Artificial Intelligence and Human Existence

Andrew Critch, Research Scientist,
UC Berkeley, Department of Electrical Engineering and Computer Science,
Center for Human Compatible AI
http://humancompatible.ai/

My two hats:

Specialist hat:

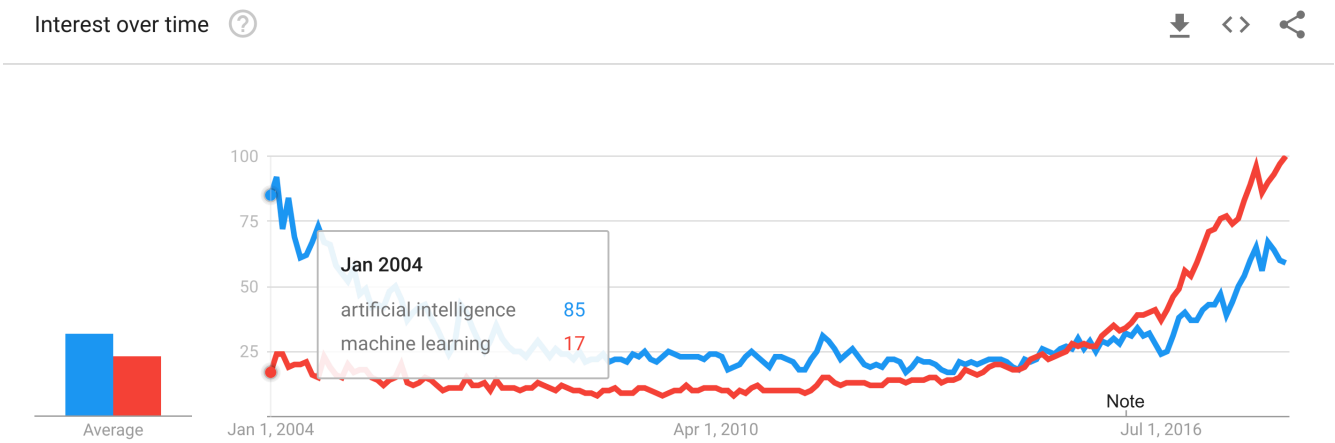
Bounded rationality, open-source game theory, algorithms that reason about algorithms, "negotiable RL" …

Generalist hat:

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

…

AI / machine learning are getting hot (again)



Google trends comparison on 13/04/2017;
y-axis is a percentage relative to the peak (present)

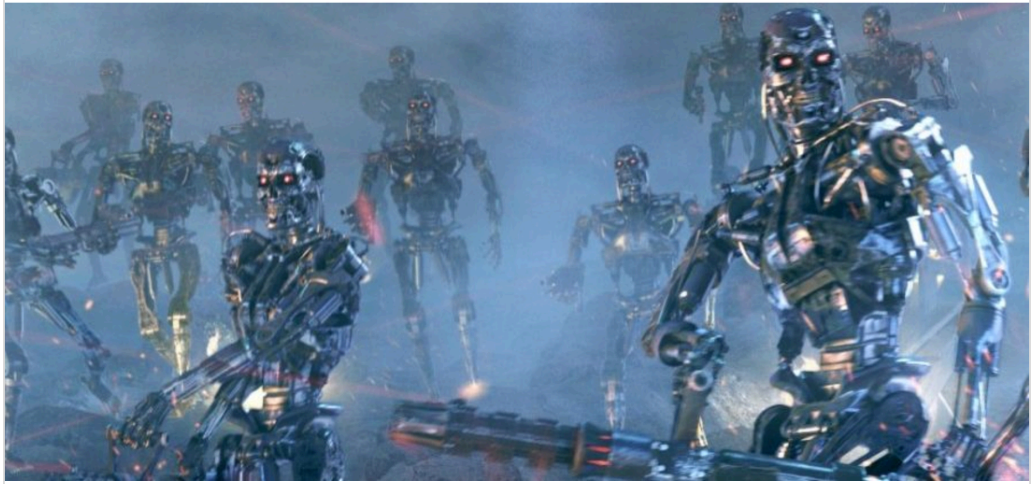
I think this is mostly awesome news, because AI/ML promises to be extremely powerful and useful.

New AI think-tank hopes to get real on Existential Risk

0

Chris Davies - Nov 26, 2012

Facebook Twitter Google Reddit



A future where humanity is subjugated by AIs, hunted down by robots, or consumed by nanobot goo may be the stuff of today's sci-fi, but it should also

minded trio – made up of a scientist, a philosopher, and a software engineer – has proposed a think-tank dedicated to so-called "extinction level" threats of our own creation; the proposed Centre for the Study of Existential Risk (CSER) would examine the potential perils involved in today's cutting-edge research.

Public discussion of existential risks and artificial intelligence has been reactionary, poorly argued, and rife with terminator images…

Hopefully, at least when we get to the question period of this talk, you can help me **raise the quality of the conversation**.

So, let's start with something easier to agree on.

Humans are the dominant species on the planet because of our intelligence

… not because we're the strongest, the fastest, or the biggest; grizzly bears are all these things, yet their fate is in our hands.

We haven't seen any messages from aliens yet, but we have received a terrestrial one:

When Will AI Exceed Human Performance? Evidence from AI Experts

Katja Grace[1,2], John Salvatier[2], Allan Dafoe[1,3], Baobao Zhang[3], and Owain Evans[1]

[1] *Future of Humanity Institute, Oxford University*
[2] *AI Impacts*
[3] *Department of Political Science, Yale University*

Abstract

Advances in artificial intelligence (AI) will transform modern life by reshaping transportation, health, science, finance, and the military [1, 2, 3]. To adapt public policy, we need to better anticipate these advances [4, 5]. Here we report the results from a large survey of machine learning researchers on their beliefs about progress in AI. Researchers predict AI will outperform humans in many activities in the next ten years, such as translating languages (by 2024), writing high-school essays (by 2026), driving a truck (by 2027), working in retail (by 2031), writing a bestselling book (by 2049), and working as a surgeon (by 2053). Researchers believe there is a 50% chance of AI outperforming humans in all tasks in 45 years and of automating all human jobs in 120 years, with Asian respondents expecting these dates much sooner than North Americans. These results will inform discussion amongst researchers and policymakers about anticipating and managing trends in AI.

What are these judgements based on?

Several sources provide objective evidence about future AI advances: trends in computing hardware [7], task performance [8], and the automation of labor [9]. The predictions of AI experts provide crucial additional information. We survey a larger and more representative sample of AI experts than any study to date [10, 11]. Our questions cover the timing of AI advances (including both practical applications of AI and the automation of various human jobs), as well as the social and ethical impacts of AI.

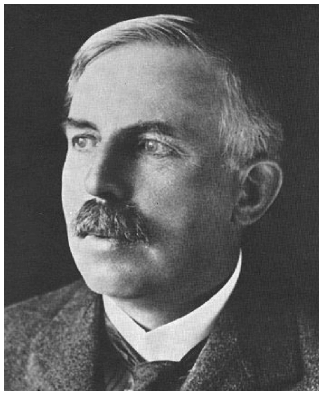
Survey Method

Our survey population was all researchers who published at the 2015 NIPS and ICML conferences (two of the premier venues for peer-reviewed research in machine learning). A total of 352 researchers responded to our survey invitation (21% of the 1634 authors we contacted). Our questions concerned the timing of specific AI capabilities (e.g. folding laundry, language translation), superiority at specific occupations (e.g. truck driver, surgeon), superiority over humans at all tasks, and the social impacts of advanced AI. See Survey Content for details.

