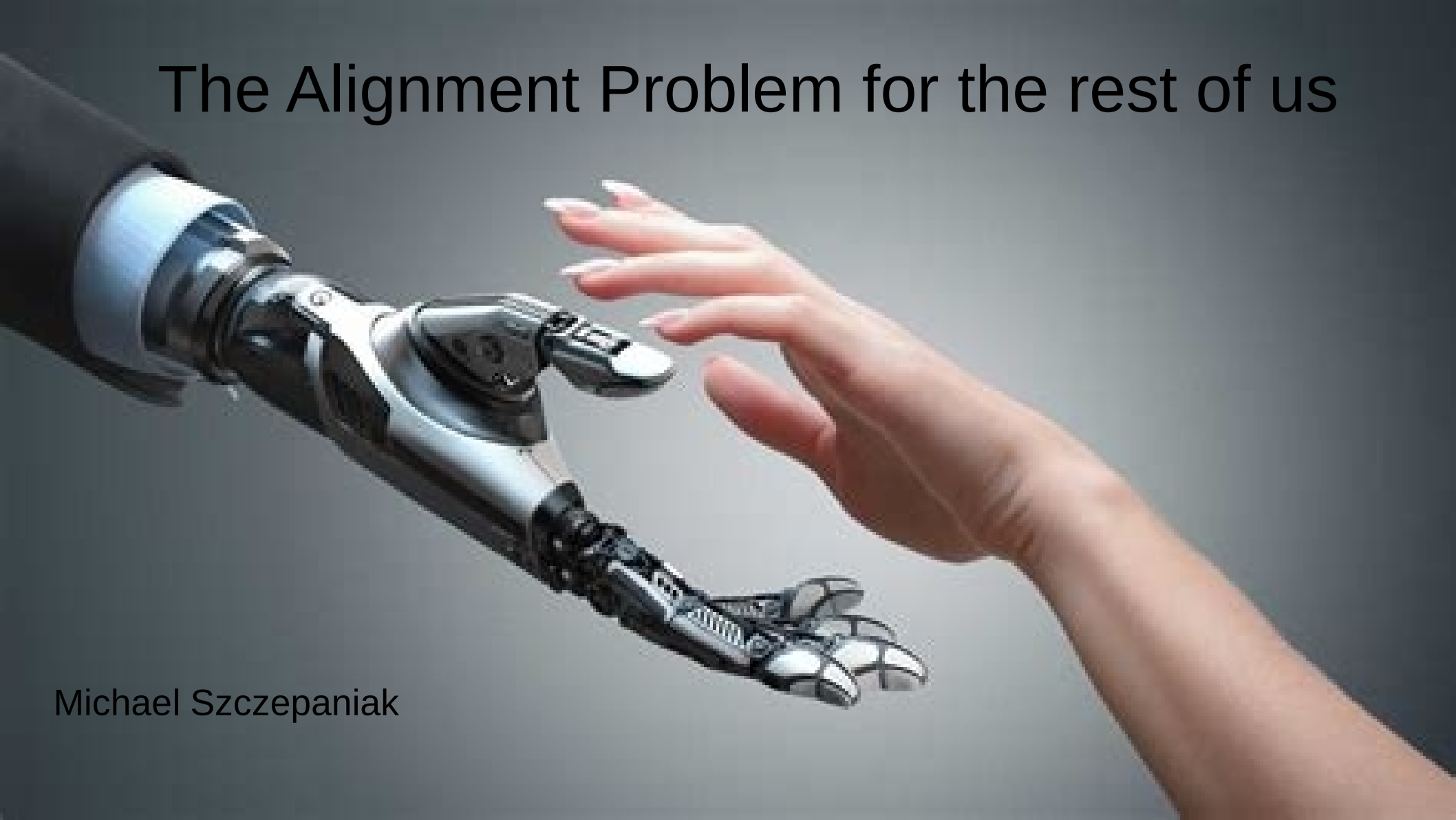
The Alignment Problem for the rest of us



