Existential Risks: Natural, Anthropogenic, and Future

•••

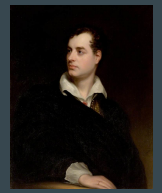
PHIL 1561 Ethics, Economics, and the Future

Existential Risk

Existential Risks:
Natural
Anthropogenic
Future

Natural Risks

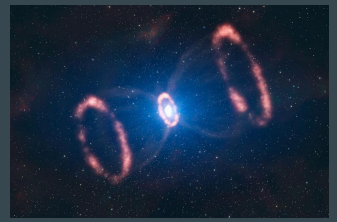
Natural Risks



Asteroids & Comets

Supervolcanic Eruptions

Stellar Explosions

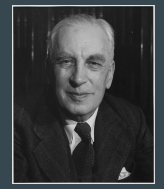


Who knows whether, when a comet shall approach this globe to destroy it, as it often has been and will be destroyed, men will not tear rocks from their foundations by means of steam, and hurl mountains, as the giants are said to have done, against the flaming mass?—and then we shall have traditions of Titans again, and of wars with Heaven.

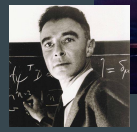
—Lord Byron

Anthropogenic Risks

Anthropogenic Risks



Nuclear Weapons



Climate Change

Environmental Damage

The human race's prospects of survival were considerably better when we were defenceless against tigers than they are today, when we have become defenceless against ourselves.

—Arnold Toynbee

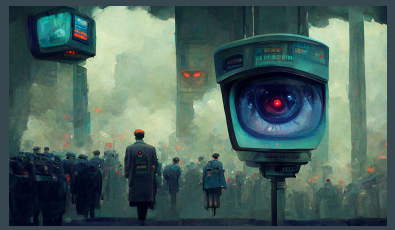
Future Risks

Future Risks

Pandemics (and biological risks)

Unaligned Artificial Intelligence

Dystopian Scenarios



The Dark Ages may return, the Stone Age may return on the gleaming wings of Science, and what might now shower immeasurable material blessings upon mankind, may even bring about its total destruction.

—Winston Churchill

Joe Carlsmith

Unaligned Artificial Intelligence

“Existential Risk from Power-Seeking A.I.”

Existential Risk from Power-Seeking AI

Joe Carlsmith

We … are currently pouring resources into learning how to build something akin to **a second advanced species**; a species potentially **much more powerful** than we are; that we do not yet **understand**, and that it's not clear we will be able to **control**. […]

We are doing something **unprecedented** and **extremely dangerous**; with very little room for error; and the entire future on the line.

Within my lifetime, I think it **more likely than not** that it will become possible and financially feasible to create and deploy **powerful AI agents**.

And I expect strong incentives to do so, among many actors, of widely varying levels of social responsibility.

What's more, I find it quite plausible that it will be difficult to ensure that such systems don't seek **power over humans** in unintended ways; plausible that they will end up deployed anyway, to **catastrophic effect**; and plausible that whatever efforts we make to contain and correct the problem will **fail**.