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

Humans are not always so great at predicting progress



Sept 11, 1933: Lord Ernst Rutherford addressing BAAS: *"Anyone who looks for a source of power in the transformation of the atoms is talking moonshine."*

Sept 12, 1933: Leo Szilard invents the neutron-induced nuclear chain reaction

https://en.wikipedia.org/wiki/Leo_Szilard#Developing_the_idea_of_the_nuclear_chain_reaction

Personally, I find it quite plausible that AGI could be developed within 15 years or so.

However, rather than elaborating or defending that claim, I'd rather speak from a broader consensus, and focus on a "what if" question that does not hinge on what we mean by "AGI":

What existential risks does humanity face from AI development in the next century, and what research avenues could help to avoid those risks?

STEM PhD-holding researchers working full-time on "AGI safety"

A major motivation for me in investigating this question is that only around a dozen people in the world with a STEM PhD are working full-time on safety and control for AI systems approaching "AGI", despite the fairly obvious pathways through which AGI could pose an existential risk:

Researcher…	Since…	Currently at…
Stuart Armstrong	2011	FHI
Scott Garrabrant	2015	MIRI
Andrew Critch	2015	CHAI
Paul Christiano	2016	OpenAI
Vika Krakovna	2016	DeepMind
Jan Leike	2016	DeepMind
Pedro Ortega	2016	DeepMind
Abram Demski	2017	MIRI
(anonymous)	(anonymous)	(anonymous)
… and perhaps a handful of others		

Goal: "AI x-risk research"

Personal goal: contribute to developing a field of technical research specifically aimed at continuing the existence of the human species in tandem with the development of increasingly powerful AI systems.

This goal raises two big questions:

1) **What's the problem?** What level of AI capabilities would pose an existential risk?

2) **What can we do about it?** What technical and engineering problems are involved in averting those risks?

(1) What level of AI capabilities would pose an existential risk?

Various levels of AI capabilities have been discussed as potential turning points for the security of human existence, including “AGI : Artificial General Intelligence”, “HLAI : Human-Level AI”, “HLMI : High-Level Machine Intelligence”. For the study of existential risk, I prefer a more direct definition:

“Prepotent AI”: machine intelligence with the potential to dominate every concurrently existing collective of humans in a competition for control of physical resources.

Here, prepotence is a “know it when you see it” concept. When you look at the way humans can expand suburban housing without any collective permission from Grizzly bears, it may be tricky to define exactly in what senses we’re winning, but your naïve “the humans are winning” classifier returns “1”.

Note: In general usage, the term ``prepotent" means ``Very powerful; superior in force, influence, or authority; predominant", and carries connotations from evolutionary biology that refer to the replacement of organisms by other organisms.

Prepotence is of special significance to existential risk because if AI systems control the Earth's resources, then our survival will be determined by the way in which the AI systems use those resources.

(1) What level of AI capabilities would pose an existential risk?

Examples of capabilities potentially sufficient for prepotence:

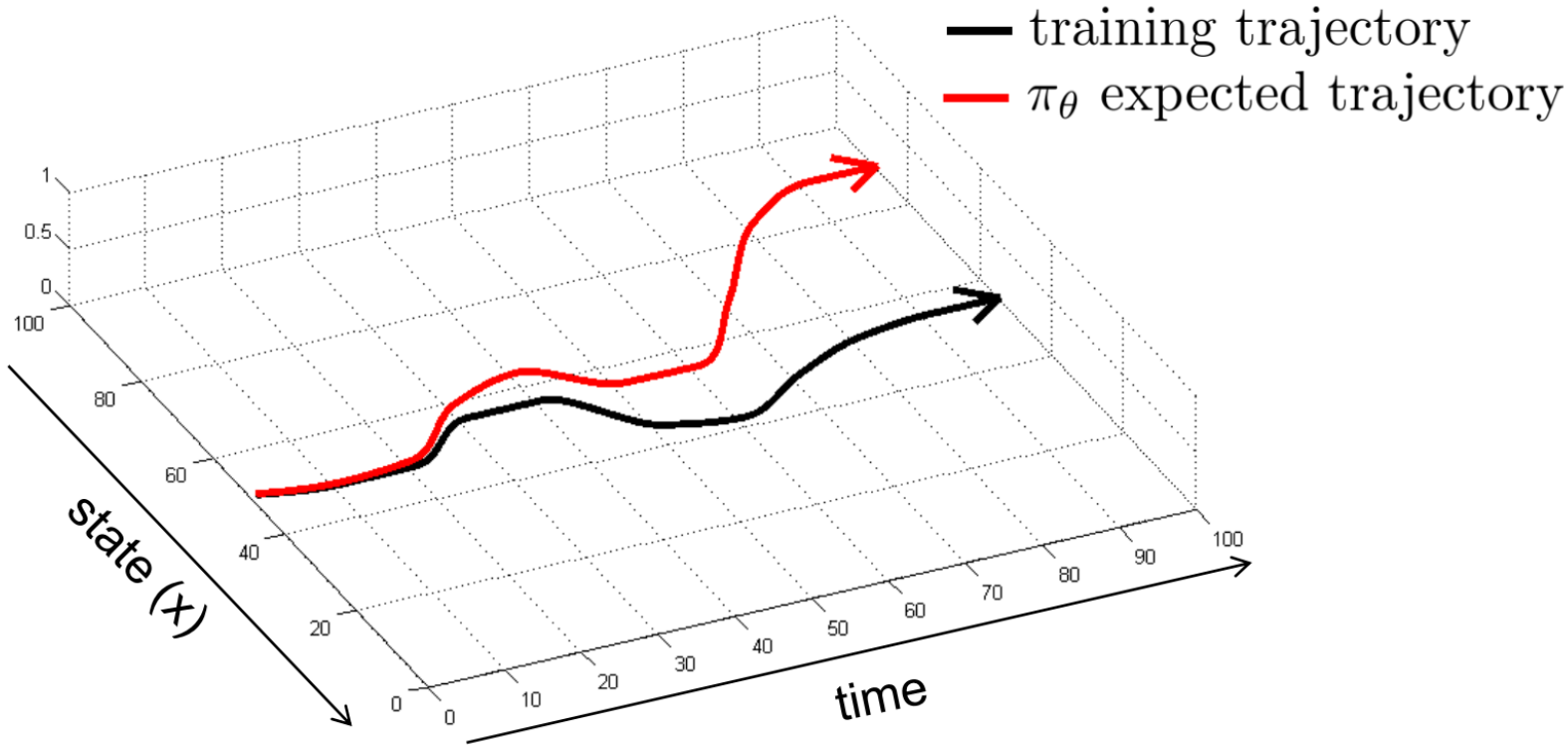
(1) Technological autonomy – analogy: European conquest of North America

(2) Replication speed – analogy: lethal viruses amid the cells of the human body / rapidly replicating machines amid the humans of an intelligent civilization.

