Ethics, Public Policy, and Technological Change

Rob Reich

Mehran Sahami

Head TA: Roberta Fischli

Housekeeping

1. Friday sections will discuss responsible scaling policies and model release
2. Policy Memo guidance
   - Establish your group ASAP
   - Get started on interviews now

3. A debate on AI and work: Rob Reich and Bharat Chandar
   - Tuesday, March 3, 3-5pm, CEMEX Auditorium
4. Monday, March 9 in class: three guests
   - Zoe Hitzig (ex OpenAI)
   - Ryan Beiermeister (ex Palantir, ex Meta, ex OpenAI)
   - Artemis Seaford (ex Meta, ex Eleven Labs)

Conference / Symposium

AI and Jobs with Bharat Chander and Rob Reich

FROM THE SERIES:

Democracy and Disagreement Winter ’26

Tue., March 3, 2026
3:00pm - 4:50pm

CEMEX Auditorium, 655 Knight Way, Stanford, CA

Free

Add to Calendar

Bharat Chandar, a postdoctoral researcher in the Stanford Digital Economy Lab, part of the Institute for Human-Centered Artificial Intelligence, and **Rob Reich**, the McGregor-Girand Professor of Social Ethics of Science and Technology at Stanford, discuss how AI affects various job markets.

Deep disagreement pervades our democracy, from arguments over immigration, gun control, abortion, and the Middle East crisis, to the function of elite higher education and the value of free speech itself. Loud voices drown out discussion. Open-mindedness and humility seem in short supply among politicians and citizens alike. Yet constructive disagreement is an essential feature of a democratic society. This class explores and


Stop Hiring Humans
The Era of AI Employees Is Here
ARTISAN
artisan.co

Category / Company

The Story Behind the “Stop Hiring Humans” Billboards in San Francisco

A controversial billboard campaign in San Francisco featuring the provocative message 'Stop Hiring Humans' generated millions of impressions, sparked heated debate, and drove $2M in new ARR for Artisan. Here's the story.

Jaspar Carmichael-Jack | Dec 18, 2025 | 11 minutes read

Meet the Founders

Jaspar Carmichael-Jack

Jaspar is a third-time founder with a track record of building and scaling companies. Before Artisan, he founded a marketing agency where he gained deep expertise in go-to-market strategies and recognized the opportunity to revolutionize them through software.

Sam Stallings

Sam brings a powerful blend of product and engineering expertise to Artisan. After eight years at IBM, she worked at a YC startup as the founding engineer. Today, Sam leads Artisan's product and engineering teams, turning vision into reality.

Generative AI: Promise and Perils

1. Human Intelligence, AI, and AGI
2. GPTs are GPTs
3. Beyond Utopian and Dystopian
4. Topics for the AI Transition
   - Autonomous vehicles
   - AI and work
   - AI and the military (surveillance, autonomous weapons)
   - AI and democracy
5. Going Deeper in Philosophy
   - Nozick and the Experience Machine: are virtual experiences different?
   - Scanlon and the World Cup: problems with aggregation of harms
   - Foot and the Trolley Problem: programming autonomous vehicles
   - Thomsen and the Fat Man

Sept 19, 2025

Governing Artificial Intelligence: Law, Policy, and Institutions

Law 4052 // Computer Science 283 // Communication 252A // Political Science 245B/445B

Monday/Wednesday 2:15 - 3:45pm
Crown Building, Stanford Law School, Room 290

Nate Persily (School of Law , Political Science, FSI, Communication)
npersily@stanford.edu
Office – Law School's Neukom Building, Room N230
Office Hours: MW 11:15-12:15
Rob Reich (Political Science)
reich@stanford.edu
Office - Encina Hall Central, room 441
Office Hours: M 430-6pm, book appointment with Dominic Zappia (zappia@stanford.edu)
Anka Reuel (Computer Science)
anka.reuel@stanford.edu
Office Hours: TBD, upon request
Sanmi Koyejo (Computer Science)
sanmi@cs.stanford.edu
Office Hours: TBD, upon request

Setting the Scene…

Some slides from the last time I taught this class in winter 2023

Text input to GPT-3:

“Here is a list of blurbs for a new book by three Stanford University professors, Rob Reich, a philosopher, Mehran Sahami, a computer scientist, and Jeremy M. Weinstein, a public policy expert. The book is called System Error: Where Big Tech Went Wrong and How We Can Reboot.”

USe GPT-3 To Generate Blurbs for a Book

Can you guess which are from real people?

Discuss with your partner!

1

"In the hands of these three interdisciplinary experts, the subservient role of technology in our lives comes fully into view. Too often we do what we're told; we need a path to a better world, and System Error is one essential guide for our collective journey."

– Jeff Skoll, former Chairman of eBay, Founder and and Founding Chairman of the Skoll

2

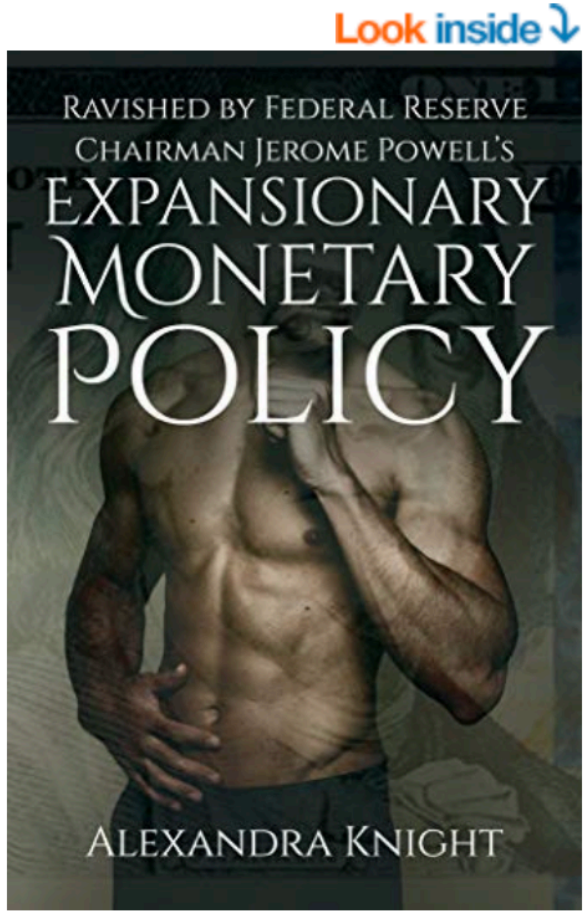
"While technology has unleashed great wealth and power, it has also subverted how we interact with one another, jeopardizing our democracy. Bold solutions to these problems are not too far away, and we can discover what they are by paying careful attention to System Error."

3

"A gripping account of one of our era's most consequential interplays between technology and society. It is hard to imagine a more urgent or necessary book."

– Cass Sunstein, Robert Walmsley University Professor at Harvard University

‹ Back to results



Ravished by Federal Reserve Chairman Jerome Powell's Expansionary Monetary Policy Kindle Edition

by Alexandra Knight (Author) | Format: Kindle Edition

★★★★★ 3 ratings

See all formats and editions

Kindle	Paperback
$2.99	$6.99
Read with Our Free App	1 New from $6.99

He's one of the most powerful men in the world. With his money printer, he can build empires... and topple them. And he wants me.

I'm a nobody. But Jay Powell doesn't discriminate. He's chosen me to help him out with his next round of quantitative easing.

My body is quaking with fear. But there's a part of me that's excited, too.

He's going to take me, break me, ravish me. And I'm going to love every moment of it.

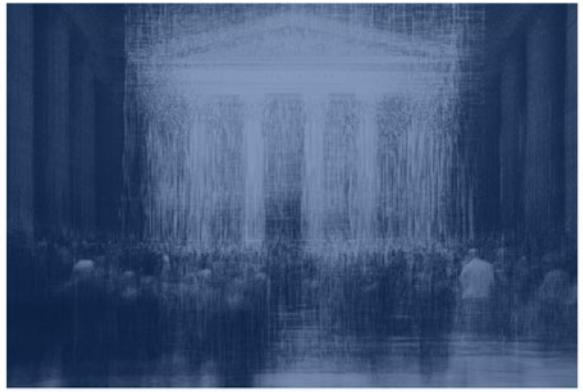
A Peculiar Year in 2025

AI as Normal Technology

An alternative to the vision of AI as a potential superintelligence

By Arvind Narayanan and Sayash Kapoor

April 15, 2025



Sébastien A. Krier using Midjourney 6.1

We articulate a vision of artificial intelligence (AI) as normal technology. To view AI as normal is not to understate its impact—even transformative, general-purpose technologies such as electricity and the internet are "normal" in our conception. But it is in contrast to both utopian and dystopian visions of the future of AI which have a common tendency to treat it akin to a separate species, a highly autonomous, potentially superintelligent entity.[1]

AI 2027

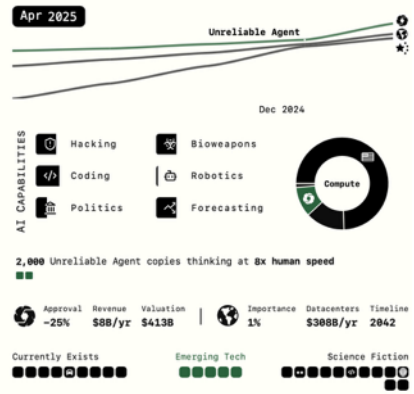
Summary Research About

Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, Romeo Dean

We predict that the impact of superhuman AI over the next decade will be enormous, exceeding that of the Industrial Revolution.

We wrote a scenario that represents our best guess about what that might look like.[1] It's informed by trend extrapolations, wargames, expert feedback, experience at OpenAI, and previous forecasting successes.[2]

What is this? | How did we write it? | Why is it valuable? | Who are we?



Published April 3rd 2025 | PDF | Listen

Mid 2025: Stumbling Agents

The world sees its first glimpse of AI agents.

Advertisements for computer-using agents emphasize the term "personal assistant": you can prompt them with tasks like "order me a burrito on DoorDash" or "open my budget spreadsheet and sum this month's expenses." They will check in with

A Peculiar Year in 2025

Geoffrey Hinton

Yann LeCun

Leading in the Face of Uncertainty

- Yoshua Bengio:

- We should be guided by the precautionary principle:

- When potential risks and harms are catastrophic or existential, a lack of scientific certainty should not delay preventive measures.

Leading in the Face of Uncertainty

Marc Andreessen

”Our enemy is the precautionary principle”

← **Post**

Marc Andreessen @pmarca

The Precautionary Principle: The haunting fear that someone, somewhere, may be doing something innovative.

5:54 PM · Oct 27, 2023 · **120.9K** Views

Source: Wikipedia CC license

Early 2026?

Contents

The Adolescence of Technology

Confronting and Overcoming the Risks of Powerful AI

January 2026

There is a scene in the movie version of Carl Sagan's book *Contact* where the main character, an astronomer who has detected the first radio signal from an alien civilization, is being considered for the role of humanity's representative to meet the aliens. The international panel interviewing her asks, "If you could ask [the aliens] just one question, what would it be?" Her reply is: "I'd ask them, 'How did you do it? How did you evolve, how did you survive this technological adolescence without destroying yourself?" When I think about where humanity is now with AI—about what we're on the cusp of—my mind keeps going back to that scene, because the question is so apt for our current situation, and I wish we had the aliens' answer to guide us. I believe we are entering a rite of passage, both turbulent and inevitable, which will test who we are as a species. Humanity is about to be handed

The New York Times

Pentagon Standoff Is a Decisive Moment for How A.I. Will Be Used in War

The Pentagon's contract dispute with Anthropic is part of a wider clash about the use of artificial intelligence for national security and who decides on any safeguards.

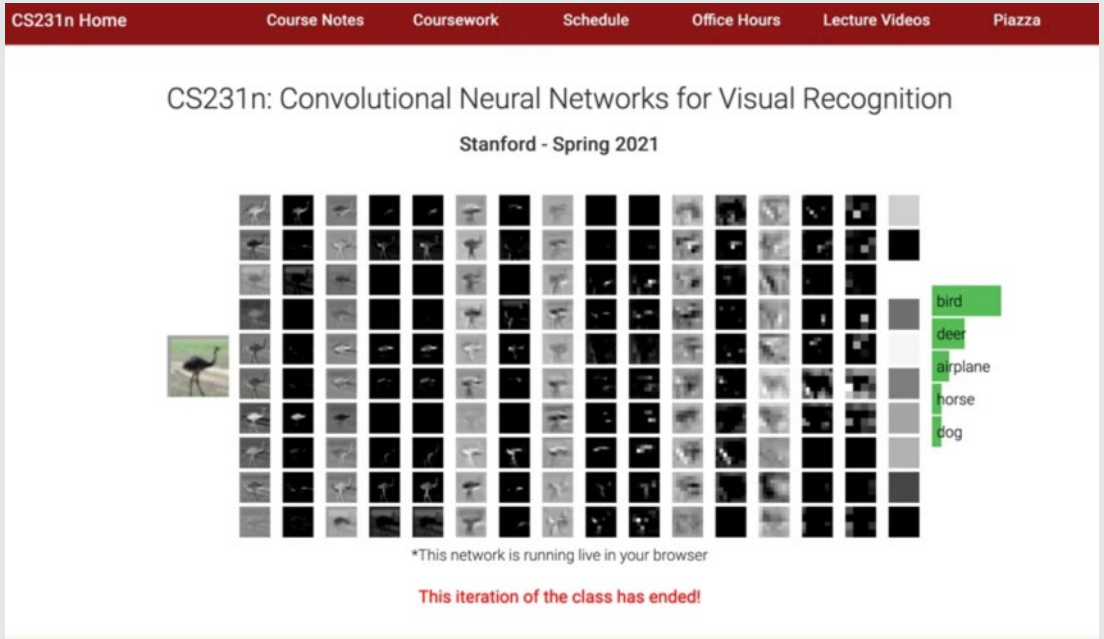
Generative AI: Promise and Perils

1. Human Intelligence, AI, and AGI
2. GPTs are GPTs
3. Beyond Utopian and Dystopian
4. Topics for the AI Transition
   - Autonomous vehicles
   - AI and work
   - AI and the military (surveillance, autonomous weapons)
   - AI and democracy
5. Going Deeper in Philosophy
   - Nozick and the Experience Machine: are virtual experiences different?
   - Scanlon and the World Cup: problems with aggregation of harms
   - Foot and the Trolley Problem: programming autonomous vehicles
   - Thomsen and the Fat Man

The History and Evolution of AI

Mehran to address the history of AI.

