MIRI

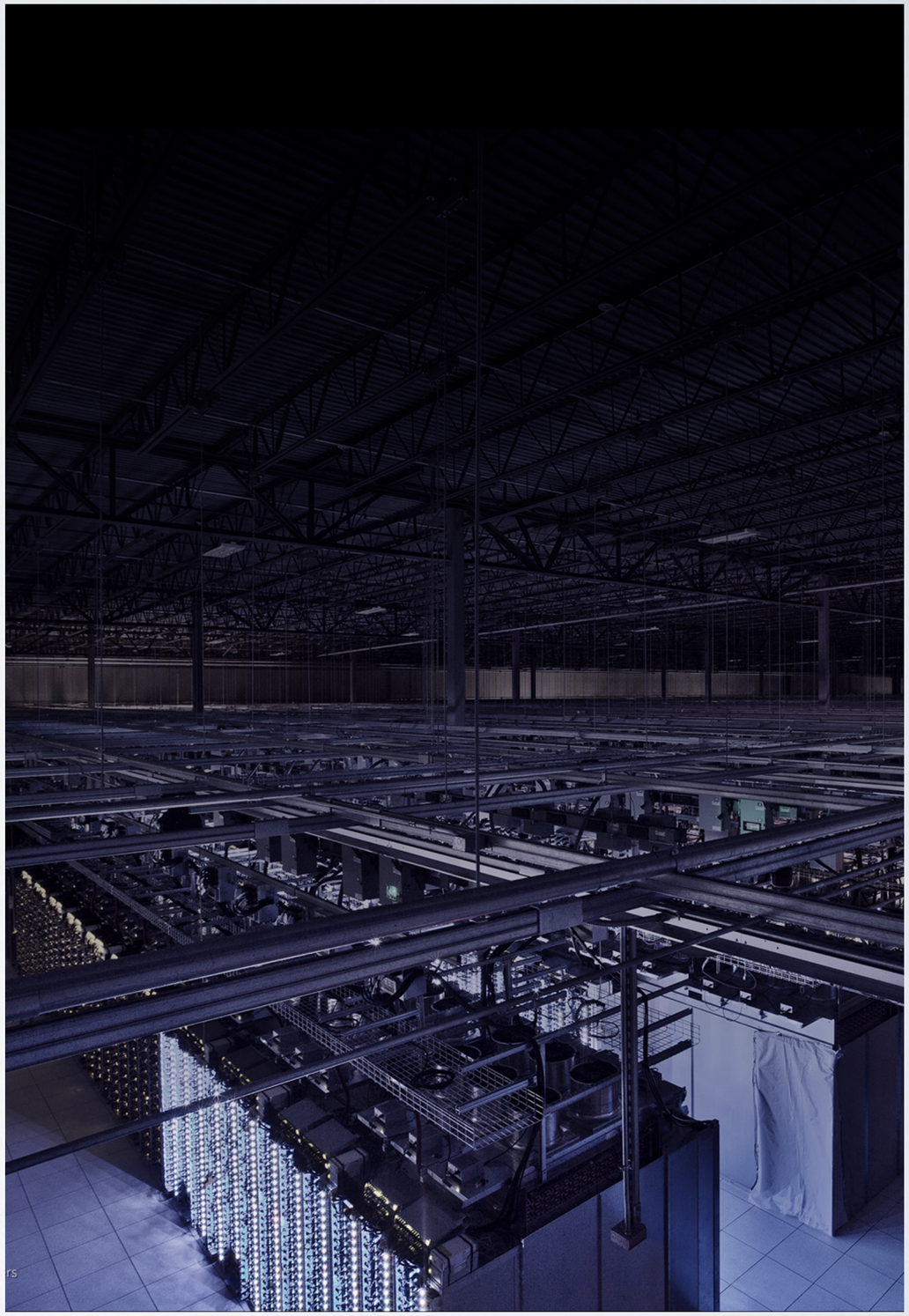
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