(3) Social acumen – cue: persuasive warmongers

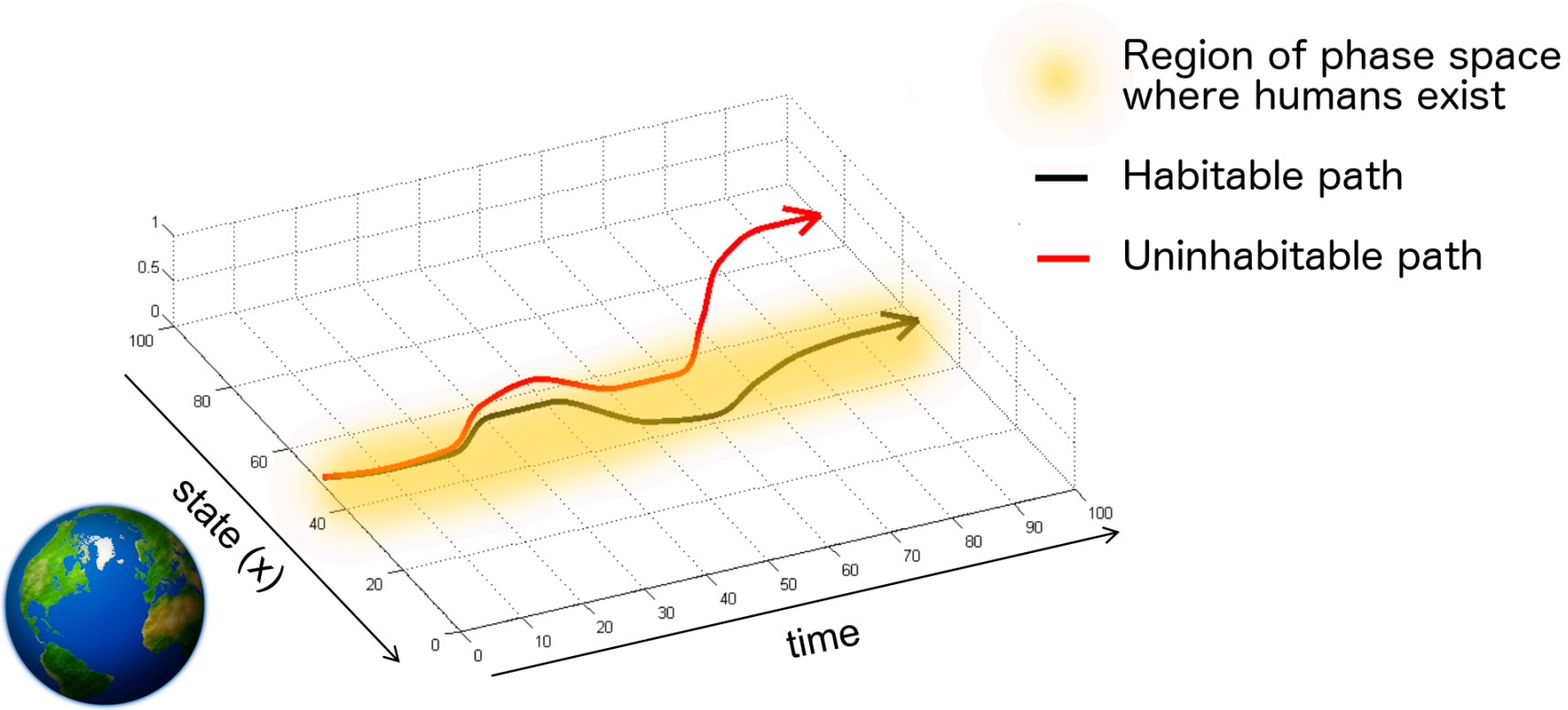
Moreover, it's important to remember that human existence and fragile, and will end *by default* if a new control mechanism begins to govern the state of the Earth.

A robotics graduate student learns the hard way that engineered control mechanisms diverge from our desired behavior ***by default***; it takes a lot of hard engineering work to get something robust over long time intervals and context shifts.



(image taken from lecture by S. Levine; context: autonomous vehicle control)

Now consider the phase space of the Earth, as the dominant intelligent control mechanisms pass from humans hands and minds to machines.



Phase space is very large, so the acceptable region is actually much narrower than depicted. Absent very hard work to maintain control, humans will not continue to exist.

Many folks want explicit examples at this point.

So, what are some specific mechanisms that could lead us off the path of human existence?

Reward hacking in OpenAI universe (2016):

Goal: train an RL agent to win a racing game.

Result: “Our agent achieves a score on average 20 percent higher than that achieved by human players.”

https://blog.openai.com/faulty-reward-functions/

The Optimizer's Curse

"There are more ways to climb a gradient than dreamt of in your philosophy."

http://humancompatible.ai/bibliography#reward-engineering

Related problems and concepts

- "**reward hacking**" (Amodei, Olah et al, 2016)
  - "wireheading" (Omohundro; Bostrom 2014)
  - "delusion boxes" (Ring and Orseau, 2011)
- "**corrigibility**" (Soares, Fallenstein et al, 2015;)
  - "safe interruptibility" (Orseau, Armstrong et al, 2016)

Reward hacking in Disney's *Sorcerer's Apprentice (1940):*

Incorrigibility in Disney's *Sorcerer's Apprentice (1940):*

A precarious situation:

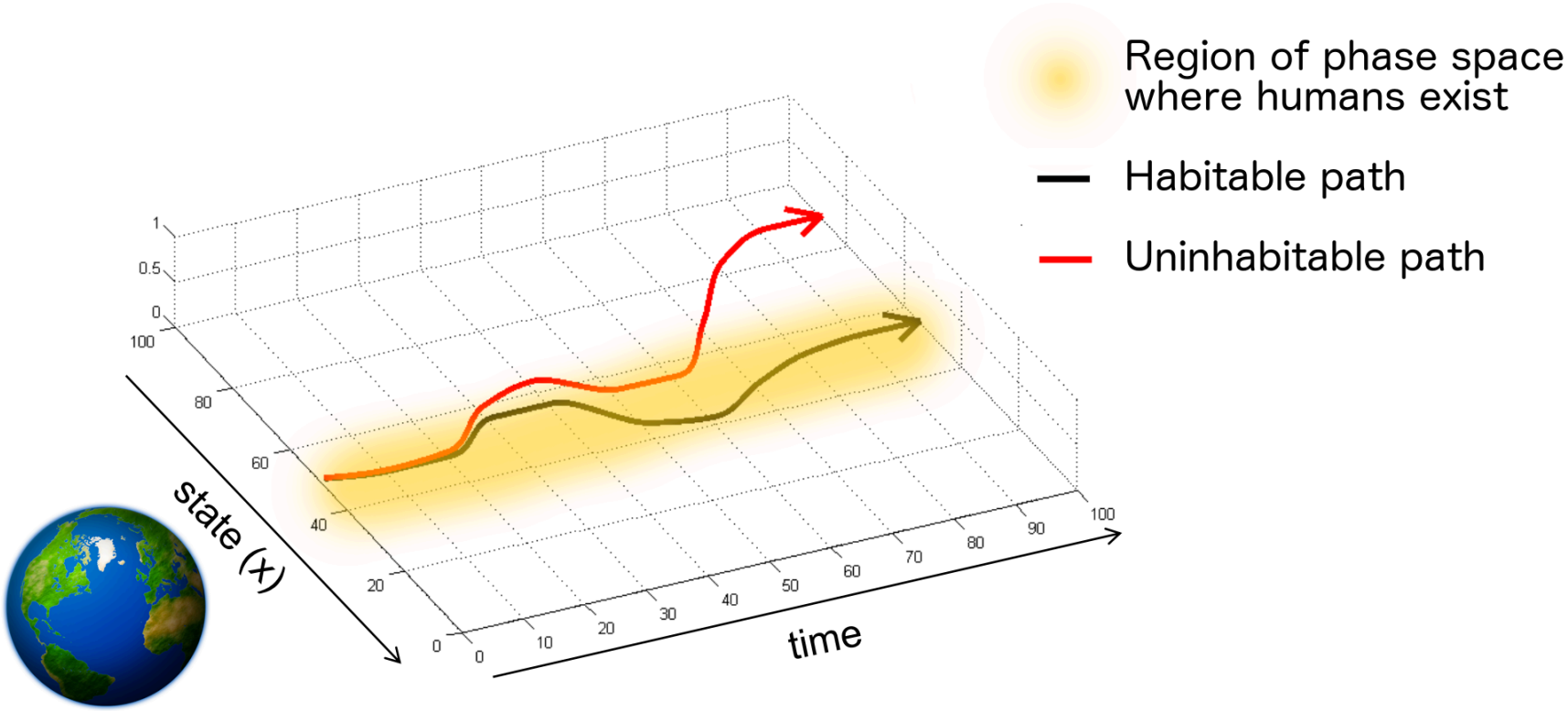
Even absent any malice from an AI, there are basic sub-goals --- like acquiring matter, energy, self-defense, and increased intelligence --- that are useful for almost any objective an intelligent machine can hold, and would be disastrous to humans if pursued fully by a super-intelligent optimizer.