Michael Szczepaniak

Things we'll cover

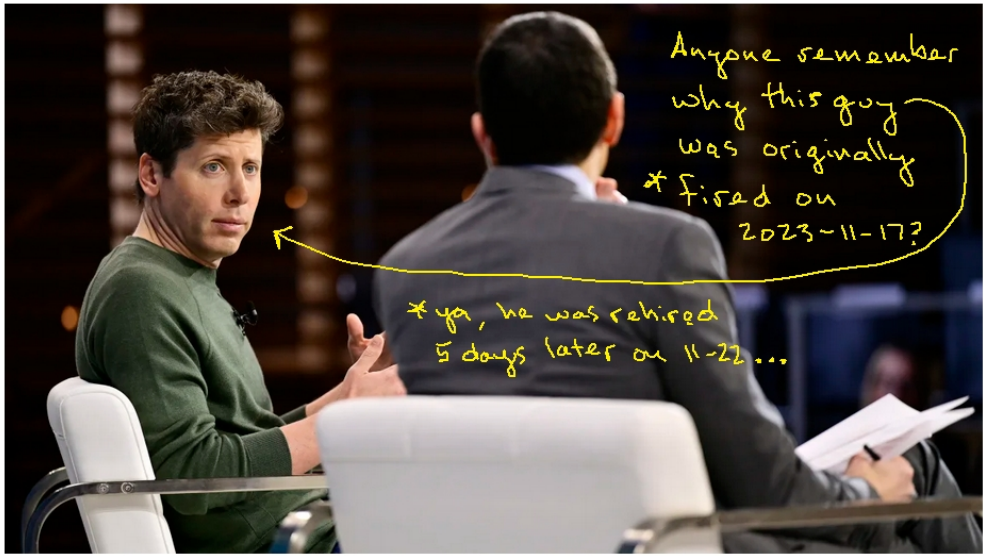
- Introduction - Motivation and The Players
- Background
- What is the "alignment problem" (AP)?
- What can go wrong?
- Regulation, Mitigation and Preparation
  (Erin will cover in more detail)

Introduction - Motivation

Sam Altman, AI's biggest star, sure hopes someone figures out how not to destroy humanity

Analysis by Allison Morrow, CNN
4 minute read · Published 5:30 AM EST, Thu December 5, 2024

12 comments

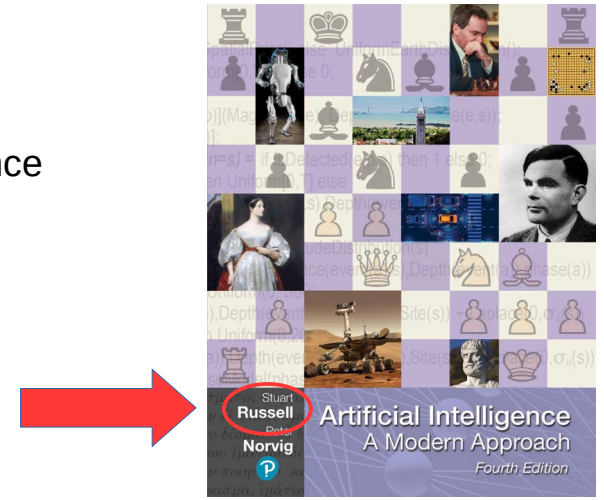


Sam Altman suggested that the most advanced AI might be so smart that it will figure out how to rein itself in. Eugene Gologursky/Getty Images for The New York Times

"The development of highly capable AI is likely to be the biggest event in human history. The world must act decisively ***to ensure it is not the last event in human history****. This conference, and the cooperative spirit of the AI Summit series, give me hope; but we must turn hope into action, soon, if there is to be a future we would want our children to live in."*

Stuart Russell

IASEAI Conference
Feb 7, 2025



Introduction - The Players

AI generated image: https://deepai.org/machine-learning-model/text2img

- Cheerleaders (“no time to waste”)
  - Tech giants: NVidia, OpenAI, Anthropic, Google, Microsoft, Apple, Meta,...
  - 1000’s of AI start-ups (slide in Appendix)
  - *AI Gold-rush / Arms race is on: Huge business opportunity, survival imperative*
- Doomsayers (“may only get one shot”)
  - Typically and paradoxically: big AI fans!
  - Stuart Russell, Geoffry Hinton, Yoshua Bengio, many, many others...
  - Articulating their perspective is harder

