

# Superintelligence alignment as the world's top research priority

Andrew Critch<sup>1,2</sup>

[critch@intelligence.org](mailto:critch@intelligence.org)

<sup>1</sup> Machine Intelligence Research Institute

<http://intelligence.org/>

<sup>2</sup> 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?



⊆



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



€



# A 30-year thought experiment:



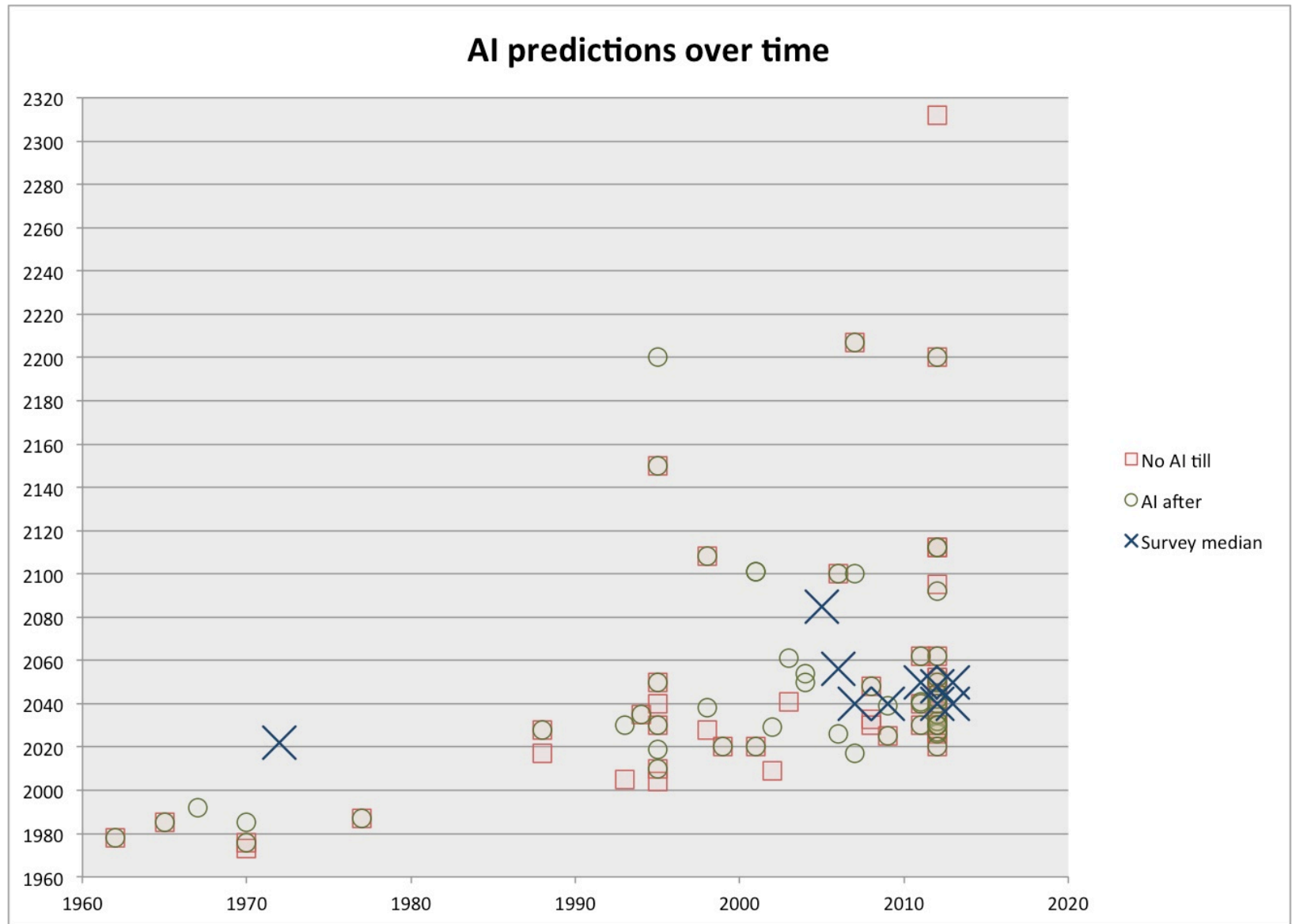
Imagine if Earth received a credible message from 30 light-years away that said:

“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/> :