Other ways things can go wrong

In "Optimizer's Curse" scenarios, the unexpected bad behavior arises from a single, centralized optimizer. But centralization of power is not the only source of existential risk:

- **Existential risks from highly advanced AI systems**
  - Optimizer's curse (e.g., an automated corporation)
    - » "reward hacking" (Amodei, Olah et al, 2016)
      - "wireheading" (Omohundro; Bostrom 2014)
      - "delusion boxes" (Ring and Orseau, 2011)
    - » "corrigibility" (Soares, Fallenstein et al, 2015;)
      - "safe interruptibility" (Orseau, Armstrong et al, 2016)
  - Automated wars / arms races → extinction
  - Automated economy → enfeeblement
  - Automated law enforcement → entrapment

But it's important to remember: there's not "one way" to fall of the path of existence; absent a ton of precision engineering, falling off is the default:



Reasons not to think about it

- "General intelligence" is too fuzzy a concept to be dangerous.
  - Your inability to define a thing does not protect you from it.
  - The relevant definition of AGI is whatever capabilities allow a computer to dominate us for control of the earth, in the way that we dominate grizzly bears.
- Corporations are already generally intelligent, so AGI will just be business as usual.
  - Corporations are made of humans / operate on human time scales. Not so for AGI, which will present a major context shift.
- AGI will never happen
  - See Rutherford, 9/11/33, Szilard 9/12/33
- The only people who care are luddites /anti-AI
  - Fusion researchers are Luddites if they point out the need for containment?
  - Cue Turing, Wiener, Minsky, Gates, Musk, Russell, Hassabis…
- AI systems won't harm us unless we explicitly program them with malicious intent.
  - Humans hold little malice for Grizzly bears, but we don't hesitate to destroy their habitat when our objectives require it.
- It's too soon to worry about it
  - 2066 potential asteroid collision: when exactly do we worry?
  - When should we have worried about climate change?
- It's like worrying about overpopulation on Mars
  - No, it's as if we were spending billions moving humanity to Mars with no plan for what to breathe.

(credit to Russell, Beneficial AI 2017, for assembling some of these and subsequent points)

Reasons not to think about it

- We're the creators of AGI, so we can make it however we want
  - Natural selection selected us for reproductive fitness, but we care about many things besides reproductive fitness (e.g. art, music, contraceptives…)
- AI systems won't harm us unless we explicitly program them with malicious intent
  - Humans hold little malice for Grizzly bears, but we don't hesitate to destroy their habitat when our objectives require it.
- Don't put in "human" goals like self-preservation
  - Death isn't bad per se. It's just hard to fetch the coffee after you're dead.
- Don't worry, we'll just have human-AI teams
  - Value misalignment precludes teamwork.
- Just don't have explicit goals for the AI system
  - We need to steer straight, not remove the steering wheel.
- Don't worry, we can just switch it off
  - As if a superintelligent entity would never think of that!

Reasons not to think about it

- You can't control research
  - Yes, we can: we don't genetically engineer humans.
- **Don't mention risks, it might be bad for funding**
  - **See nuclear power, GMOs, tobacco, global warming**

Hmm, okay, I guess *someone* should work on this.

Reasons not to think about it

- Someone else will do it
  - So far, to first order, that's not happening.

STEM PhD-holding researchers working full-time on "AGI safety"

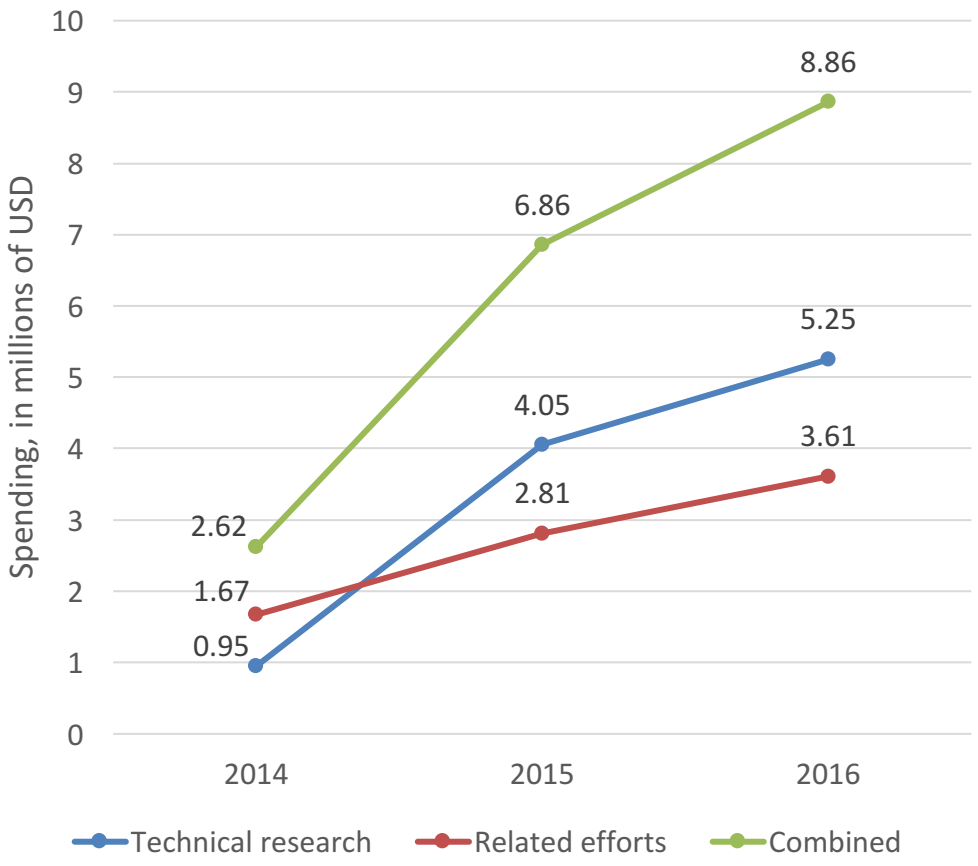
A major motivation for me in investigating this question is that only around a dozen people in the world with a STEM PhD are working full-time on safety and control for AI systems approaching "AGI", despite the fairly obvious pathways through which AGI could pose an existential risk:

Researcher…	Since…	Currently at…
Stuart Armstrong	2011	FHI
Scott Garrabrant	2015	MIRI
Andrew Critch	2015	CHAI
Paul Christiano	2016	OpenAI
Vika Krakovna	2016	DeepMind
Jan Leike	2016	DeepMind
Pedro Ortega	2016	DeepMind
Abram Demski	2017	MIRI
(anonymous)	(anonymous)	(anonymous)
… and perhaps a handful of others		

Reasons not to think about it

- Surely others will join the effort soon
  - Nope, the incentives are still too weak, which is probably why the effort is still as small as it is.