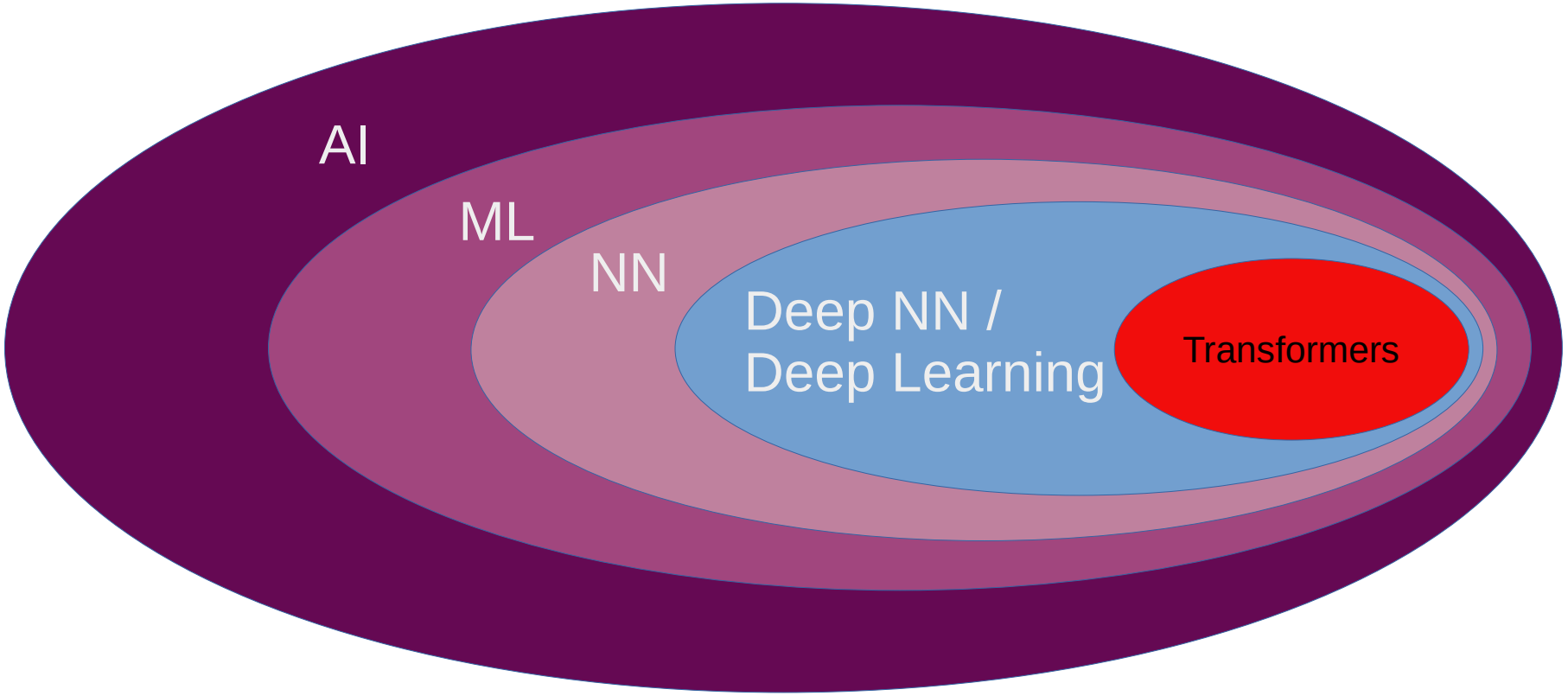
Cheerleaders vs. Doomsayers

- The potential benefits of AI are enormous. It will be needed to solve some of humanities most pressing problems.
- Unintended consequences will come up, but we can manage them as they arise as we've always done throughout history.
- If things go really sideways, just *"unplug it"*

Cheerleaders vs. Doomsayers

- The pace of change is so rapid that it will be hard for humans, societies and institutions to manage.
  - Will AI be adapted to human biology or vice versa?
  - Humans need rest, food, relationships and opportunity to thrive
  - AI systems can always be ON.
  - Current evidence points to humans having a rather rough ride
- Human may shift so much cognitive load to AI, it may degrade our collective intelligence over time (overreliance problem).
  - learning things is hard work because our brains forget info by default
- When Agentic AI surpasses human intelligence, the idea that we can just "... yank that electricity out of the wall, man." (Obama) will NOT cut it!