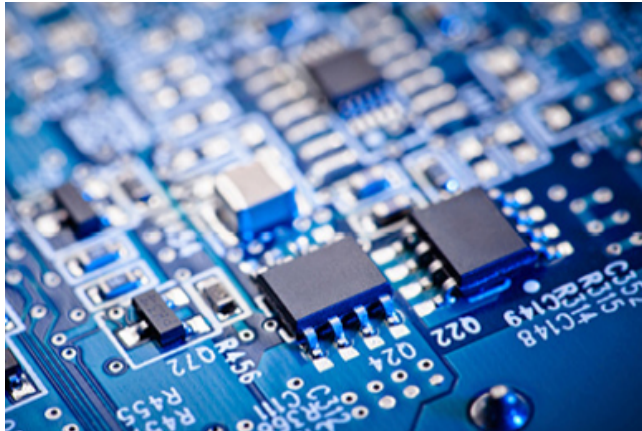


Brain model	CPU demand (FLOPS)	\$1MM availability via supercomputer / commodity computer	Memory (TB)	\$1MM availability
<b>analog network population model</b>	$10^{14}$	2008 / 2023	$10^2$	present
<b>spiking neural network</b>	$10^{18}$	2019 / 2042	$10^4$	present
<b>electrophysiology</b>	$10^{22}$	2033 / 2068	$10^4$	2019
<b>states of protein complexes</b>	$10^{27}$	2052 / 2100	$10^8$	2038
<b>stochastic behavior of single molecules</b>	$10^{43}$	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](http://www.fhi.ox.ac.uk/reports/2008-3.pdf)



# hardware performance $\neq$ software performance



$\neq$



but our source code isn't so long...



$<$

750 MB

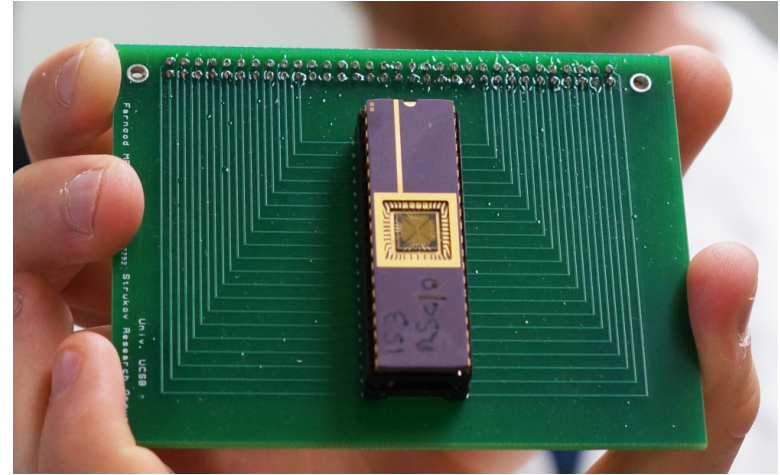
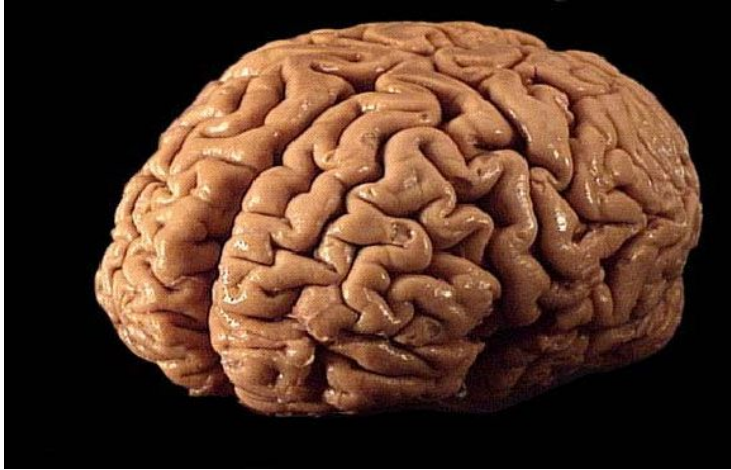
$<$



# HLAI → SHAI



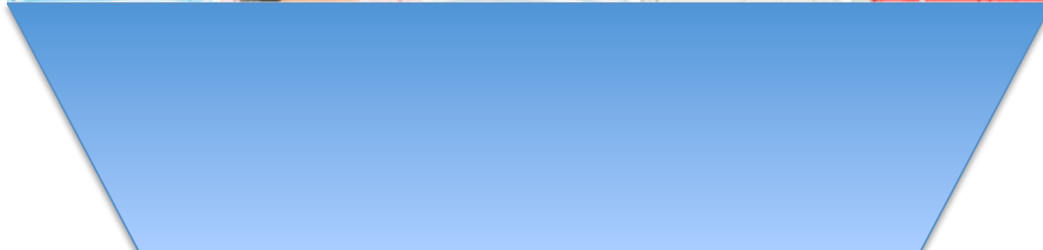
# HLAI → SHAI



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."