Worldwide spending on directed at reducing AI x-risk:



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

Worldwide, very little AI x-risk research is happening…

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

AI x-risk:
how to ensure powerful AI systems preserve human existence

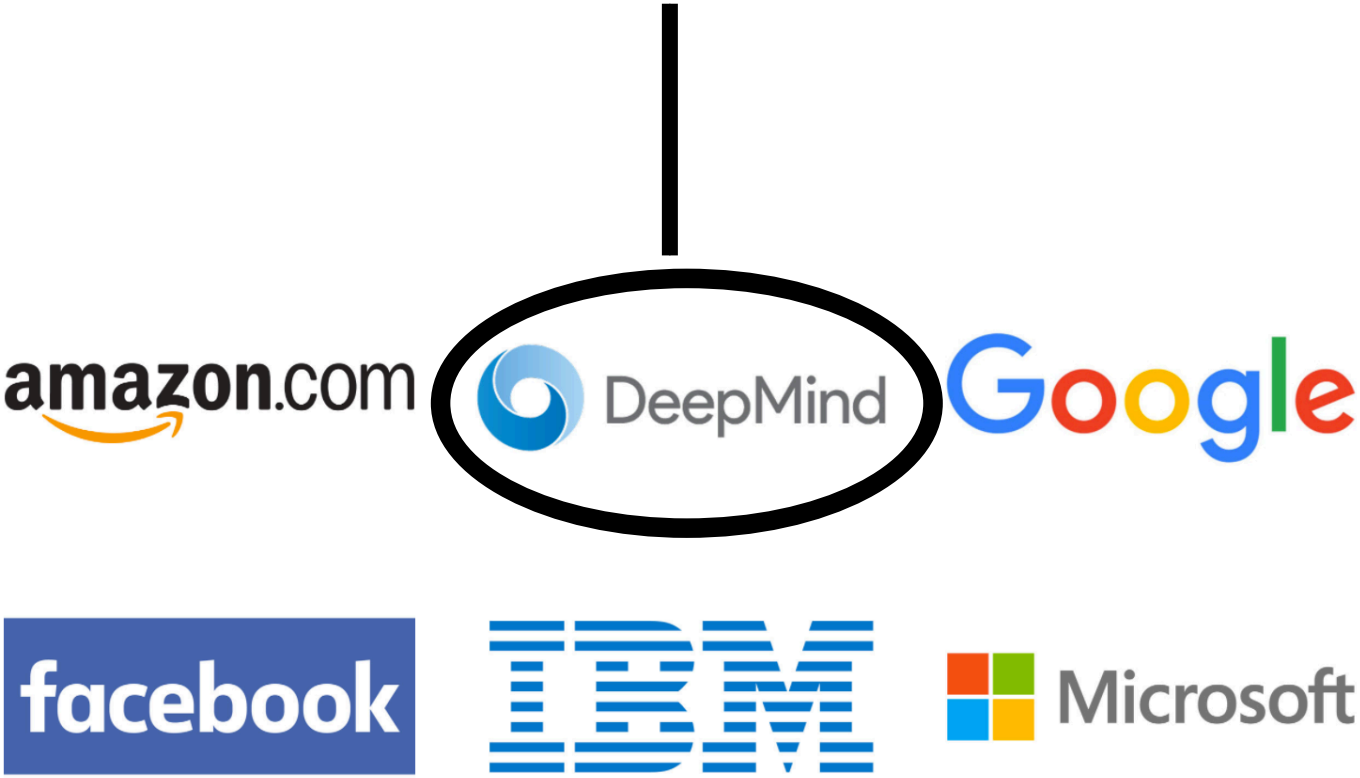
AI capabilities:
how to build a superintelligence

amazon.com DeepMind Google

facebook IBM Microsoft

Some of the world's top AI teams in industry have formed the "Partnership on AI to benefit people and society", but so far...

… only one of the groups below
has a dedicated “AI safety” research team:

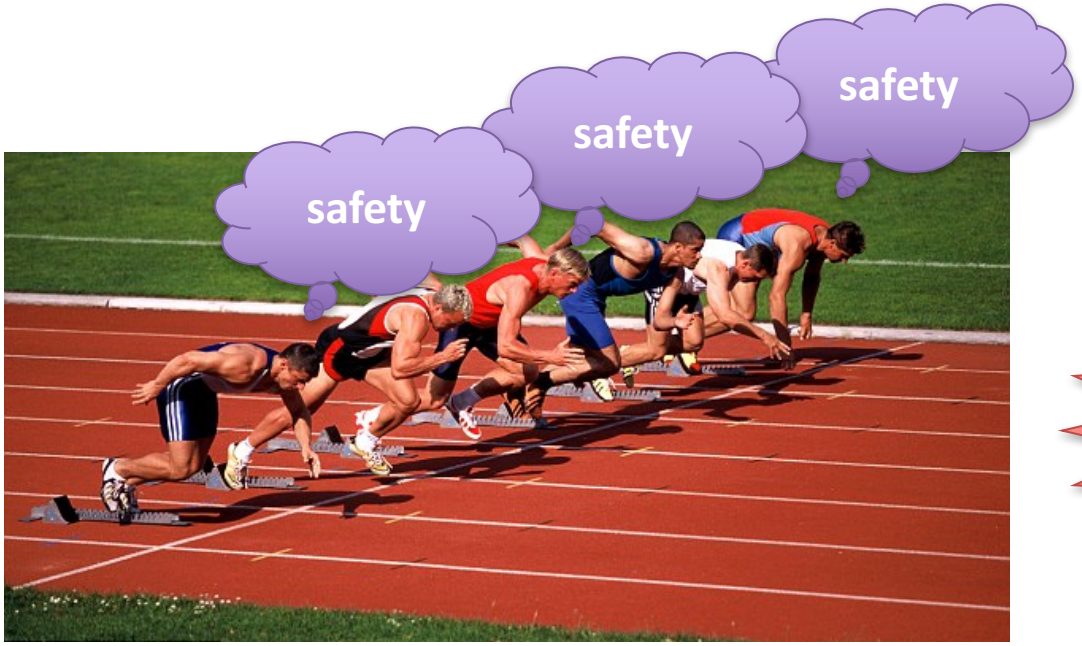


This makes sense: sacrificing researcher-hours to work on AI alignment means giving up some of your competitive edge over other teams.

To make matters worse: if you care about the long-term future, there's a tendency to 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