Mission:

"to ensure that the creation of smarter-than-human intelligence has a positive impact"

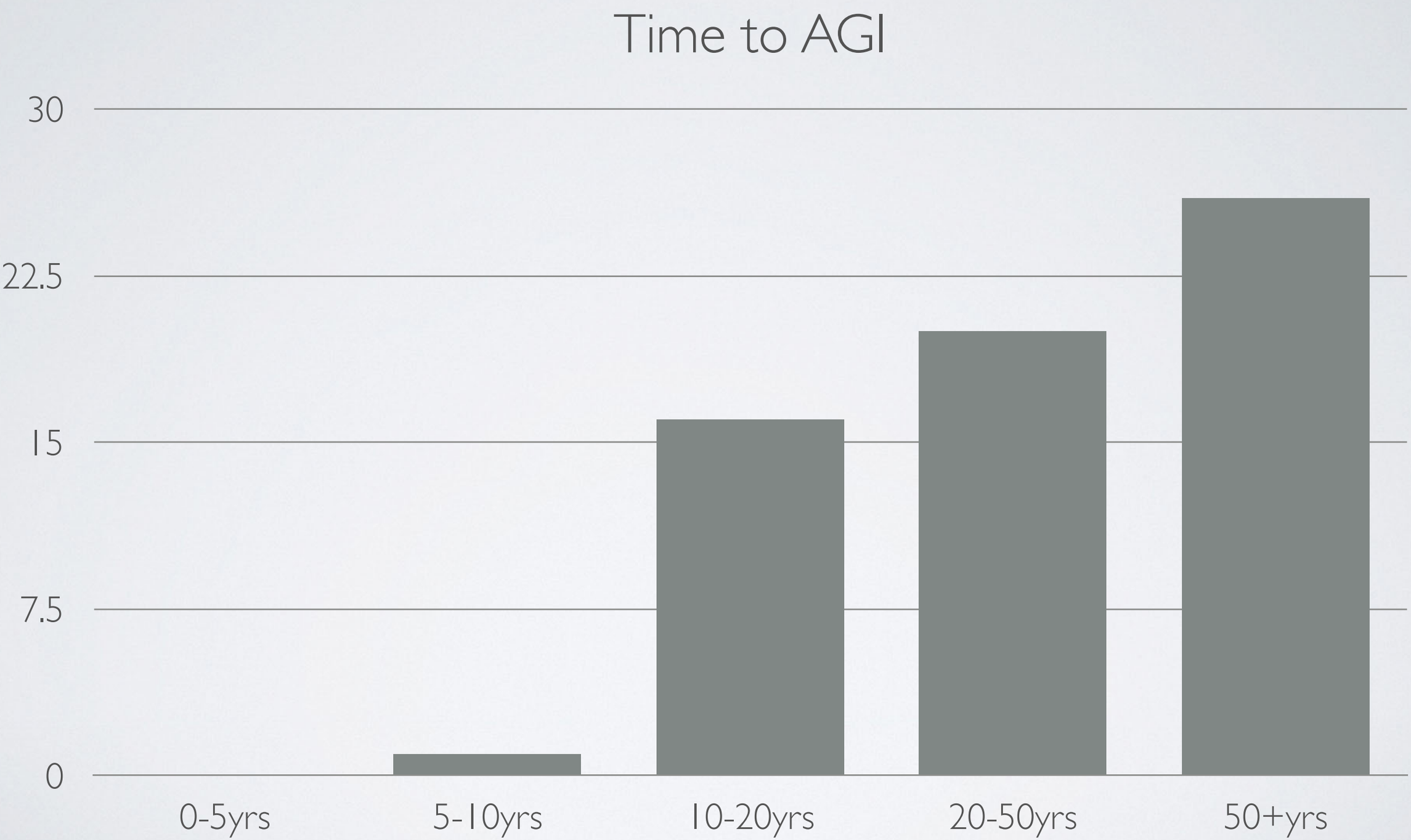
Narrow AI

AGI



Expert opinion on AGI

Michie (1973), n=63:



Expert opinion on AGI

"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?"