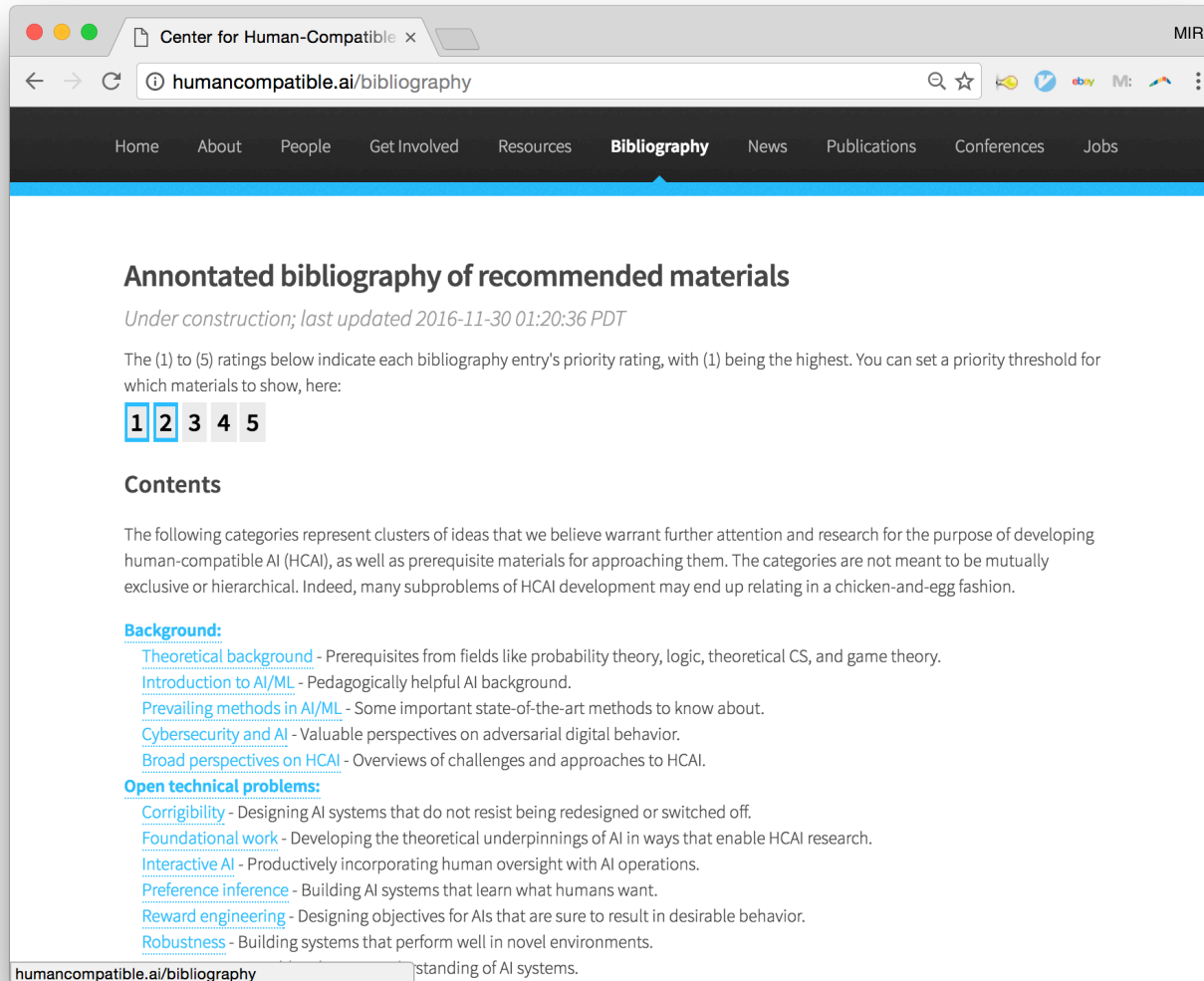
What percentage of the **world's greatest minds** would be worth allocating to prepare for that scenario?

