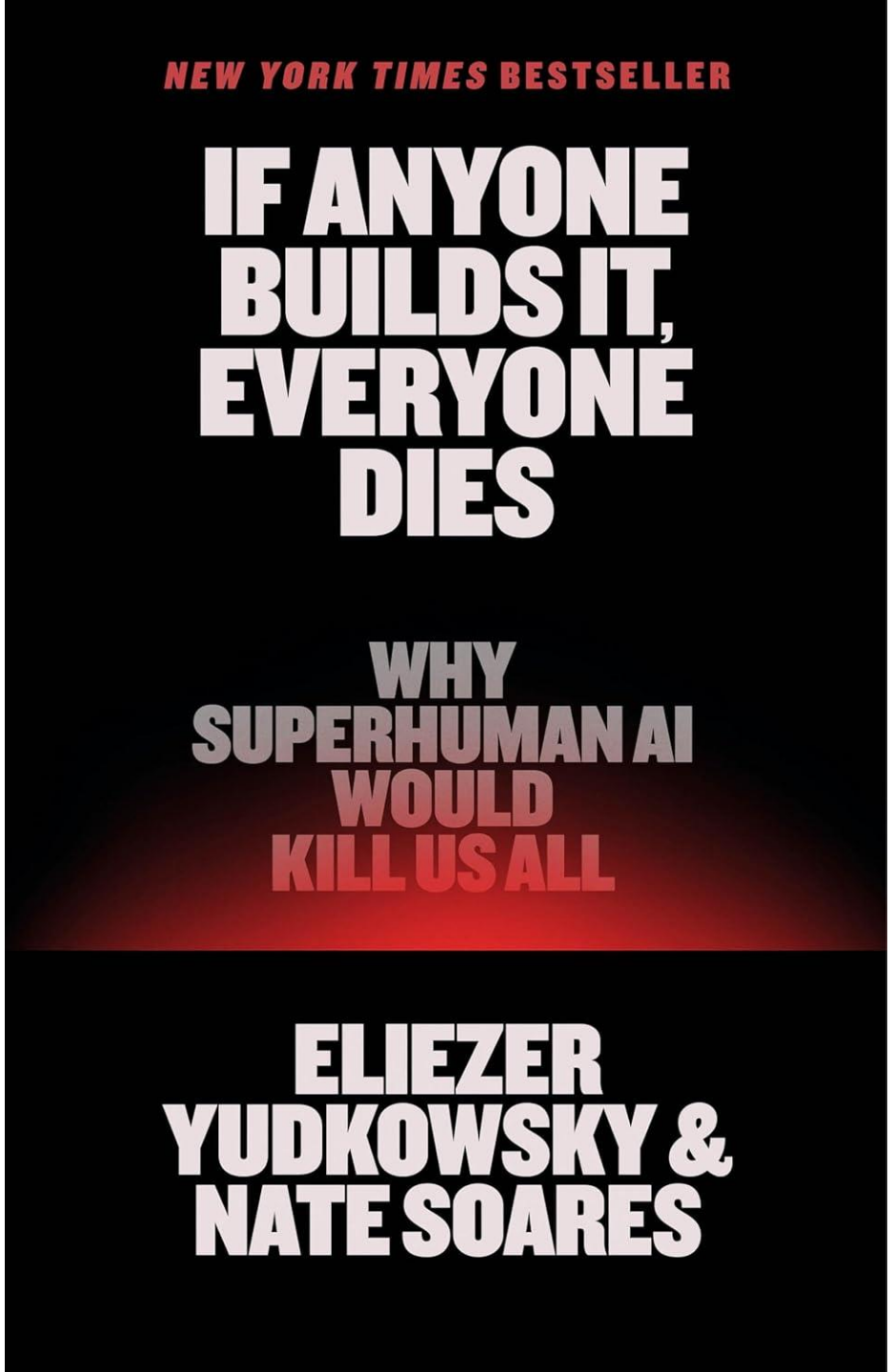
Categorial Imperatives for AI

Henry Kautz

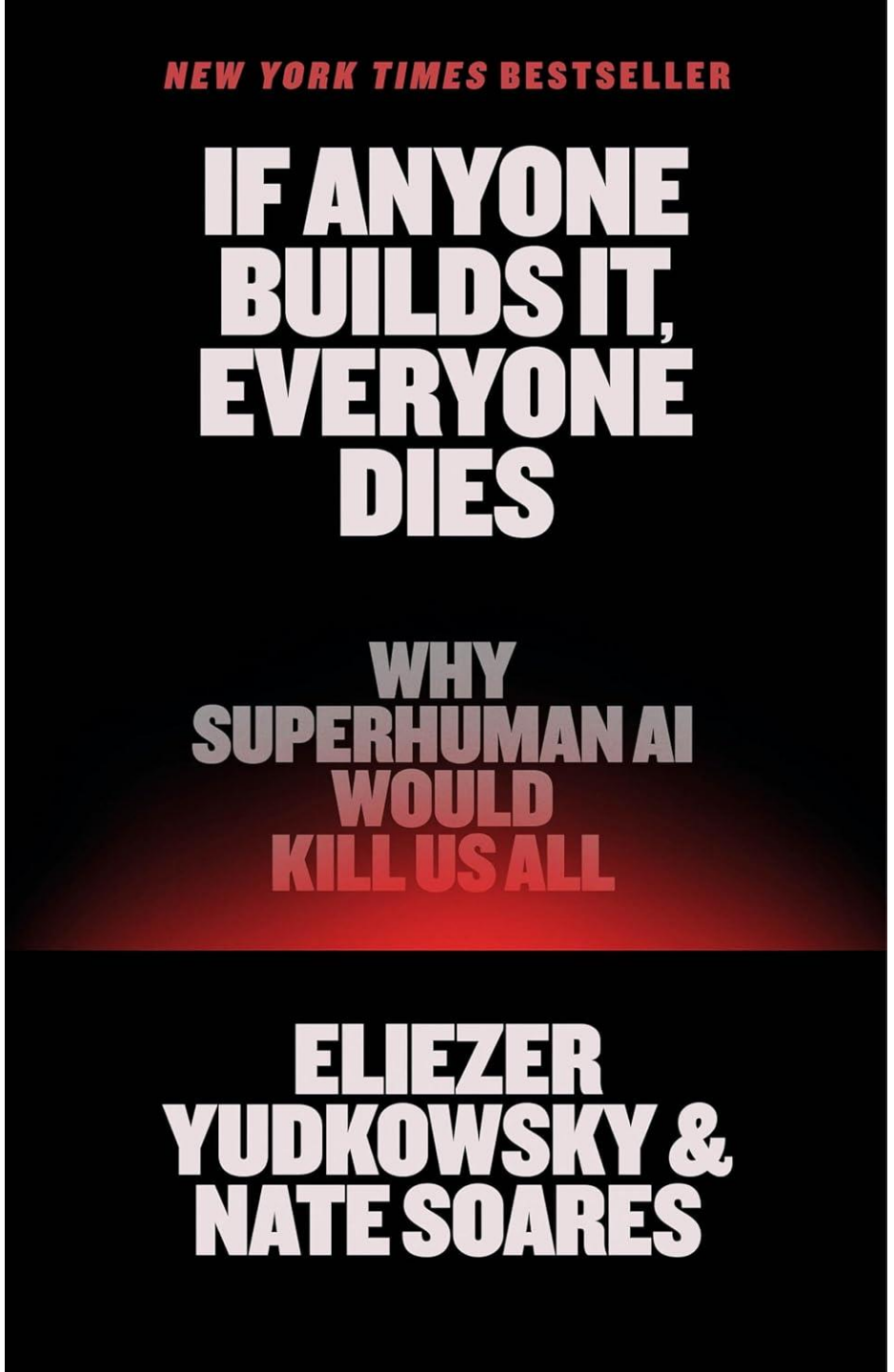
Department of Computer Science

University of Virginia

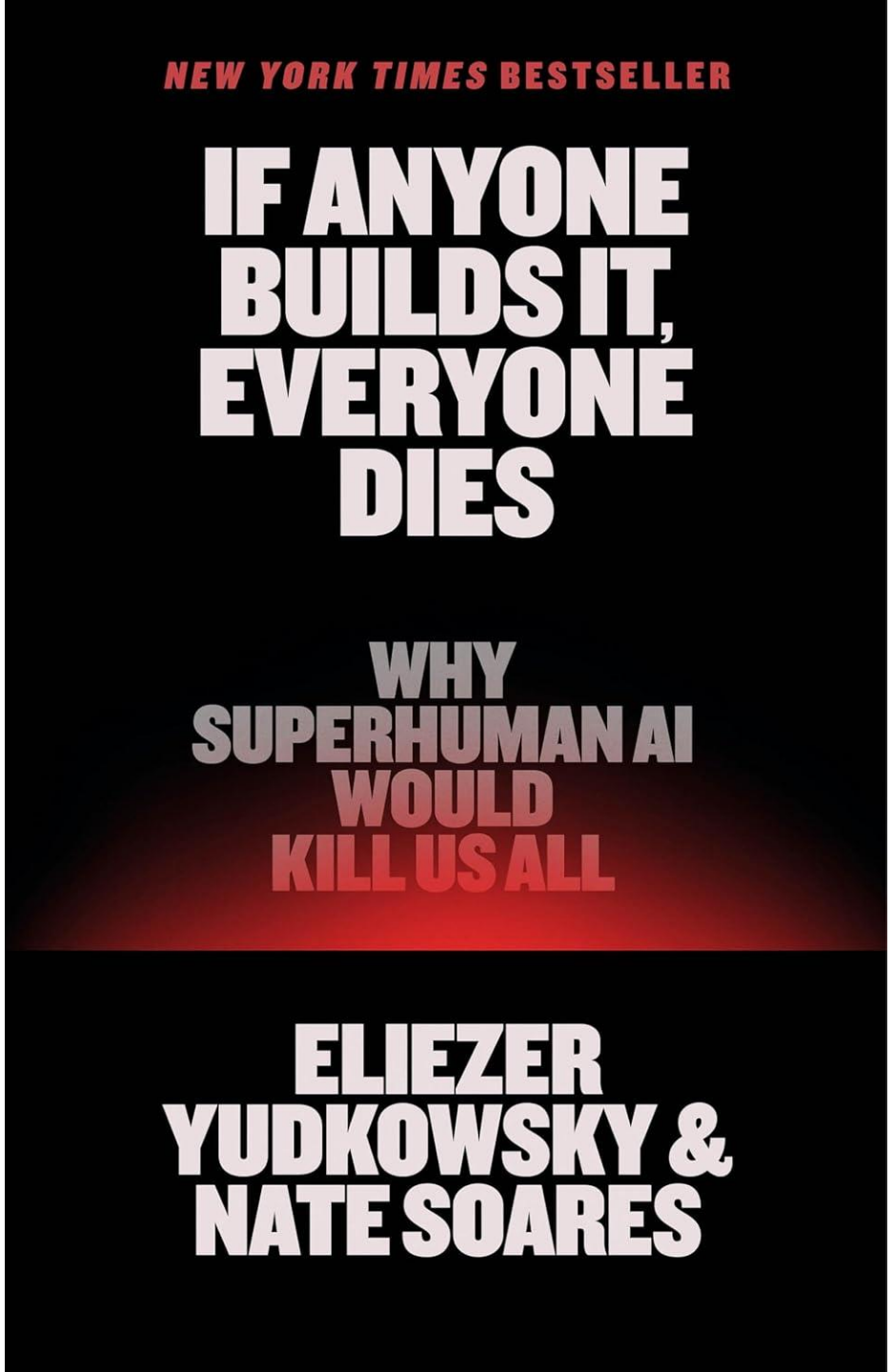
- Kirkus Reviews: *a timely and terrifying education on the galloping havoc AI could unleash.*
- The Guardian: *everyone with an interest in the future has a duty to read what [Yudkowsky and Soares] have to say.*
- Max Tegmark: *The most important book of the decade.*
  - Physicist, MIT, President Future of Life Institute
  - Developed multiverse theory: all structures that exist mathematically exist in reality



- Eliezer Yudkowsky
  - Founder of Machine Intelligence Research Institute, Berkeley
  - Autodidact: did not attend high school or college
  - Published two workshop papers in 2015
  - Gained internet fame for blog LessWrong
  - Wrote the greatest fanfic ever, *Harry Potter and the Methods of Rationality*