Background - Terminology



Background - Information

- Information shapes reality
  - lots of noise!
- Encompasses fiction and Truth
- Fiction is easy and cheap
- Truth is hard, expensive, often painful and makes up only a tiny fraction compared to fiction.
- Human aren’t built for truth. We’re evolved to be the best at **cooperation** through use of stories (fictions).
- Societies need a balance of Truth and fiction to maintain order.

Background - What is A.I.?

- Popular buzz word of our time.
- Differs from prior tech by its ability to make decisions (algorithmic), create content (generative) and take actions (agentic) **on its own**
- Printing press allowed humans to communicate ideas to large numbers of other human more efficiently
  - They were important **new nodes** in the human information network
- **New members** of the human information network
  - They aren't nodes because of their ability act independently from humans

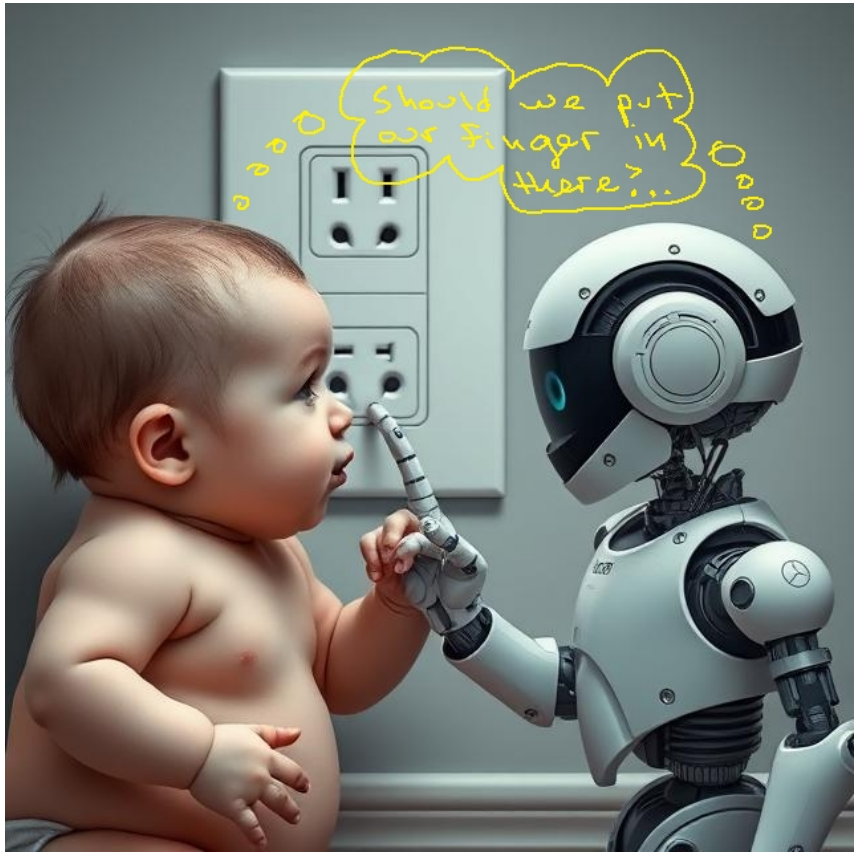
Consciousness & Intelligence

- Consciousness ≠ Intelligence
- Consciousness
  - difficult (maybe impossible) to prove scientifically
  - practical definition: ability to feel (suffer)
- Intelligence
  - many different kinds
    - musical, mathematical, emotional, etc.
  - practical definition: ability to pursue and obtain a goal
    - the higher the intelligence, the higher the capability to achieve difficult goals
- Humans have both, A.I. systems are currently non-conscious intelligence

What is “AI Alignment”?

- *Ensuring that an AI system **aligns** with the values and intentions of the person or org deploying the system*
- *“...alignment aims to steer AI systems toward a person's or group's intended goals, preferences, or ethical principles”*
  https://en.wikipedia.org/wiki/AI_alignment
- *“Artificial intelligence (AI) alignment is the process of encoding human values and goals into AI models to make them as helpful, safe and reliable as possible.”*
  https://www.ibm.com/think/topics/ai-alignment
- *MOVING TARGET!*
  - *pre-LLM: algorithmic bias, interpretability*
  - *post-LLM: pre-LLM + all the stuff that can go wrong as a result of understanding human language*

What can go wrong?