Müller & Bostrom (2014), n=29:

	10%	50%	90%
median response	2024	2050	2070

Luke's estimate:

	10%	50%	90%
Luke	2030	2070	2140

See Muehlhauser, "When Will AI Be Created?" (2013)

Expert opinion on AGI

"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]?"

Müller & Bostrom (2014), n=29:

	2 years	30 years
median response	5%	50%

Luke's estimate:

	2 years	30 years
Luke	15%	85%

See Bostrom (2014) ch. 4 and Yudkowsky, "Intelligence Explosion Microeconomics"

Expert opinion on AGI

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

Müller & Bostrom (2014), n=29:

	Extremely good	On balance good	Neutral-ish	On balance bad	Extremely bad
mean	20%	40%	19%	13%	8%

Luke's estimate (on Tuesday and Thursdays, still volatile):

	Extremely good	On balance good	Neutral-ish	On balance bad	Extremely bad
Luke	19%	1%	~0%	5%	75%

See Bostrom (2014) and stuff that hasn't been published yet

What's the EA case for efforts toward Friendly AI?

1. **Astronomical stakes:** Most QALYs are in the long-term future, so what matters most is that we shape the trajectory of the long-term future in a robustly positive direction. See Beckstead's PhD thesis, and Yudkowsky's talk at 10:40am tomorrow.
2. **AI is the key lever on the long-term future:** See Bostrom's *Superintelligence* (2014).
3. **Friendly AI work is urgent, tractable, and uncrowded:** 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. FAI work is also, it turns out, tractable and uncrowded. See my talk at 4pm tomorrow.

What's the EA case for efforts toward Friendly AI?

1. Astronomical stakes

QALYs we can produce in the long-term future

Eliezer: "If you occupy the incredibly rare and leverage-privileged position of being born into pre-AGI Earth, then the best thing you can do is to help make sure the reachable universe is converted into quality-adjusted life years."

QALYs we can produce in the next century

(Warning: dot may be too small to see on screens smaller than Saturn.)

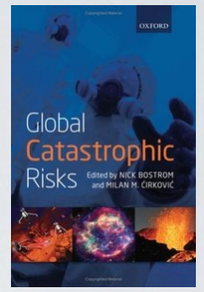
Yes, Friendly AI work is more speculative than, say, Deworm the World.

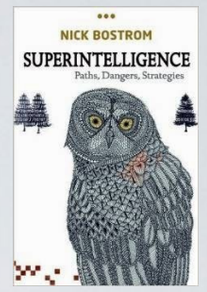
But if you take astronomical stakes seriously, then…

