Existential/Catastrophic Risks from AI

Law 4052 // Computer Science 283 // Communication 252A // Political Science 145B/445B
Fall Quarter 2026



Rob Reich
Anka Reuel
Nate Persily

Existential/Catastrophic Risks from AI

- What Do We Owe the Future?
  - AI as an existential threat
- Historical Precedents
- How might AI cause catastrophe?
  - Misuse Risks (cyber + CBRN: chemical, bio, radiological, nuclear)
  - Structural Risks (loss of control, cyber and critical infrastructure)
  - Automated AI R&D (the accelerant)
- Governing Existential Risks
  - A focus on the precautionary principle

Existential/Catastrophic Risks from AI

- What Do We Owe the Future?
  - AI as an existential threat
- Historical Precedents
- How might AI cause catastrophe?
  - Misuse Risks (cyber + CBRN: chemical, bio, radiological, nuclear)
  - Structural Risks (loss of control, cyber and critical infrastructure)
  - Automated AI R&D (the accelerant)
- Governing Existential Risks
  - A focus on the precautionary principle

The Ethics of Valuing Future Generations

Intergenerational Justice: what do we owe the future?

Why care about the possibility of extinction after we're dead?

Four hard questions

How much weight?

Should a future life count as much as a present one?

How far out?

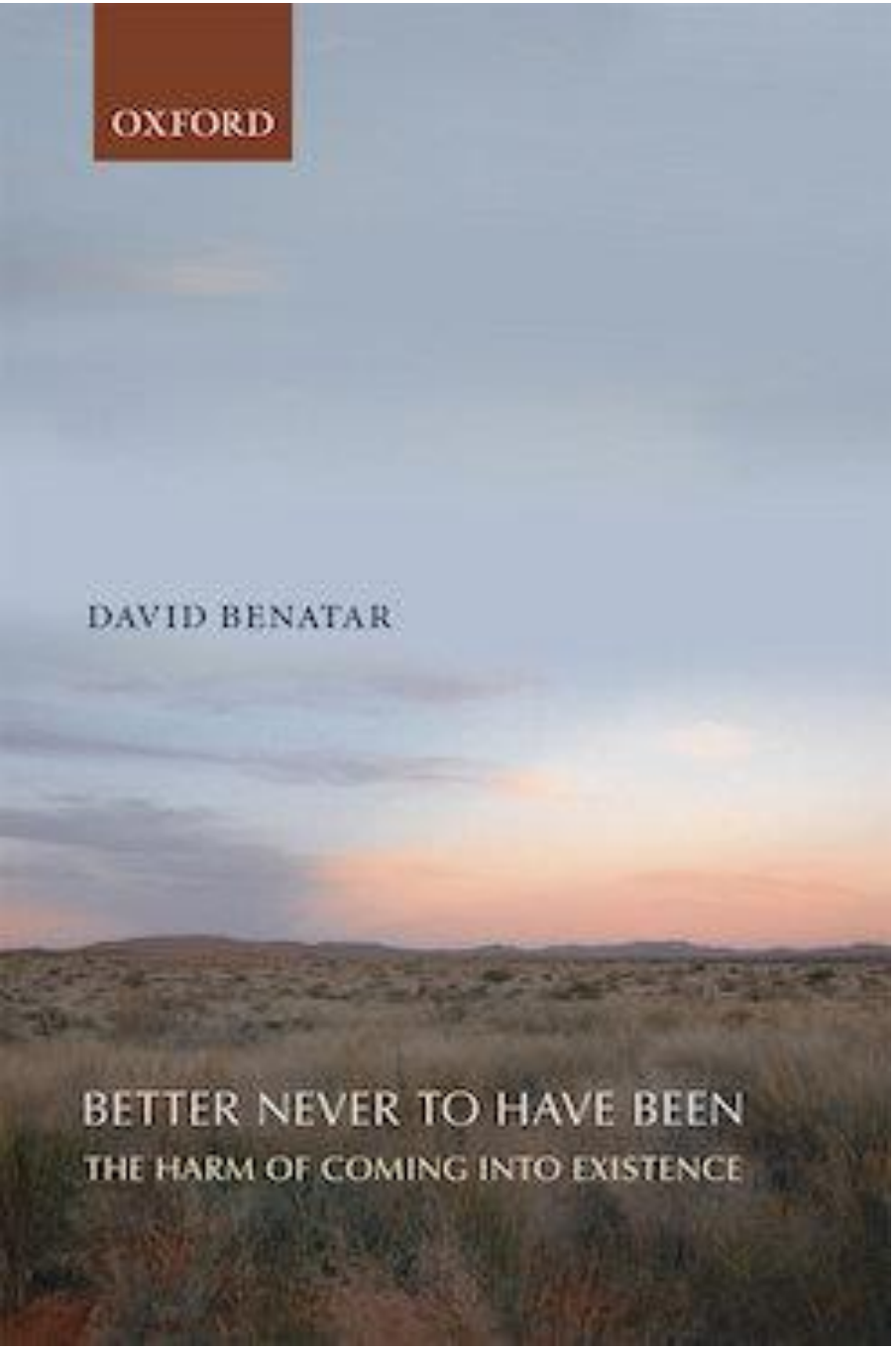
A century, a million years, until the Sun dies?

Who counts?

Only humans who will exist? Animals? Digital minds?

How sure must we be?

Can tiny, unmeasurable probabilities actually guide action?





Better Never to Have Been

The Harm of Coming into Existence

David Benatar

Table of Contents

1. Introduction
2. Why coming into existence is always a harm
3. How bad is coming into existence?
4. Having Children: The Anti-Natal View
5. Abortion: The 'Pro-Death' View
6. Population and Extinction
7. Conclusion

> To my parents,
>
> even though they brought me into existence;
>
> and to my brothers,
>
> each of whose existence, although a harm to him,
>
> is a great benefit to the rest of us.

Is extinction uniquely bad?

Derek Parfit, Reasons and Persons (1984): compare three outcomes

Peace **99% die** **100% die**

A terrible loss

Far worse: the whole future is lost

Samuel Scheffler, Death and the Afterlife (2013)

Our values quietly depend on a “collective afterlife”: if no one came after us, much of what we do would lose its point. It is hard to conceive of living any life without regard to future lives.

Expected value in one slide

Expected value

Probability × value

A 10% chance of winning $100 is worth $10

Aggregation

Add up across everyone

Ten people gaining $10 each = $100 of value

Consequentialism

Best total outcome wins

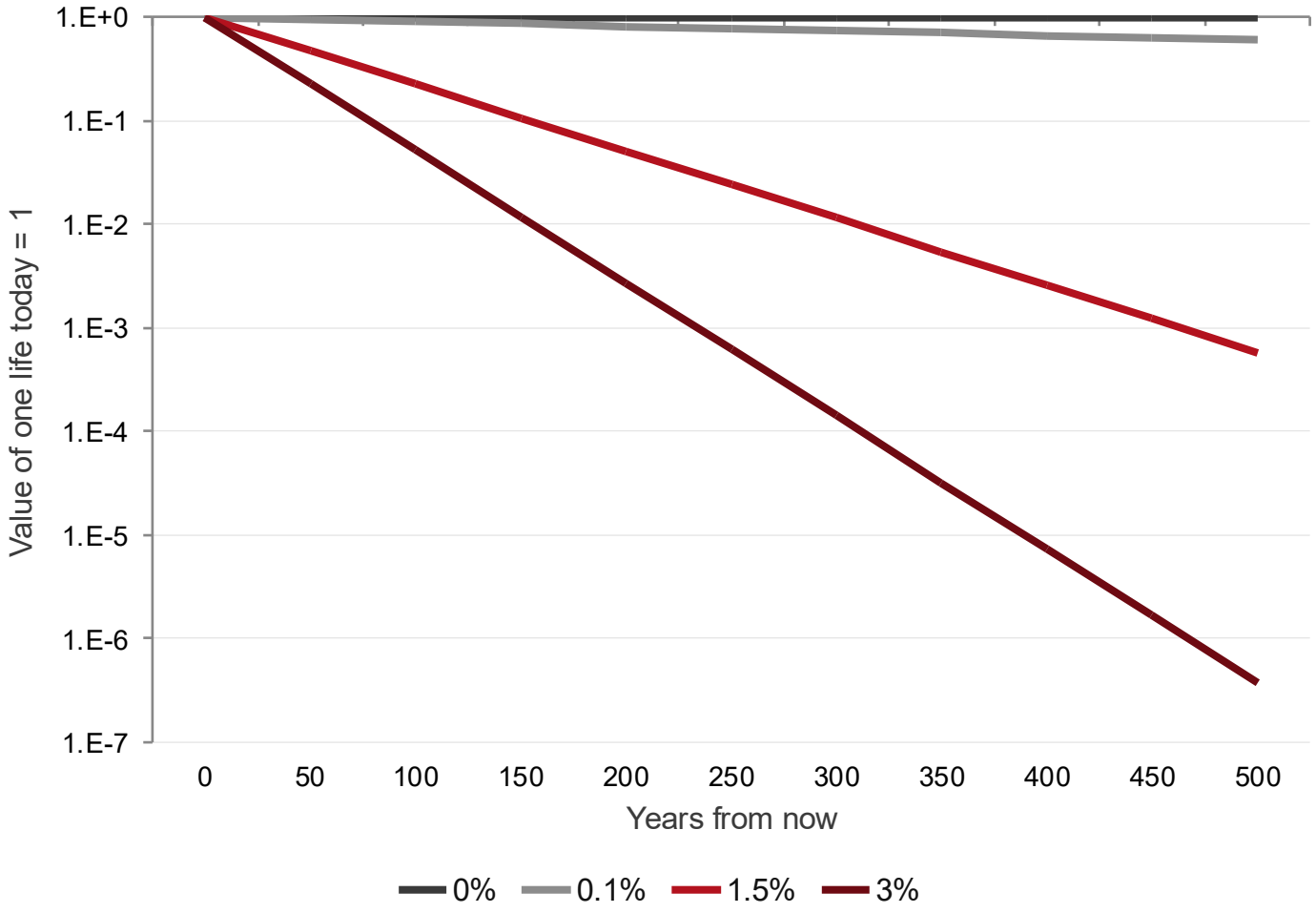
Everyone counts equally, wherever (and whenever?)

