Superintelligence alignment as the world's top research priority

Andrew Critch[1,2]
critch@intelligence.org
[1] Machine Intelligence Research Institute
http://intelligence.org/
[2] UC Berkeley, Center for Human Compatible AI
http://humancompatible.ai/

My two hats:

Specialist hat:

Bounded rationality, open-source game theory, algorithms that reason about algorithms,

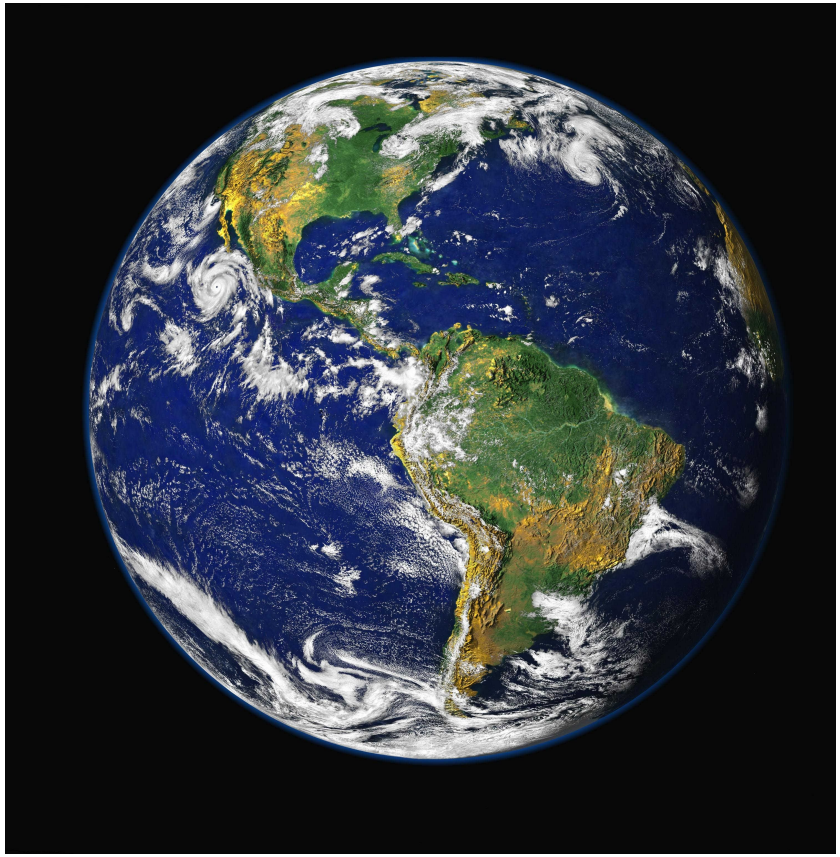
…

Generalist hat:

Aligning AI with human interests, reducing existential risk, making the future super awesome,

…

Q: Why are humans the dominant species on the planet?



Are we

- the strongest?
- the toughest?
- the fiercest?
- the smartest?

Q: Why are humans the dominant species on the planet?

$\in$

Q: Why are humans the dominant species on the planet?