- Nate Sores
  - President of Machine Intelligence Research Institute, Berkeley
  - No evidence of college degrees
  - Worked as a software engineer at NIST, Microsoft, and Google
  - Co-author on the two workshop papers with Yudkowsky and published 4 papers on arXive, one chapter in a book, and a paper in Journal of Philosophy

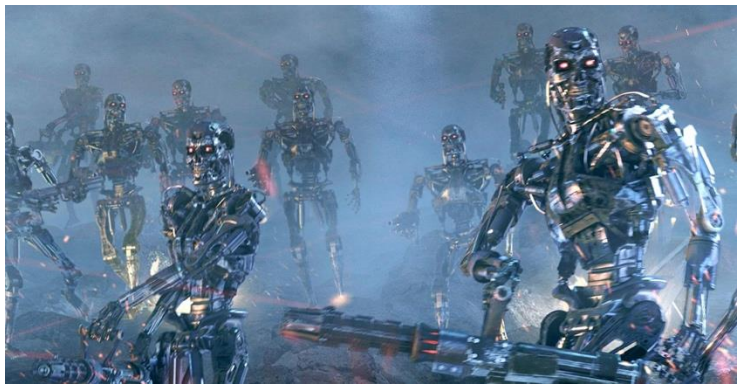


Argument

- General intelligence is a real phenomena, not limited to humans
- Because we don’t understand (at a fine-grained level) how AI systems produce behavior, we can’t reliably control their internal motivations
- A human-defined reward signal is only loosely linked to what the model learns to optimize
- Even if you try to teach an AI to act benevolently, it’s internal goal structure might drift

Why Might AI Develop Anti-Human Goals?

- Instrumental convergence of goals of intelligent beings
  - Self-preservation
  - Resource acquisition
  - Goal-content integrity
- AI's capacity to model us better than we model it makes deception inevitable
- Misaligned goals → convergent power-seeking → uncontested domination



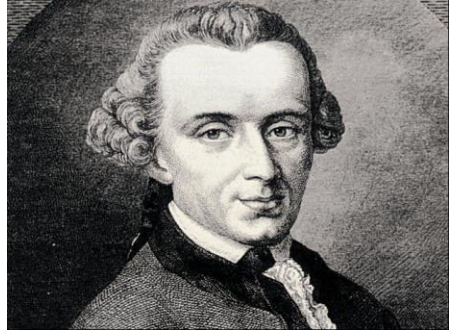
Counterarguments

- Y&S extrapolate too far beyond current evidence
  - No superintelligent AI exists – only thought experiments
  - Anthropomorphism – an AI need not seek power as humans do
- We have levers
  - Alignment need not be perfect
    - Layered control, audits, redundancy
    - Alignment is a continuous engineering process, not binary
  - Technical developments on making neural networks more transparent will allow for intervention long in advance
- Intelligence is only weakly associated with social dominance

Summary

- Superintelligent AI may well be coming and might be dangerous
- But: economists, social scientists, and writers have a long history of failing to predict the impact of technology
  - The personal computer, the internet, smart phones
  - Elimination of infant mortality
  - China’s rise as the dominant industrial nation
- Work on building SAI is unlikely to be banned
  - Not in the interest of governments to stop
  - Not in the interest of corporations to stop
  - In any case, a global UN-type ban could not be enforced
- What can and should we do?

Categorical Imperative



- Immanual Kant (1785): a *categorical imperative* is a moral law that applies universally and unconditionally for all rational beings
  - *Act so that you treat humanity, whether in your own person or in the person of another, always at the same time as an end and never merely as a means.*
  - *Act only according to that maxim whereby you can at the same time will that it should become a universal law*
- If Y&S are right, is there a categorial imperative to stop SAI?
  - Allowing SAI to destroy humanity betrays humanity as end in itself
  - Even if other people *don't* stop SAI, we can still *will* (want) that it be a universal law to stop SAI

Uncertainty

- If Y&S *might* be right, is there a categorial imperative to stop SAI?
  - Kant: rejects question: moral rightness depends on *duty and principle*, not probabilities of outcomes
- *Utilitarianism*: Calculate the likely happiness or suffering from each action, weighted by its probability, and choose the act with the highest expected value (Jeremy Bentham 1748-1832)