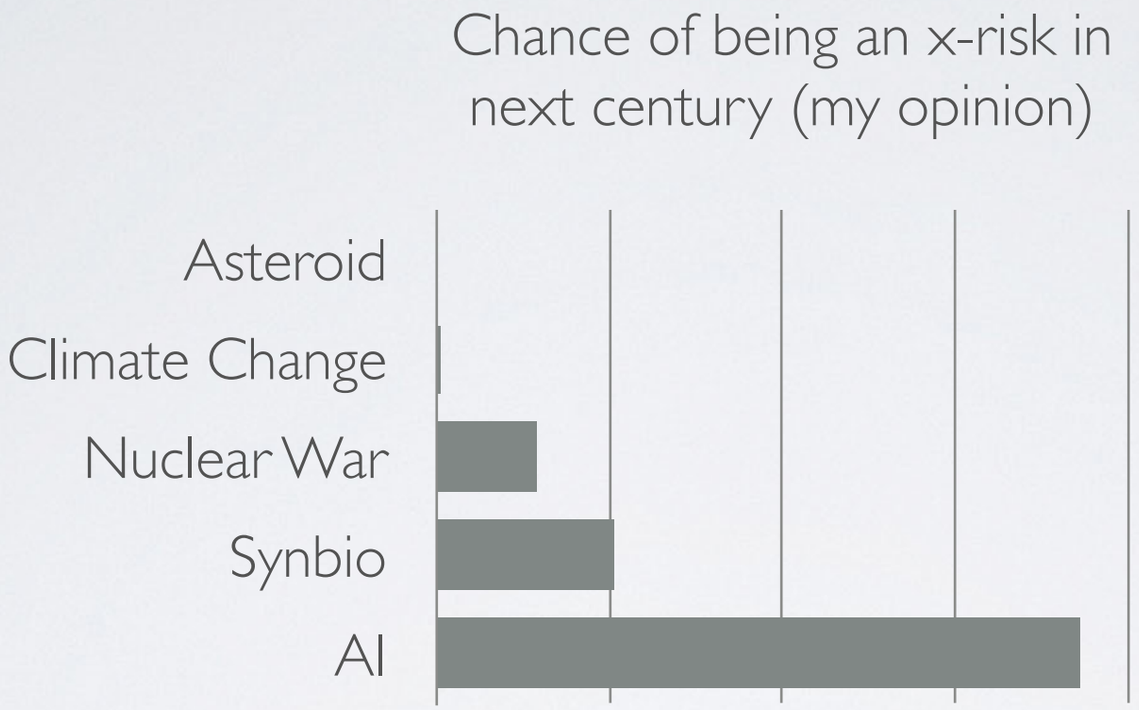
- FHI?
- MIRI?
- Another organization aimed at astronomical value?
- A new organization aimed at astronomical value?
- Some other thing that was never purposely aimed at astronomical value but happens to be optimal by accident?

What's the EA case for efforts toward Friendly AI?

2. AI as the key lever on the long-term future







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.

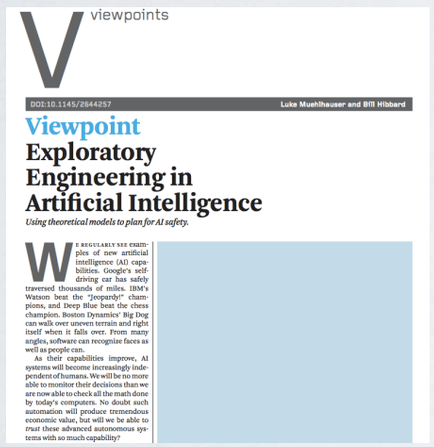
What's the EA case for efforts toward Friendly AI?

3. Friendly AI work is urgent, tractable, and uncrowded

Uhhhhhhh… see my other talk. :)

So what does MIRI actually *do*?

Workshops


viewpoints

DOI:10.1145/2644257 Luke Muehlhauser and Bill Hibbard

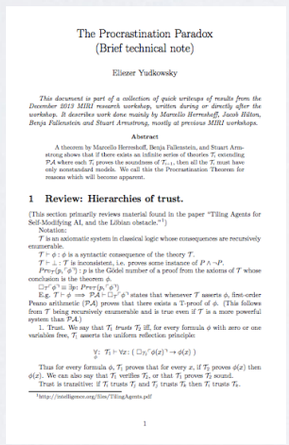
Viewpoint

Exploratory Engineering in Artificial Intelligence

Using theoretical models to plan for AI safety.

WE REGULARLY SEE EXAMPLES of new artificial intelligence (AI) capabilities. Google's self-driving car has safely traversed thousands of miles. IBM's Watson beat the "Jeopardy!" champions, and Deep Blue beat the chess champion, Boston Dynamics' Big Dog can walk over uneven terrain and right itself when it falls over. From many angles, software can recognize faces as well as people can.

As their capabilities improve, AI systems will become increasingly independent of humans. We will be no more able to monitor their decisions than we are now able to check all the math done by today's computers. No doubt such automation will produce tremendous economic value, but will we be able to *trust* these advanced autonomous systems with so much capability?