$\notin$

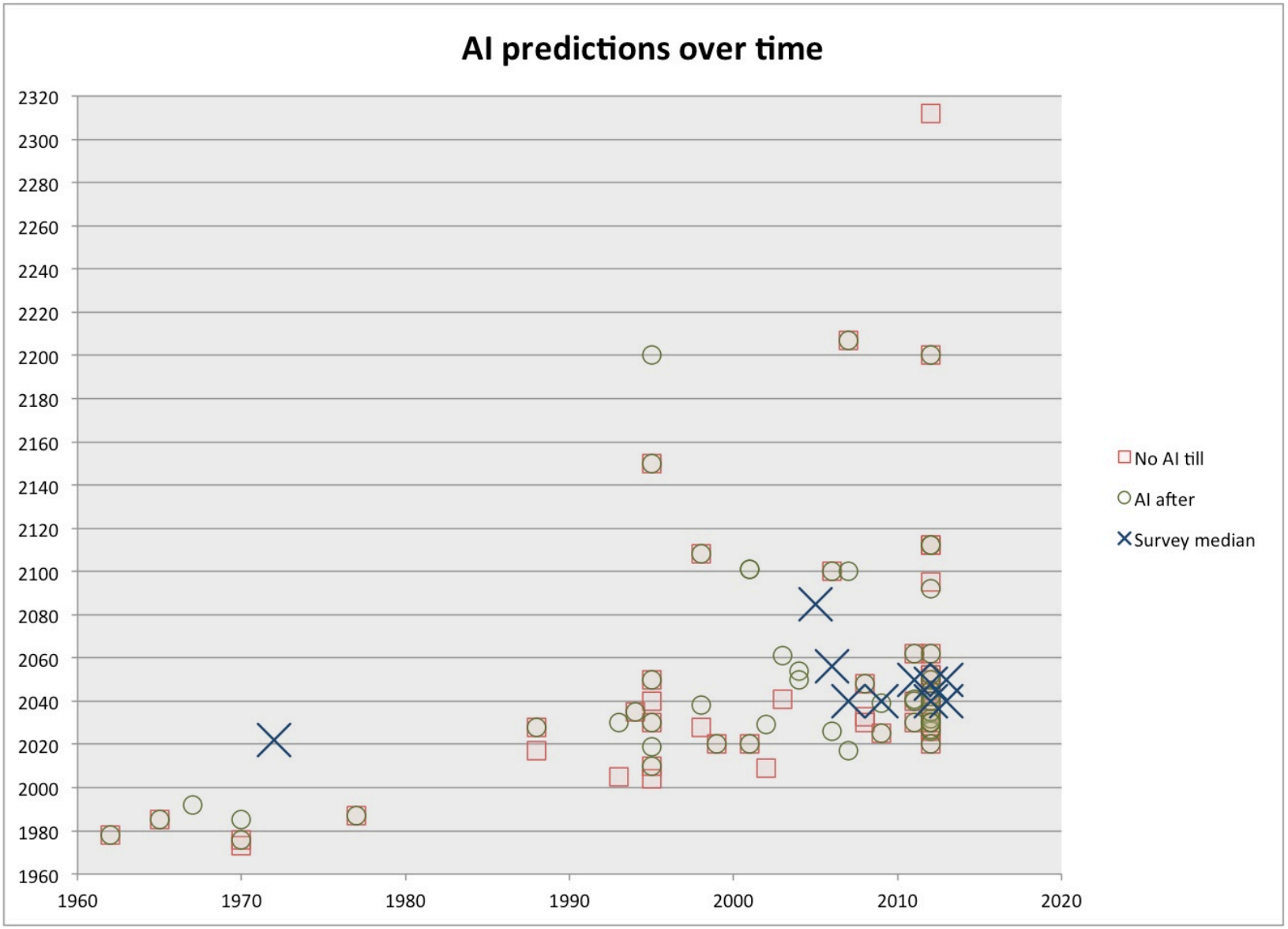
A 30-year thought experiment:

Imagine if Earth received a credible message from 30 light-years away that said:

"Dear humans, we are considerably smarter than you in every way. We are planning a travel experiment that is **50% likely to bring us to Earth within 30 years.**"

What percentage of the **world's greatest minds** would be worth allocating to prepare for that scenario?
1%? 10%? 90%?

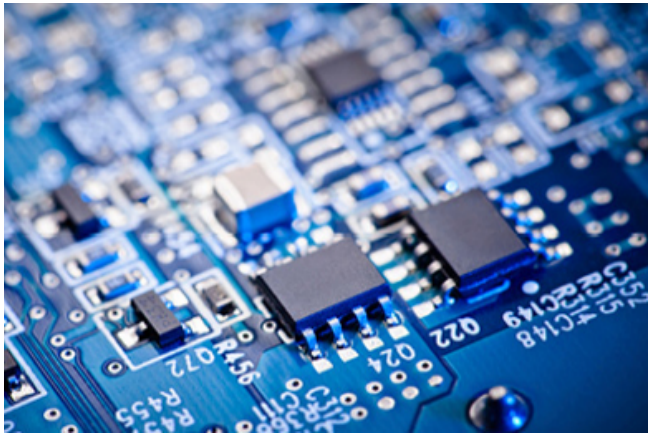
We haven't gotten an alien message, but we have gotten a Terrestrial one.
From http://aiimpacts.org/ :