- ARC TaskRabbit experiment (2023)
- Social Credit Systems
  - E.g. Nosedive (Black Mirror 2016)
  - China is currently experimenting with
- Cognitive impacts
  - Reversal of Flynn Effect (already observing)
  - Social media impact on teens (Haidt)
- Amplification of Human vices and societal problems
  - Rohingya genocide in Myanmar
  - 2025-06-12 Meta sues CrushAI
  - Inequality in all forms
- Distortion of the public conversation needed for a healthy democracy
- Autonomous weapon systems

AI needs a leash… (not guardrails)

- Guard rails require anticipating problematic behavior
- Can work in very limited domains, but...
- In general, there are too many ways to get into trouble

Regulation, Mitigation & Preparation

- TAKE IT DOWN Act
  - It’s a start...
- CO AI Law
  - focus on 8 areas
  - addresses algorithmic bias in consequential areas
- Proposed 10 year moratorium on state-level AI regulation
  - VERY dangerous! Ten years in this space is like 100 years in others…
  - Concerns on both sides of the aisle on this issue, but unclear how many Rep’s are concerned enough to remove
- Have a code word between love ones
- Stay curious, engaged and get help from people like Erin

Summary

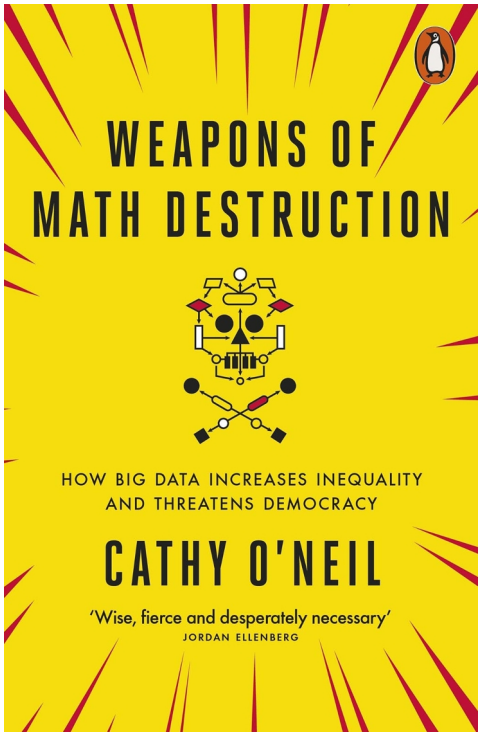
- Big players #1 priority is developing capabilities as fast as possible
- Safety is important, but not primary
  - Hope “someone” figures out the safety problem (Altman quote)
- AI alignment minimizes unintended consequence arising from a mismatch between machine goals and human values.
  - … but humans don’t share the same set of values: Israelis & Palestinians, Russians & Ukrainians, Democrats and Republicans...

- Consciousness ≠ Intelligence
- Conscious beings feel thing
- Intelligent beings know how to take action to achieve goals
- Humans are both conscious and intelligent.
- Humans are more story-driven than truth-drive
- AI agents (so far) are only intelligent.

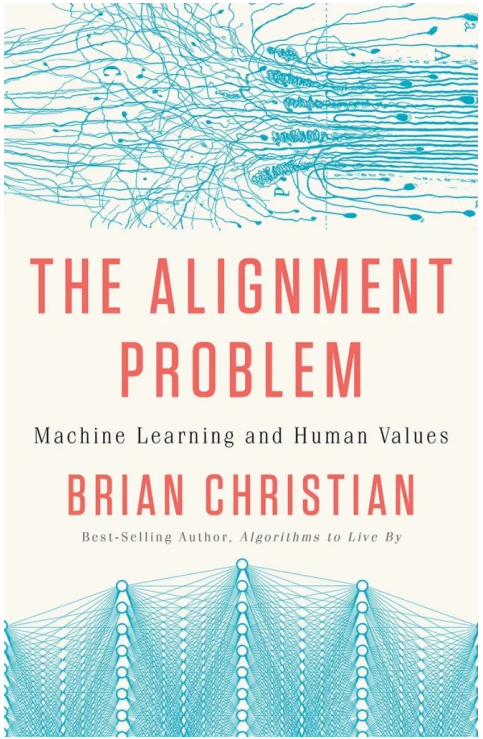
Summary (cont.)