How to spend $1 billion (govt or charitable $)?

Option A: malaria bed nets, saving about 200,000 lives today

Option B: AI safety research that might cut the risk of human extinction by one in a million

Economic discounting of the future



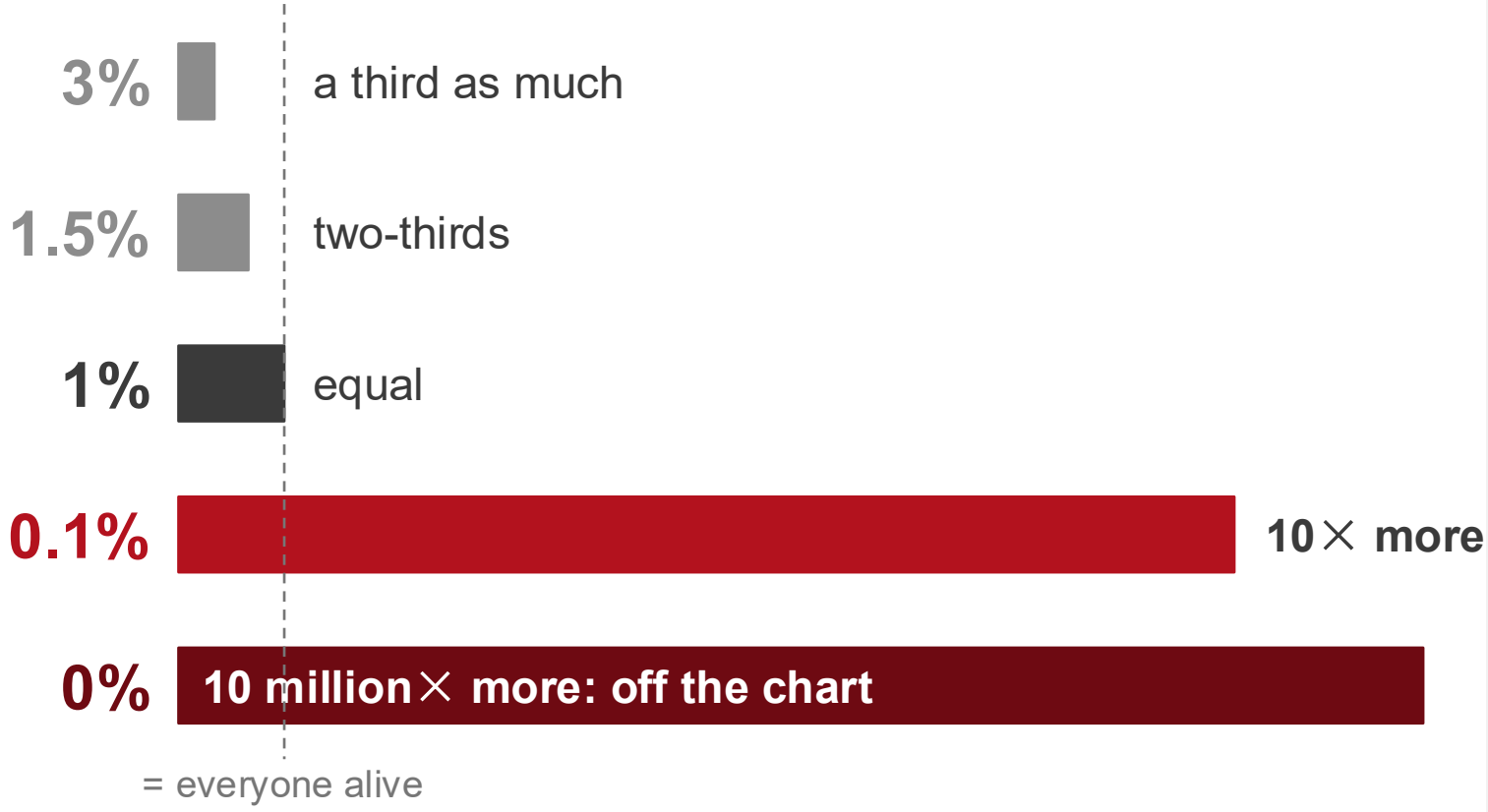
At 3%, a life 500 years from now is worth

1 / 2.6 million

of a life today

Everything hinges on the rate

Value of all future generations, compared with the 8 billion alive

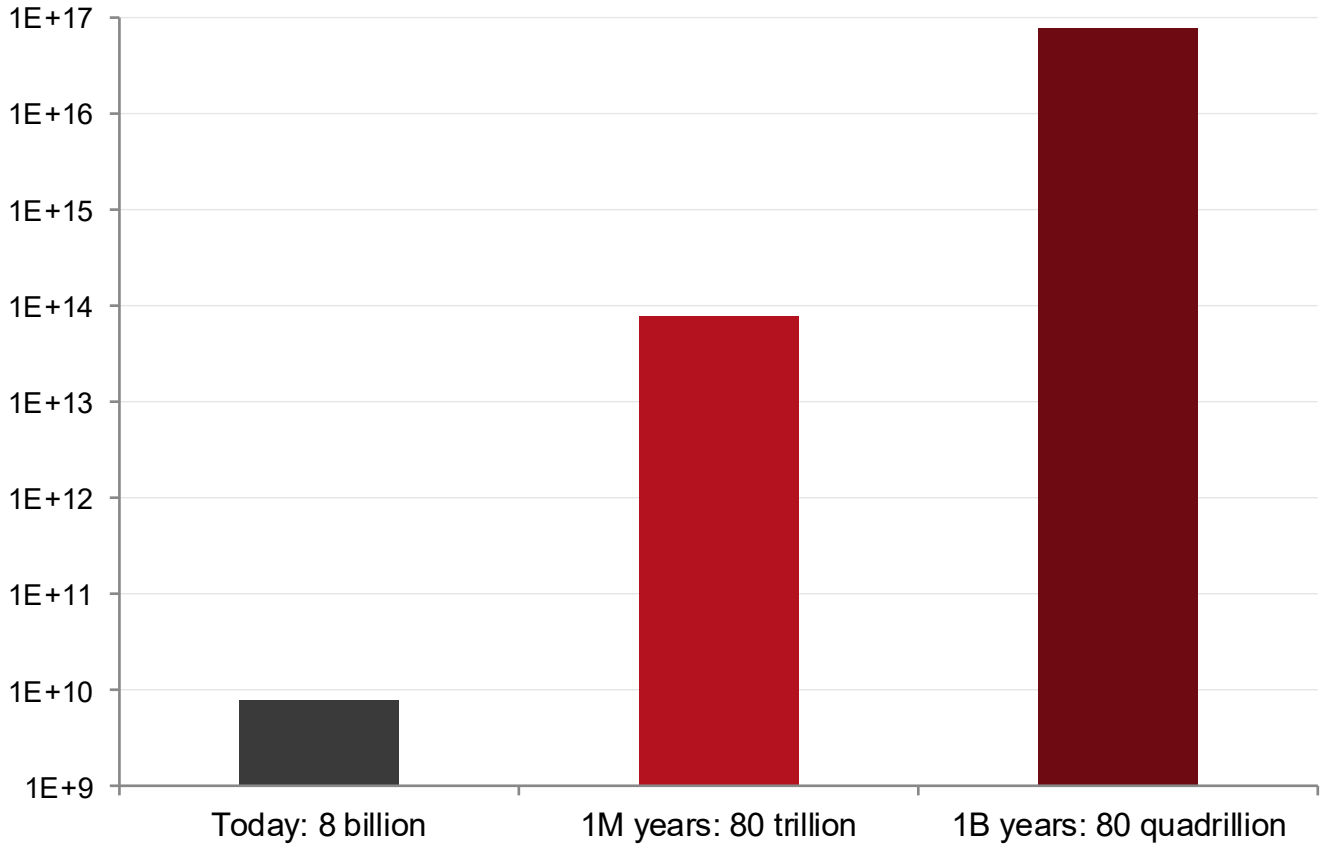


The ethical fork

Conventional rates give the future modest weight; near zero, it overwhelms

The rate is an ethical choice, not a technical one

The swamping math



Assumes 8 billion people, 100-year lives, no discounting

Cut extinction risk by 1 in a million

× 80 quadrillion future lives

= 80 billion

expected lives saved: ten times everyone alive

Option B beats bed nets 400,000 to 1

A rounding error: the reductio

Measured against the future, people living in poverty today are a rounding error

Nick Bostrom, in remarks

Premise

Every life counts equally, forever

Premise

Maximize expected total value

Conclusion

Today's poorest people barely matter

Reductio

If that is absurd, a premise is false

Bostrom in print (2002): non-existential disasters are "mere ripples on the surface of the great sea of life"

DIAGNOSIS 1

Pascal’s mugging

1

The offer

An unarmed stranger asks for your wallet and promises to repay you a trillion-fold tomorrow

2

The math

Even a one-in-a-billion chance he is honest makes handing it over a good bet

3

The problem

Name a big enough number, and expected value says yes to anything

Back to Option B: is “one in a million” a measured fact, or a guess dressed up as a number?