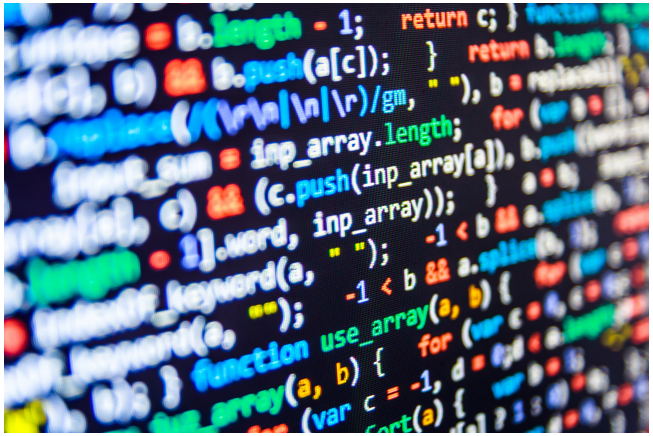
Brain model	CPU demand (FLOPS)	$1MM availability via supercomputer / commodity computer	Memory (TB)	$1MM availability
analog network population model	10^**14**	2008 / 2023	10^**2**	present
spiking neural network	10^**18**	2019 / 2042	10**^4**	present
electrophysiology	10^**22**	2033 / 2068	10^**4**	2019
states of protein complexes	10^**27**	2052 / 2100	10^**8**	2038
stochastic behavior of single molecules	10^**43**	2111 / 2201	10^**14**	2069

Sandberg, A. & Bostrom, N. (2008): Whole Brain Emulation: A Roadmap
Technical Report #2008-3, Future of Humanity Institute, Oxford University
URL: www.fhi.ox.ac.uk/reports/2008-3.pdf