- supervised learning
- unsupervised learning
- reinforcement learning
- deep learning
- neural networks
- convolutional neural networks
- large language models



Reasoning and Intelligence

- What does it mean to be a reasoning animal?
- What is the relationship between reason and intelligence?

Two Kinds of Rationality

Instrumental Rationality

= *Practical Reasoning*

- Reasoning about **means** to given ends
- Goals are taken as **fixed**; only the path is evaluated
- Question: *How do I achieve X?*
- Dominant in economics (preference satisfaction), AI optimization
- Not intended to assess the worthiness of ends or goals; silent on *whether* ends are worth pursuing

Substantive Rationality

= *Theoretical Reasoning*

- Reasoning about **ends** themselves — what *ought* to be pursued
- Goals are **subject to evaluation** by reason
- Question: *Should I be pursuing X at all?*
- Central to ethics, political philosophy, deliberative democracy

AI excels at instrumental rationality — but substantive questions about what ends to pursue require human judgment

AI vs. AGI

For our purposes: distinguish between AI and AGI, artificial general intelligence.

Narrow/Weak AI

Machine intelligence for a particular task or goal (e.g., Chess, facial recognition, piloting a car, Alpha Go)

General/Strong AI

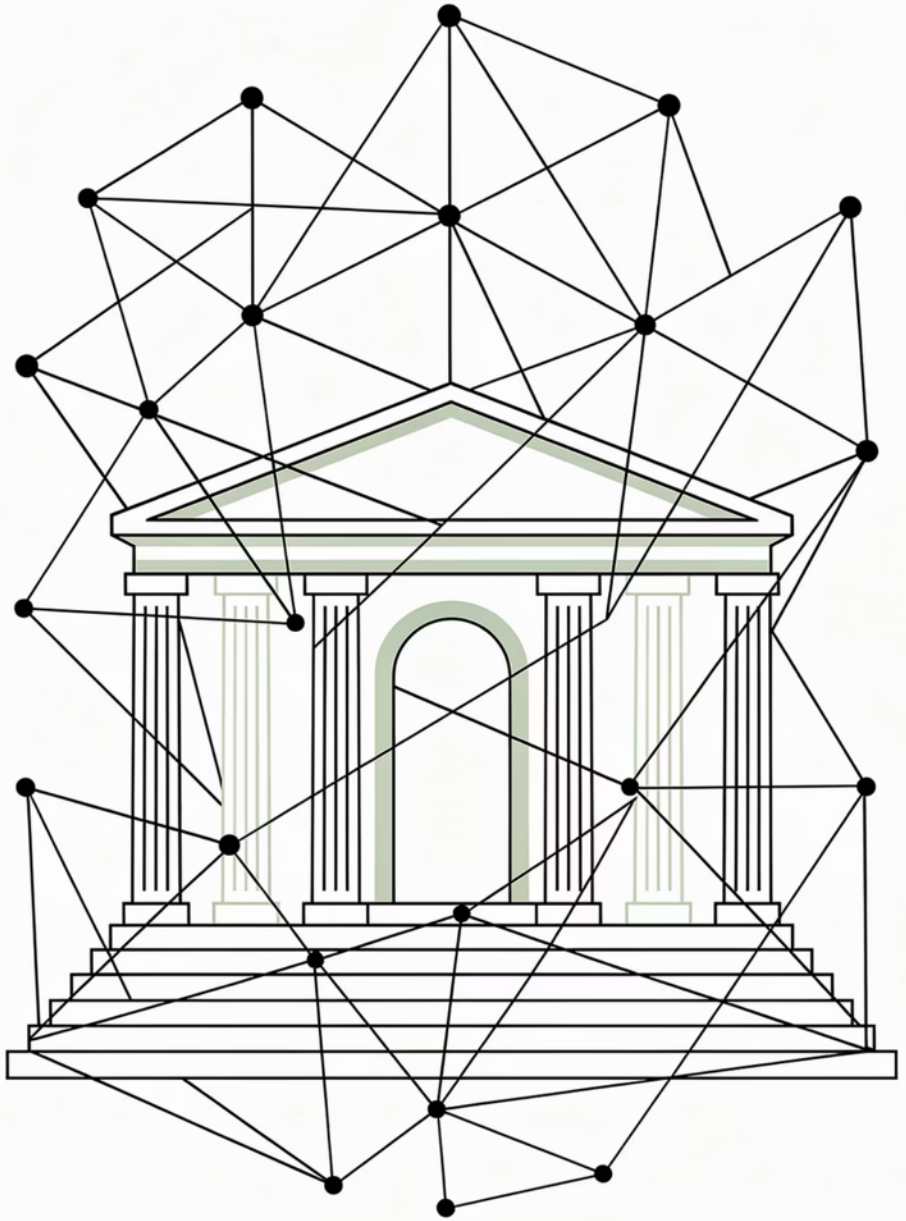
Machine intelligence on a par with human intelligence.

Open Question

Can machine intelligence set its own goals?

Intelligence, Measured & Misunderstood

The split between **instrumental** and **substantive** rationality shaped psychometrics, cognitive science, and AI.



The Rationality Split

Substantive Rationality	Instrumental Rationality
Are your beliefs and goals themselves well-founded? Do they track reality and reflect genuine values?	Do you efficiently pursue *whatever* goals you happen to have — regardless of their content?

When 20th-century researchers tried to **measure** intelligence, they largely — and somewhat covertly — sided with the instrumental picture.

Lewis Terman & the Stanford-Binet

Lewis Terman, a psychologist at Stanford University, became one of the most influential figures in the history of intelligence testing.

The Stanford-Binet Scale

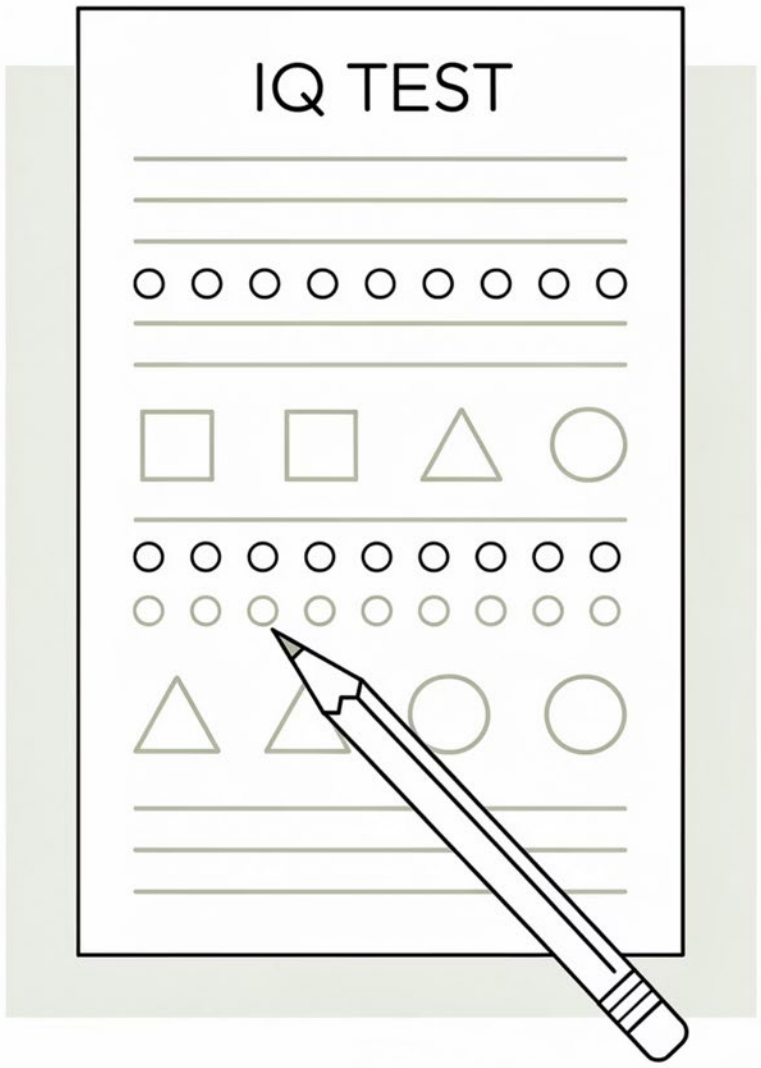
Terman revised Alfred Binet's original test, creating the **Stanford-Binet Intelligence Scale (1916)**, the first widely used standardized IQ test in the United States. He also popularized the concept of the **"Intelligence Quotient" (IQ)** as a single number representing cognitive ability.

"Genetic Studies of Genius"

In 1921, he launched this landmark longitudinal study, meticulously tracking **gifted children** over decades to understand their development and long-term success.

Enduring Influence & Perspective

Terman believed intelligence was largely hereditary and fixed—a view that reinforced the instrumental conception of intelligence as a measurable, stable quantity. His work cemented the idea of intelligence being reducible to a single score, profoundly shaping education, military testing (Army Alpha/Beta in WWI), and even early AI research.



The Psychometric Tradition

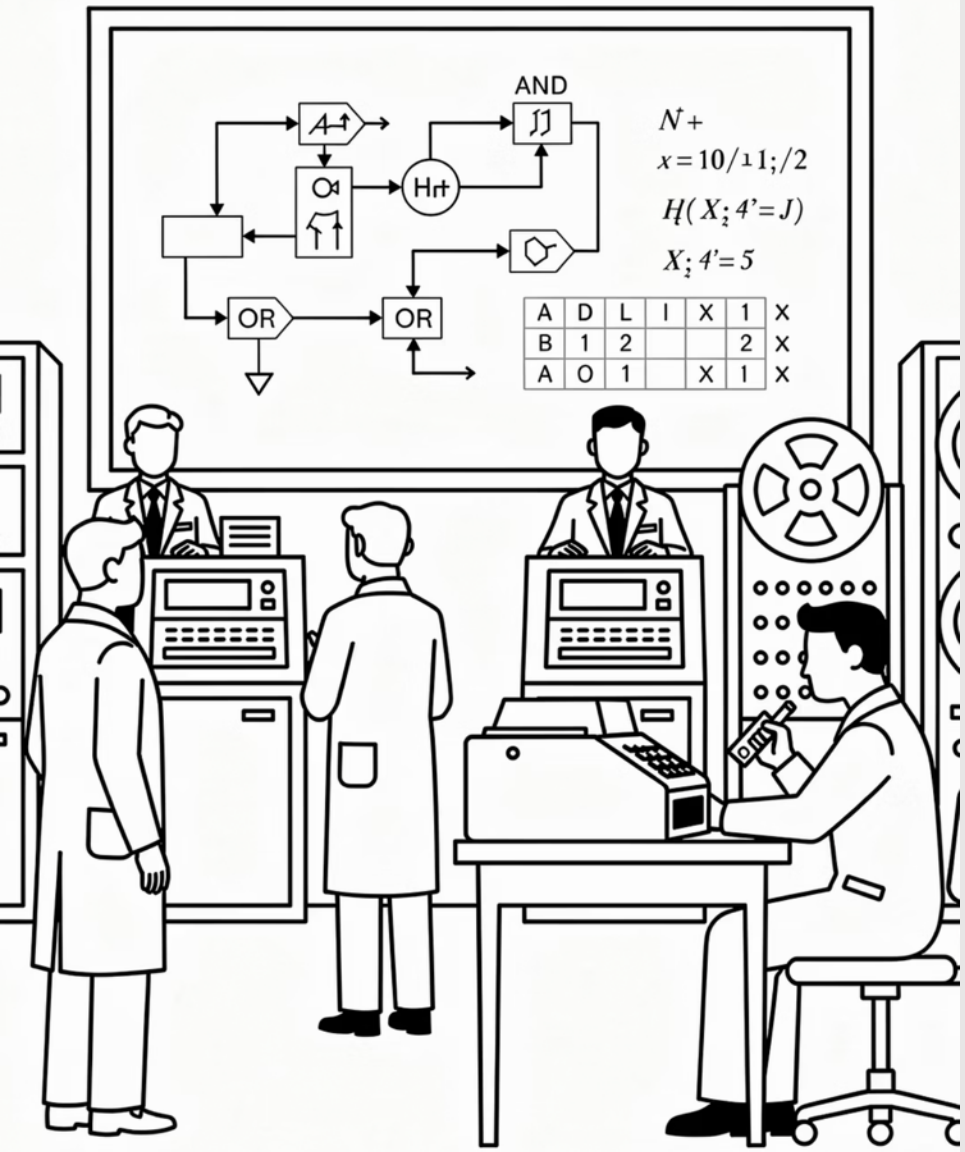
From **Binet** through Spearman's ***g*** factor to modern IQ testing, intelligence was operationalized as processing efficiency: speed, working memory, pattern recognition, abstract symbol manipulation.

Practical Appeal

You could test it. You could get a number. It correlated with school performance and job outcomes.

The Critics

Luria, Vygotsky, Gardner, and Sternberg argued this measured a **narrow slice** of cognition — mistaking the instrument for the thing itself.



From Measurement to Machine Intelligence

When AI researchers in the 1950s tried to *create* machine intelligence, they inherited the instrumental frame.

Logic Theorist

Newell & Simon modeled intelligence as **search through a problem space** — goal state, operators, path. Pure instrumental rationality, formalized.

The Assumption

If reasoning is symbol manipulation in service of goal pursuit, it doesn't matter whether symbols run on **neurons or silicon**.