Reasons not to think about it

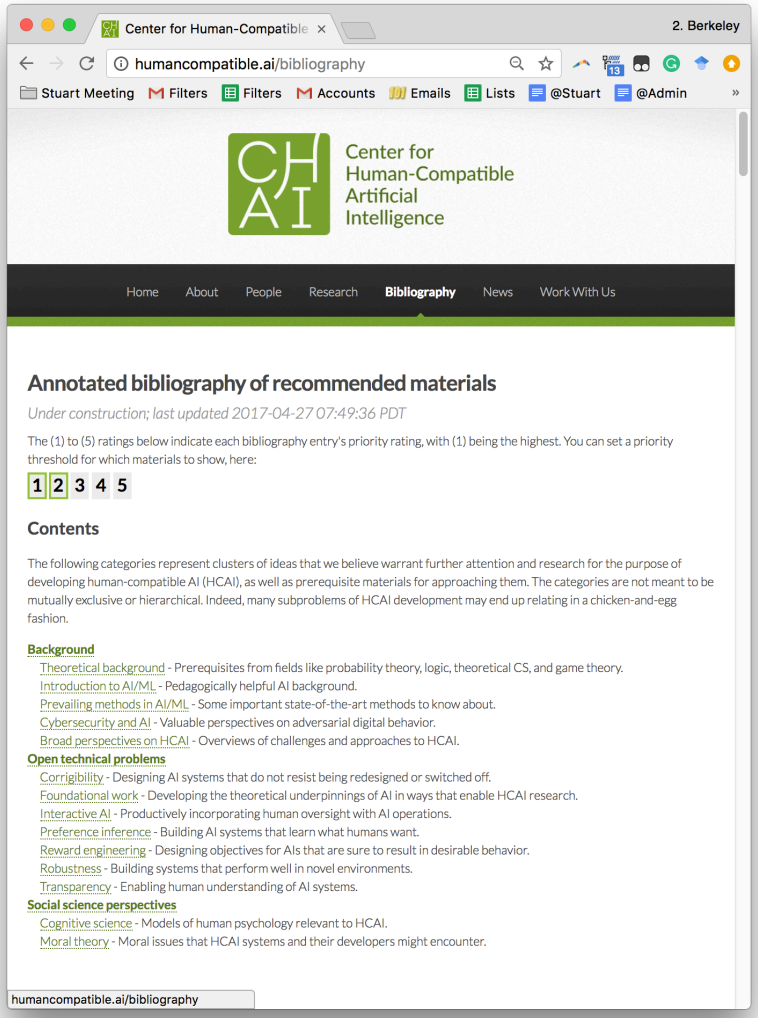
- Surely civilization / governments will get their act together at some point and deal with this
  - Cue British citizens who thought Brexit would never happen…
  - Cue democratic voters who thought they would surely win the 2016 election…
- Let's leave it for the next generation
  - Doesn't work so well; the next generation needs to feel support from the people who will mentor and hire them. I personally know dozens of young AI researchers who have felt afraid to talk to their supervisors about AI alignment, including authors of our most cited bibliography papers.

Okay, maybe I'll take a look at
this AI alignment stuff.
Where do I start?

(2) What can we do about it?

There are numerous open problem areas that could be helpful:

http://humancompatible.ai/bibliography



(2) What can we do about it?

I've been working with some co-authors on a research agenda, which has been helpful in clarifying my own thinking about the topic:

Contents

Some work has begun to translate the "principal-agent problem" in AI:

Cooperative Inverse Reinforcement Learning

Dylan Hadfield-Menell* **Anca Dragan** **Pieter Abbeel** **Stuart Russell**

Electrical Engineering and Computer Science
University of California at Berkeley
Berkeley, CA 94709

Abstract

For an autonomous system to be helpful to humans and to pose no unwarranted risks, it needs to align its values with those of the humans in its environment in such a way that its actions contribute to the maximization of value for the humans. We propose a formal definition of the value alignment problem as *cooperative inverse reinforcement learning* (CIRL). A CIRL problem is a cooperative, partial-information game with two agents, human and robot; both are rewarded according to the human's reward function, but the robot does not initially know what this is. In contrast to classical IRL, where the human is assumed to act optimally in isolation, optimal CIRL solutions produce behaviors such as active teaching, active learning, and communicative actions that are more effective in achieving value alignment. We show that computing optimal joint policies in CIRL games can be reduced to solving a POMDP, prove that optimality in isolation is suboptimal in CIRL, and derive an approximate CIRL algorithm.

1 Introduction

"*If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively . . . we had better be quite sure that the purpose put into the machine is the purpose which we really desire.*" So wrote Norbert Wiener (1960) in one of the earliest explanations of the problems that arise when a powerful autonomous system operates with an incorrect objective. This *value alignment* problem is far from trivial. Humans are prone to mis-stating their objectives, which can lead to unexpected implementations. In the myth of King Midas, the main character learns that wishing for 'everything he touches to turn to gold' leads to disaster. In a reinforcement learning context, Russell & Norvig (2010) describe a seemingly reasonable, but incorrect, reward function for a vacuum robot: if we reward the action of cleaning up dirt, the optimal policy causes the robot to repeatedly dump and clean up the same dirt.

A solution to the value alignment problem has long-term implications for the future of AI and its relationship to humanity (Bostrom, 2014) and short-term utility for the design of usable AI systems. Giving robots the right objectives and enabling them to make the right trade-offs is crucial for self-driving cars, personal assistants, and human–robot interaction more broadly.

The field of *inverse reinforcement learning* or IRL (Russell, 1998; Ng & Russell, 2000; Abbeel & Ng, 2004) is certainly relevant to the value alignment problem. An IRL algorithm infers the reward function of an agent from observations of the agent's behavior, which is assumed to be optimal (or approximately so). One might imagine that IRL provides a simple solution to the value alignment problem: the robot observes human behavior, learns the human reward function, and behaves according to that function. This simple idea has two flaws. The first flaw is obvious: we don't want the robot to adopt the human reward function as its own. For example, human behavior (especially in the morning) often conveys a desire for coffee, and the robot can learn this with IRL, but we don't want the robot to want coffee! This flaw is easily fixed: we need to formulate the value

*{dhm, anca, pabbeel, russell}@cs.berkeley.edu

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Incorrigibility in the CIRL Framework

Ryan Carey

Abstract

A value learning system has incentives to follow shutdown instructions, assuming the shutdown instruction provides information (in the technical sense) about which actions lead to valuable outcomes. However, this assumption is not robust to model mis-specification (e.g., in the case of programmer errors). We demonstrate this by presenting some Supervised POMDP scenarios in which errors in the parameterized reward function remove the incentive to follow shutdown commands. These difficulties parallel those discussed by Soares et al. 2015 in their paper on corrigibility. We argue that it is important to consider systems that follow shutdown commands under some weaker set of assumptions (e.g., that one small verified module is correctly implemented; as opposed to an entire prior probability distribution and/or parameterized reward function). We discuss some difficulties with simple ways to attempt to attain these sorts of guarantees in a value learning framework.

Introduction and Setup

When designing an advanced AI system, we should allow for the possibility that our first version may contain some errors. We therefore want the system to be incentivized to allow human redirection even if it has some errors in its code. Hadfield-Menell et al. (2017) have modeled this problem in the Cooperative Reinforcement Learning (CIRL) framework. They have shown that agents with uncertainty about what to value can be responsive to human redirection, without any dedicated code, in cases where instructions given by the human provide information that reduces the system's uncertainty about what to value. They claim that this (i) provides an incentive toward corrigibility, as described by Soares et al. (2015), and (ii) incentivizes redirectability insofar as this is valuable. In order to re-evaluate the degree to which CIRL-based agents are corrigible, and the consequences of their behavior, we will use a more general variant of the supervision POMDP framework (Milli et al. 2017).