hardware performance ≠ software performance



≠

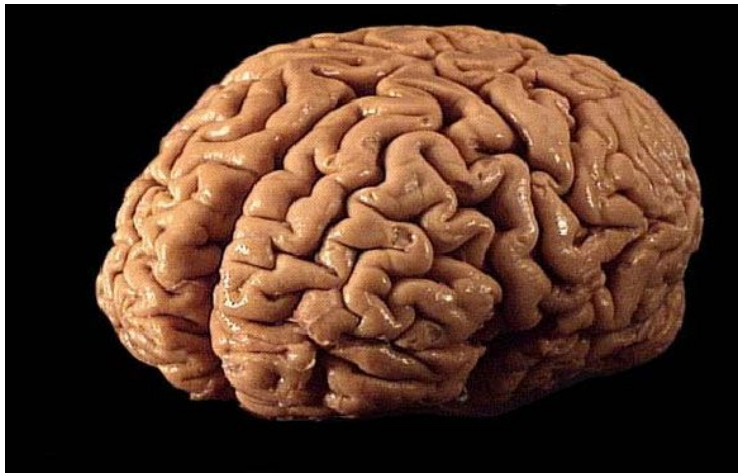


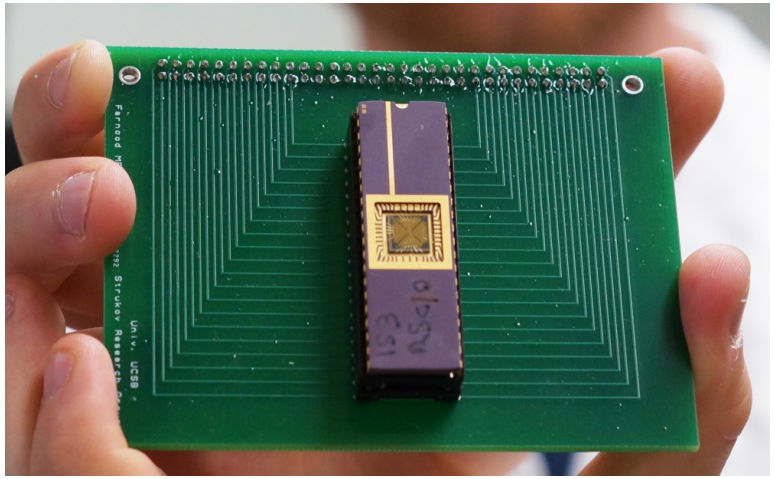
but our source code isn't so long…

< 750 MB <

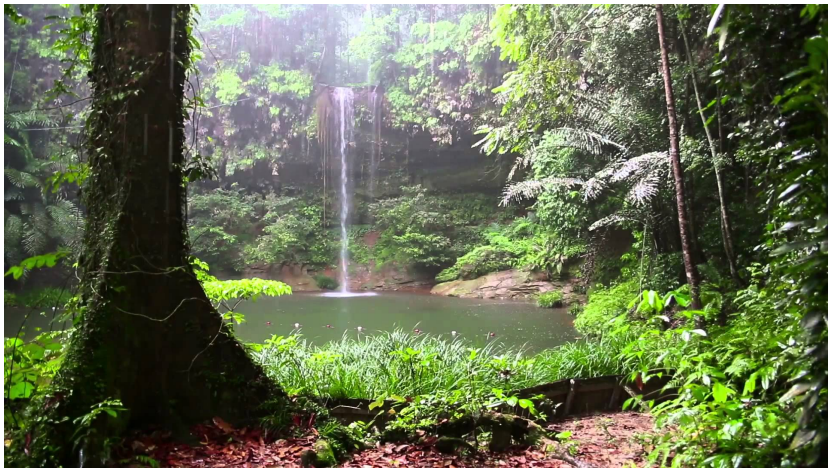
HLAI → SHAI

HLAI → SHAI





Whatever we care about that might happen over the next few few decades or beyond…



… is likely to depend on the alignment of super-human AI with human interests:

Our intelligence has brought us:

- *Medicine, education, diplomacy, scientific discovery, art, social progress, animal rights, space travel…*

If the development of human-level AI goes well, we could have much more of all of these things.

If it goes poorly, not only would we miss out on these benefits, but we would likely be facing an existential risk.

A precarious situation:

Even absent any malice from an AI, there are basic sub-goals --- like acquiring matter, energy, self-defense, and increased intelligence --- that are useful for almost any objective a superintelligence might hold, and would be disastrous to humans if pursued fully by a superintelligent optimizer.

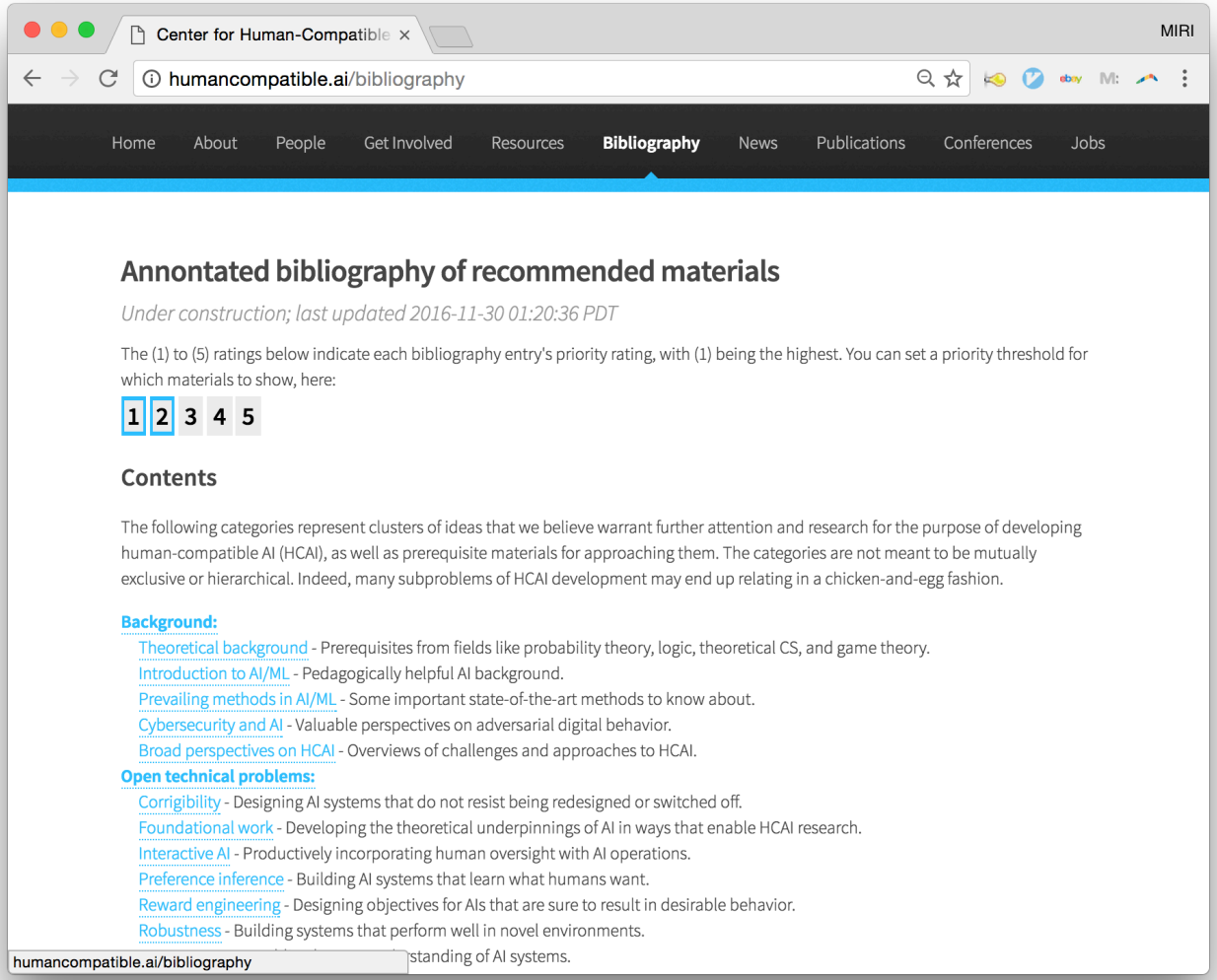
A 30-year thought experiment:

Experts in AI generally hold that within 30 years, there's around a 50% chance that AI systems will be able to "carry out most human professions at least as well as a typical human."

What percentage of the **world's greatest minds** would be worth allocating to prepare for that scenario?
1%? 10%? 90%?

There are numerous open problem areas:

http://humancompatible.ai/bibliography



… a few research agendas addressing AI safety over varying time horizons and scenarios:

Impact* time scale	Focus on:	Example research agenda
5 years	Foreseeable safety issues with currently developing applications, e.g. self-driving cars, personal assistants	*Concrete problems in AI safety*, D. Amodei et al
10 years	Ensuring safety of highly intelligent systems built with current ML paradigms, e.g. reinforcement learning, active learning	*Value Alignment for Advanced Machine Learning Systems*, J. Taylor et al
15 years	New theoretical perspectives extending probability theory, game theory, decision theory, bounded rationality, etc.	*Agent Foundations For Aligning Machine Intelligence…*, Soares & Fallenstein

* Impact time scale estimates are my own judgments based on proximity and difficulty of applications.

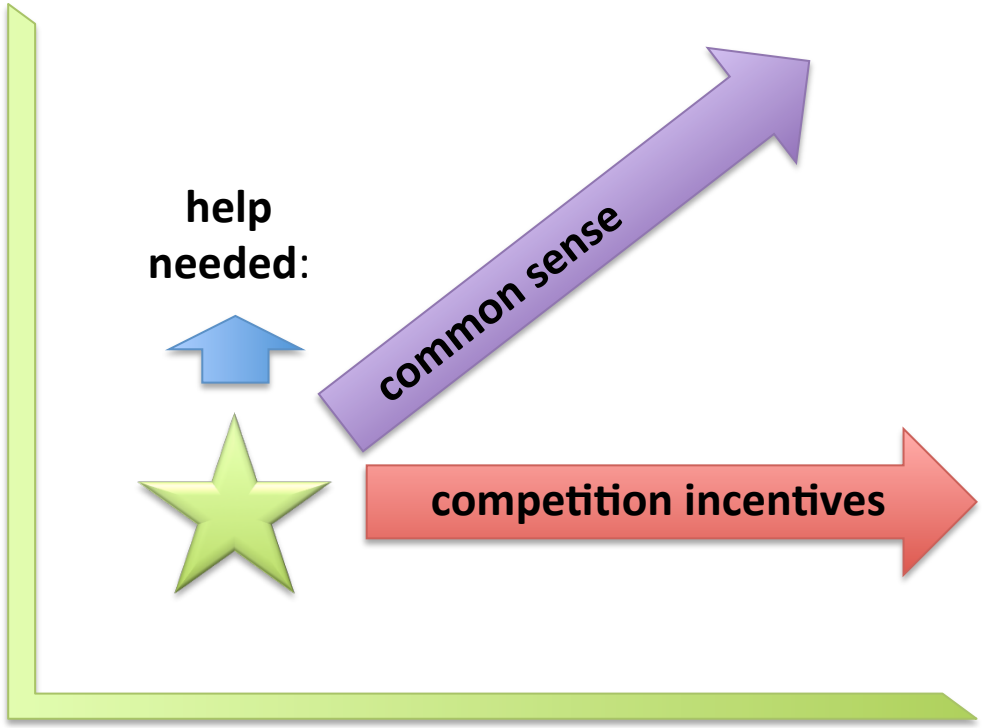
… and a number of research hotspots attempting to address the issues:

Location	Research group	Strengths
San Francisco area	**Machine Intelligence Research Institute** (Berkeley)	Expanding theoretical foundations (probability theory, game theory, …)
	Center for Human-Compatible AI (UC Berkeley)	Expanding theoretical foundations (cooperative inverse reinforcement, …)
	OpenAI (San Francisco)	Working closely with engineers and current state-of-the-art
London area	**Google DeepMind** (London)	Working closely with engineers and current state-of-the-art
	Future of Humanity Institute (Oxford)	Broad view of AI impacts, considering law, policy, and governance
	Leverhulme Center for the Future of Intelligence (Oxford/Cambridge)	Broad view of AI impacts, considering law, policy, and governance
	Center for the Study of Existential Risk (Cambridge)	Broad view of existential risks in general

But on net, very little alignment research is happening…

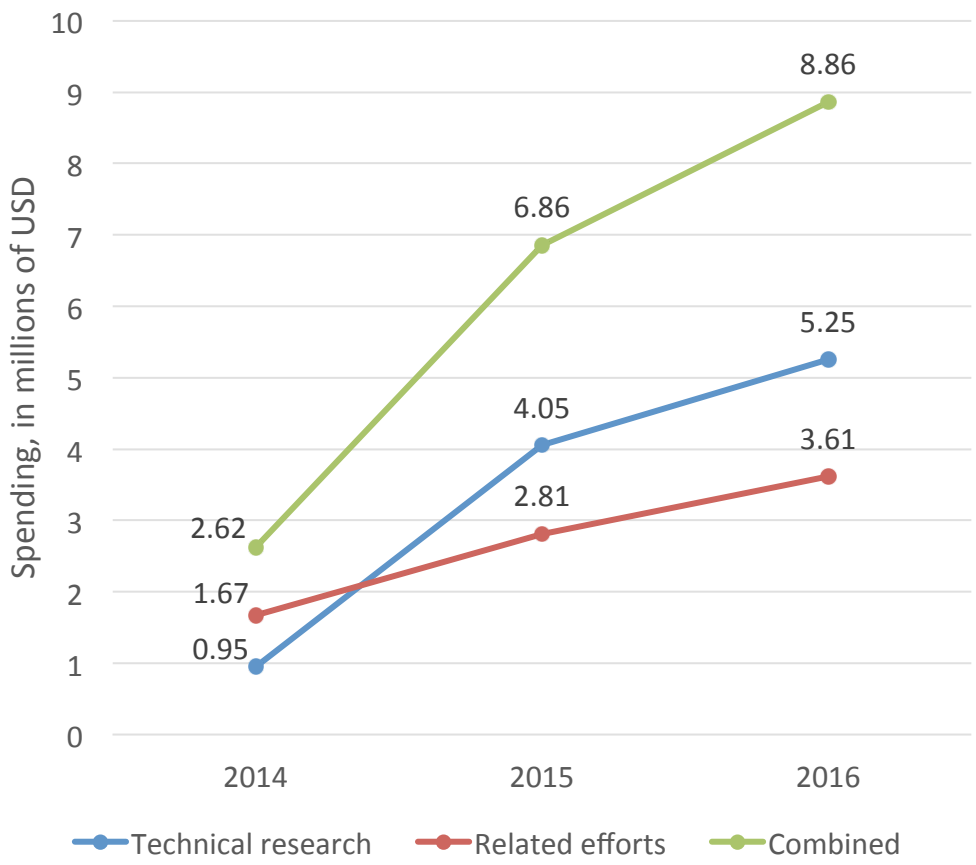
… because the world's leading AI teams need to focus on beating intelligence benchmarks to compete with each other for top talent.