**1%? 10%? 90%?**



# There are numerous open problem areas:

<http://humancompatible.ai/bibliography>



The screenshot shows a web browser window with the URL [humancompatible.ai/bibliography](http://humancompatible.ai/bibliography). The page has a dark navigation bar with links for Home, About, People, Get Involved, Resources, Bibliography (highlighted), News, Publications, Conferences, and Jobs. The main content area features the heading "Annotated bibliography of recommended materials" with a sub-note "Under construction; last updated 2016-11-30 01:20:36 PDT". Below this is a paragraph explaining the (1) to (5) priority ratings and a filter interface with buttons for 1, 2, 3, 4, and 5. The "Contents" section lists several categories with brief descriptions: Background (Theoretical background, Introduction to AI/ML, Prevailing methods in AI/ML, Cybersecurity and AI, Broad perspectives on HCAI), Open technical problems (Corrigibility, Foundational work, Interactive AI, Preference inference, Reward engineering, Robustness).

Center for Human-Compatible AI x MIRI

← → ↻ ⓘ humancompatible.ai/bibliography 🔍 ☆ 📧 🌐 📺 M: 🌈 ☰

Home About People Get Involved Resources **Bibliography** News Publications Conferences Jobs

## Annotated bibliography of recommended materials

*Under construction; last updated 2016-11-30 01:20:36 PDT*

The (1) to (5) ratings below indicate each bibliography entry's priority rating, with (1) being the highest. You can set a priority threshold for which materials to show, here:

**1** 2 3 4 5

### Contents

The following categories represent clusters of ideas that we believe warrant further attention and research for the purpose of developing human-compatible AI (HCAI), as well as prerequisite materials for approaching them. The categories are not meant to be mutually exclusive or hierarchical. Indeed, many subproblems of HCAI development may end up relating in a chicken-and-egg fashion.

#### Background:

- [Theoretical background](#) - Prerequisites from fields like probability theory, logic, theoretical CS, and game theory.
- [Introduction to AI/ML](#) - Pedagogically helpful AI background.
- [Prevailing methods in AI/ML](#) - Some important state-of-the-art methods to know about.
- [Cybersecurity and AI](#) - Valuable perspectives on adversarial digital behavior.
- [Broad perspectives on HCAI](#) - Overviews of challenges and approaches to HCAI.

#### Open technical problems:

- [Corrigibility](#) - Designing AI systems that do not resist being redesigned or switched off.
- [Foundational work](#) - Developing the theoretical underpinnings of AI in ways that enable HCAI research.
- [Interactive AI](#) - Productively incorporating human oversight with AI operations.
- [Preference inference](#) - Building AI systems that learn what humans want.
- [Reward engineering](#) - Designing objectives for AIs that are sure to result in desirable behavior.
- [Robustness](#) - Building systems that perform well in novel environments.

humancompatible.ai/bibliography standing of AI systems.