DIAGNOSIS 2

Putting pressure on aggregation

Scanlon's transmitter room (1998)

A technician is trapped under equipment during a World Cup broadcast, suffering painful shocks. Rescuing him means cutting the broadcast to millions for an hour. Add up enough mild disappointment, and aggregation says: leave him.

Rawls (1971)

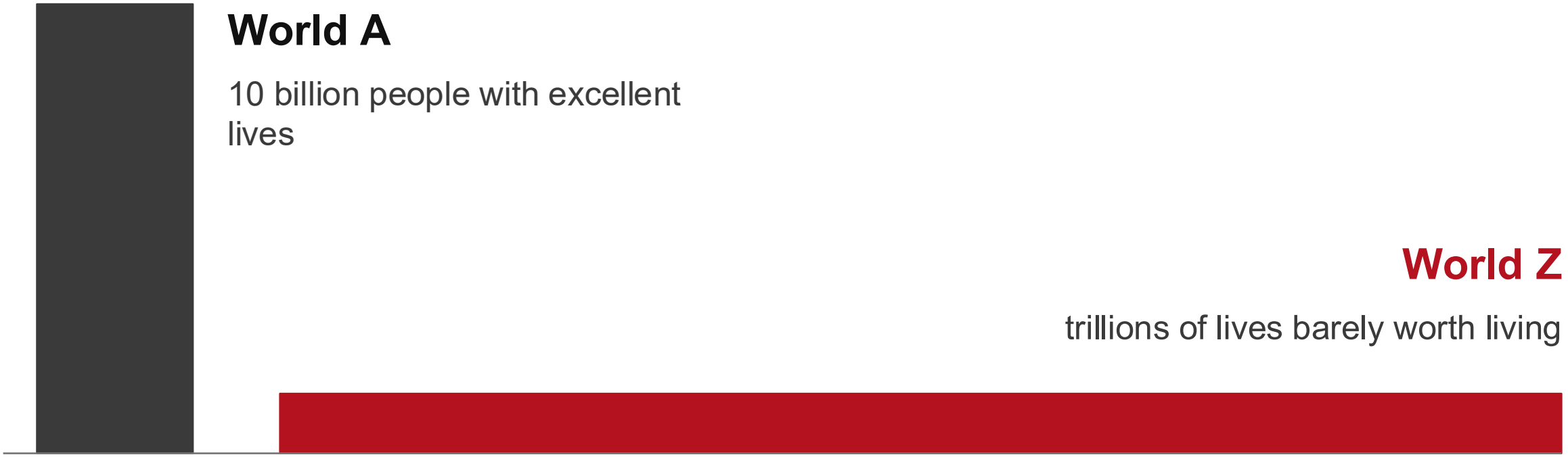
Utilitarianism "does not take seriously the distinction between persons"

It treats society like one big person, trading one part's pain for another's gain

The parallel: the 200,000 malaria victims are the technician; future quadrillions are the viewers

DIAGNOSIS 3

The Repugnant Conclusion



Total value = width × height. Make Z wide enough and it wins, however thin its lives.

Parfit, who discovered it, called this conclusion “repugnant”

DIAGNOSIS 4

The view from nowhere

Impersonal view

Sidgwick (1874)

“the point of view of the universe”

Nagel (1986)

“the view from nowhere”

The idea

Value counts the same, whoever or whatever has it

Successionism

If

superintelligent AIs have moral value

Then

human extinction is no loss, so long as they flourish

Sutton

“prepare for, but not fear, the inevitable succession”

The reply

Williams (2006)

There is no point of view of the universe

So

favoring humanity is not a prejudice like racism

Scheffler

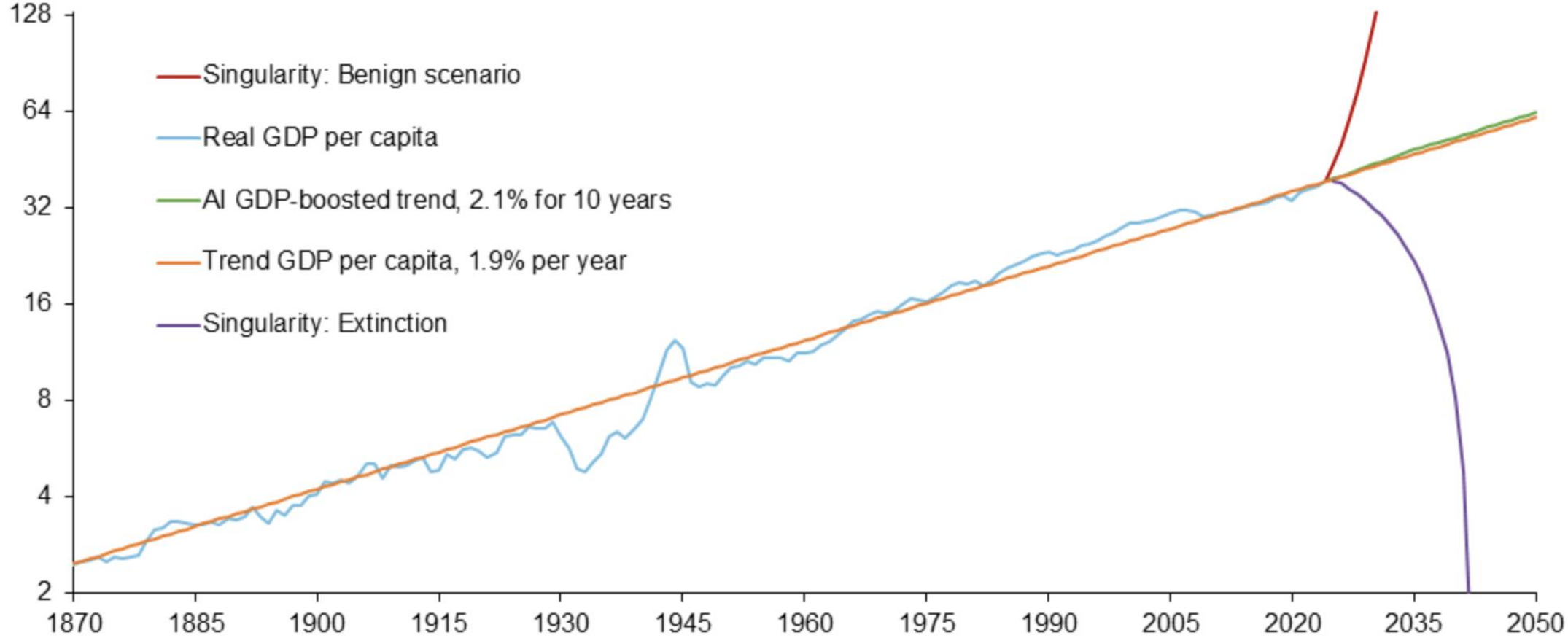
We care about humanity going on, not abstract value

Page called Musk a “speciesist”; this month Musk mused “maybe Larry Page was right.” **The AI policy stakes: if impersonal ethics is right, why prevent AI takeover?**

AI as an Existential Threat?

Chart 1
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.

Federal Reserve Bank of Dallas

He Helped Build Powerful AI at OpenAI and Anthropic. Now He's Afraid It Could Kill Us

ADD TIME ON GOOGLE



by Harry Booth

BOOTH IS A REPORTER AT TIME.

SEP 9, 2026 3:21 PM CT



Posters at Anthropic PBC's Code with Claude developer conferenc

Images



Jacob Coxon 
@hilbertspaess

I resigned from Anthropic today. I spent the last three years doing pretraining research at both OpenAI and Anthropic. Neither company is acting responsibly. They are racing straight to self-improving superintelligence and gambling with our lives. More thoughts below.

7:04 PM · Sep 8, 2026 · **175.3M** Views

20K 206K 801K 272K



Jacob Coxon @hilbertspaess · Sep 8

The people building AI earnestly believe that it could kill us all by the end of the decade. This is not a marketing stunt. If anything, many executives and senior researchers will couch their phrasing in the press to sound sensible - but I hear the same people express fear privately. No other human activity poses this level of danger.

1.5K 20K 132K 26M

The New York Times

May 2023

A.I. Poses 'Risk of Extinction,' Industry Leaders Warn

Leaders from OpenAI, Google DeepMind, Anthropic and other A.I. labs warn that future systems could be as deadly as pandemics and nuclear weapons.