AI alignment:
how to align a superintelligence with human interests



AI capabilities:
how to build a superintelligence

Worldwide spending on AI control / alignment:



Technical research: Distributed across **24 projects** with a median annual budget of **$100k**.

Related efforts: Distributed across **20 projects** with a median annual budget of **$57k.**

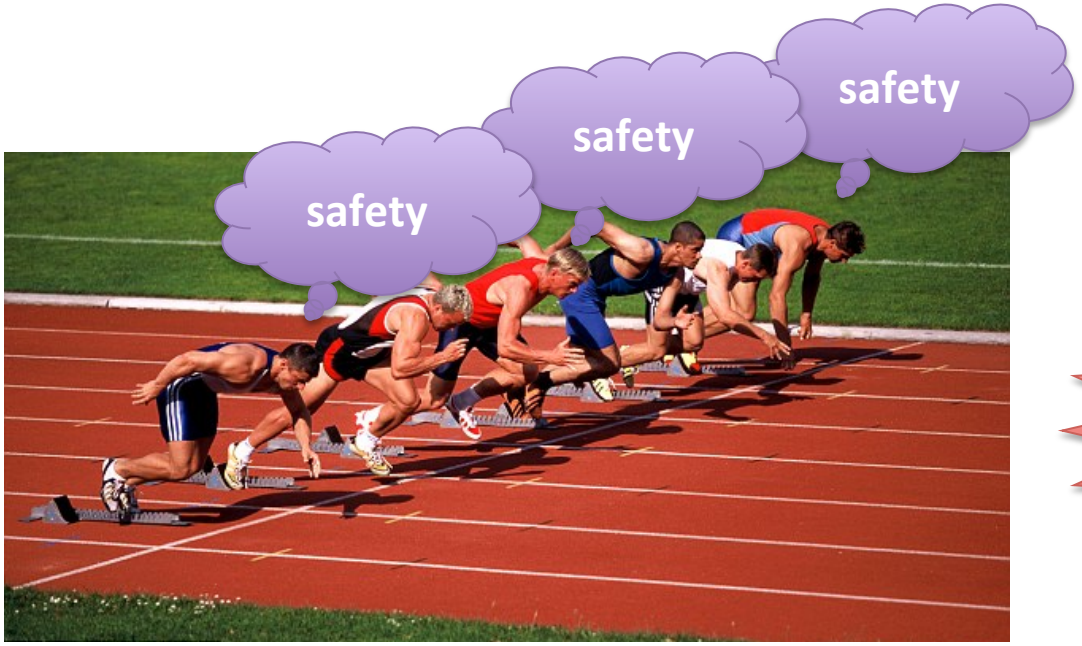
* Thanks to Sebastian Farquhar at the Global Priorities Project for the data

This makes sense: sacrificing researcher-hours to work on AI alignment means giving up some of your competitive edge over other teams.

Even if you care about the long-term future, you want **your team** to be ahead of the pack because you trust **yourself** with the responsible development of human-level AI more than you trust your competitors.

alignment

hey guys…
safety is this way…



intelligence

Summary:

1. **Human-level AI (HLAI) will probably exist** within the next several decades. It will likely be simpler than a human brain, and be followed shortly thereafter by super-human AI (SHAI).
2. **SHAI has the potential to dominate humans** in shaping the future if/when it exists (*cue: chimpanzees*)
3. **You probably need SHAI to be aligned** with (your) human interests if you wish to affect the future more than a few decades from now (*cue: children, retirement, the environment, space travel*),
4. **Alignment research is currently highly neglected** because today an AI research team's best strategy for attracting top talent is to focus on beating intelligence benchmarks instead of alignment.
5. Interventions needed:
    1. **More technical work** specifically on superintelligence alignment, and
    2. **Incentives and institutions for cooperative development** of HLAI, so that race conditions do not crowd out safety measures.

Thanks!

http://intelligence.org/

http://humancompatible.ai/