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

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

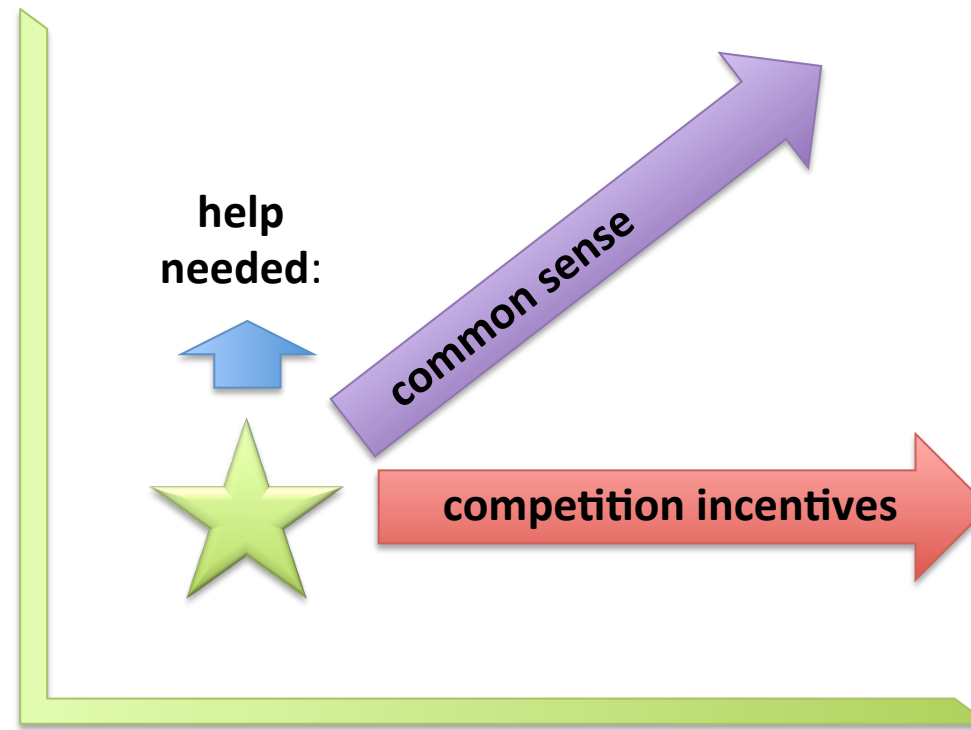
Location	Research group	Strengths
San Francisco area	<b>Machine Intelligence Research Institute</b> (Berkeley)	Expanding theoretical foundations (probability theory, game theory, ...)
	<b>Center for Human-Compatible AI</b> (UC Berkeley)	Expanding theoretical foundations (cooperative inverse reinforcement,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

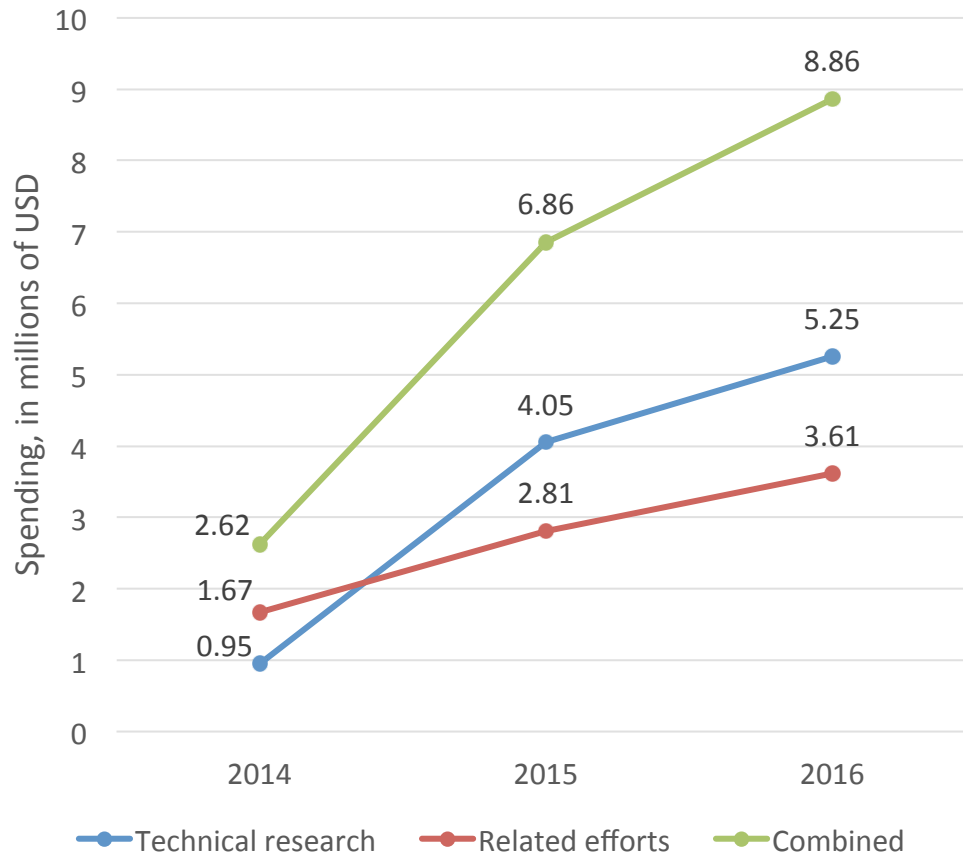
how to align a superintelligence with human interests



## 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 HLAIs, so that race conditions do not crowd out safety measures.

# Thanks!

<http://intelligence.org/>

<http://humancompatible.ai/>

