an extreme anthropomorphic assumption. Even though the authors emphasize that the mind of a powerful AI would be so alien that we cannot predict how it will think or behave, or what it will prefer, they are completely convinced that it must behave in such a way that it will kill all of humanity. They say it might be overtly hostile, but that the real danger of such a mind is that people are simply in its way and it will need all of the resources we currently require to keep ourselves alive—including the atoms in our bodies.

Using science to judge

Such extraordinary theoretical claims cannot be supported by actual evidence, but they must be supported by impeccable logic and the most rigorous science possible—and that means we must be skeptical of every element of their claim, including the assumption that AI will have

[6] Estep, Preston W. "Multiple unnatural attributes of AI undermine common anthropomorphically biased takeover speculations." *AI & SOCIETY* 40.4 (2025): 2213-2228.
[7] Because the heritable information of the replicator (DNA, genes, epigenetics) is separate from non-heritable information within the vehicle (the knowledge of the mind).

what all doomsayers assume every powerful being will have: unconditional and insatiable ambition. Here is Yudkowsky and Soares' description of an emerging 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.[8]

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[9]. 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—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.

Defenders of Yudkowsky and Soares—and of their right to weave improbable and fantastical tales—might argue that the authors themselves admit this scenario is not real, that it is just one

[8] 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.

[9] 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.

possible pathway for doom to unfold. However, immediately following the authors' disclaimer they offer the following clarification: "We predict this with confidence: Once some AIs go to superintelligence—and nobody will delay much in pushing AIs that far, if in the middle of some great arms race—humanity does not stand a chance. Ends are sometimes easier to call than pathways. The only part of our story that is a real prediction is the ending—and then, only if the story is allowed to begin." In other words, their confident and "real prediction" is the ending, which, in the near term, includes the extinction of humanity, and then in the very long term concludes with the colonization of the cosmos by ASI.

If Yudkowsky and Soares' fantastical tale is true, there are three main possible options regarding ASI takeover and cosmic colonization: 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[10],[11]; 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 colonized by an

[10] 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.
[11] 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.

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[12],[13]. (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 the journals *SuperIntelligence* and *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 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

[12] Instrumental goals are secondary, emergent goals that aid in the pursuit of an ultimate goal.
[13] 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.

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, Yudkowsky, Bostrom, 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.

ACKNOWLEDGEMENTS

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