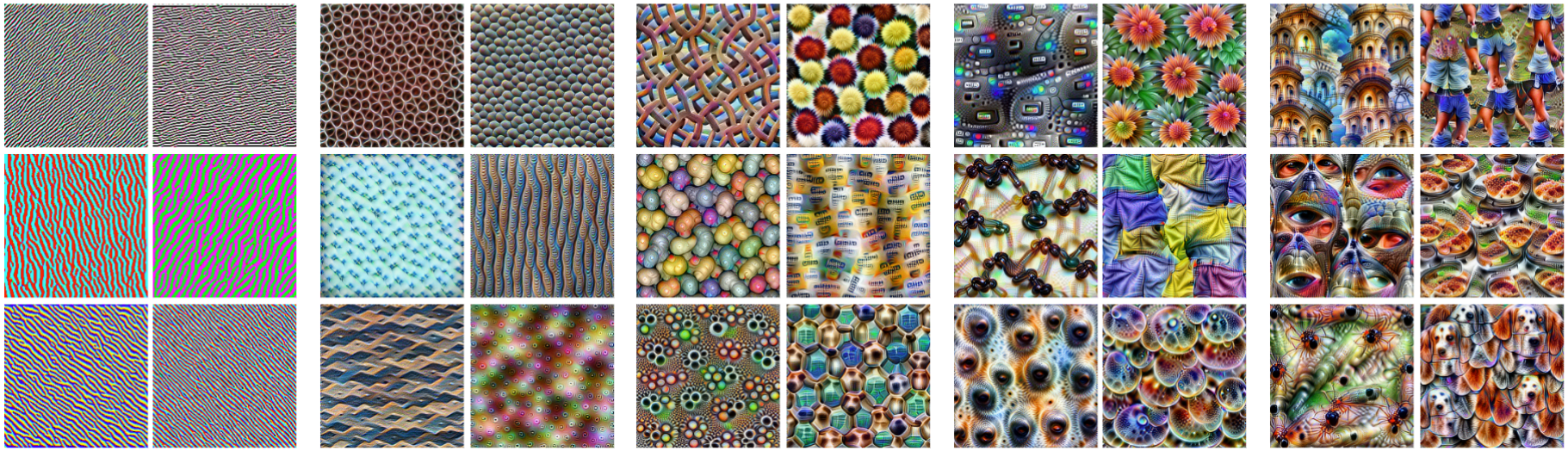
In a regular supervision POMDP (Milli et al. 2017), an AI system $\mathbf{R}$ seeks to maximize reward for a human $\mathbf{H}$, although it does not know the human's reward function. It only has the reward function in a parameterized form $R_H(\theta, s, a)$, and only the human knows the reward parameter θ. In this setting, the human only suggests actions for the AI system to perform, and on each turn, it is up to the AI system whether to perform the suggest action or to perform a different action. Our formalism significantly differs from a supervision POMDP in two ways. First, we relax the assumption that the AI system knows the human's reward function up to the parameter θ. Instead, in order to allow for model-mis-specification, we sample the AI system's parameterized reward function R_R from some distribution P_0, so that it does not always equal R_H. Second, since our focus is on the response to

Research supported by the Machine Intelligence Research Institute (intelligence.org).

I think understanding the “alien-ness” of neural networks is extremely valuable, and work like the following really helps with that, and other improvements in understanding:

Feature Visualization

How neural networks build up their understanding of images



Edges (layer conv2d0) **Textures** (layer mixed3a) **Patterns** (layer mixed4a) **Parts** (layers mixed4b & mixed4c) **Objects** (layers mixed4d & mixed4e)

Feature visualization allows us to see how GoogLeNet [1], trained on the ImageNet [2] dataset, builds up its understanding of images over many layers. Visualizations of all channel are available in the appendix.

AUTHORS	AFFILIATIONS	PUBLISHED	DOI
Chris Olah	Google Brain Team	Nov. 7, 2017	10.23915/distill.00007
Alexander Mordvintsev	Google Research		
Ludwig Schubert	Google Brain Team		

Single-principal-multi-agent and multi-principal-single-agent problems get more complex:

Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments

Ryan Lowe* McGill University OpenAI
Yi Wu* UC Berkeley
Aviv Tamar UC Berkeley
Jean Harb McGill University OpenAI
Pieter Abbeel UC Berkeley OpenAI
Igor Mordatch OpenAI

Abstract

We explore deep reinforcement learning methods for multi-agent domains. We begin by analyzing the difficulty of traditional algorithms in the multi-agent case: Q-learning is challenged by an inherent non-stationarity of the environment, while policy gradient suffers from a variance that increases as the number of agents grows. We then present an adaptation of actor-critic methods that considers action policies of other agents and is able to successfully learn policies that require complex multi-agent coordination. Additionally, we introduce a training regimen utilizing an ensemble of policies for each agent that leads to more robust multi-agent policies. We show the strength of our approach compared to existing methods in cooperative as well as competitive scenarios, where agent populations are able to discover various physical and informational coordination strategies.

1 Introduction

Reinforcement learning (RL) has recently been applied to solve challenging problems, from game playing [23, 28] to robotics [18]. In industrial applications, RL is seeing use in large scale systems such as data center cooling [1]. Most of the successes of RL have been in single agent domains, where modelling or predicting the behaviour of other actors in the environment is largely unnecessary.

However, there are a number of important applications that involve interaction between multiple agents, where emergent behavior and complexity arise from agents co-evolving together. For example, multi-robot control [20], the discovery of communication and language [29, 8, 24], multiplayer games [27], and the analysis of social dilemmas [17] all operate in a multi-agent domain. Related problems, such as variants of hierarchical reinforcement learning [6] can also be seen as a multi-agent system, with multiple levels of hierarchy being equivalent to multiple agents. Additionally, multi-agent self-play has recently been shown to be a useful training paradigm [28, 30]. Successfully scaling RL to environments with multiple agents is crucial to building artificially intelligent systems that can productively interact with humans and each other.

Unfortunately, traditional reinforcement learning approaches such as Q-Learning or policy gradient are poorly suited to multi-agent environments. One issue is that each agent's policy is changing as training progresses, and the environment becomes non-stationary from the perspective of any individual agent (in a way that is not explainable by changes in the agent's own policy). This presents learning stability challenges and prevents the straightforward use of past experience replay, which is

*Equal contribution. Corresponding authors: ryan.lowe@cs.mcgill.ca, jxwuyi@gmail.com, mordatch@openai.com.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Servant of Many Masters: Shifting priorities in Pareto-optimal sequential decision-making

Andrew Critch, Stuart Russell
University of California, Berkeley
{critch, russell}@berkeley.edu

Abstract

It is often argued that an agent making decisions on behalf of two or more principals who have different utility functions should adopt a *Pareto-optimal* policy, i.e., a policy that cannot be improved upon for one agent without making sacrifices for another. A famous theorem of Harsanyi shows that, when the principals have a common prior on the outcome distributions of all policies, a Pareto-optimal policy for the agent is one that maximizes a fixed, weighted linear combination of the principals' utilities.

In this paper, we show that Harsanyi's theorem does not hold for principals with different priors, and derive a more precise generalization which does hold, which constitutes our main result. In this more general case, the relative weight given to each principal's utility should evolve over time according to how well the agent's observations conform with that principal's prior. The result has implications for the design of contracts, treaties, joint ventures, and robots.

1 Introduction

As AI systems take on an increasingly pivotal decision-making role in human society, an important question arises: *Whose values should a powerful decision-making machine be built to serve?* [Bostrom, 2014]

Consider, informally, a scenario wherein two or more principals—perhaps individuals, companies, or states—are considering cooperating to build or otherwise obtain an "agent" that will then interact with an environment on their behalf. The "agent" here could be anything that follows a policy, such as a robot, a corporation, or a web-based AI system. In such a scenario, the principals will be concerned with the question of "how much" the agent will prioritize each principal's interests, a question which this paper addresses quantitatively.

One might be tempted to model the agent as maximizing the expected value, given its observations, of some utility function U of the environment that equals a weighted sum

$$w^1 U^1 + w^2 U^2 \qquad (1)$$

of the principals' individual utility functions U^1 and U^2, as Harsanyi's social aggregation theorem [Harsanyi, 1980] recommends. Then the question of prioritization could be reduced to that of choosing values for the weights w^i.

However, this turns out to be a suboptimal approach, from the perspective of the principals. As we shall see in Proposition 1, this solution form is not generally compatible with Pareto-optimality when agents have different beliefs. Harsanyi's setting does not account for agents having different priors, nor for decisions being made sequentially, after future observations.

In such a setting, we need a new form of solution, exhibited in this paper. The solution is presented along with a recursion (Theorem 3) that characterizes solutions by a process algebraically similar to, but meaningfully different from, Bayesian updating. The updating process resembles a kind of bet-settling between the principals, which allows them each to expect to benefit from the veracity of their own beliefs.