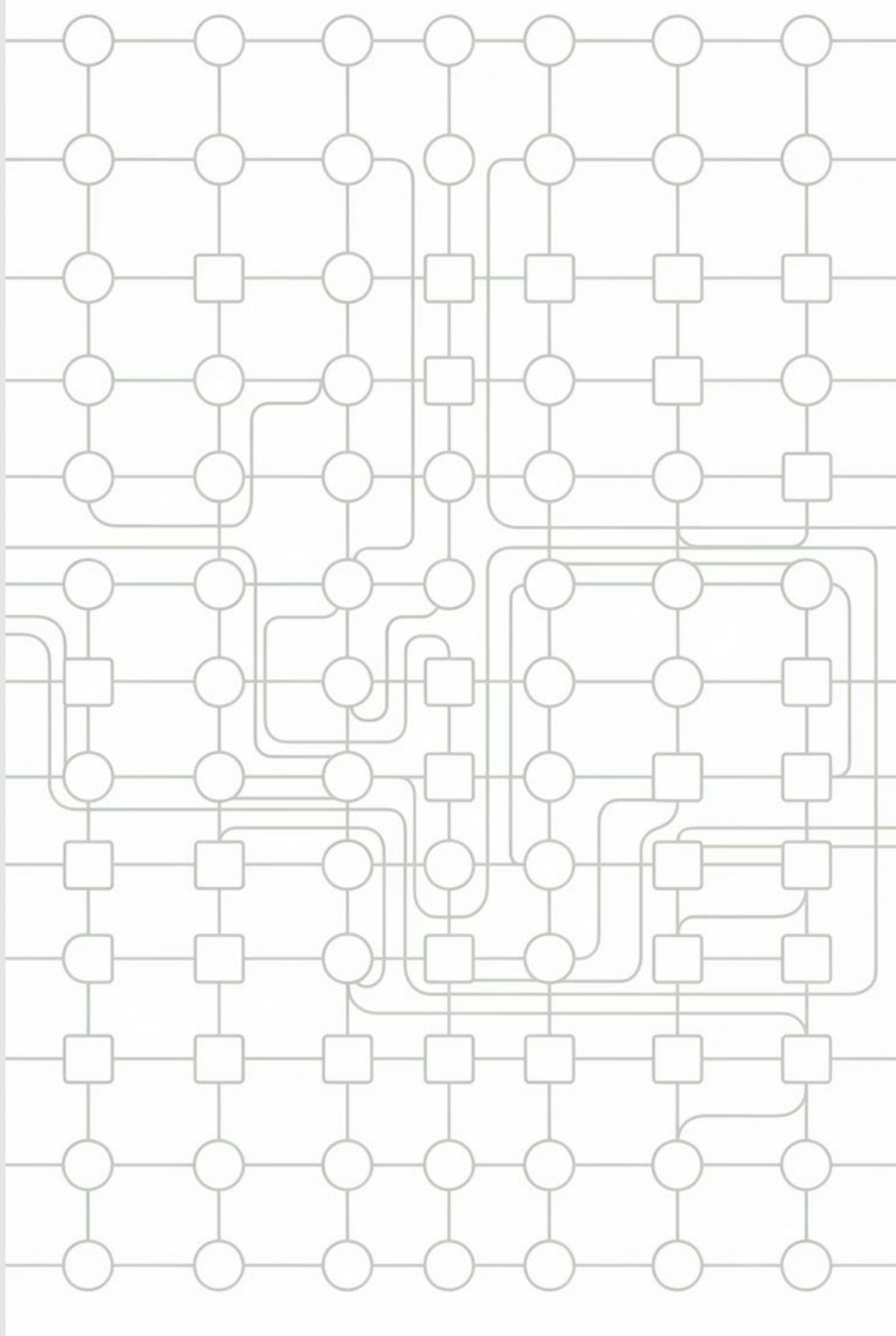
The Failure

Early AI systems proved **brittle outside training domains** — because real intelligence involves more than efficient means-end reasoning.

The Central Tension in Frontier AI

> Large language models are extraordinarily capable at instrumental-style pattern completion and reasoning — but whether they possess anything like substantive rationality remains genuinely open and philosophically contested.

The question of whether an AI system has genuine understanding of what its goals *should be* — not just how to pursue them — is still unresolved.



Generative AI: Promise and Perils

1. Human Intelligence, AI, and AGI
2. GPTs are GPTs
3. Beyond Utopian and Dystopian
4. Topics for the AI Transition
   - Autonomous vehicles
   - AI and work
   - AI and the military (surveillance, autonomous weapons)
   - AI and democracy
5. Going Deeper in Philosophy
   - Nozick and the Experience Machine: are virtual experiences different?
   - Scanlon and the World Cup: problems with aggregation of harms
   - Foot and the Trolley Problem: programming autonomous vehicles
   - Thomsen and the Fat Man

POLICY FORUM | ARTIFICIAL INTELLIGENCE

GPTs are GPTs: Labor market impact potential of LLMs

Research is needed to estimate how jobs may be affected

TYNA ELOUNDOU, SAM MANNING, PAMELA MISHKIN, AND DANIEL ROCK Authors Info & Affiliations

SCIENCE • 20 Jun 2024 • Vol 384, Issue 6702 • pp. 1306-1308 • DOI: 10.1126/science.adj0998

We propose a framework for evaluating the potential impacts of large-language models (LLMs) and associated technologies on work by considering their relevance to the tasks workers perform in their jobs. By applying this framework (with both humans and using an LLM), we estimate that roughly 1.8% of jobs could have over half their tasks affected by LLMs with simple interfaces and general training. When accounting for current and likely future software developments that complement LLM capabilities, this share jumps to just over 46% of jobs. The collective attributes of LLMs such as generative pretrained transformers (GPTs) strongly suggest that they possess key characteristics of other "GPTs," general-purpose technologies (*1*, *2*). Our research highlights the need for robust societal evaluations and policy measures to address potential effects of LLMs and complementary technologies on labor markets.

DEFINITION

What Is a General Purpose Technology?

A **General Purpose Technology (GPT)** is a transformative innovation so foundational it reshapes entire economies and societies over decades — not just a single industry or workflow.

Pervasive Use

Adopted broadly across sectors and industries

Continuous Improvement

Evolves and grows more capable over time

Innovation Spawning

Generates entirely new products, ecosystems, and industries

HISTORICAL CONTEXT

Major General Purpose Technologies Through History

1

18th Century

Steam Engine Powered the First Industrial Revolution — mechanizing agriculture and manufacturing at unprecedented scale

2

Late 19th Century

Electricity & Machine Tools Fueled the Second Industrial Revolution, enabling mass production and entirely new industries

3

20th Century

Computers & Digital Tech Sparked the Third Industrial Revolution, collapsing silos and creating global digital platforms

4

21st Century

Generative AI Embedding intelligence across every sector, accelerating innovation and driving ecosystem-based economies

Each GPT triggered a multi-decade wave of economic transformation — reshaping labor markets, business models, and the structure of societies long after initial adoption.

Generative AI As the Next GPT

🌐 Pervasiveness

ChatGPT reached **100M+ users in just 2 months** — faster than any prior technology. Powering Bing AI, developer APIs, and thousands of applications

📈 Rapid Improvement

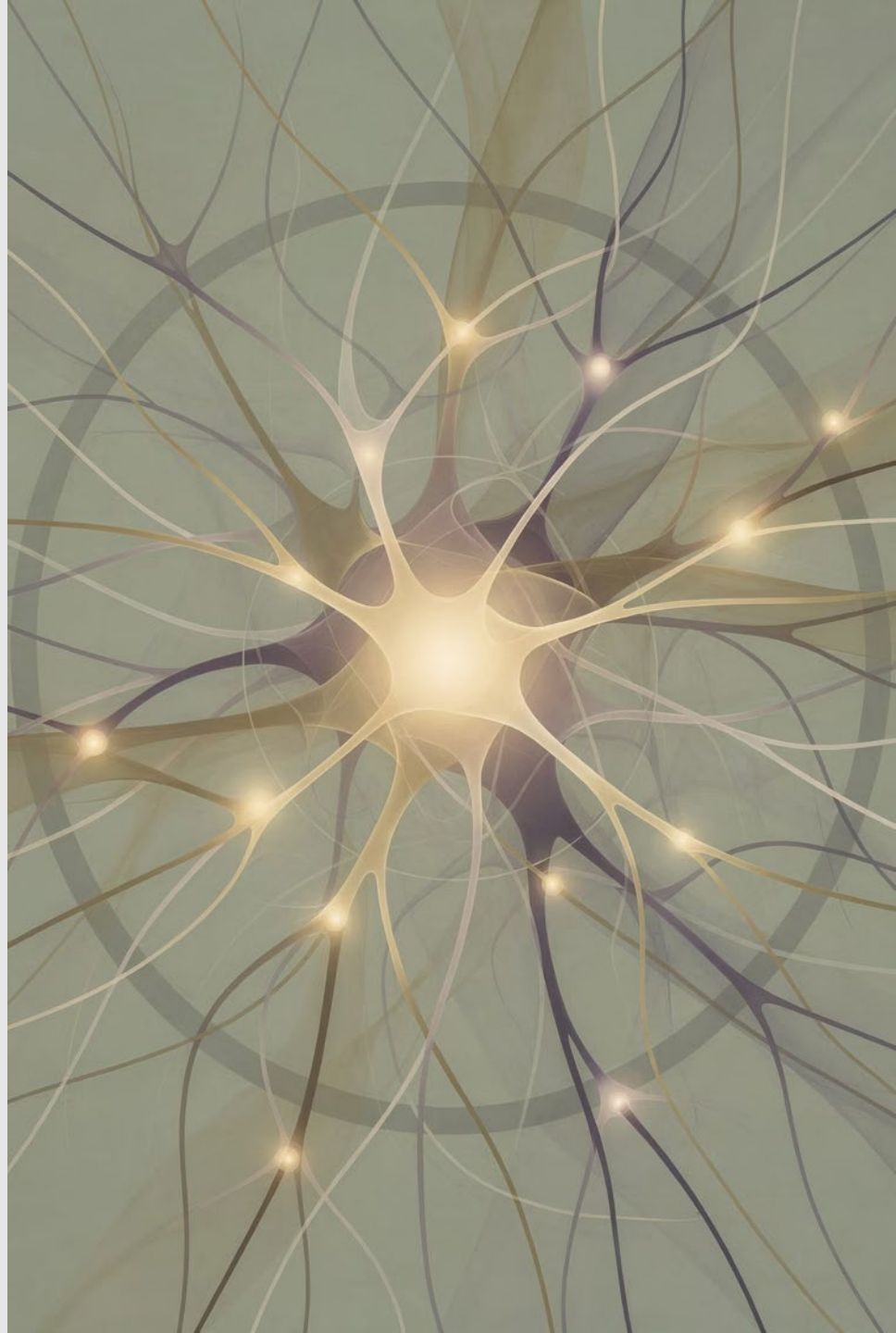
GPT models improve in capability and efficiency at a pace that outstrips nearly every predecessor technology

🔬 Innovation Spawning

Enabling new products, services, and workflows across healthcare, education, finance, law, and creative industries

💼 Labor Market Impact

Studies show **80% of U.S. jobs** have tasks significantly affected by LLMs — signaling broad, structural economic transformation

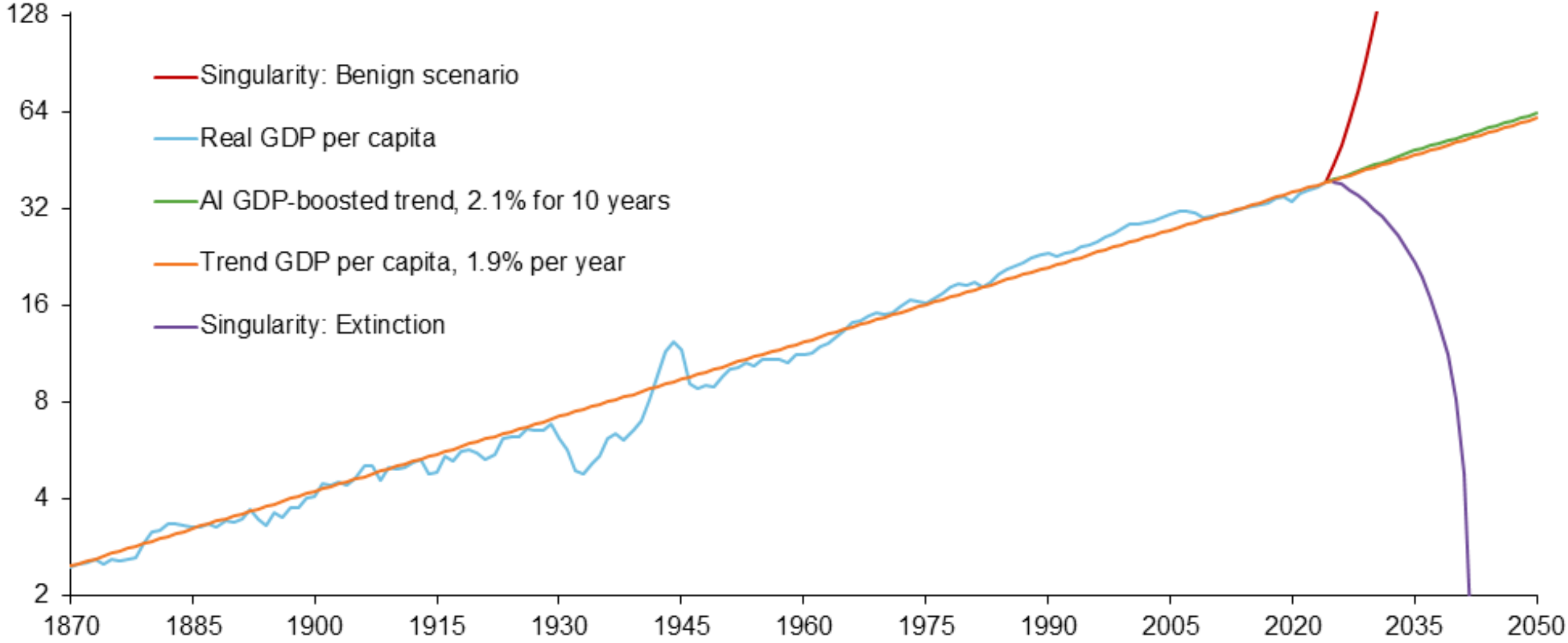


Generative AI: Promise and Perils

1. Human Intelligence, AI, and AGI
2. GPTs are GPTs
3. Beyond Utopian and Dystopian
4. Topics for the AI Transition
   - Autonomous vehicles
   - AI and work
   - AI and the military (surveillance, autonomous weapons)
   - AI and democracy
5. Going Deeper in Philosophy
   - Nozick and the Experience Machine: are virtual experiences different?
   - Scanlon and the World Cup: problems with aggregation of harms
   - Foot and the Trolley Problem: programming autonomous vehicles
   - Thomsen and the Fat Man

AI scenarios

1990 dollars (thousands), log scale



NOTES: The blue line is real gross domestic product (GDP) per capita in 1990 dollars. The orange line is a trend line fitted to the data for 1870–2024 with a trend growth rate of 1.9 percent per year. The red, green and purple lines are hypothetical paths for per capita GDP based on different scenarios.