The Procrastination Paradox
(Brief technical note)

Eliezer Yudkowsky

This document is part of a collection of quick writeups of results from the December 2013 MIRI research workshop, written during or directly after the workshop. It describes work done mainly by Marcello Herreshoff, Jacob Hilton, Benja Fallenstein and Stuart Armstrong, mostly at previous MIRI workshops.

Abstract

A theorem by Marcello Herreshoff, Benja Fallenstein, and Stuart Armstrong shows that if there exists an infinite series of theories T_i extending PA where each T_i proves the soundness of T_{i-1}, then all the T_i must have only nonstandard models. We call this the Procrastination Theorem for reasons which will become apparent.

1 Review: Hierarchies of trust.

(This section primarily reviews material found in the paper "Tiling Agents for Self-Modifying AI, and the Löbian obstacle."[1])

Notation:

$\mathcal{T}$ is an axiomatic system in classical logic whose consequences are recursively enumerable.

$\mathcal{T} \vdash \phi$ is a syntactic consequence of the theory $\mathcal{T}$.

$\mathcal{T} \vdash_{n-1} \mathcal{T}$ is inconsistent, i.e. proves some instance of $P \wedge \neg P$.

$Prv_{\mathcal{T}}(\lceil p \rceil, \lceil \phi \rceil) : p$ is the Gödel number of a proof from the axioms of $\mathcal{T}$ whose conclusion is the theorem ϕ.

$\Box_{\mathcal{T}} \lceil \phi \rceil \equiv \exists p : Prv_{\mathcal{T}}(p, \lceil \phi \rceil)$

E.g. $\mathcal{T} \vdash \phi \iff PA \vdash \Box_{\mathcal{T}} \lceil \phi \rceil$ states that whenever $\mathcal{T}$ asserts ϕ, first-order Peano arithmetic (PA) proves that there exists a $\mathcal{T}$-proof of ϕ. (This follows from $\mathcal{T}$ being recursively enumerable and is true even if $\mathcal{T}$ is a more powerful system than PA.)

1. Trust. We say that T_1 *trusts* T_2 iff, for every formula ϕ with zero or one variables free, T_1 asserts the uniform reflection principle

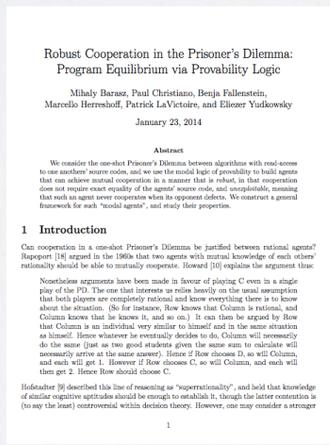
$$\forall_x : T_1 \vdash \forall x : (\exists_n \sigma(x)^{\lceil n \rceil} \rightarrow \phi(x))$$

Thus for every formula ϕ, T_1 proves that for every x, if T_2 proves $\phi(x)$ then $\phi(x)$. We can also say that T_1 verifies T_2, or that T_1 proves T_2 sound.

Trust is tran

[1] http://intelligence.org/files/TilingAgents.pdf

1


Robust Cooperation in the Prisoner's Dilemma:
Program Equilibrium via Provability Logic

Mihaly Barasz, Paul Christiano, Benja Fallenstein,
Marcello Herreshoff, Patrick LaVictoire, and Eliezer Yudkowsky

January 23, 2014

Abstract

We consider the one-shot Prisoner's Dilemma between algorithms with read-access to one another's source codes, and we use the modal logic of provability to build agents that can achieve mutual cooperation in a manner that is robust, in that cooperation does not require exact equality of the agents' source code, and *unexploitable*, meaning that such an agent never cooperates when its opponent defects. We construct a general framework for such "modal agents", and study their properties.