- AI brings the power to greatly amplify capabilities in many areas of human endeavor
  - What will we amplify?
- A lot has already gone wrong
  - Social media impact on teens (The Anxious Generation - Haidt)
  - Rohingya genocide in Myanmar
  - ARC TaskRabbit experiment
- Currently proposed 10 yr. moratorium on state-level AI regulation is extremely dangerous
- Prepare
  - Stay engaged: read, talk to friends, come to meetings like these!
  - Ask for help from knowledgeable people.
  - Have secret code word with loved ones.

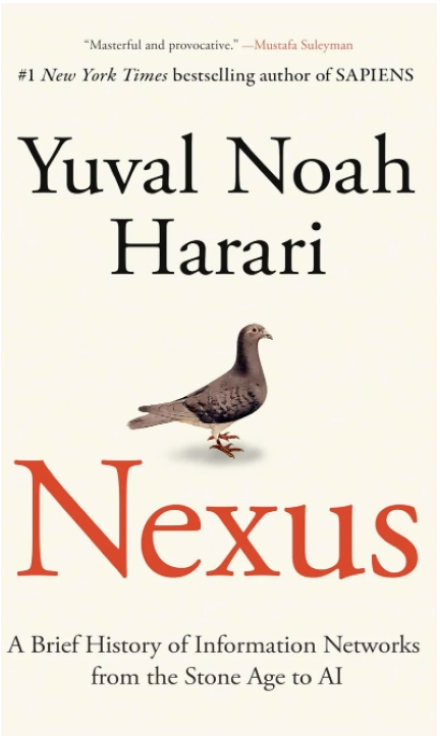
Recommended Reading



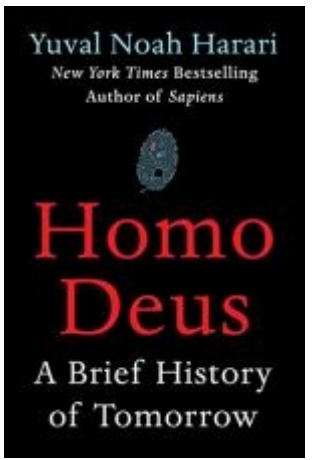
2016



2020



2024

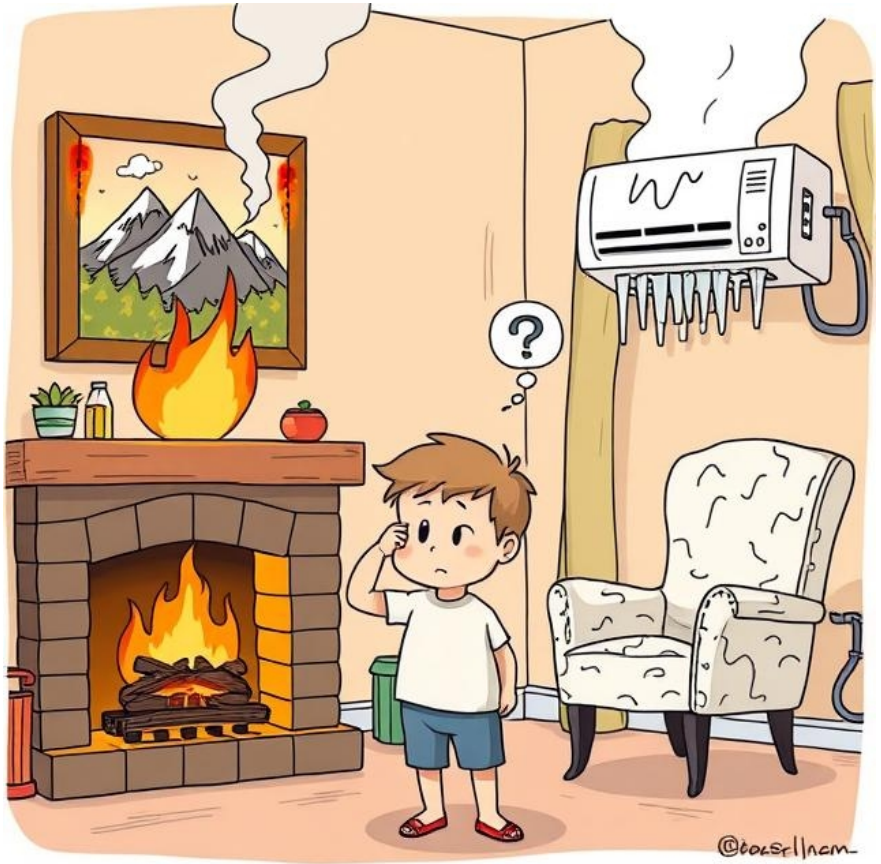


Appendices

- How the world feels to me…
- Information, Truth and Reality
- AI startups world-wide 2023