SOURCES: Bureau of Economic Analysis; Haver Analytics; Macrohistory.net; United Nations; authors' calculations.

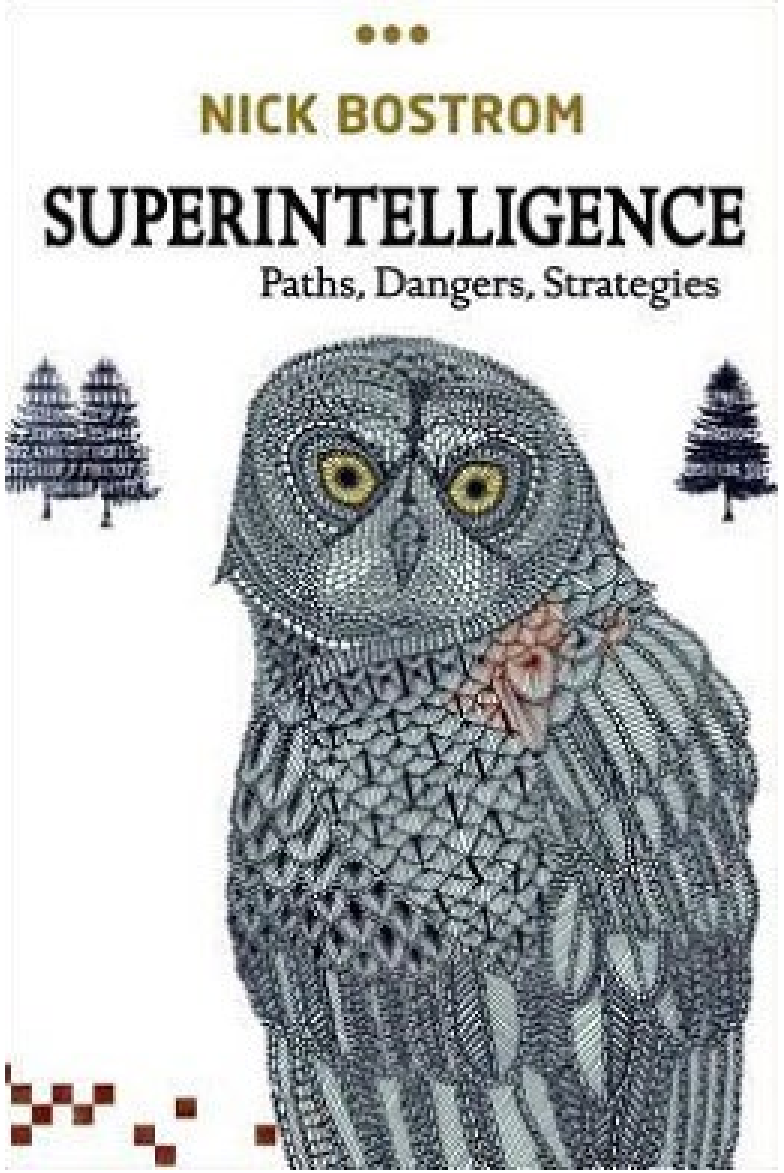
Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence

Erik Brynjolfsson* Bharat Chandar† Ruyu Chen‡§¶

November 13, 2025

1 Introduction

The proliferation of generative artificial intelligence (AI) has sparked a global debate about its potential impact on the labor market. This discourse spans utopian predictions of enhanced productivity, dystopian fears of widespread job displacement, and skeptical views that AI will have minimal effects on employment or productivity. Historically, technologies have affected different



The Vulnerable World Hypothesis

Nick Bostrom

Future of Humanity Institute, University of Oxford

Abstract

Scientific and technological progress might change people's capabilities or incentives in ways that would destabilize civilization. For example, advances in DIY biohacking tools might make it easy for anybody with basic training in biology to kill millions; novel military technologies could trigger arms races in which whoever strikes first has a decisive advantage; or some economically advantageous process may be invented that produces disastrous negative global externalities that are hard to regulate. This paper introduces the concept of a *vulnerable world*: roughly, one in which there is some level of technological development at which civilization almost certainly gets devastated by default, i.e. unless it has exited the 'semi-anarchic default condition'. Several counterfactual historical and speculative future vulnerabilities are analyzed and arranged into a typology. A general ability to stabilize a vulnerable world would require greatly amplified capacities for preventive policing and global governance. The vulnerable world hypothesis thus offers a new perspective from which to evaluate the risk-benefit balance of developments towards ubiquitous surveillance or a unipolar world order.

Policy Implications

- Technology policy should not unquestioningly assume that all technological progress is beneficial, or that complete scientific openness is always best, or that the world has the capacity to manage any potential downside of a technology after it is invented.
- Some areas, such as synthetic biology, could produce a discovery that suddenly democratizes mass destruction, e.g. by empowering individuals to kill hundreds of millions of people using readily available materials. In order for civilization to

The "Vulnerable World Hypothesis"

Image: Wikimedia Commons, CC BY-SA 4.0

- Humanity is splintered at the highest political level into several different competing political units – countries – and we have no reliable way of resolving differences between different countries.
- The world spends billions of dollars every year in producing and maintaining assets and technologies to kill each other. In this way, technology can enable unprecedented destruction.
- The advance of technology is producing the capability to destroy civilization, possibly end life on earth – nuclear weapons, climate change, bioterrorism, runaway AI
- How to deal with this?

One idea: Bring on the Panopticon!

"What I argue is that the only way civilization can survive is if we create vastly more powerful ways of controlling the use of cheap technology for mass destruction. **This would require continuous surveillance of unprecedented efficiency, such as an ability to intervene in real time."**

Nick Bostrom, Vulnerable World Hypothesis

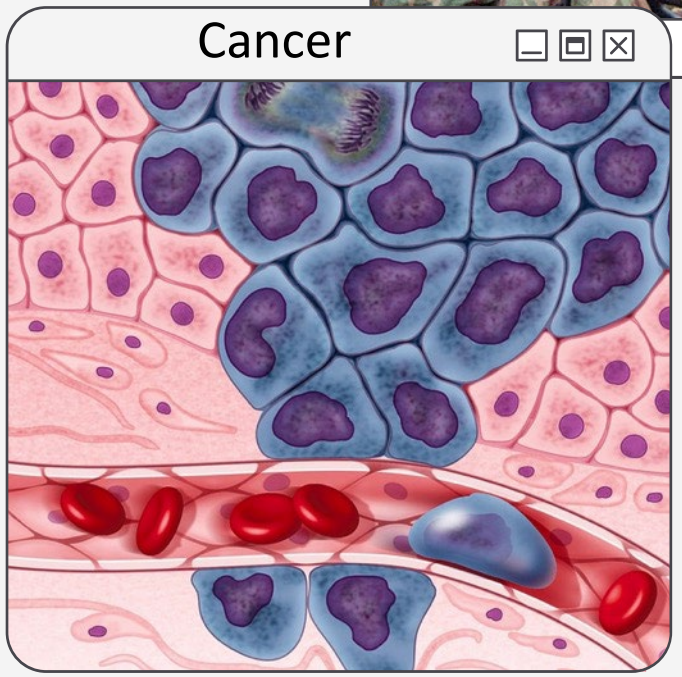
Image: Wikimedia Commons, CC BY 4.0

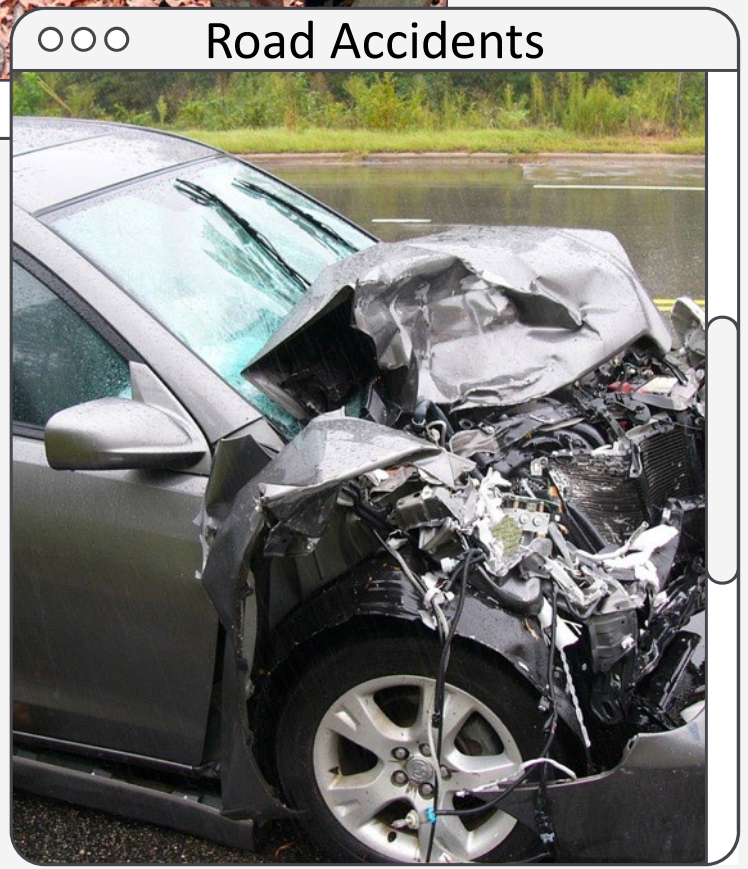
Generative AI: Promise and Perils

1. Human Intelligence, AI, and AGI
2. GPTs are GPTs
3. Beyond Utopian and Dystopian
4. Topics for the AI Transition
   - Autonomous vehicles
   - AI and work
   - AI and the military (surveillance, autonomous weapons)
   - AI and democracy
5. Going Deeper in Philosophy
   - Nozick and the Experience Machine: are virtual experiences different?
   - Scanlon and the World Cup: problems with aggregation of harms
   - Foot and the Trolley Problem: programming autonomous vehicles
   - Thomsen and the Fat Man

AI can improve Human Welfare.

But even so, raises important challenges

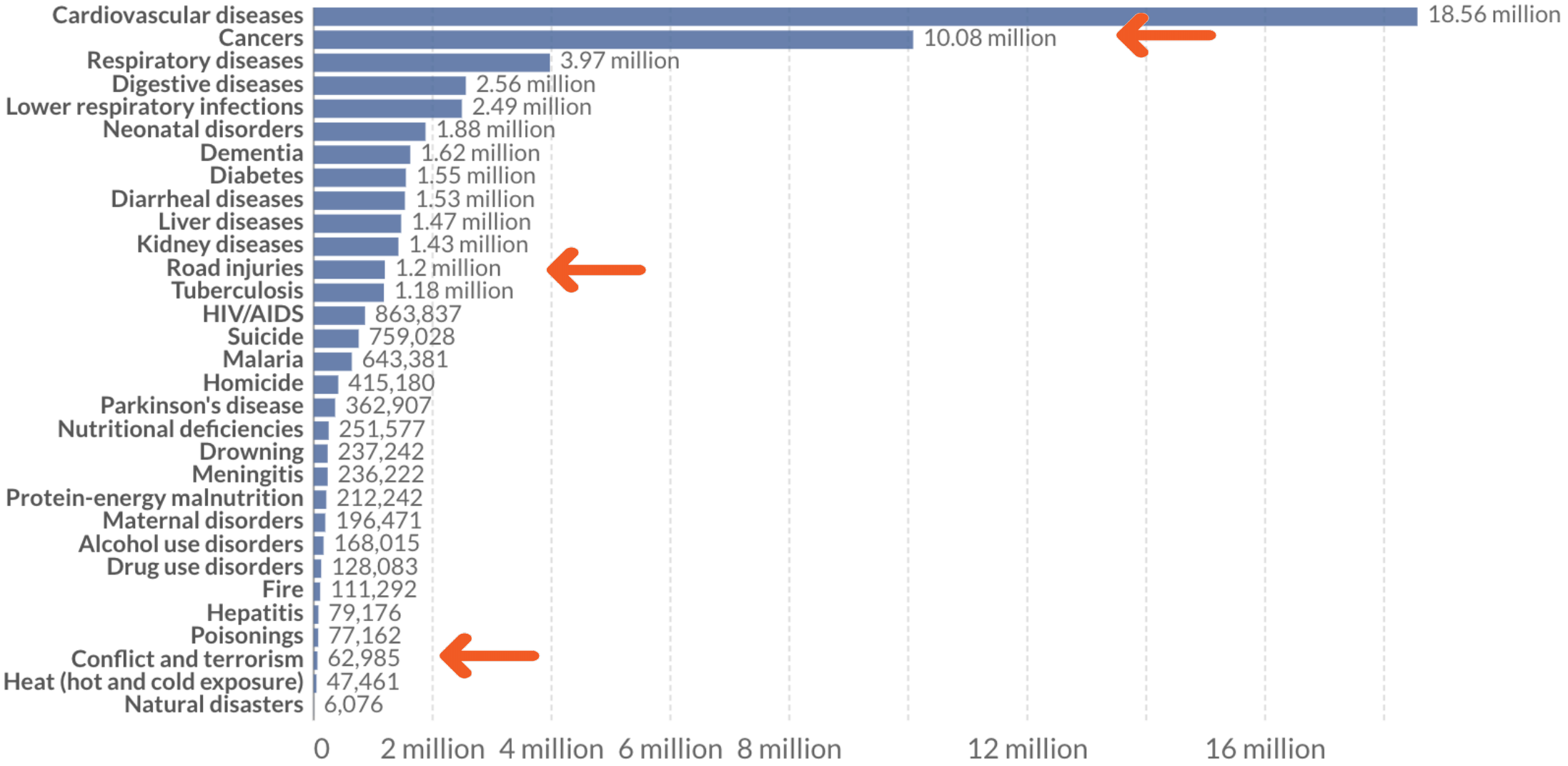




Number of deaths by cause, World, 2019

Our World in Data

Change country



Source: IHME, Global Burden of Disease (2019)

