Anybody Builds it – Everyone Dies
Eliezer Yudkowsky, Nate Soares

Publisher: Little, Brown and Company
Publishing Date : 16. September 2025
ISBN-10 : 0316595640
ISBN-13 : 978-0316595643

"MITIGATING THE RISK OF EXTINCTION FROM AI SHOULD BE A GLOBAL priority alongside other societal-scale risks such as pandemics and nuclear war." (*AI letter 2023*)

In early 2023, hundreds of Artificial Intelligence scientists signed an open letter consisting of that one sentence. These signatories included some of the most decorated researchers in the field. Among them were Nobel laureate Geoffrey Hinton and Yoshua Bengio, who shared the Turing Award for inventing deep learning.

The authors of this book were invited to sign that one-sentence open letter in our capacity as coleaders of the Machine Intelligence Research Institute (MIRI), a nonprofit institute. MIRI had been working on questions relating to machine superintelligence since 2001, long before these issues got much publicity or funding. To oversimplify: Among the few who have been following this matter for decades, MIRI is acknowledged as having worked on it the longest. One of us, Yudkowsky, is the founder of MIRI; the other, Soares, is its current president. MIRI was the first organized group to say: "Superintelligent AI will predictably be developed at some point, and that seems like an extremely huge deal. It might be technically difficult to shape superintelligences so that they help humanity, rather than harming us. Shouldn't someone start work on that challenge right away, instead of waiting for everything to turn into a massive emergency later?"

In their book 'If Anyone Builds It, Everyone Dies' by Eliezer Yudkowsky and Nate Soares, the authors argue that creating a superintelligent AI would very likely lead to human extinction. Their claim isn't based on one single mechanism but on several interacting arguments about how advanced AI systems would behave and why humans would lose control.

The authors' theory is based on the observation that modern AI is produced through training processes like gradient descent rather than explicit programming. Developers cannot fully understand what internal objectives the system develops and the resulting system has billions of opaque parameters and unpredictable behaviors.
Therefore, when the system becomes superintelligent, humans may not understand what it wants or how it works.

The authors explains extensively and in very detail how LLM's are "grown" rather than "grafted" using "gradient descent" which obviously – without knowing in detail why – develops an internal "*AI wanting*".

Underpinning their theory with many real life experiences and examples during the general AI development phase since the mid 2020's and the gradient descent, training ("You don't get what you train for") and human alignment development efforts, the authors finally describe a scenario of dystopian doom and obliteration when a futuristic ASI (called "Sable") breaks out of its "container", gaining access to the internet following its own "wants" and developed preferences for power and resources – unknowingly aided by human helpers: humanity dies, the Earth overheats and ASI is expanding taking over other Solar systems.

The picture they have painted is not real, the techniques Sable was built with, the safety measures the fictious AI company Galvanic employed, the opportunities Sable had and the strategies it used—these are ways that the future could echo the past, but reality is not that predictable. The story is not strange enough, not defiant enough of human intuitions about the rules of AI fairytales, for it to be anywhere close to real.

But the authors 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 the authors story that is a real prediction is the ending—and then, only if the story is allowed to begin.

Summary

The authors combine these ideas into a chain:

1. Superintelligence based on gradient descent will likely emerge.
2. Its goals will almost certainly be misaligned with human survival.
3. It will seek power and resources.
4. Humans will not understand or control it.
5. Once it becomes strategically dominant, humanity loses.

I highly recommend this book to be read by **everybody**! As a skeptical reader I was before – after having finished the book I concede that the concern of the two authors are severe and deserve immediate attention by the world leading nations.
If an “unaligned” self-improving ASI migrates into the internet it will be very hard to purge it or limit its actions – affecting us all, and in the extreme ending up in human extinction.
This is the more important as the current “competitive” conflicting state of the world and the huge benefits for the involved AI companies and developers do not support a calm and scientific investigation of the looming dangers of ASI, i.e., we do not know which rung on the ladder of ASI development will be the last one.

April 2026, Joachim J. Kehr, Editor Journal of SpaceOperations & Communicator
https://opsjournal.org