Steering the Future of Artificial Intelligence

Luke Muehlhauser

MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

This talk assumes…

1. You understand narrow AI vs. AGI vs. superintelligence

2. You understand astronomical stakes

3. You see why some people think AI is the key lever on the long-term future

4. You know that Friendly AI research is uncrowded, and you're open to the idea that it's tractable

This talk focuses on:

5. **Friendly AI work is urgent:** Most AGIs do not stably optimize for desirable values, Friendly AI is strictly (much) harder than AGI, and today AGI progress is vastly outpacing Friendly AI progress.

"Assume that human scientific activity continues without major negative disruption. By what year would you see a (10% / 50% / 90%) probability for [AGI] to exist?"

	10%	50%	90%
AI scientists, median	2024	2050	2070
Luke	2030	2070	2140

"Assume… that [AGI] will at some point exist. How likely do you then think it is that within (2 years / 30 years) thereafter there will be machine [superintelligence]?"

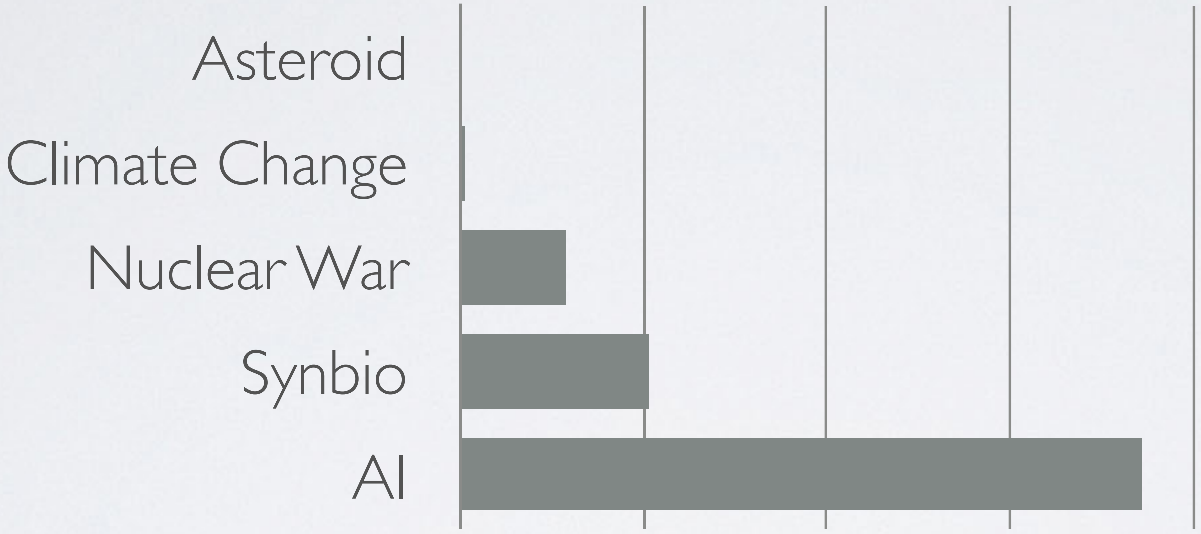
	2 years	30 years
AI scientists, median	5%	50%
Luke	15%	85%

"Assume… that [AGI] will at some point exist. How positive or negative would be overall impact on humanity, in the long run?"

	Extremely good	good	Neutral-ish	bad	Extremely bad
AI scientists, mean	20%	40%	19%	13%	8%
Luke (volatile)	19%	1%	~0%	5%	75%

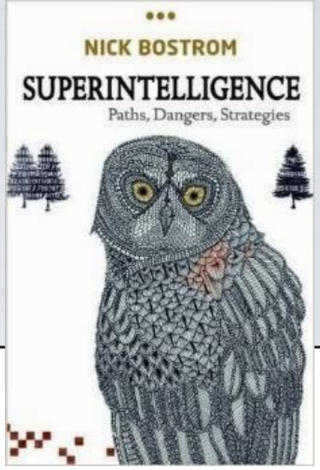
AI as the key lever on the long-term future

Chance of being an x-risk in next century (my opinion)