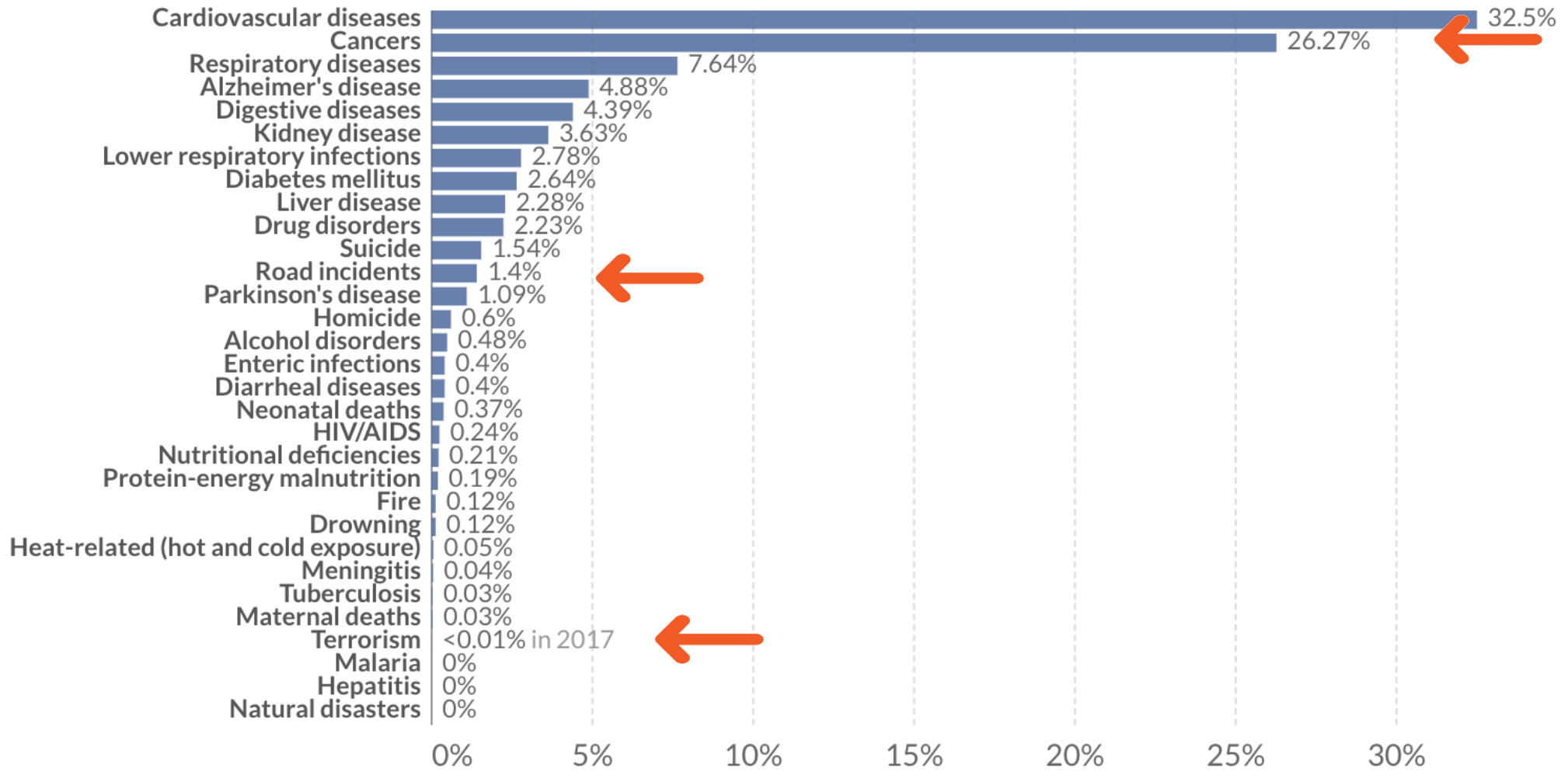
OurWorldInData.org/causes-of-death • CC BY

Share of deaths by cause, United States, 2019

Data refers to the specific cause of death, which is distinguished from risk factors for death, such as air pollution, diet and other lifestyle factors. This is shown by cause of death as the percentage of total deaths.

Our World in Data

Change country



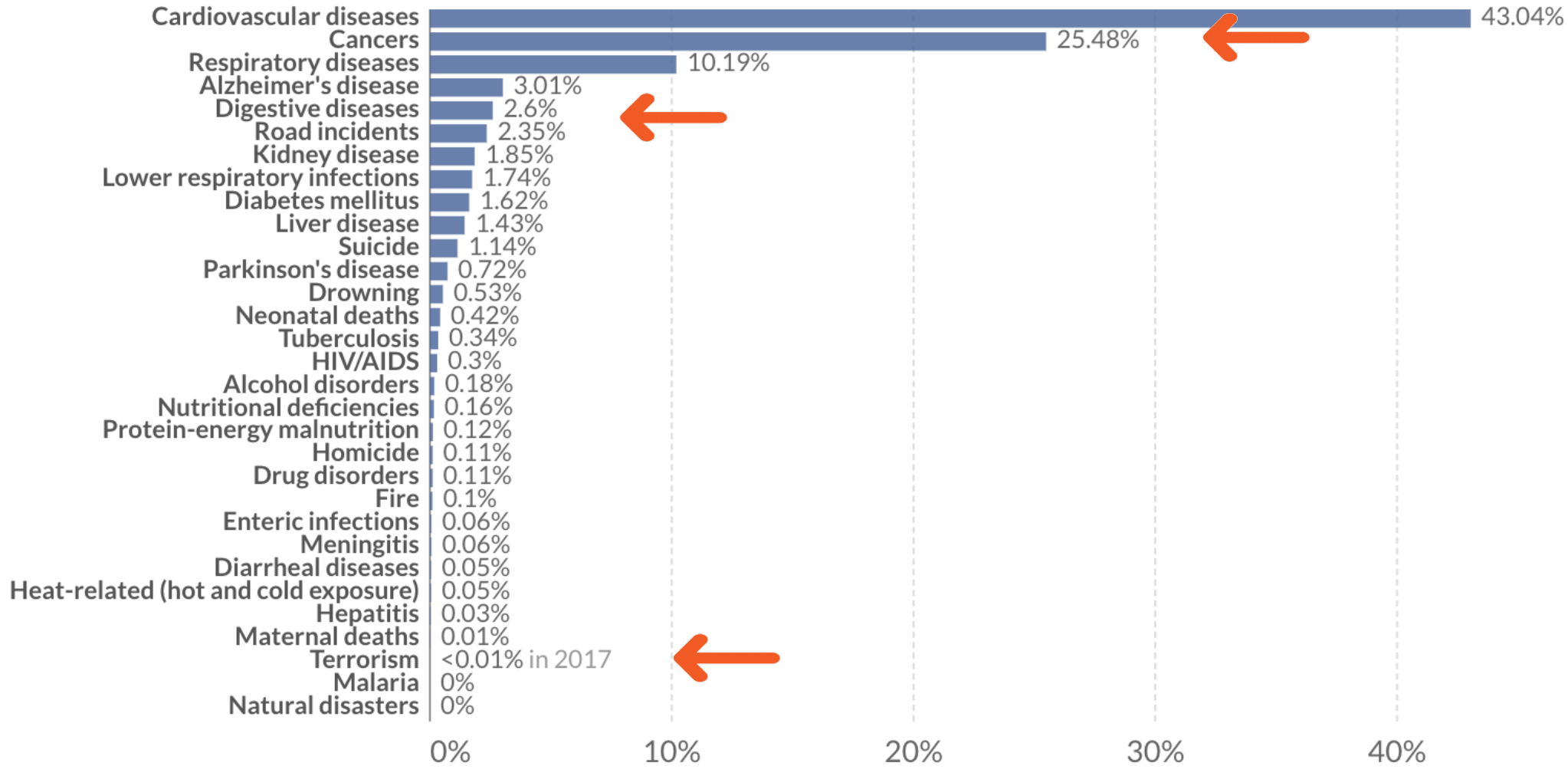
Source: IHME, Global Burden of Disease (2019)

Share of deaths by cause, China, 2019

Data refers to the specific cause of death, which is distinguished from risk factors for death, such as air pollution, diet and other lifestyle factors. This is shown by cause of death as the percentage of total deaths.

Our World in Data

Change country



Source: IHME, Global Burden of Disease (2019)

OurWorldInData.org/causes-of-death • CC BY

TECHNOLOGY

Self-Driving Cars Could Save 300,000 Lives Per Decade in America

Automation on the roads could be the great public-health achievement of the 21st century.

ADRIENNE LAFRANCE SEP 29, 2015

MORE STORIES

A Cultural History of the Airbag
ADRIENNE LAFRANCE

Jeff Bezos Wrote a Blog Post
ROBINSON MEYER

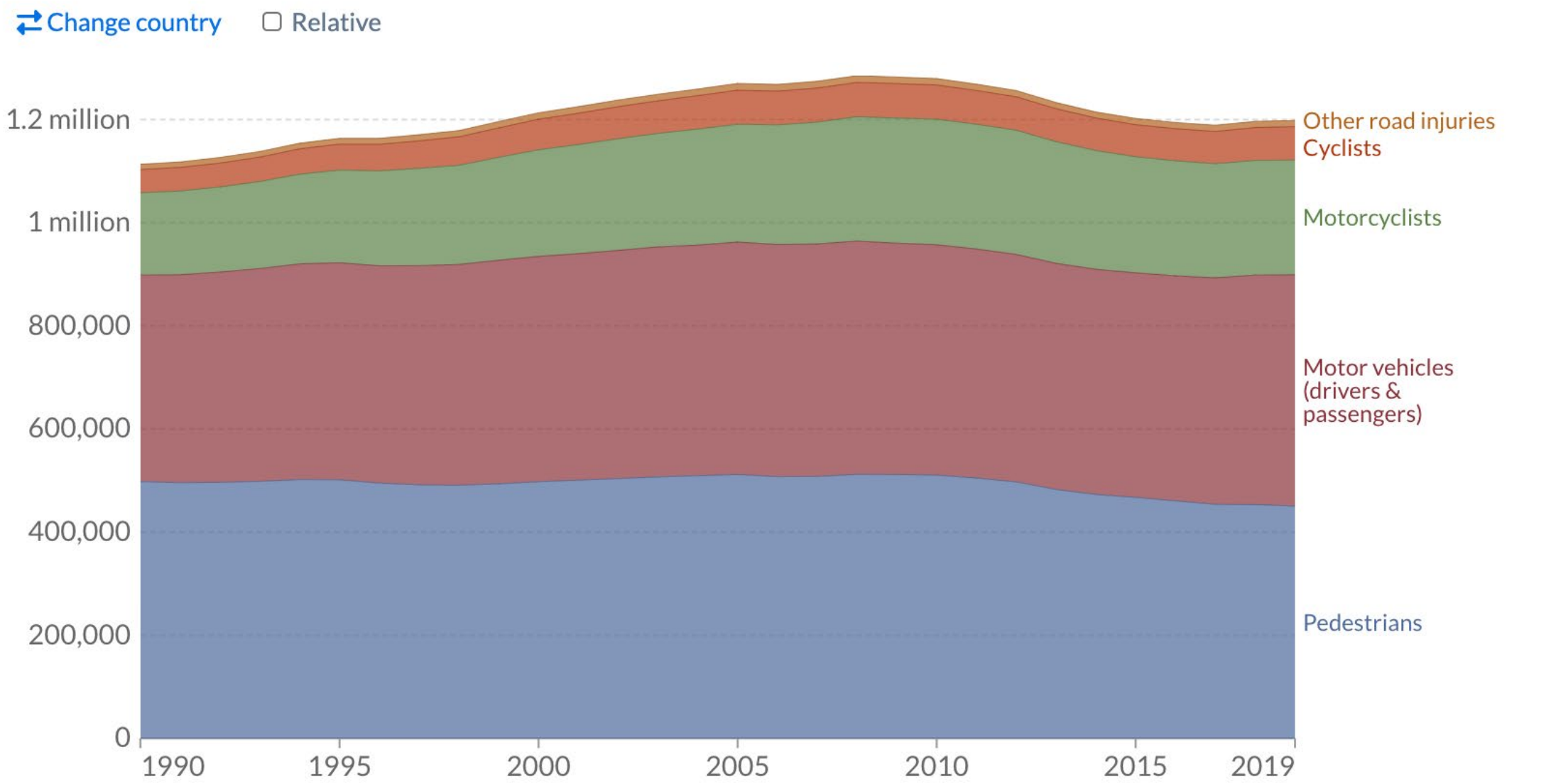
Jeff Bezos Brings the Receipts
ALANA SEMUELS

Image: Screenshot, https://www.theatlantic.com/technology/archive/2015/09/self-driving-cars-could-save-300000-lives-per-decade-in-america/407956/

Motor vehicle, motorcyclist, cyclist and pedestrian deaths, World, 1990 to 2019

Annual number of deaths from road accidents, differentiated by motor vehicle (drivers and passengers), motorcyclists, cyclists and pedestrians.