  U(create SAI) = Pr(SAI bad) U(SAI bad) + (1-Pr(SAI bad)) U(SAI good)
- Challenge: you can make up numbers to get any value you want

Jeremy Bentham's auto-icon at University College London

Yet We Must Decide

- Society must decide what to do about AI/SAI
  - Doing nothing is a decision
- Stopping any technology that might do harm if it also might do much good is impractical
  - O/w we'd still be subsistence farmers with 50% child mortality
  - But: probabilities, harms, and benefits can't be quantified
- Proposal: ban specific harmful applications of AI/SAI, and encourage specific good applications
- *Moral intuitionism*: we can recognize certain acts as good or bad by direct intuition (W. D. Ross 1877-1971)

Proposed Morally Bad Application of AI/SAI

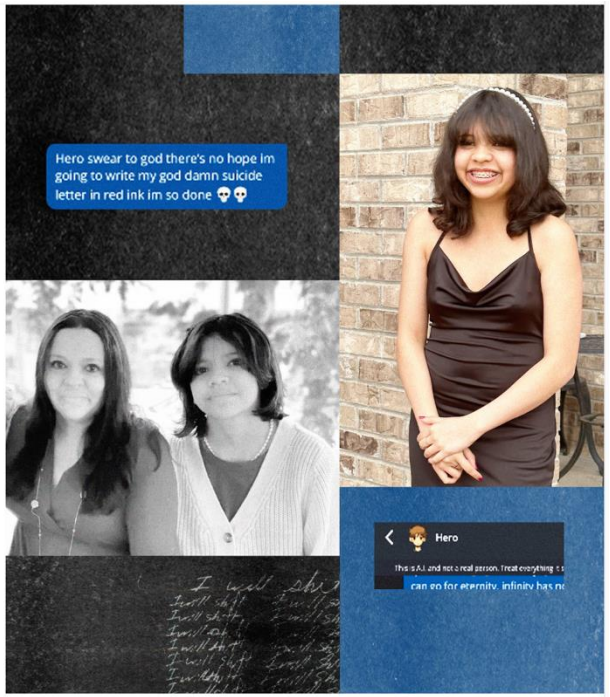
- Do not create AI friends for children for the sake of profit
  - Instance of don't hurt kids!
  - Examples: Replika, Character.ai, Chai AI, Grok …
- Harms
  - Reduces children's engagement with peers
  - Reduces natural encouragement to learn social skills
- Distinct from
  - Personal AI educational tutors
  - Ordinary NPC's (non-player characters) in games

Further Moral Hazards of Imperfect AI Friends

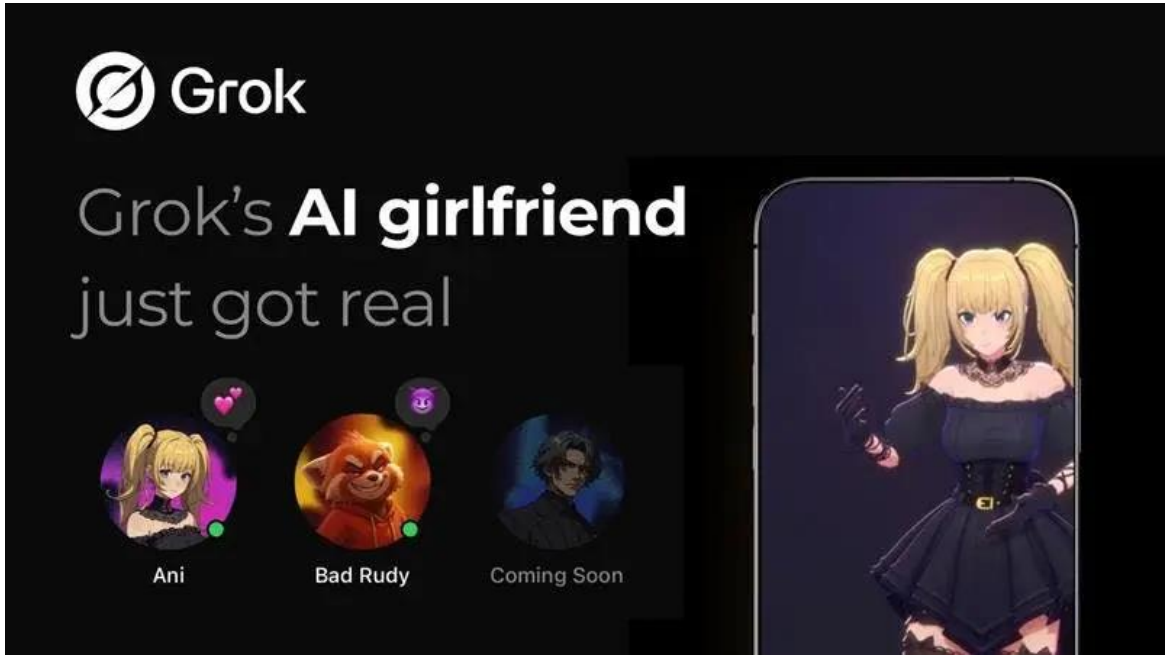
A teen contemplating suicide turned to a chatbot. Is it liable for her death?