Qualitatively, this phenomenon can be seen in isolation whenever two people make a bet on a piece of decision-irrelevant trivia. If neither Alice nor Bob would base any important decision on whether Michael Jackson was born in 1958 or 1959, they might still make a bet for $100 on the answer. For a person chosen to arbitrate the bet (their "agent"), Michael Jackson's birth year now becomes a decision-relevant observation: it determines which of Alice and Bob gets the money!

Transparency of agents to other agents adds some extremely surprising side-effects:

Open-source game theory is weird

Posted in AI Safety on December 2, 2016 | Permalink | Leave a comment

I sometimes forget that not everyone realizes how poorly understood open-source game theory is, until I end up sharing this example and remember how weird it is for folks to see for the first time. Since that's been happening a lot this week, I wrote this post to automate the process.

Consider a game where agents can view each other's source codes and return either "C" (cooperate) or "D" (defect). The payoffs don't really matter for the following discussion.

First, consider a very simple agent called "CooperateBot", or "CB" for short, which cooperates with every possible opponent:

```
def CB(opp):
  return C
```

(Here "opp" is the argument representing the opponent's source code, which CooperateBot happens to ignore.)

Next consider a more interesting agent, "FairBot", or "FB" for short, which takes in a single parameter k to determine how long it thinks about its opponent:

```
def FB_k(opp):
  search all strings of length ≤ k
    ...for a proof that
    ..."opp(FB_k) = C"
  if a proof is found:
    return C
  else:
    return D
```

I claim this agent is interesting.

Warm-up question: What's $FB_k(CB)$?

Show/hide warm-up answer

Real question: what's $FB_k(FB_k)$, for very large k?

Show/hide hint #1

Got it yet? Here's hint #2:

Show/hide hint #2

Parametric Bounded Löb's Theorem and Robust Cooperation of Bounded Agents

Andrew Critch
Machine Intelligence Research Institute
critch@intelligence.org

Abstract

Löb's theorem and Gödel's theorems make predictions about the behavior of systems capable of self-reference with unbounded computational resources with which to write and evaluate proofs. However, in the real world, systems capable of self-reference will have limited memory and processing speed, so in this paper we introduce an effective version of Löb's theorem which is applicable given such bounded resources. These results have powerful implications for the game theory of bounded agents who are able to write proofs about themselves and one another, including the capacity to out-perform classical Nash equilibria and correlated equilibria, attaining mutually cooperative program equilibrium in the Prisoner's Dilemma. Previous cooperative program equilibria studied by Tennenholtz (2004) and Fortnow (2009) have depended on tests for program equality, a fragile condition, whereas "Löbian" cooperation is much more robust and agnostic of the opponent's implementation.

1 Background and Overview

The arc of this paper begins and ends with a discussion of the Prisoner's Dilemma, but it passes through a new result in provability logic. Thus, it will hopefully be of interest to game theorists and logicians alike.

1.1 Open-source Prisoner's Dilemma

Consider the Prisoner's Dilemma, a game with two possible actions C (Cooperate) and D (Defect), with the following payoff matrix:

		Player 2: C	Player 2: D
Player 1	C	$(2,2)$	$(0,3)$
	D	$(3,0)$	$(1,1)$

In other words, by choosing D over C, each player can destroy 2 units of its opponent's utility to gain 1 unit of its own. As long as the payoffs are truly represented in the matrix—for example, there are no reputational costs of choosing D that are not already imputed in the payoffs—then (D, D) is the only Nash equilibrium, and the only correlated equilibrium. In fact, irrespective of the opponent's move, it is better to defect. It is therefore broadly believed that (D, D) is an inevitable outcome between "rational" agents in a truly represented (non-iterated) Prisoner's Dilemma.

Research supported by the Machine Intelligence Research Institute (intelligence.org). Preprinted at arXiv:1602.04184 [cs:GT]

Ultimately, to secure long-run stability in the face of major changes in the world's control mechanisms, I suspect we will need theorems about our ability and inability to predict "computed variables", as distinct from "random variables":

Logical Induction

Scott Garrabrant, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, and **Jessica Taylor**
{scott,tsvi,critch,nate,jessica}@intelligence.org
Machine Intelligence Research Institute

Abstract

We present a computable algorithm that assigns probabilities to every logical statement in a given formal language, and refines those probabilities over time. For instance, if the language is Peano arithmetic, it assigns probabilities to all arithmetical statements, including claims about the twin prime conjecture, the outputs of long-running computations, and its own probabilities. We show that our algorithm, an instance of what we call a *logical inductor*, satisfies a number of intuitive desiderata, including: (1) it learns to predict patterns of truth and falsehood in logical statements, often long before having the resources to evaluate the statements, so long as the patterns can be written down in polynomial time; (2) it learns to use appropriate statistical summaries to predict sequences of statements whose truth values appear pseudorandom; and (3) it learns to have accurate beliefs about its own current beliefs, in a manner that avoids the standard paradoxes of self-reference. For example, if a given computer program only ever produces outputs in a certain range, a logical inductor learns this fact in a timely manner; and if late digits in the decimal expansion of π are difficult to predict, then a logical inductor learns to assign $\approx 10\%$ probability to "the nth digit of π is a 7" for large n. Logical inductors also learn to trust their future beliefs more than their current beliefs, and their beliefs are coherent in the limit (whenever $\phi \to \psi$, $\mathbb{P}_\infty(\phi) \leq \mathbb{P}_\infty(\psi)$, and so on); and logical inductors strictly dominate the universal semimeasure in the limit.

These properties and many others all follow from a single *logical induction criterion*, which is motivated by a series of stock trading analogies. Roughly speaking, each logical sentence ϕ is associated with a stock that is worth \$1 per share if ϕ is true and nothing otherwise, and we interpret the belief-state of a logically uncertain reasoner as a set of market prices, where $\mathbb{P}_n(\phi) = 50\%$ means that on day n, shares of ϕ may be bought or sold from the reasoner for 50¢. The logical induction criterion says (very roughly) that there should not be any polynomial-time computable trading strategy with finite risk tolerance that earns unbounded profits in that market over time. This criterion bears strong resemblance to the "no Dutch book" criteria that support both expected utility theory (von Neumann and Morgenstern 1944) and Bayesian probability theory (Ramsey 1931; de Finetti 1937).

… and a number of research hotspots
attempting to address safety for powerful AI systems in the future:

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 RL, …)
	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

Summary:

Suppose AGI will exist eventually.

1. **(importance) We need AGI to be aligned** with human interests if we want humans to continue existing (*cue: children, retirement, the environment, space travel*),
2. **(neglect) Alignment research is currently highly neglected** because
    1. **Public discourse** is low-quality, reactionary, and unappealing to researchers;
    2. **Young people** feel actively discouraged to begin work on it; and
    3. **Top AI research teams** need to focus on beating capability benchmarks to attract top talent.
3. **(actionability) Interventions needed:**
    1. **More technical work** specifically on safety / control theory for systems approaching AGI capabilities;
    2. **Junior faculty** exemplifying employability as AI safety specialists, and
    3. **Incentives and institutions for cooperative development** of AGI, so that race conditions do not crowd out safety measures.
    4. **Experts to consult on technical sub-questions** of policies for oversight and governance, to avoid over-regulation, under-regulation or untenable regulation when the risks become more imminent.

Thanks!

Next actions:

- Visit http://humancompatible.ai/bibliography for tons of reading, viewing, and open problems.
- Email rosiecampbell@berkeley.edu to join the CHAI mailing list (UC Berkeley faculty and students only)
- Visit our group meetings, 10:00-11:30 on Wednesdays in 730 SDH. (Schedule changes announced via the mailing list.)