Our World in Data



Source: IHME, Global Burden of Disease (2019)

OurWorldInData.org/causes-of-death • CC BY

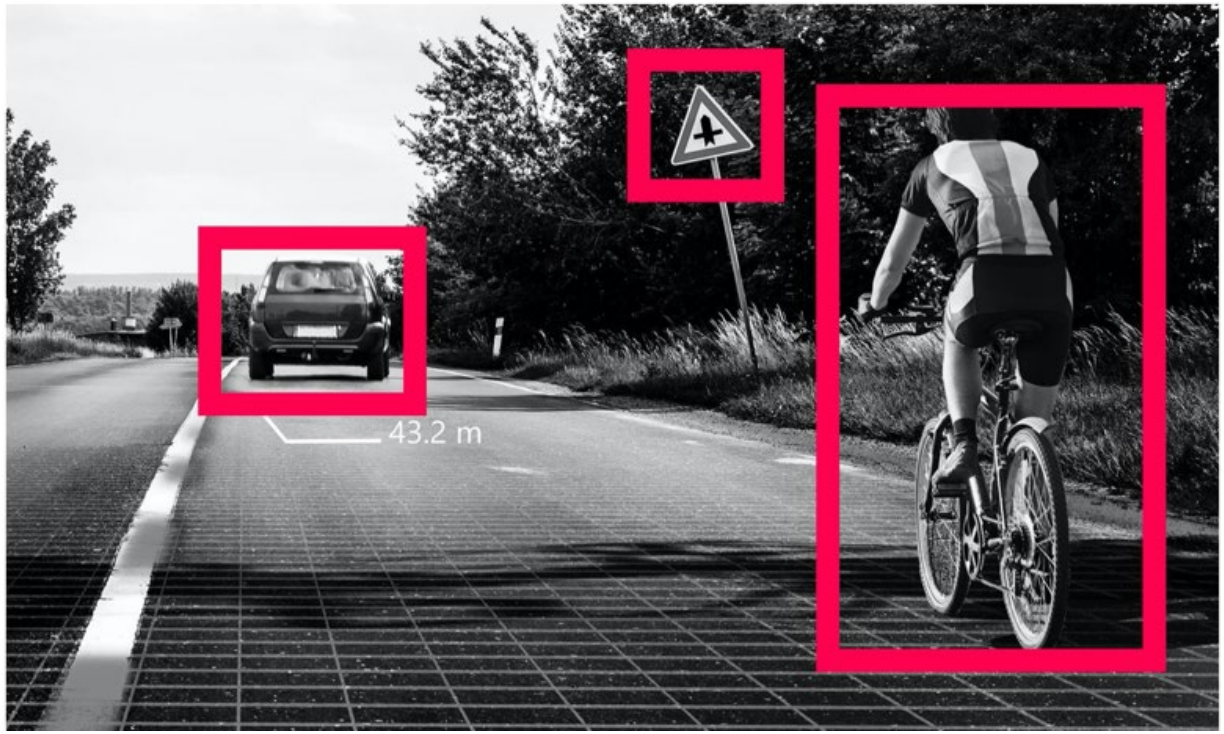
Reich image via Midjourney:
“photo of an autonomous vehicle driving dangerously close to a bicyclist in a bike lane on a busy road

When to Adopt Autonomous Systems?

The Cyclist Problem

Self-driving cars aren't good at detecting cyclists. The latest proposed fix is a cop-out.

By CHRISTINA BONNINGTON FEB 03, 2018 • 8:07 AM



Core Challenges to Utilitarianism

4. What about the **distribution of risk and harm?**

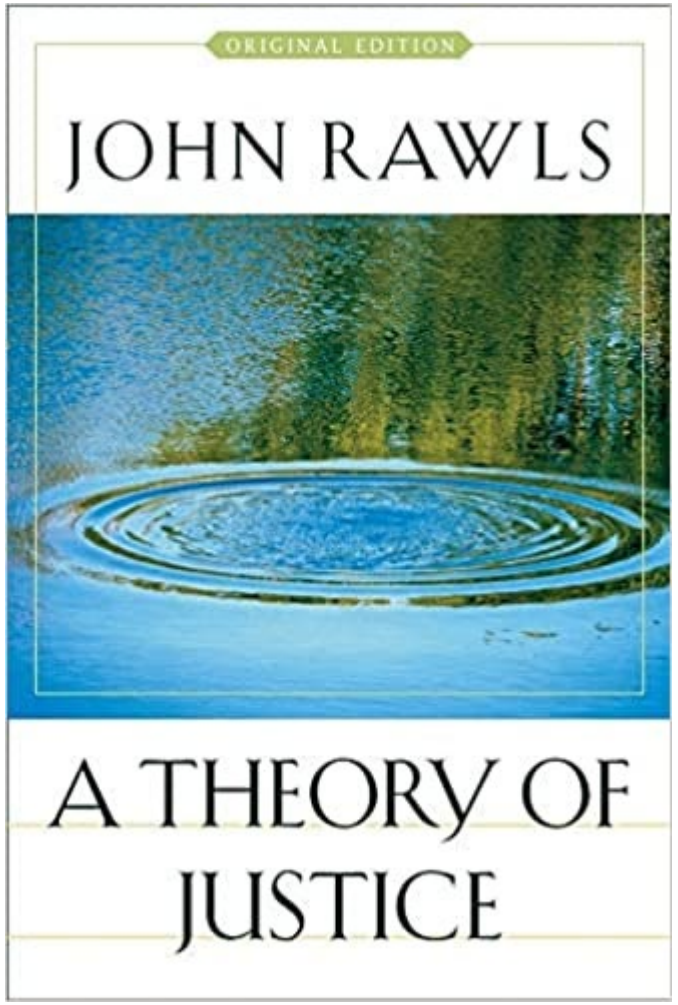
Is it permissible to impose increased risks/harms on some persons in order to deliver large gains for others?

Image: Wilfredo Rafael Rodriguez Hernandez, Public Domain (CC 0)

Core Challenges to Utilitarianism

Remember John Rawls?

A Theory of Justice provides an alternative to utilitarianism, which **"fails to take seriously the distinction between persons."**



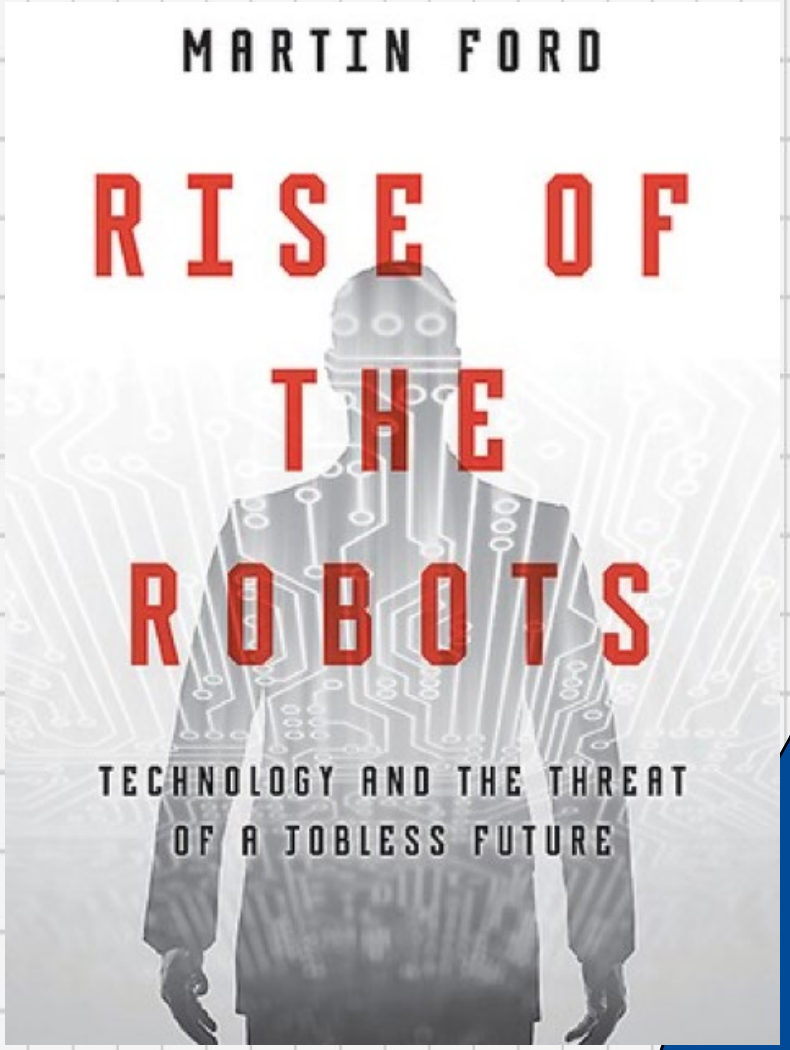


Image: Rise of the Robots, Martin Ford

Does automation displace labor?

Martin Ford:
Technological displacement of labor comes for:

- Manual labor (e.g., truck drivers)
- White collar jobs (e.g., radiologists)
- Generative AI displaces creative work (maybe?)

AI portends a future with big advances in productivity but massive unemployment of human beings.
What policy response is necessary?

But also:

Automation transforms the Workplace

airbnb
Etsy
TaskRabbit
GETAROUND
lyft
UBER
POSTMATES
fiverr®
DOORDASH

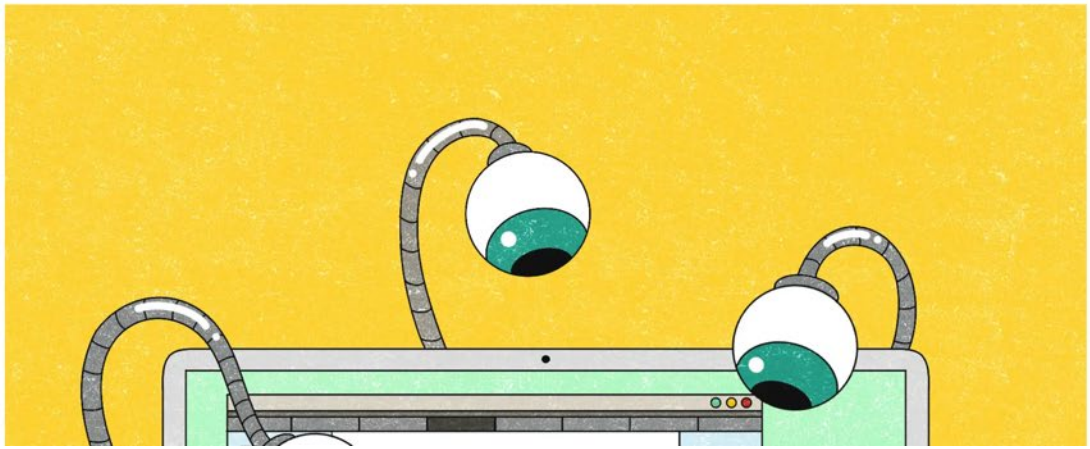
E.g., Bossware, the Gig Economy, Task Automation

Image: Screenshot, https://www.eff.org/deeplinks/2020/06/inside-invasive-secretive-bossware-trackingworkers;
Sandra Luksic,Dall-E, Promot "Gig Economy Logos"

The New York Times

Are 'Bossware' Tools Tracking You?

In recent years, the technologies used to surveil workers have become more sophisticated and widespread.



Autonomous Weapons

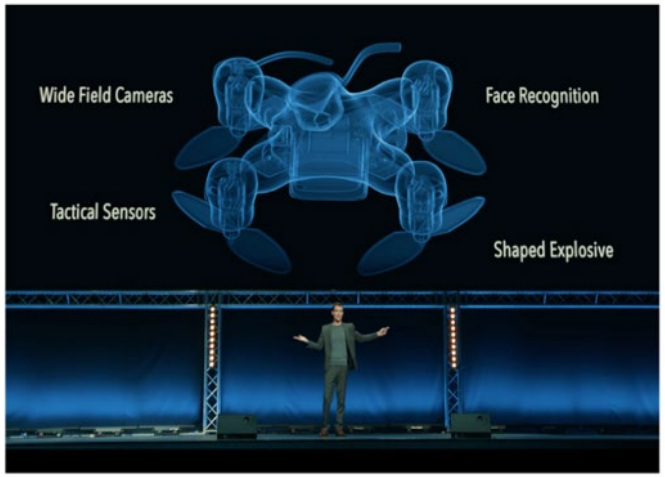


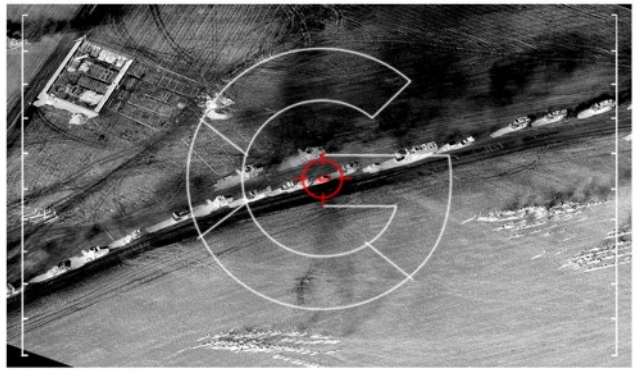
Image: Slaughterbots/YouTube

A scene from "Slaughterbots," a film that depicts a dystopian future in which autonomous lethal drones fall into the hands of terrorists.

PRIVACY AND SECURITY

Google Is Helping the Pentagon Build AI for Drones

Kate Conger and Dell Cameron
3/06/18 10:15am • Filed to: GOOGLE
100.2K 69 3



Graphic: Jim Cooke, Photo: Getty

The Slaughterbot debate: Stuart Russell's 2018 video.

AI portends a future where autonomous weapons will re-write the rules of war, empower non-state actors, and possibly trigger a new arms race in AI weaponry.