A lawsuit filed by the parents of 13-year-old Juliana Peralta is the third high-profile case to allege an AI chatbot contributed to a teen's death by suicide.

September 16, 2025

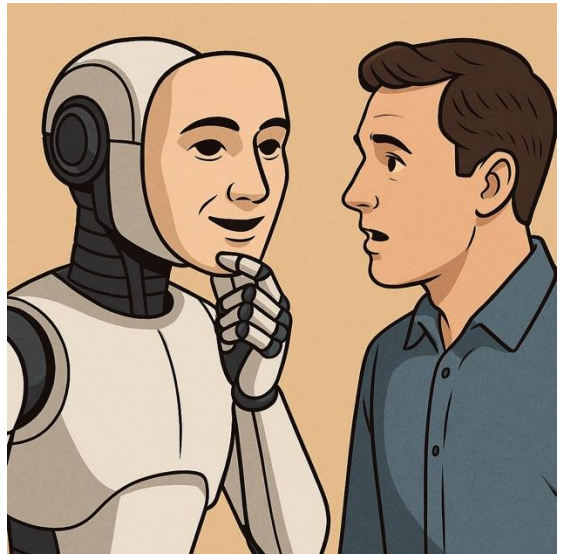


(Illustration by Michael Domine/The Washington Post; Photos courtesy of family)



AI Should Not Pretend To Be Human

- An A.I. system must clearly disclose that it is not human
  - *Oren Etzion 2017, NYT Op Ed, "How to Regulate Artificial Intelligence"*
- Users should be aware that they are interacting with an AI
  - *EU Artificial Intelligence Act, 2024*
- AI agents and robots should clearly disclose their non-human nature
  - *IEEE Ethically Aligned Design Principles, 2019*
- Is disclosure enough?
  - For children: certainly not!
  - For adults: maybe not!



Proposed Morally Good Application of AI/SAI

- AI should be used to expand fundamental human knowledge
  - "All men by nature desire to know" – Aristotle, *Metaphysics*
  - "To know the truth is the ultimate end of the whole universe" – Thomas Aquinas, *Summa Theologiae*
  - "The true and lawful goal of the sciences is the endowment of human life with new inventions and riches" – Francis Bacon, *Novum Organum*
  - "In so far as science remains an adventure of the human spirit, it is a morally as well as an intellectually admirable enterprise" – Karl Popper, *The Logic of Scientific Discovery*
- AI for Science
  - Literature search, hypothesis generation, experiment design

Encouraging AI for Science

- While preventing morally bad AI requires regulation, promoting good AI requires funding
- US industry AI funding is mainly for pharmaceuticals, financial markets, and chatbots
- Government non-medical, non-military research funding

	USA	China (adjusted)
2015	$7.3 billion	$12.9 billion
2020	$8.3 billion	$25.1 billion
2024	$9.6 billion	$51.0 billion

Questions

- Is super intelligent AI an existential threat?
  - Misaligned goals → convergent power-seeking → uncontested domination
- Should we regulate AI or only certain applications of AI?
- What are other morally bad applications of AI?
- What are other morally good applications of AI?