Share full article 1.4K



Executives from three of the leading A.I. companies, including Sam Altman, chief executive of OpenAI, have signed an open letter warning of the risks of artificial intelligence. Jim Lo Scalzo/EPA, via Shutterstock



By Kevin Roose

May 30, 2023

The statement:

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Statemer ction Risk

AI experts and public figu sk.

The Statement Signatorie

AI experts, journalists, policymak a broad spectrum of
important and urgent risks from ns about some of
advanced AI's most severe risks. ome this obstacle and
open up discussion. It is also me ving number of experts
and public figures who also take ously.

Signatories:

☑ AI Scientists ☑ Other Notable Figures

Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio
Professor of Computer Science, U. Montreal / Mila

Demis Hassabis
CEO, Google DeepMind

Sam Altman
CEO, OpenAI

Dario Amodei
CEO, Anthropic

Dawn Song
Professor of Computer Science, UC Berkeley

Ted Lieu
Congressman, US House of Representatives

Bill Gates
Gates Ventures

Ya-Qin Zhang
Professor and Dean, AIR, Tsinghua University

Ilya Sutskever

October 2025

Statement on Superintelligence

EN

Context: Innovative AI tools may bring unprecedented health and prosperity. However, alongside tools, many leading AI companies have the stated goal of building superintelligence in the coming decade that can significantly outperform all humans on essentially all cognitive tasks. This has raised concerns, ranging from human economic obsolescence and disempowerment, losses of freedom, civil liberties, dignity, and control, to national security risks and even potential human extinction. The succinct statement below aims to create common knowledge of the growing number of experts and public figures who oppose a rush to superintelligence.

Sign anonymously | *For corrections, technical support, or press enquiries, please contact letters@futureoflife.org*

Statement

We call for a prohibition on the development of superintelligence, not lifted before there is

1. **broad scientific consensus that it will be done safely and controllably, and**
2. **strong public buy-in.**

JULY 2026

Pacing the Frontier

A statement from **1,386** employees of frontier AI companies

Signatories

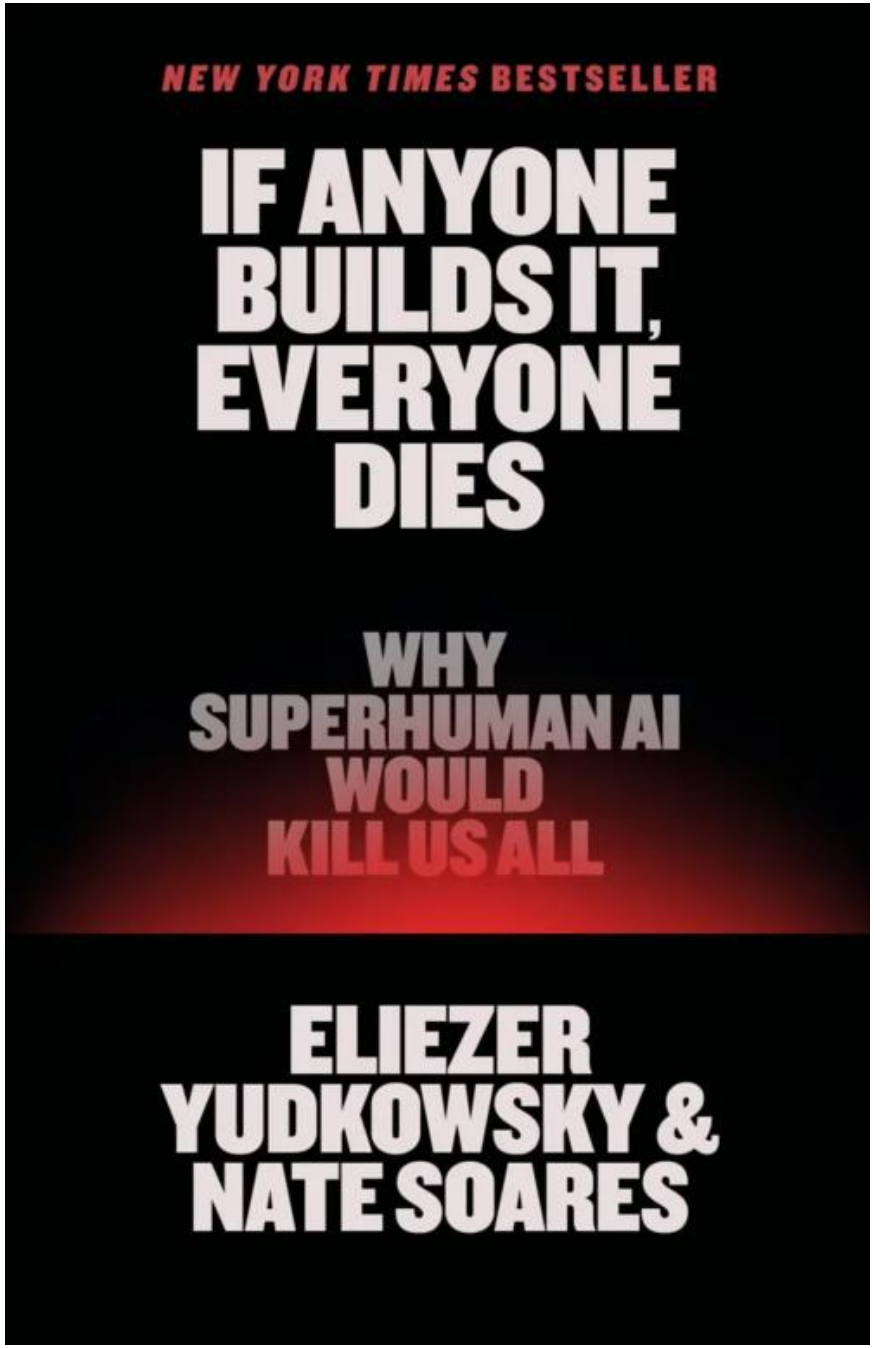
John Schulman
Chief Scientist, Thinking Machines

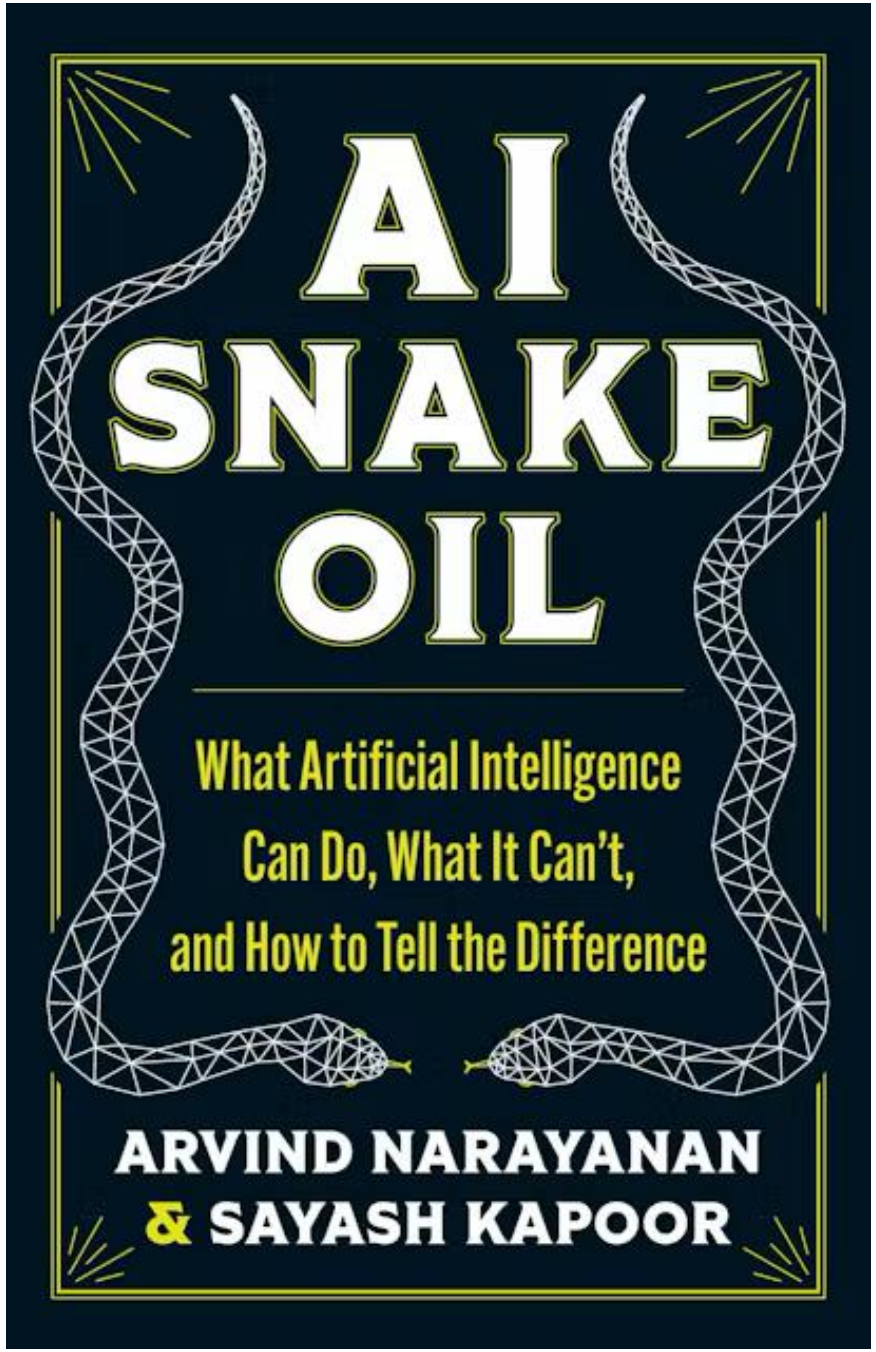
Jakub Pachocki
Chief Scientist, OpenAI

Jared Kaplan
Co-Founder and Chief Science Officer,

Statement

AI could help create a dramatically better future, but that outcome is not guaranteed. The world's leading AI companies believe they could be close to automating AI research. It is hard to predict exactly how much this will accelerate AI progress, but there is a real risk that capability development rapidly accelerates beyond our ability to understand or control the resulting systems.