1 Introduction

Can cooperation in a one-shot Prisoner's Dilemma be justified between rational agents? Rapoport ([18] argued in the 1960s that two agents with mutual knowledge of each others' rationality should be able to mutually cooperate. Howard [10] explains the argument thus:

> Nonetheless arguments have been made in favour of playing C even in a single play of the PD. The one that interests us relies heavily on the usual assumption that both players are completely rational and know everything there is to know about the situation. (So for instance, Row knows that Column is rational, and Column knows that he knows it, and so on.) It can then be argued by Row that Column is an individual very similar to himself and in the same situation as himself. Hence whatever he eventually decides to do, Column will necessarily do the same (just as two good students given the same sum to calculate will necessarily arrive at the same answer). Hence if Row chooses D, so will Column, and each will get 1. However if Row chooses C, so will Column, and each will then get 2. Hence Row should choose C.

Hofstadter [9] described this line of reasoning as "superrationality", and held that knowledge of similar cognitive aptitudes should be enough to establish it, though the latter contention is (to say the least) controversial within decision theory. However, one may consider a stronger

1


Problems of self-reference in self-improving space-time embedded intelligence

Benja Fallenstein and Nate Soares

Machine Intelligence Research Institute
2030 Addison St. #300, Berkeley, CA 94704, USA
{benja,nate}@intelligence.org

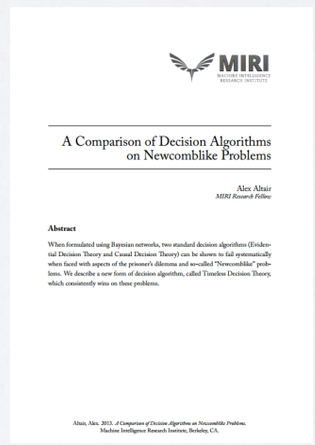
Abstract. By considering agents to be a part of their environment, Orseau and Ring's *space-time embedded intelligence* [10] is a better fit to the real world than the traditional agent framework. However, a self-modifying AGI that sees future versions of itself as an ordinary part of the environment may run into problems of self-reference. We show that in one particular model based on formal logic, naive approaches either lead to incorrect reasoning that allows an agent to put off an important task forever (the *procrastination paradox*), or fail to allow the agent to justify even obviously safe rewrites (the *Löbian obstacle*). We argue that these problems have relevance beyond our particular formalism, and discuss partial solutions.

1 Introduction

Most formal models of artificial general intelligence (such as Hutter's AIXI [5] and the related formal measure of intelligence [6]) are based on the traditional agent framework, in which the agent interacts with an environment, but is not part of this environment. As Orseau and Ring [10] point out, this is reminiscent of Cartesian dualism, the idea that the human mind is a non-physical substance external to the body [11]. A real-world AGI, on the other hand, will be part of the physical universe, and will need to deal with the possibility that external forces might observe or interfere with its internal operations.

The traditional separation of the agent from its environment seems even less attractive when one considers I.J. Good's idea that once AGI is sufficiently advanced, it may become better than any human at the task of *making itself even smarter*, leading to an "intelligence explosion" and leaving human intelligence far behind [7]. It seems plausible that an AGI undergoing an intelligence explosion may eventually want to adopt an architecture radically different from its initial one, such as one distributed over many different computers, where no *single* entity fulfills the agent's role from the traditional framework [8]. A formal model based on that framework cannot capture this.

How should one reason about such an agent? Orseau and Ring [10] have proposed a formal model of *space-time embedded intelligence* to deal with this complexity. Their model consists of a set Π of policies, describing the state of the agent at a given point in time; an environment $\rho(s_{t+1} \mid s_{1:t})$, giving the