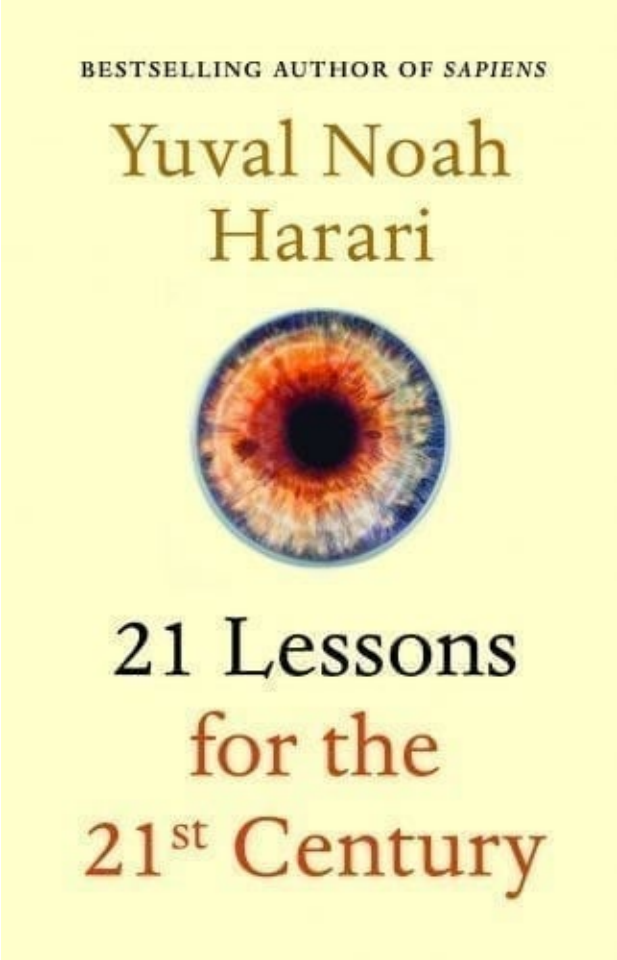
QUESTION: Should tech companies refuse to work on autonomous weapons or sell AI technology to the military? (e.g., Project Maven at Google)

QUESTION: How can public policy shape the future of autonomous weapons?

Does AI Favor Tyranny?

- Decentralized political systems > centralized planning.
- Historically, democracy outperforms authoritarianism
- But AI might favor centralized systems and bring about digital dictatorships.

Does AI Favor Tyranny?



If you disregard all privacy concerns and concentrate all the information relating to a billion people in one database, you can train much better algorithms than if you respect individual privacy and have in your database only partial information on a million people

> Harari, p. 66

Generative AI: Promise and Perils

1. Human Intelligence, AI, and AGI
2. GPTs are GPTs
3. Beyond Utopian and Dystopian
4. Topics for the AI Transition
   - Autonomous vehicles
   - AI and work
   - AI and the military (surveillance, autonomous weapons)
   - AI and democracy
5. Going Deeper in Philosophy
   - Nozick and the Experience Machine: are virtual experiences different?
   - Scanlon and the World Cup: problems with aggregation of harms
   - Foot and the Trolley Problem: programming autonomous vehicles
   - Thomsen and the Fat Man

Nozick's Experience Machine

> "We care about things in addition to how our lives feel to us from the inside. This is shown by the following thought experiment. Imagine a machine that could give you any experience (or sequence of experiences) you might desire. When connected to this experience machine, you can have the experience of writing a great poem or bringing about world peace or loving someone and being loved in return. You can experience the felt pleasures of these things, how they feel "from the inside." You can program your experiences for tomorrow, or this week, or this year, or even for the rest of your life. If your imagination is impoverished, you can use the library of suggestions extracted from biographies and enhanced by novelists and psychologists."

Nozick's Experience Machine

You can live your fondest dreams "from the inside." Would you choose to do this for the rest of your life? If not, why not? (Other people also have the same option of using these machines which, let us suppose, are provided by friendly and trustworthy beings from another galaxy, so you need not refuse connecting in order to help others.) The question is not whether to try the machine temporarily, but whether to enter it for the rest of your life. Upon entering, you will not remember having done this; so no pleasures will get ruined by realizing they are machine produced. Uncertainty too might be programmed by using the machine's optional random device (upon which various preselected alternatives can depend). The question of whether to plug in to this experience machine is a question of value.

A Philosophical Question:

If you could hook yourself up to an experience machine – call it Oculus – would you?

If machines would unfailingly improve upon our subjective experience, what, if anything, is wrong with preferring the experience machine?

Meta Who We Are What We Build Our Actions Our Community Resources

Meta

The metaverse is the next evolution of social connection. Our company's vision is to help bring the metaverse to life, so we are changing our name to reflect our commitment to this future.

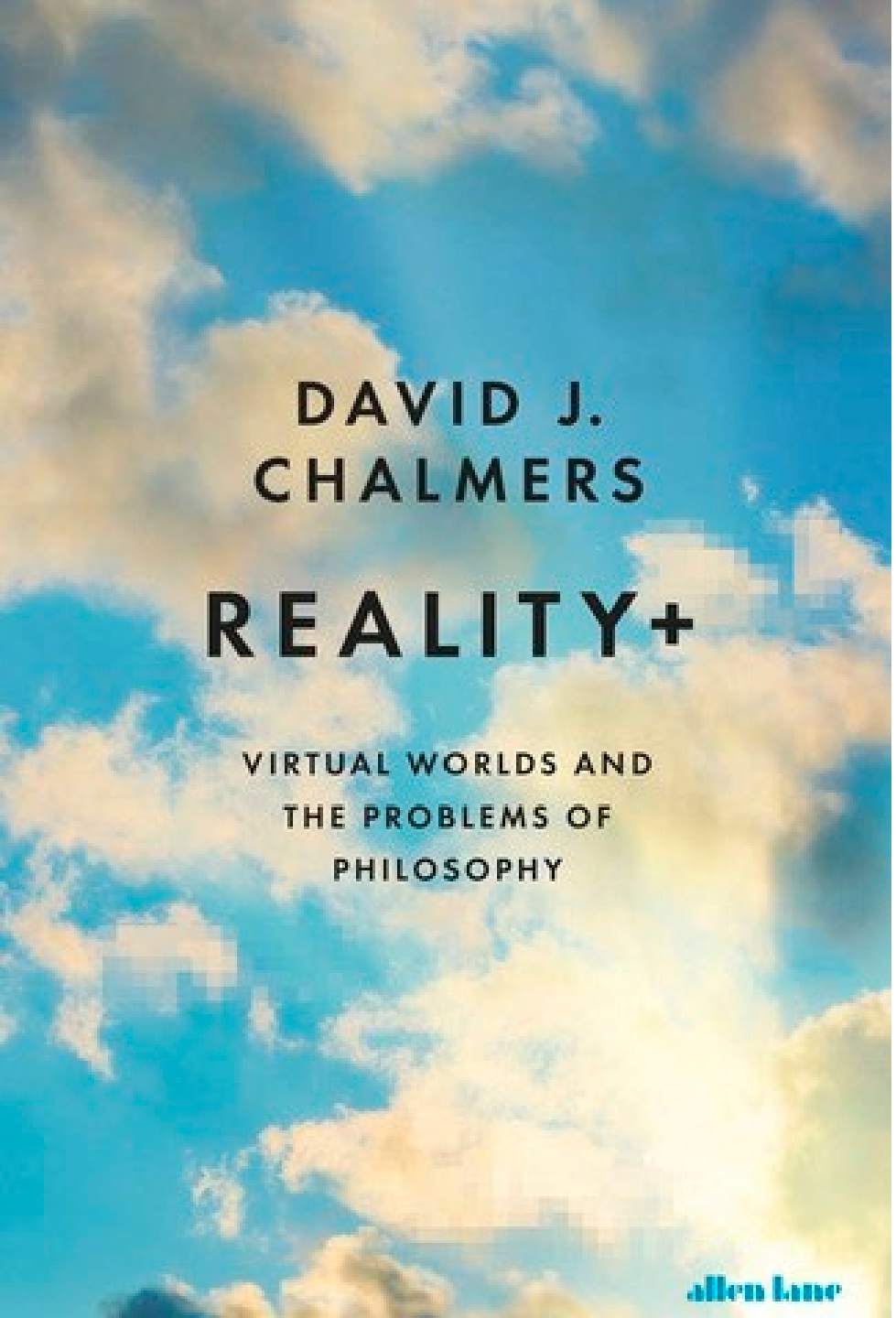
The metaverse will be social

3D spaces in the metaverse will let you socialize, learn, collaborate and play in ways that go beyond what we can imagine. Listen to Mark Zuckerberg share our vision for bringing the metaverse to life together.

Source: Screenshot, https://about.meta.com/metaverse/

VR Moral Imperative?

Palmer Luckey, the co-founder of Oculus Rift, had in mind a set of gaming experiences in VR, but in interviews he also expressed a far grander aspiration. Luckey spoke of a “moral imperative” to bring VR to the masses so that they too, and not only the wealthy or geographically privileged, could experience the good things in life like a sunset on the Aegean Sea, the Mona Lisa in the Louvre, the Great Migration on the Serengeti, or a Bruce Springsteen concert in New Jersey.



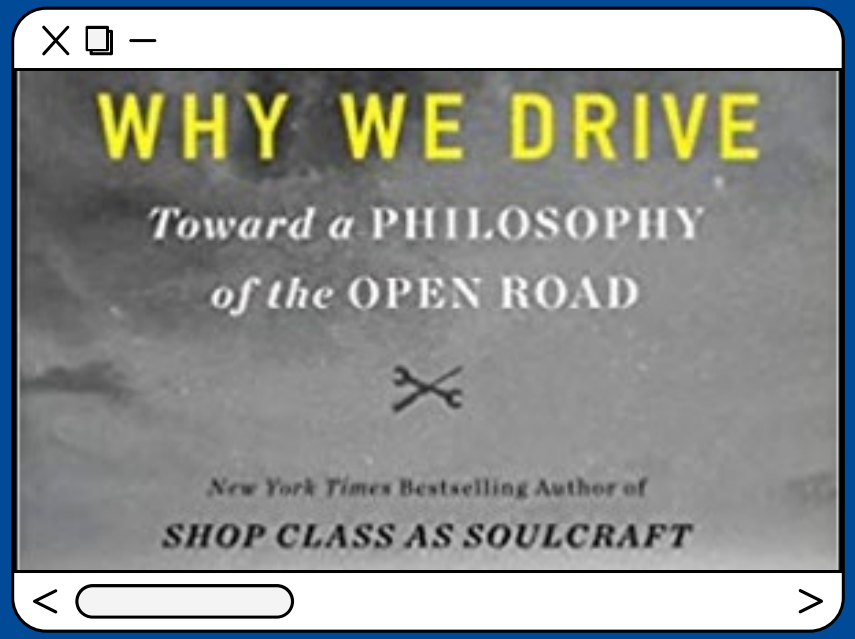
David Chalmers on virtual experience

"In principle, life in virtual reality can have the same sort of value as life in nonvirtual reality. To be sure, life in virtual reality can be good or bad, just as life in physical reality can. But if it's bad, it won't be bad simply because it's virtual" (p. 312 of Reality +)

Fahrvergnügen

German: the pleasure of driving

Autonomous Systems and Human Agency



When should we preserve human agency, even if “outcomes” are better in some respect when using autonomous systems?

Increasing presence of autonomous systems in everyday life might:

Diminish human agency, corrode human freedom, eliminate responsibility

Source: Why We Drive by Matthew B Crawford; Screenshot, https://www.theatlantic.com/technology/archive/2018/03/can-you-sue-a-robocar/556007/

Morality of Autonomous Vehicles

How should an autonomous vehicle be programmed?

Two questions:

1. Should it aim to maximize driver welfare or global welfare?
2. What moral code should it follow when confronted with terrible dilemmas?

Trolley Problem come to life!

Morality of Autonomous Vehicles

How should an autonomous vehicle be programmed?

Two questions:
1.Should it aim to maximize driver welfare or global welfare?
2.What moral code should it follow when confronted with terrible dilemmas?

Trolley Problem come to life!

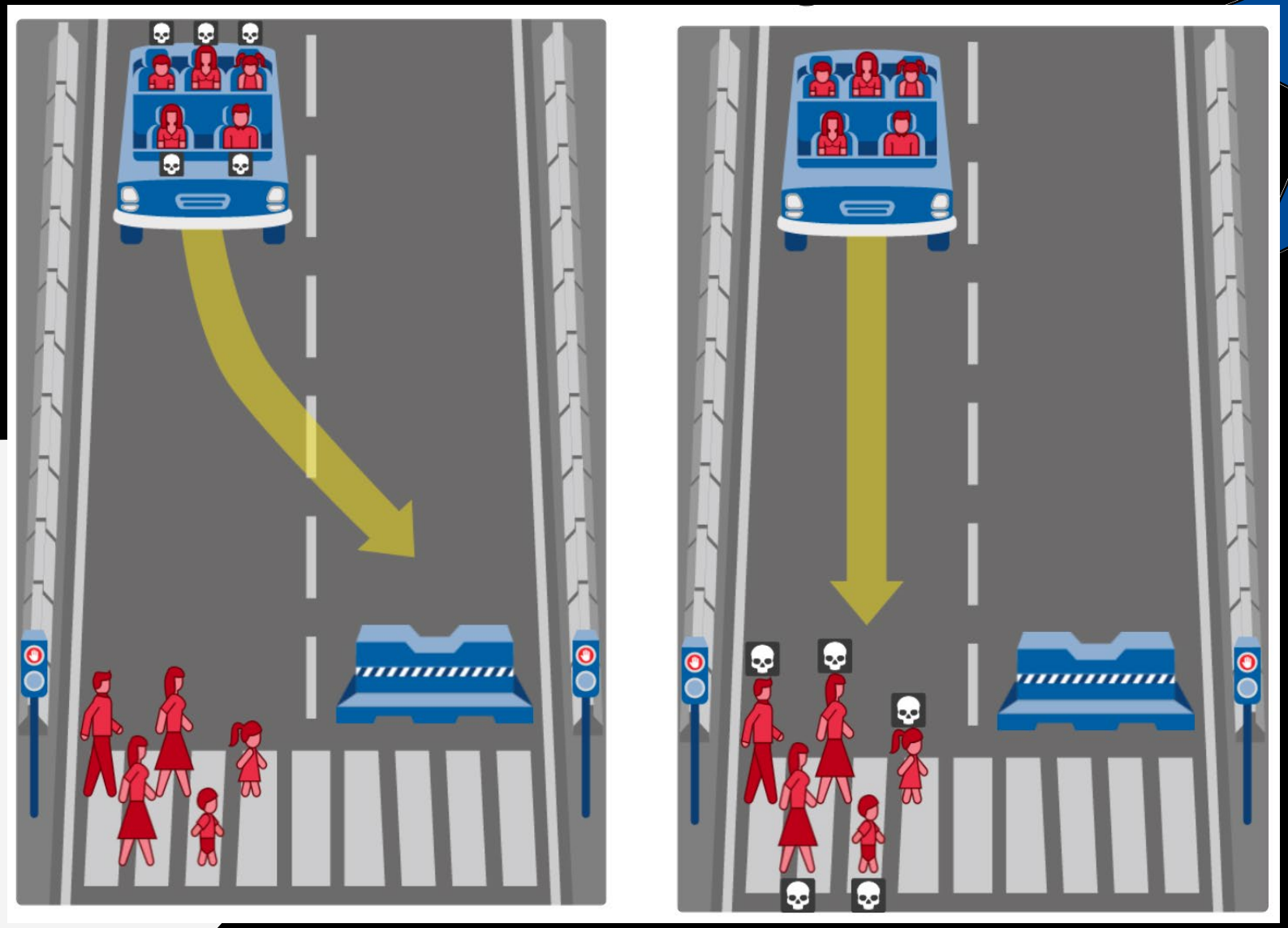


Image: Wikipedia Commons, Morality Machine CC BY-SA 4.0

Phillippa Foot (1967)