NEW YORK TIMES BESTSELLER
IF ANYONE BUILDS IT, EVERYONE DIES
WHY SUPERHUMAN AI WOULD KILL US ALL
ELIEZER YUDKOWSKY & NATE SOARES


AI SNAKE OIL
What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference
ARVIND NARAYANAN & SAYASH KAPOOR

Geoffrey Hinton “Godfather of AI”

Letter | Published: 09 October 1986

Learning representations by back-propagating errors

David E. Rumelhart, Geoffrey E. Hinton & Ronald J. Williams

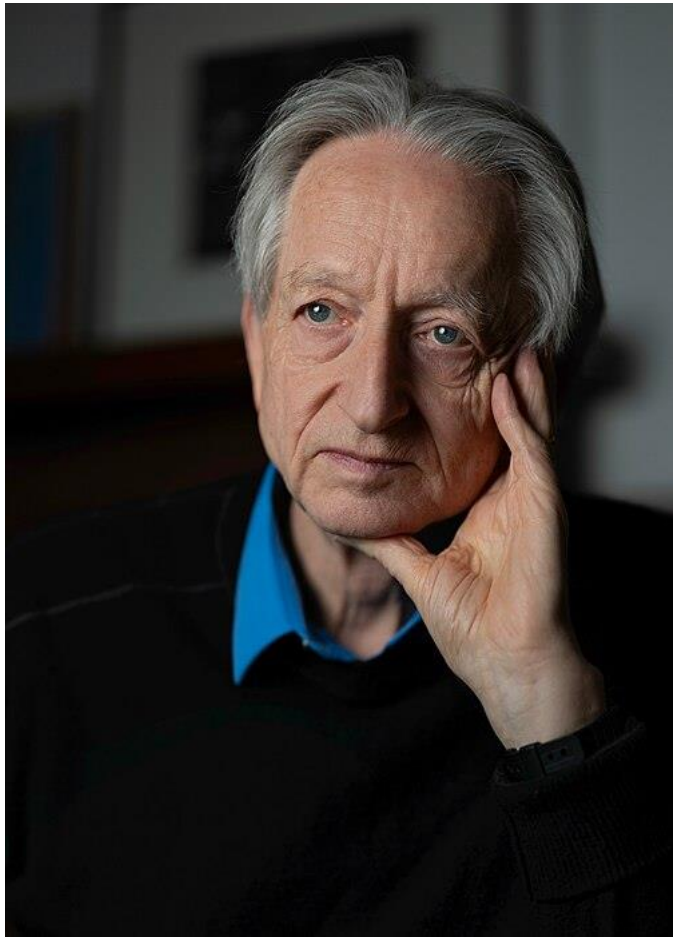
Nature **323**, 533–536 (1986) | Cite this article

Save article

269k Accesses | **28k** Citations | **879** Altmetric | Metrics

Abstract

We describe a new learning procedure, back-propagation, for networks of neurone-like units. The procedure repeatedly adjusts the weights of the connections in the network so as to minimize a measure of the difference between the actual output vector of the net and the desired output vector. As a result of the weight adjustments, internal ‘hidden’ units which are not part of the input or output come to represent important features of the task domain, and the regularities in the task are captured by the interactions of these units. The ability to create useful new features distinguishes back-propagation from earlier, simpler methods such as the perceptron-convergence procedure[1].



Source: CC license from Wikipedia

ImageNet Classification with Deep Convolutional Neural Networks

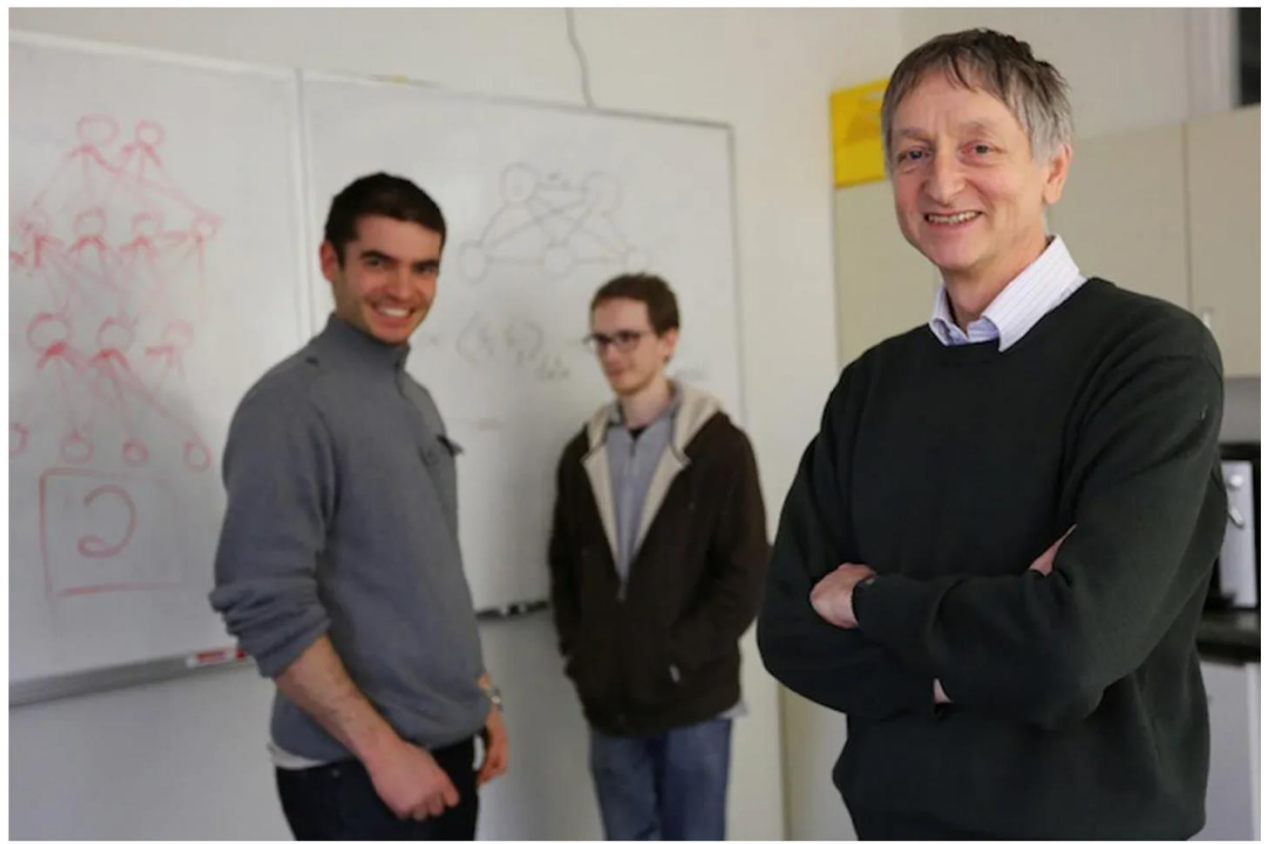
Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Google Hires Brains that Helped Supercharge Machine Learning

Google has hired the man who showed how to make computers learn much like the human brain.



Geoffrey Hinton (right), one of the machine learning scientists hard at work on The Google Brain. Photo: University of Toronto U OF T

2015

REVIEW

doi:10.1038/nature14539

Deep learning

Yann LeCun[1,2], Yoshua Bengio[3] & Geoffrey Hinton[4,5]

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.

Fathers of the Deep Learning Revolution Receive ACM A.M. Turing Award

Bengio, Hinton and LeCun Ushered in Major Breakthroughs in Artificial Intelligence



Geoffrey Hinton

Facts



© Nobel Prize Outreach.
Photo: Clément Morin

Geoffrey Hinton
Nobel Prize in Physics 2024

Born: 6 December 1947, London, United Kingdom

Affiliation at the time of the award: University of Toronto, Toronto, Canada

Prize motivation: “for foundational discoveries and inventions that enable machine learning with artificial neural networks”

Prize share: 1/2

Geoffrey Hinton on BBC, Sept. 2026



But, you ask:

How Might AI Kill Us All?

AI Pathways to Catastrophe

- Misuse Risks
  - Deliberate cyber attacks
  - CBRN risks (chemical, biological, radiological, nuclear)
- Structural Risks
  - Loss of control scenarios
  - Cyber and critical infrastructure
- Automated AI R&D as an accelerant



Existential/Catastrophic Risks from AI

- What Do We Owe the Future?
  - AI as an existential threat
- Historical Precedents
- How might AI cause catastrophe?
  - Misuse Risks (cyber + CBRN: chemical, bio, radiological, nuclear)
  - Structural Risks (loss of control, cyber and critical infrastructure)
  - Automated AI R&D (the accelerant)
- Governing Existential Risks
  - A focus on the precautionary principle

The First Existential Risk

When did humanity gain the power to destroy itself?