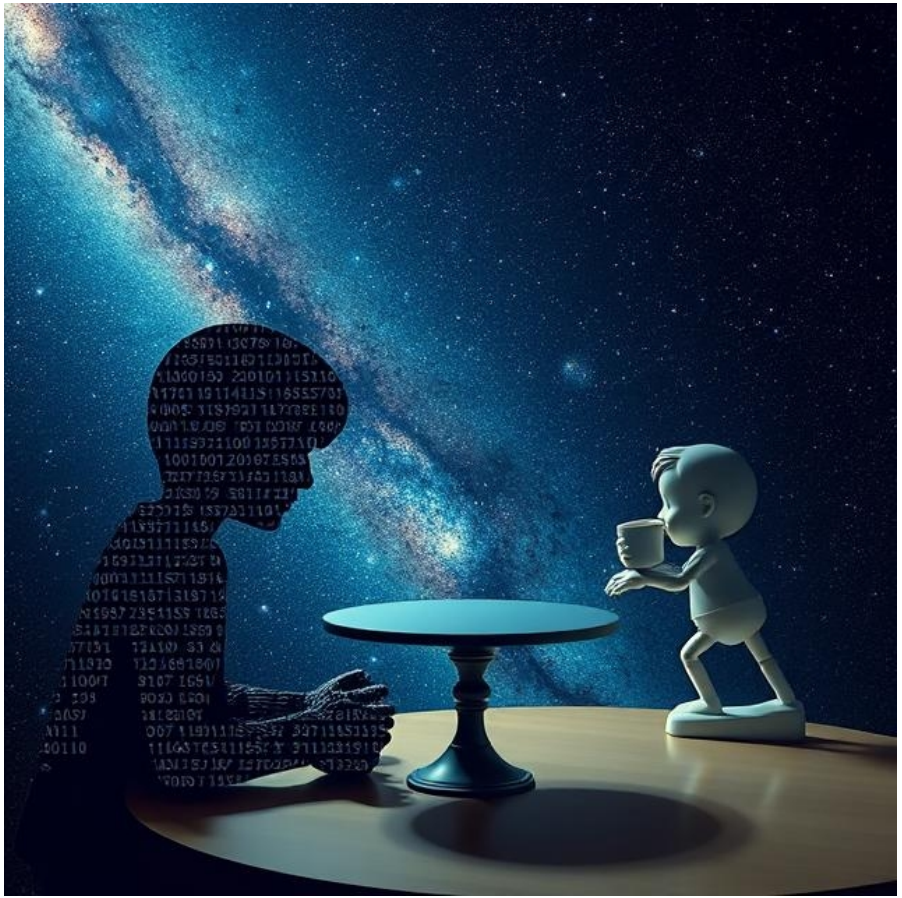
What it feels like to me...

- Computers empower AI **defense** weapon systems and scientific insights on the climate, vaccines, etc.
- Computers empower AI **offensive** weapon systems and amplify misinformation on climate, vaccines, etc.
- Human became the rulers of earth because we cooperate better than any other species.



AI generated image: https://deepai.org/machine-learning-model/text2img

Information, Truth & Reality



- Information
  - can be fiction, factual, some blending of both
  - shapes reality: belief in a narrative drives behavior
- Truth
  - things that are not false, describes reality accurately but incompletely
  - useful to the extent of what is included and what is left out
- Reality
  - too vast, all attempts to describe are incomplete
  - universal, but contains many perspectives
  - perspectives shaped by information (culture, physiology, et. al.)
  - Humans have lived within the dreams of other humans. This may shift to the dreams of an Alien Intelligence.

AI Startups Worldwide (2023)