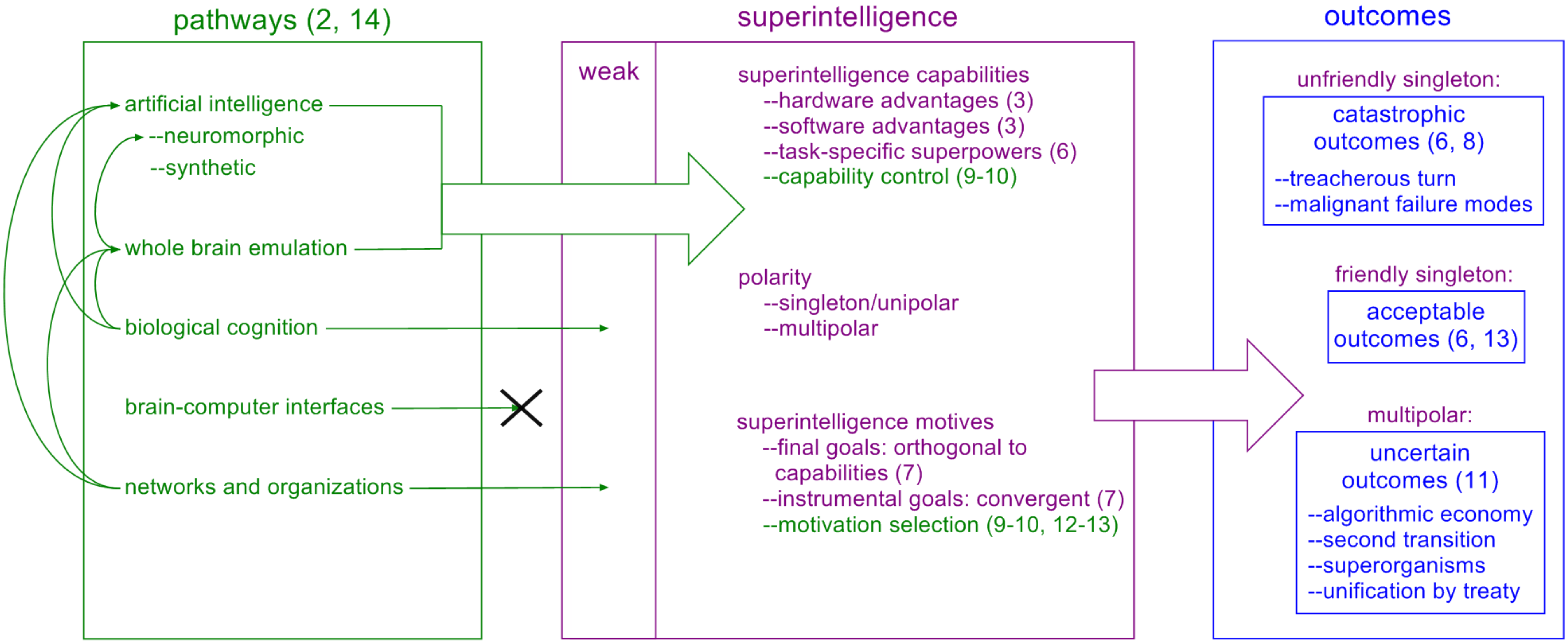


Asymmetry #1:
FAI helps us mitigate other risks,
but solving climate change, asteroids,
etc. doesn't help us much with other risks.

Asymmetry #2:
FAI is the only technology that lets us
convert the reachable universe
into quality-adjusted life years.



A Visualization of Nick Bostrom's *Superintelligence*



pathways (2, 14)

- artificial intelligence
  - --neuromorphic
  - --synthetic
- whole brain emulation
- biological cognition
- brain-computer interfaces
- networks and organizations

superintelligence

weak

superintelligence capabilities
--hardware advantages (3)
--software advantages (3)
--task-specific superpowers (6)
--capability control (9-10)

polarity
--singleton/unipolar
--multipolar

superintelligence motives
--final goals: orthogonal to capabilities (7)
--instrumental goals: convergent (7)
--motivation selection (9-10, 12-13)

outcomes

unfriendly singleton:
catastrophic outcomes (6, 8)
--treacherous turn
--malignant failure modes

friendly singleton:
acceptable outcomes (6, 13)

multipolar:
uncertain outcomes (11)
--algorithmic economy
--second transition
--superorganisms
--unification by treaty

🟩 human interventions

🟪 superintelligence properties

🟦 outcomes

(#) chapter reference

Amanda E House

Why am I pessimistic?

- AGI presents a "no turning back" point. We're good at iterating with testing and feedback, but we're terrible at getting something exactly right the first time.
- Very strong incentives to build AGI even given known large risk.
- An arms race, incentivizing speed of development over safety of development, seems likely.
- Progress may be rapid right when novel control problems become relevant.
- Moore's law of mad science + AI more difficult to control than nuclear fissile materials.
- Good outcomes seem to require as-yet unobserved philosophical success.

Why is Bostrom pessimistic?

“Before the prospect of [superintelligence], we humans are like small children playing with a bomb. Such is the mismatch between the power of our plaything and the immaturity of our conduct…

“For a child with an undetonated bomb in its hands, a sensible thing to do would be to put it down gently, quickly back out of the room, and conduct the nearest adult. Yet what we have here is not one child but many, each with access to an independent trigger mechanism…

“Nor can we attain safety by running away, for the blast of [superintelligence] would bring down the entire firmament. Nor is there a grown-up in sight.”

Yudkowsky’s summary case for Friendly AI work

- **Astronomical stakes**
- **Orthogonality of system goals and capability**
- **Convergent instrumental goals:** Self-preservation, goal-content integrity, self-improvement, resource acquisition.
- The resource acquisition goal implies **infrastructure profusion**.
- **Intelligence explosion**
- **Complexity + fragility of human value** implies **unforseen instantiation** (and remember, “the genie knows but doesn’t care”)
- Therefore, **indirect normativity**
- Therefore, **bounded extra difficulty of Friendliness,** which needs to be built in from the ground up
- Therefore, **Friendly AI is a technical problem, less so a favorite-political-faction problem**

Superintelligence control methods

Capability control

- Boxing methods
- Stunting
- Incentive methods
- Tripwires

Motivation selection

- Direct specification
- Domesticity
- Augmentation (doesn’t work for AI paths)
- Indirect normativity

Four "castes" of superintelligence

An **oracle** is a motivated question-answering system, a kind of "domesticity" solution. Might be useful for building Friendly AI, but probably can't halt all progress toward less domesticated AGI.

A **tool** is non-motivated. Might be useful for building Friendly AI, but the incentives for someone else to build a motivated agent remain huge. Also, it's not clear one can get to a *superintelligent* tool or oracle if the AI isn't helping humans build itself (recursive self-improvement from AGI-level to superintelligence-level).

A **genie** or a **sovereign** has all the usual difficulties.

What can be done?

- More forecasting & strategic analysis
- Build capacity / consensus behind safety efforts
- Direct technical work on the design challenges (MIRI’s specialty)
- Regulation?