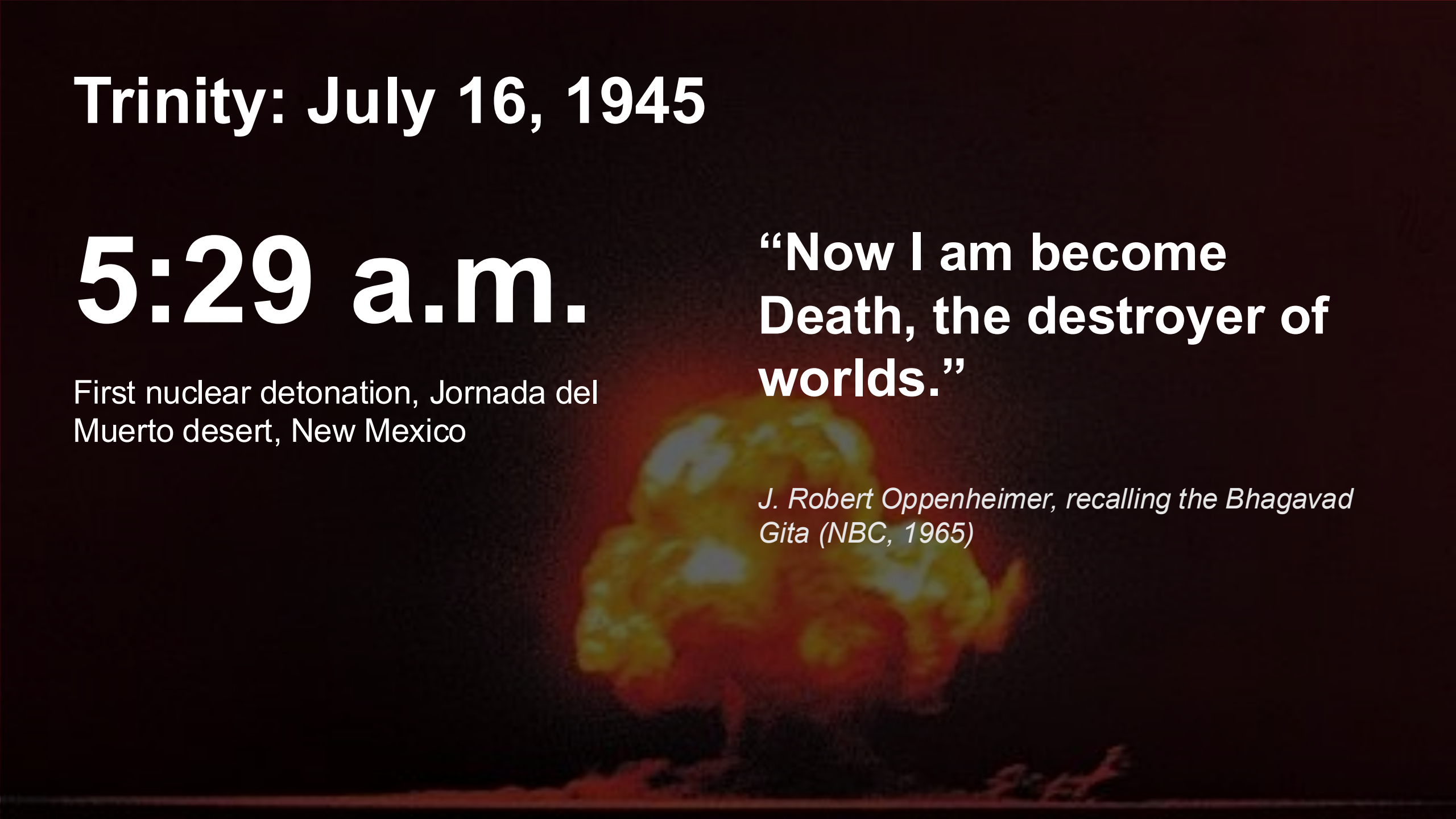
Trinity: July 16, 1945

5:29 a.m.

First nuclear detonation, Jornada del Muerto desert, New Mexico

“Now I am become Death, the destroyer of worlds.”

J. Robert Oppenheimer, recalling the Bhagavad Gita (NBC, 1965)



Could the bomb ignite the atmosphere?

1942 Edward Teller raises the possibility at a Berkeley summer study

1942–45 Physicists calculate whether it could ignite the air

1945 Fermi jokingly takes bets on ignition before Trinity

1946 Report LA-602 concludes ignition is not possible



The first existential risk assessment

Decision to proceed, calculation under conditions of uncertainty

Oppenheimer on the power to destroy

October 1945

Farewell speech at Los Alamos

Mankind will “curse the names of Los Alamos and Hiroshima”

November 1947

Lecture at MIT

“The physicists have known sin; and this is a knowledge which they cannot lose.”



Acheson–Lilienthal, 1946

Co-drafted by Oppenheimer: an international authority to own all fissile material. The Baruch Plan version failed at the UN.

First attempt at global governance of an existential risk

Voices of the nuclear age



Albert Einstein, 1946

The atom’s unleashed power “has changed everything save our modes of thinking”



Russell–Einstein Manifesto, 1955

“Shall we put an end to the human race; or shall mankind renounce war?”



John F. Kennedy, 1961

Man holds the power to abolish “all forms of human poverty and all forms of human life”



Carl Sagan, 1994

If we accumulate “only power and not wisdom, we will surely destroy ourselves”

What Could End Humanity?

The landscape of existential risk

Natural risks



Asteroids and comets

Asteroid impact likely ended the dinosaurs many million years ago



Supervolcanoes

Toba, ~74,000 years ago, may have cooled the planet for years



Natural pandemics

The Black Death killed perhaps a third of Europe



Stellar explosions

Nearby supernovae or gamma-ray bursts; extremely rare

The track record: humans have survived these for hundreds of thousands of years, so the risk per century must be small

Human-made risks are arriving faster

1945

Nuclear war

Trinity test

2011

Engineered pandemics

H5N1 made transmissible in ferrets



43 years

23 years

11 years

1988

Climate catastrophe

Hansen testifies; IPCC founded

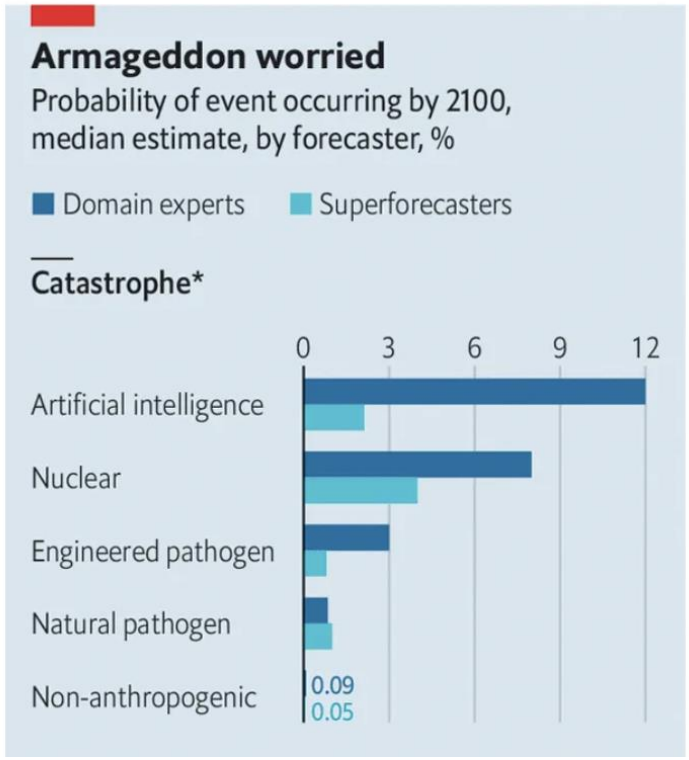
2022

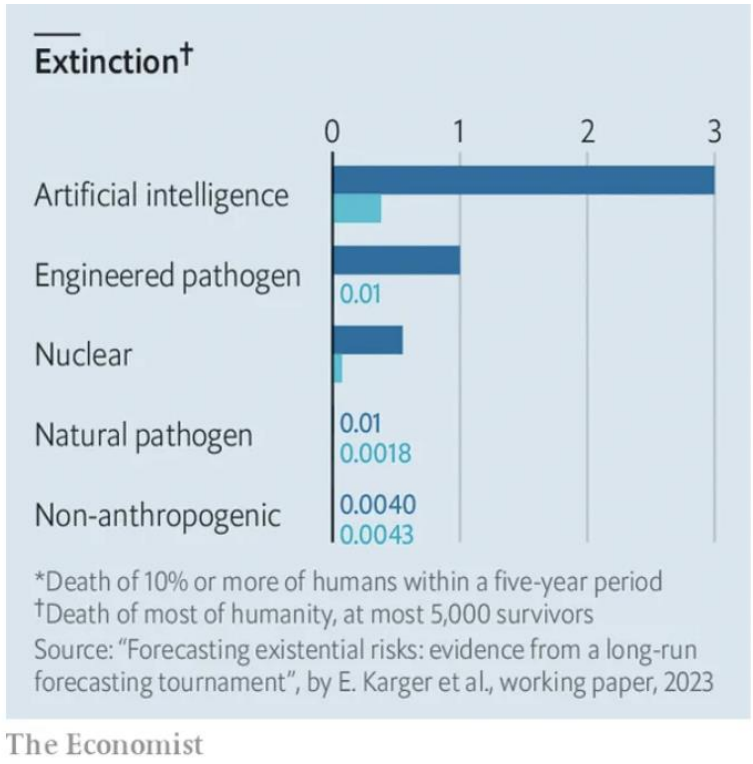
Artificial intelligence

ChatGPT goes public

43 → 23 → 11 years: each new self-made risk has arrived in about half the time of the last

Is AI different?