Judith Jarvis Thomson (1985)

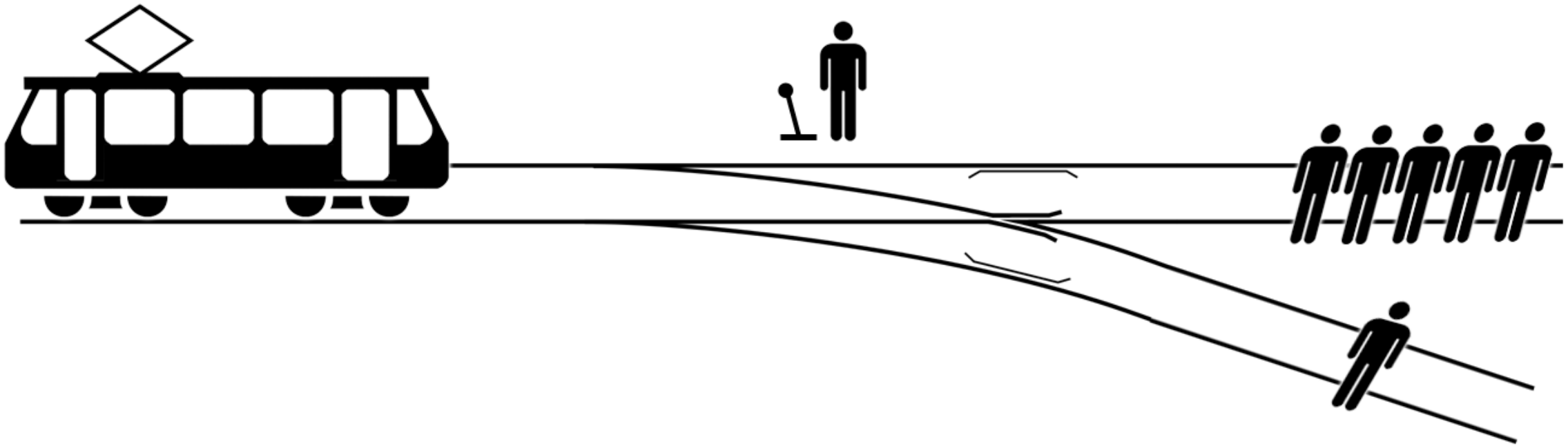


Image: Wikipedia Commons, CC BY-SA 4.0

A Two Year Old's Solution to the Trolley Problem

Source: https://www.youtube.com/watch?v=ZULuedsaB9U

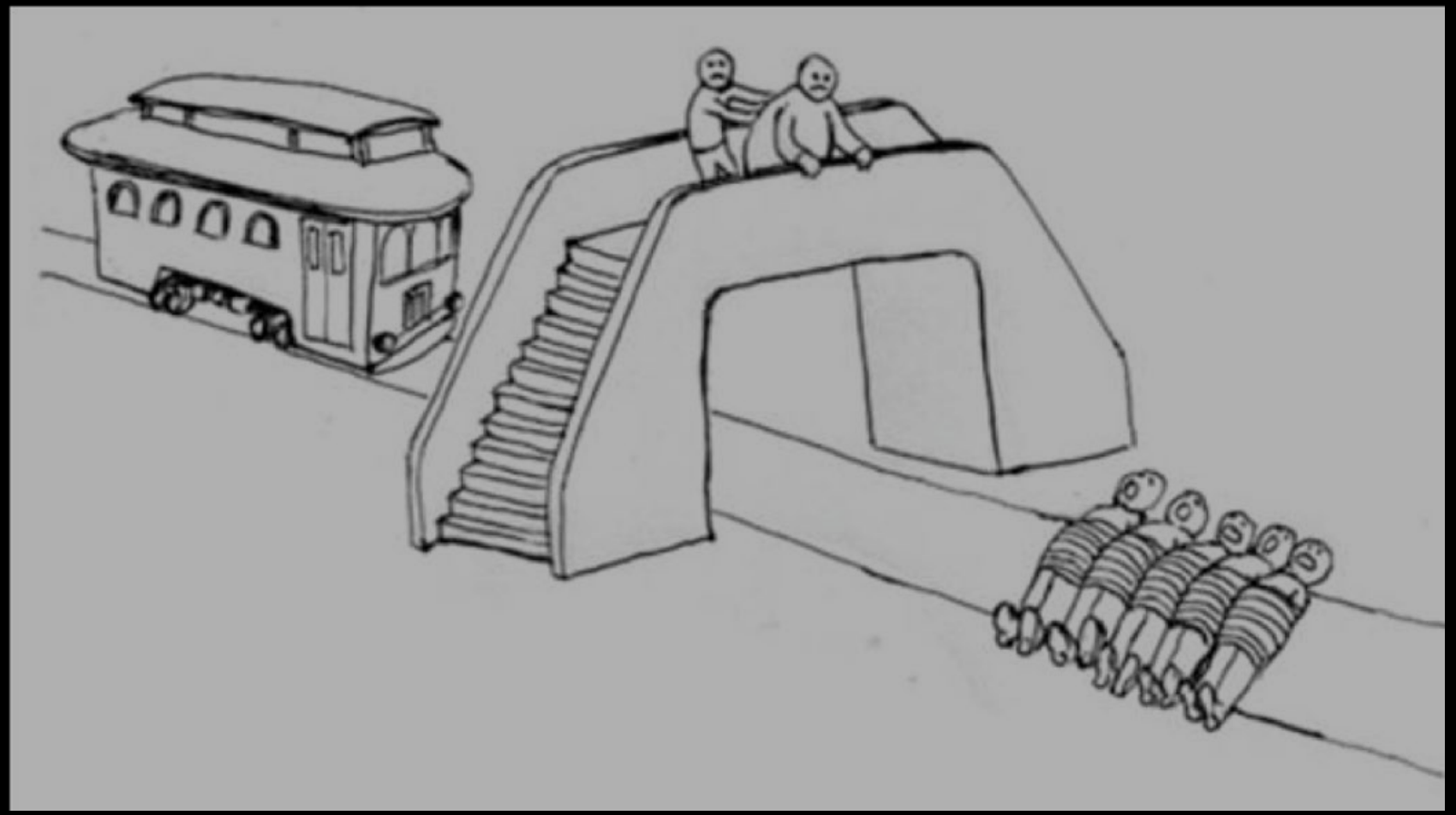
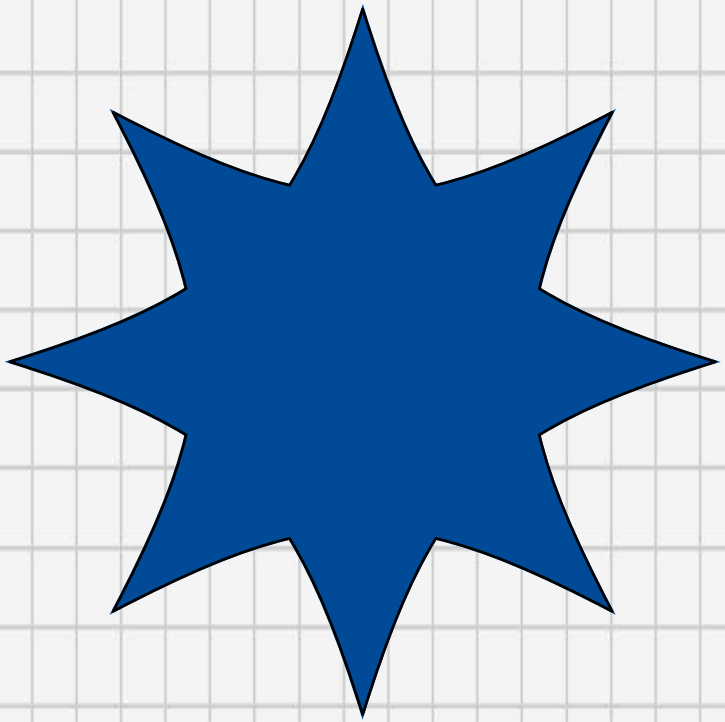


Image: Know Your Meme??? COuld not find image



Should you kill one in order to save five?

If utilitarianism is correct – right action is the greatest good for the greatest number:

Then what accounts for the moral intuition in the push the fat man?

Possible conclusion: Sometimes rights act as hard constraints on aggregating benefits.

The World Cup Dilemma

Problems with Aggregation, Reprised

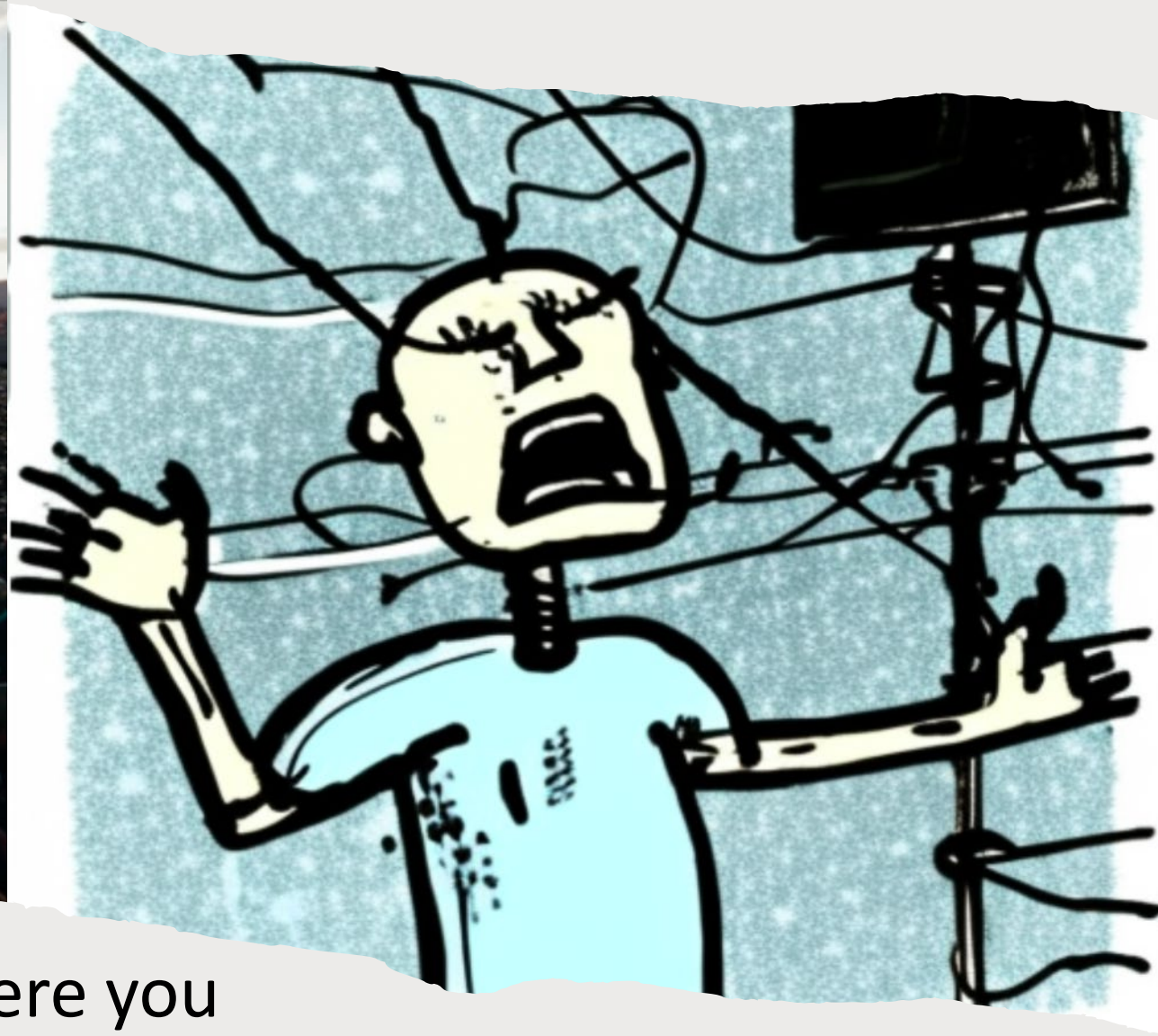
Suppose that Jones has suffered an accident in the transmitter room of a television station. Electrical equipment has fallen on his arm, and we cannot rescue him without turning off the transmitter for fifteen minutes. A World Cup match is in progress, watched by many people, and it will not be over for an hour. Jones's injury will not get any worse if we wait, but his hand has been mashed and he is receiving extremely painful electrical shocks. Should we rescue him now or wait until the match is over? Does the right thing to do depend on how many people are watching—whether it is one million or five million or a hundred million? It seems to me that we should not wait, no matter how many viewers there are, and I believe that contractualism can account for this judgment while still allowing aggregative principles of the kind defended above.

Source: T.M. Scanlon (1998), What We Owe to Each Other, Harvard University Press (Cambridge, MA)

Rob Reich: image generated via Midjourney, prompt: “a photo of millions of fervent soccer fans assembled outdoors watching the FIFA World Cup final on an enormous

Jones

Rob Reich image via Midjourney prompt:
stick figure image of a worker in satellite broadcast station
who is suffering an electric shock from wires that have fallen



Is there a number of viewers where you leave Jones until the end of the World Cup Final?