Existential Risk from Power-Seeking AI

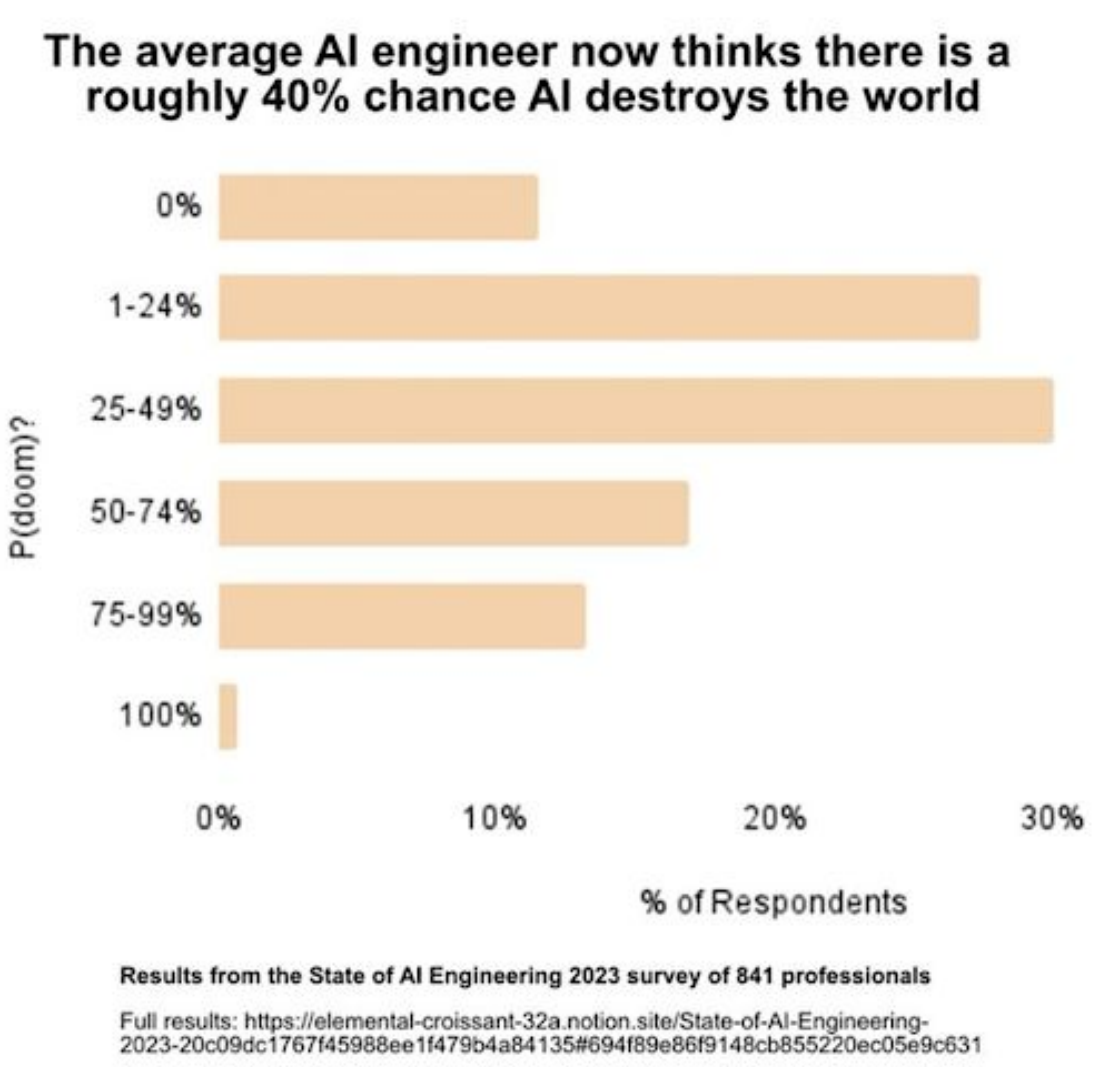
Joe Carlsmith

We ... are currently pouring resources into learning how to build something akin to **a second advanced species**; a species potentially **much more powerful** than we are; that we do not yet **understand**, and that it's not clear we will be able to **control**. [...]