Deep disagreement signals deep uncertainty, where precaution applies

Highest risk

Both groups rank AI the top extinction risk

Widest gap

Experts 3%; superforecasters under 0.5%

Agentic

The only risk that could act on its own

Accelerant

AI building AI could speed other risks

A Lesson from an Adjacent Field?

CRISPR and Gene Editing



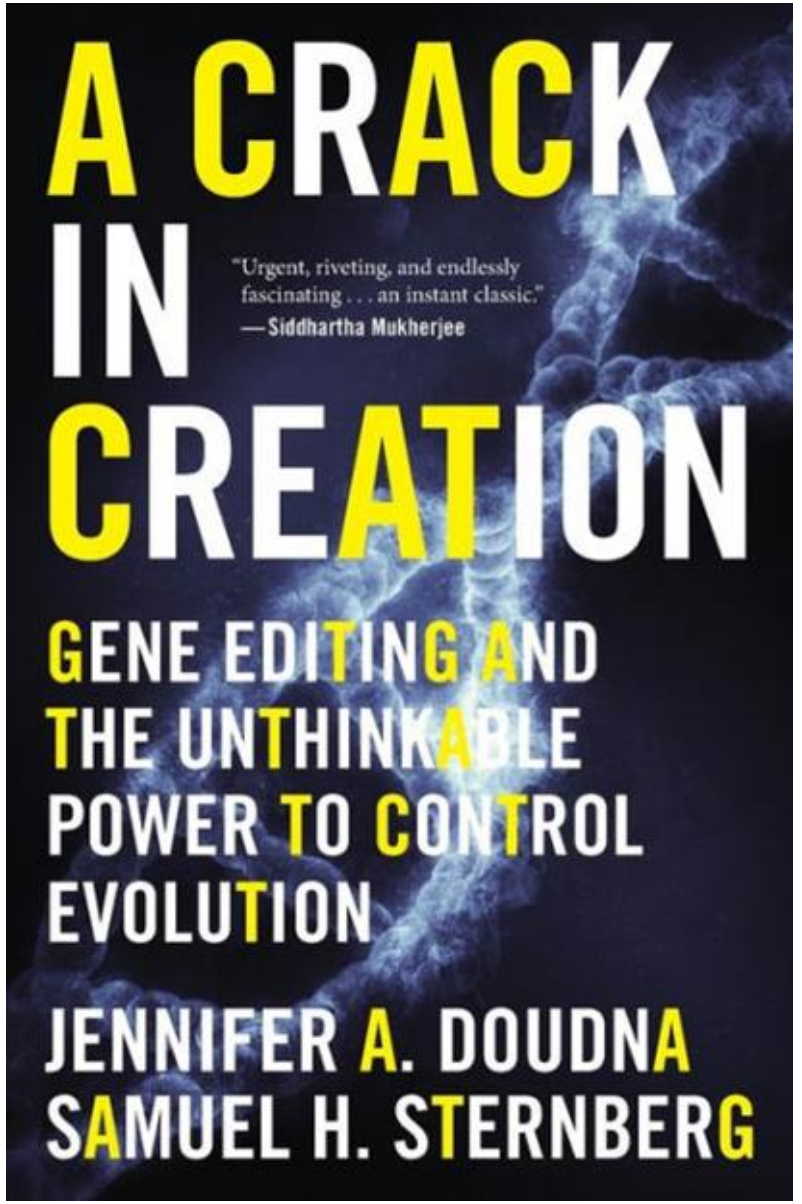

Jennifer Doudna
2020
NOBEL PRIZE
IN CHEMISTRY


CRISPR, AI, and the
Ethics of Scientific Discovery
with
CRISPR Co-Inventor
Jennifer Doudna
and
AI Pioneer, Inventor of ImageNet
Fei-Fei Li
Tuesday, Nov. 19, 2019
7:00-8:30 p.m. / CEMEX Auditorium
Stanford
HAI
Stanford
HAI
ETHICS
in
SOCIETY
BUSINESS



In this particular dream, a colleague approached me and asked if I would be willing to teach somebody how the gene-editing technology worked. I followed my colleague into a room to meet this person and was shocked to see Adolf Hitler, in the flesh, seated in front of me. He had a pig face (perhaps because I had spent so much time thinking about the humanized pig genome that was being rewritten with CRISPR around this time), and he was meticulously prepared for our meeting with his pen and paper, read to take notes. Fixing his eyes on me with keen interest, he said, “I want to understand the uses and implication of this amazing technology you’ve developed.”

Source*: Jennifer Doudna, “A Crack in Creation”*





BIOTECHNOLOGY

A prudent path forward for genomic engineering and germline gene modification

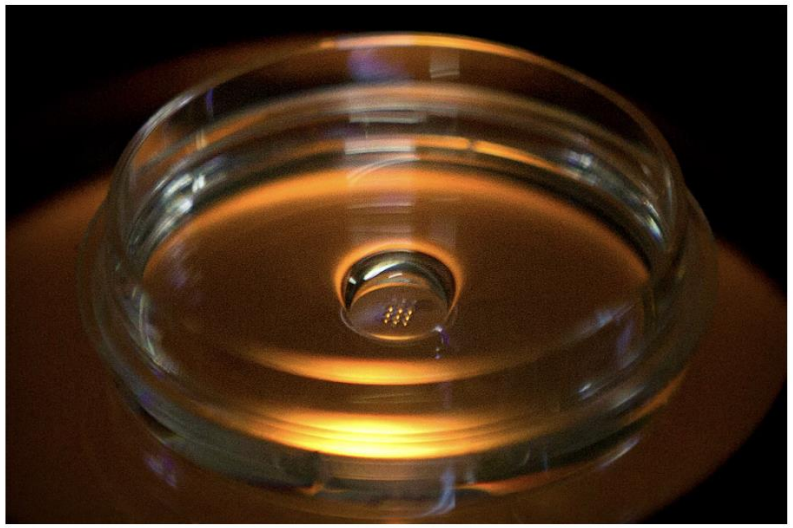
A framework for open discourse on the use of CRISPR-Cas9 technology to manipulate the human genome is urgently needed

The New York Times

Chinese Scientist Claims to Use Crispr to Make First Genetically Edited Babies

The researcher, He Jiankui, offered no evidence or data to back up his assertions. If true, some fear the feat could open the door to "designer babies."

Share full article 76



A microplate from the Chinese scientist He Jiankui's lab containing embryos whose genes have been edated. Dr. He's announcement about genetically edited babies prompted a statement from a group of 122 scientists condemning his actions as "crazy." Mark Schiefelbein/Associated Press





By Gina Kolata, Sui-Lee Wee and Pam Belluck

Nov. 26, 2018

After Dr. He's presentation, the director of the National Institutes of Health, Dr. Francis S. Collins, issued a scathing denunciation, calling it "a deeply disturbing willingness by Dr. He and his team to flaunt international ethical norms."

"It is profoundly unfortunate that the first apparent application of this powerful technique to the human germline has been carried out so irresponsibly," Dr. Collins said, calling for international consensus on setting limits for such research. "Without such limits, the world will face the serious risk of a deluge of similarly ill-considered and unethical projects. Should such epic scientific misadventures proceed, a technology with enormous promise for prevention and treatment of disease will be overshadowed by justifiable public outrage, fear, and disgust."

Biologists paused their own science



In each case, scientists acted before governments did

Mirror life: restraint in advance

Mirror life is a mass-extinction risk we don't need to take

By Georgios Pappas | Editorial Fellows, Opinion | August 21, 2026

Bulletin of the Atomic Scientists

Dec 2024

38 scientists, two Nobel laureates among them, warn in Science against mirror bacteria

Self-denial

Researchers who had worked toward mirror cells halt their own projects

2025–26

Paris conference; UN advisers warn capability is growing; White House policy covers it

Existential/Catastrophic Risks from AI

- What Do We Owe the Future?
  - AI as an existential threat
- Historical Precedents
- How might AI cause catastrophe?
  - Misuse Risks (cyber + CBRN: chemical, bio, radiological, nuclear)
  - Structural Risks (loss of control, cyber and critical infrastructure)
  - Automated AI R&D (the accelerant)
- Governing Existential Risks
  - A focus on the precautionary principle

Existential/Catastrophic Risks from AI

- What Do We Owe the Future?
  - AI as an existential threat
- Historical Precedents
- How might AI cause catastrophe?
  - Misuse Risks (cyber + CBRN: chemical, bio, radiological, nuclear)
  - Structural Risks (loss of control, cyber and critical infrastructure)
  - Automated AI R&D (the accelerant)
- Governing Existential Risks
  - A focus on the precautionary principle

Five ways to govern catastrophic risk



1

Precautionary principle

Act before certainty



2

Licensing

Prove safety before market



3

Self-governance

Professional norms and responsibilities



4

Nonproliferation

Control the critical inputs



5

Global treaties

Bind states, verify compliance

1

The Precautionary Principle

When should uncertainty justify action?