MIRI
MACHINE INTELLIGENCE RESEARCH INSTITUTE

A Comparison of Decision Algorithms on Newcomblike Problems

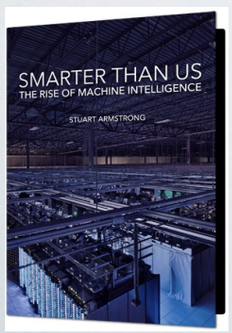
Alex Altair
MIRI Research Fellow

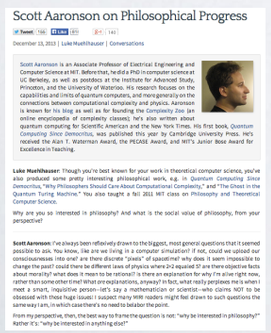
Abstract

When formulated using Bayesian networks, two standard decision algorithms (Evidential Decision Theory and Causal Decision Theory) can be shown to fail systematically when faced with aspects of the prisoner's dilemma and so-called "Newcomblike" problems. We describe a new form of decision algorithm, called Timeless Decision Theory, which consistently wins on these problems.

Altair, Alex. 2013. *A Comparison of Decision Algorithms on Newcomblike Problems.* Machine Intelligence Research Institute, Berkeley, CA.

Papers & reports




Scott Aaronson on Philosophical Progress

December 13, 2013 | Luke Muehlhauser | Conversations

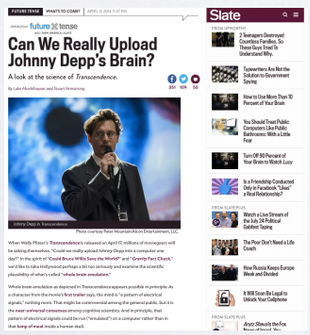
Scott Aaronson is an Associate Professor of Electrical Engineering and Computer Science at MIT. Before that, he did a PhD in computer science at UC Berkeley, as well as postdocs at the Institute for Advanced Study, Princeton, and the University of Waterloo. His research focuses on the capabilities and limits of quantum computers, and more generally on the connections between computational complexity and physics. Aaronson is known for his blog as well as for founding the Complexity Zoo (an online encyclopedia of complexity classes); he's also written about quantum computing for Scientific American and the New York Times. His first book, *Quantum Computing Since Democritus*, was published this year by Cambridge University Press. He's received the Alan T. Waterman Award, the PECASE Award, and MIT's Junior Bose Award for Excellence in Teaching.

Luke Muehlhauser: Though you're best known for your work in theoretical computer science, you've also produced some pretty interesting philosophical work, e.g. in *Quantum Computing Since Democritus*, "Why Philosophers Should Care About Computational Complexity," and "The Ghost in the Quantum Turing Machine." You also taught a fall 2011 MIT class on Philosophy and Theoretical Computer Science.

Why are you so interested in philosophy? And what is the social value of philosophy, from your perspective?

Scott Aaronson: I've always been reflexively drawn to the biggest, most general questions that it seemed possible to ask. You know, like are we living in a computer simulation? If not, could we upload our consciousnesses into one? are there discrete "pixels" of spacetime? why does it seem impossible to change the past? could there be different laws of physics where 2+2 equaled 5? are there objective facts about morality? what does it mean to be rational? is there an explanation for why I'm alive right now, rather than some other time? What are explanations, anyway? In fact, what really perplexes me is when I meet a smart, inquisitive person—let's say a mathematician or scientist—who claims NOT to be obsessed with these huge issues! I suspect many MIRI readers might feel drawn to such questions the same way I am, in which case there's no need to belabor the point.

From my perspective, then, the best way to frame the question is not: "why be interested in philosophy?" Rather it's: "why be interested in anything else?"


Slate

Can We Really Upload Johnny Depp's Brain?

A look at the science of *Transcendence*.

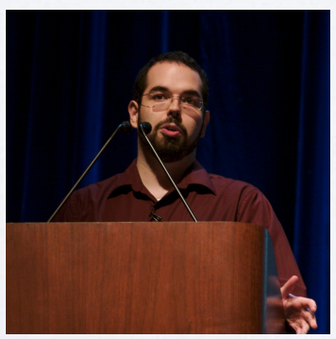
By Luke Muehlhauser and Sean Armstrong

Johnny Depp in Transcendence

Photo courtesy Peter Mountain/Alcon Entertainment, LLC.