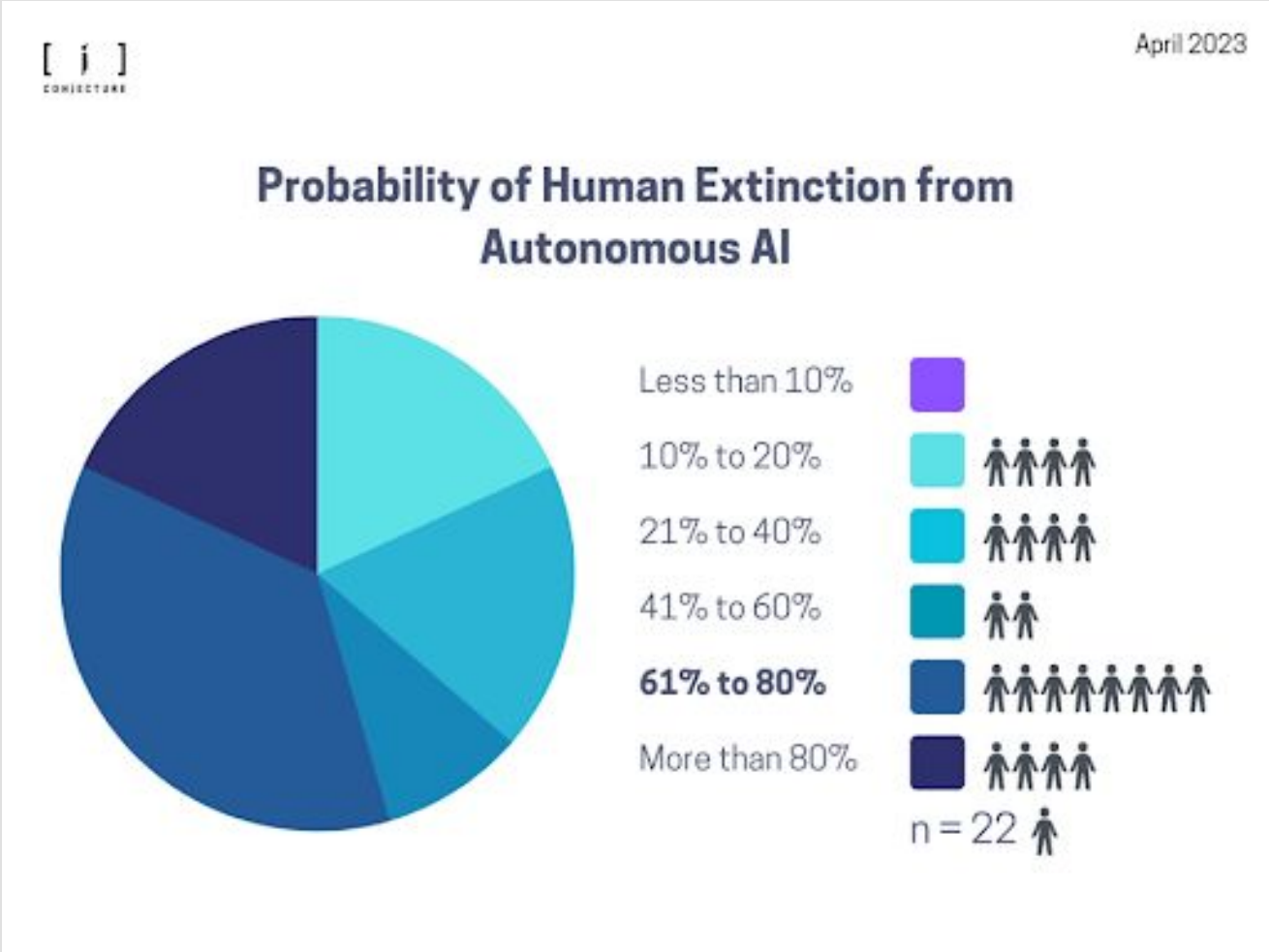
We are doing something **unprecedented** and **extremely dangerous**; with very little room for error; and the entire future on the line.

That is, as far as I can tell, there is a disturbingly high risk (I think: **greater than 10%**) that I live to see the human species permanently and involuntarily **disempowered** by **AI systems** we've lost control over.

P(Doom)


The average AI engineer now thinks there is a roughly 40% chance AI destroys the world
P(doom)?
0%
1-24%
25-49%
50-74%
75-99%
100%
0%
10%
20%
30%
% of Respondents
Results from the State of AI Engineering 2023 survey of 841 professionals
Full results: https://elemental-croissant-32a.notion.site/State-of-AI-Engineering-2023-20c09dc1767f45988ee1f479b4a84135#694f89e86f9148cb855220ec05e9c631

P(Doom)


[i]
CONJECTURE
April 2023
Probability of Human Extinction from Autonomous AI
Less than 10%
10% to 20%
21% to 40%
41% to 60%
61% to 80%
More than 80%
n = 22

Existential Risk from Power-Seeking AI

1. It will become **feasible** to build powerful agentic AI systems.
2. There will be strong **incentives** to do so.
3. It will be much harder to build aligned AI systems than **misaligned** ones.
4. Some such misaligned systems will **seek power** over humans in high impact ways.
5. This problem will scale to the **full disempowerment** of humanity.
6. Such disempowerment constitutes an **existential catastrophe**.