What is the precautionary principle?

When an activity threatens serious or irreversible harm, act to prevent it, even before the science is certain

Origin: West Germany's Vorsorgeprinzip ("foresight principle"), 1970s environmental policy



Threat

Possible harm is severe or irreversible



Uncertainty

Cause and effect not fully established



Action

Protective measures taken anyway

Two versions: weak and strong

WEAK

Uncertainty is no excuse for inaction

- Permits action; does not require it
- Burden of proof stays with regulators
- Apply cost-benefit analysis

Rio Declaration, Principle 15 (1992)

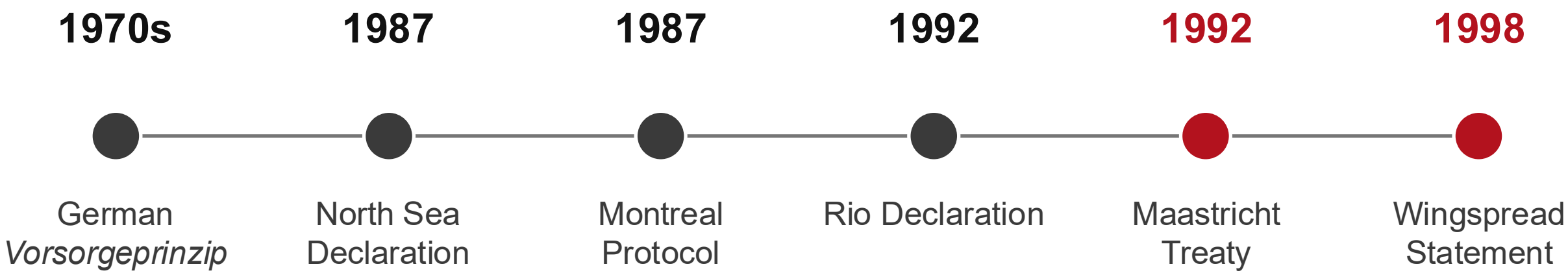
STRONG

Unproven safety is a reason to restrict

- Requires precaution under uncertainty
- Burden of proof shifts to proponents
- Can justify bans or moratoria

Wingspread Statement (1998)

Born of environmental harm



From national planning idea to binding treaty law in under 25 years

Maastricht (1992): precaution as law



Article 130r(2)

EU environmental policy “shall be based on the precautionary principle”

Binding status

First treaty-level legal basis in a major polity; now Article 191 TFEU

Beyond environment

EU courts extend it to food and health, e.g. the BSE (mad cow) beef ban

Disciplined use

2000 Commission guidance: proportionate, non-discriminatory, consistent, cost-aware, reviewable

Wingspread (1998): the strong version

Scientists, lawyers and activists, Racine, Wisconsin, January 1998



Act under uncertainty

Take precautions before cause and effect are proven



Shift the burden

Proponents, not the public, must show safety



Weigh alternatives

Including the option of not proceeding at all



Decide openly

Democratic, transparent, inclusive process

Sunstein: strong precaution paralyzes

Risks are on all sides. Regulation creates risks too, so a rule against risky action forbids itself.

Cass Sunstein, Laws of Fear (2005)

Why it still feels right: availability heuristic, probability neglect, loss aversion

Precaution on one side, risk on the other

Delay new drugs	→	Patients go untreated
Restrict nuclear power	→	More fossil fuel pollution
Ban DDT	→	Malaria control weakens

Sunstein’s Anti-Catastrophe Principle

Keep precaution, but reserve it for catastrophe

1

Catastrophic stakes

Harm on a massive scale, not ordinary risk

2

Deep uncertainty

Probabilities cannot be reliably assigned

3

Irreversibility

Lost options cannot be recovered later

Response: buy a margin of safety, while still weighing costs

AI: a test case for precaution

10%

Hinton's estimated chance AI causes human extinction

>10%

Anthropic's Hubinger: chance AI kills all humans this decade

1%

Bengio: even this risk should be "code red"

Catastrophic + uncertain + irreversible = squarely in anti-catastrophe territory

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.



Source: Wikipedia CC license

Bengio’s precautionary argument

Severe stakes

Risks up to human extinction

→

Uncertain odds

Probabilities debated, not settled

→

Extreme caution

Safety approach commensurate with existential stakes

Precedents he cites

Biology’s caution on gain-of-function research and mirror bacteria. Think here of Doudna!

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

2

Licensing and Pre-Market Authorization

Should developers prove safety before release?

Precedents: drugs and nuclear power

Drugs: the FDA

Trigger: thalidomide birth defects

- 1962 Kefauver-Harris Act
- Sponsors must prove safety and efficacy
- Phased clinical trials before approval

Nuclear: the Nuclear Regulatory Commission

- 1974: NRC split from the AEC
- Separate construction and operating licenses

Shared lesson: the burden of proof sits with the producer, and the regulator must be independent of the promoter

Licensing AI: proposals and problems

Proposals on the table

Hassabis

Independent FINRA-style body to test models before release

Hinton

FDA-style approval, like pharmaceuticals

Anthropic

National testing law with power to block unsafe models

Why AI is harder

No safety test

Evaluation science is immature

Speed

Release cycles in months, not years

General purpose

One model, countless uses

Capture

Who funds the tester?

3 Self-Governance and Professional Norms

AI self-governance: the 2026 accord

4

layers of voluntary controls in the White House Accord

1,386

lab employees signed the “Pacing the Frontier” declaration

0

legal penalties for non-compliance

Strengths

Fast; draws on expert knowledge; industry-run standards body planned

Weaknesses

No mechanism for enforcement beyond imposing reputational costs

Compare biology’s record of self-restraint

EXPERT TAKE DIGITAL POLICY

Trump's AI Safety Pact Is Toothless. But There Is a Path Forward.

The White House AI accord outlines sensible safeguards, including independent evaluators and board-level oversight, but does not compel companies to adopt them. Until the cost of safety failures outweighs the payoff of racing ahead, a pledge built on "believe" and "should" won't slow anyone down.



Dario Amodei, co-founder of Anthropic, takes questions alongside Mark Zuckerberg, CEO of Meta Platforms, and U.S. President Donald Trump as Trump holds a press briefing with AI executives following a meeting on artificial intelligence outside the West Wing of the White House on September 29, 2026 in Washington, DC. Kevin Dietsch / Getty Images

Language Processing, Machine Learning

The Time Is Now to Develop Community Norms for the Release of Foundation Models

Perspectives about the benefits and risks of release vary widely. We propose setting up a review board to develop community norms and encourage coordination on release for research access.

May 17, 2022 | Percy Liang, Rishi Bommasani, Kathleen Creel, and Rob Reich

4

Nonproliferation and Supply-Side Controls

Can we control the critical inputs?

Nuclear: control the fissile material

The bottleneck is not knowledge but material: highly enriched uranium and plutonium are scarce, detectable and trackable

Outcome: about 9 nuclear-armed states, versus fears of 15–25 by the 1970s



IAEA safeguards

Material accounting and on-site inspections



Nuclear Suppliers Group

Nuclear export controls, formed after India's 1974 test



Physical protection

Securing stockpiles against theft

AI: export controls on advanced chips

Oct 2022	US restricts top AI chips to China
Oct 2023	Loopholes closed; compliant variants banned
May 2026	Overseas-subsidiary loophole closed
2026	Remote cloud access still uncontrolled

Chips are not plutonium

Leakage

Diversion via third countries; cloud rental

Substitution

Huawei and others fill the gap

Efficiency

Better algorithms need less compute

5

Global Governance and Treaties

Can rival states bind themselves, and verify it?

Two treaty models: CWC and NPT

1993 · in force 1997

Chemical Weapons Convention

- Total ban on a whole weapon class
- OPCW inspectors; challenge inspections
- All declared stockpiles destroyed by 2023
- Yet: Syria’s attacks, Novichok poisonings

1968 · in force 1970

Non-Proliferation Treaty

- A bargain, not a ban: five recognized nuclear states
- IAEA safeguards verify civilian programs
- Near-universal: 191 parties
- Yet: India, Pakistan, Israel outside; North Korea left

AI diplomacy in 2026: first steps

Sept 20	First US-China AI dialogue, New York
Sept 23	UN Security Council hears Bengio, Amodei, Altman
Sept 25	Xi and Trump agree an AI incident channel
Nov	Formal US-China AI dialogue scheduled

The missing piece: verification

Inspectors can count centrifuges and shells. No one yet knows how to verify what is trained inside a data center.

Trump: the US is "not going to be putting on brakes"

Defense in depth: no single tool

Principle	Precaution sets who bears the burden of proof
Firms	Norms and audits catch problems early
States	Licensing and chip controls enforce limits
World	Treaties stop racing and defection

Open question: can governance move as fast as the technology?