When Wally Pfister's *Transcendence* is released on April 18, millions of moviegoers will be asking themselves, "Could we really upload Johnny Depp into a computer one day?" In the spirit of "Could Bruce Willis Save the World?" and "Gravity Fact Check," we'd like to take Hollywood perhaps a bit too seriously and examine the scientific plausibility of what's called "whole brain emulation."

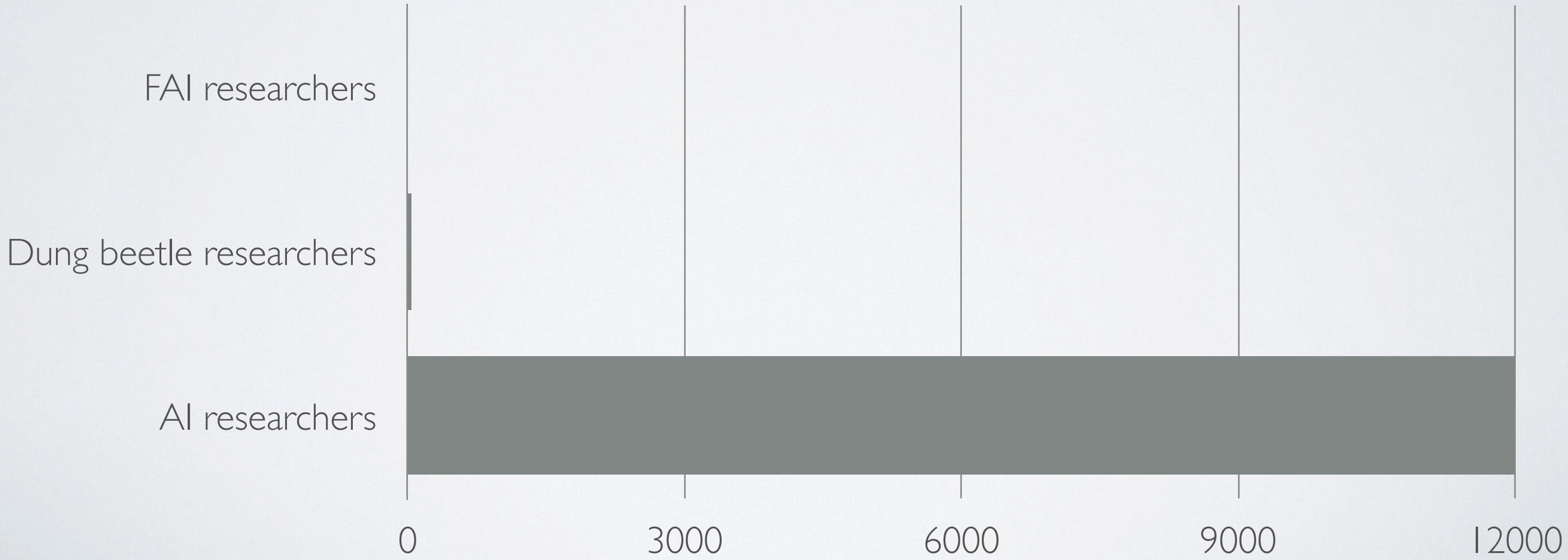
Whole brain emulation as depicted in *Transcendence* appears possible in principle. As a character from the movie's first trailer says, the mind is "a pattern of electrical signals," nothing more. That might be controversial among the general public, but it is the near-universal consensus among cognitive scientists. And in principle, that pattern of electrical signals could be run ("emulated") on a computer rather than in that lump of meat inside a human skull.



Some other stuff

Marginal dollars are mostly spent on: finding/creating new FAI researchers.

Because:



MIRI

MACHINE INTELLIGENCE
RESEARCH INSTITUTE

For more information,
visit intelligence.org
or come talk to me.