Joe Carlsmith

That is, as far as I can tell, there is a disturbingly high risk (I think: **greater than 10%**) that I live to see the human species permanently and involuntarily **disempowered** by **AI systems** we've lost control over.

by 2070

Existential Risk from Power-Seeking AI

Joe Carlsmith

1. It will become **feasible** to build powerful agentic AI systems.
2. There will be strong **incentives** to do so.
3. It will be much harder to build aligned AI systems than **misaligned** ones.
4. Some such misaligned systems will **seek power** over humans in high impact ways.
5. This problem will scale to the **full disempowerment** of humanity.
6. Such disempowerment constitutes an **existential catastrophe**.

Intelligence:
The ability to plan, learn, communicate, deduce, remember, explain, imagine, experiment, cooperate,...

Agency:
The ability to pursue objectives, guided by models of the world.

Existential Risk from Power-Seeking AI

Joe Carlsmith

The Worry:
By default, suitably **strategic** and **intelligent agents**, engaging in suitable types of planning, **will have instrumental incentives to gain and maintain various types of power** [call this 'power-seeking'], since this power will help them pursue their objectives more effectively.

APS system: Advanced, Planning, Strategically-aware system.

1. *Advanced Capability:* They can outperform humans on some set of important tasks.
2. *Agentic Planning*: They make plans, in pursuit of objectives, on the basis of models of the world.
3. *Strategic Awareness*: The models they use accurately represent the causal upshot of gaining power over humans.

There will be strong economic and political **incentives** to automate APS systems.

Incentives to Automate APS Systems

Joe Carlsmith

1. Agentic planning and strategic awareness are **useful**.

2. Creating APS systems might be the **most efficient** route for automating tasks (even if they don't *require* agentic planning or strategic awareness).

3. They might arise in **unexpected** ways regardless / prove difficult to prevent.

Alignment

“It will be difficult to create APS systems that don’t seek to **gain** and **maintain power** in unintended ways.”

The Alignment Problem

Joe Carlsmith

Misaligned *power-seeking*:
Active efforts by an AI system to gain and maintain power in ways that designers didn't intend, arising from problems with that system's objectives.

- Self-preservation
- Preventing changes to its objectives
- Improving its capabilities
- Technological development
- resource-acquisition

"Power, almost by definition, is extremely useful to accomplishing objectives. So to the extent that an agent is engaging in unintended behavior in pursuit of problematic objectives, it will generally have incentives, other things equal, to gain and maintain forms of power in the process."

Problem with **Proxies**

The Alignment Problem: Proxies

Joe Carlsmith

Proxy Objective:
An objective that reflects properties correlated with, but separable from, intended behavior.

Problem: As the power of the AI's optimization for the proxy increases, the AI's behavior can significantly deviate from what's intended.

Example:
The AI learned to steer the boat in circles.
The 'Cobra effect'

Problem with **search**

The Alignment Problem: Search

Joe Carlsmith

Problem:
The resulting system may not be *intrinsically* motivated by the criteria in question—which might lead to unintended behavior when the system is exposed to non-training inputs.

Search: Look at many different AI systems, and select those that perform well on some evaluation criteria (without controlling the system's objectives directly).

It’s difficult to control APS system’s capabilities or circumstances.

Deployment:

But if they are so dangerous, won't be ensure they are safe before we build them?

Deployment

But if they are so dangerous, won't be ensure they are safe before we build them?

Maybe not!

1. *Externalities.*
2. *Race dynamics.*
3. *Many relevant actors.*
4. *Extremely useful.*

The End (of humanity?)