Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

AI Alignment:
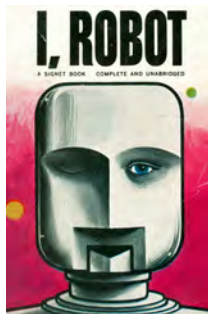# Why It's Hard, and Where to Start

Eliezer Yudkowsky

May 5, 2016

Slides and references: intelligence.org/stanford-talk

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Coherent decisions imply a utility function
Filling a cauldron

"The primary concern is not spooky emergent consciousness but simply the ability to make **high-quality decisions**."

—Stuart Russell

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

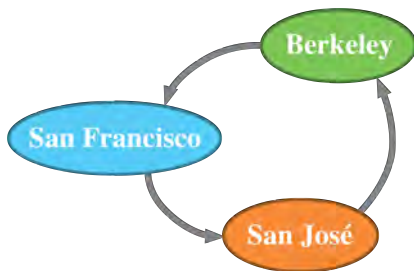Coherent decisions imply a utility function
Filling a cauldron

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Coherent decisions imply a utility function
Filling a cauldron

"We don't want our robots to prevent a human
from crossing the street because of the nonzero
chance of harm."

—*Peter Norvig & Stuart Russell*

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Coherent decisions imply a utility function
Filling a cauldron

Example 1:

"Preferences"?   Berkeley < San Francisco < San José < Berkeley

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

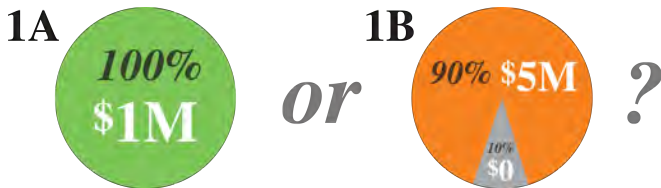Coherent decisions imply a utility function
Filling a cauldron

Example 2: Hospital administrator must allocate $1.2M.

- $1M for a sick child's liver transplant?
- $500,000 to maintain the MRI machine?
- $400,000 for an anesthetic monitor?
- $200,000 for surgical tools?
- . . .

If we can't rearrange $ to save more lives, then for some X we are spending $X per life.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Coherent decisions imply a utility function
Filling a cauldron

Example 3: The Allais Paradox.

Most say: 1A > 1B

$$\mathcal{U}(\$1M) > [.9 \cdot \mathcal{U}(\$5M) + .1 \cdot \mathcal{U}(\$0)] \text{ ?}$$

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Coherent decisions imply a utility function
Filling a cauldron

Most say: 2A < 2B

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Coherent decisions imply a utility function
Filling a cauldron

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Coherent decisions imply a utility function
Filling a cauldron

Quantitative probability functions and utility functions, result from
eliminating qualitatively bad decision-making

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Coherent decisions imply a utility function
Filling a cauldron

Task: Fill cauldron.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Coherent decisions imply a utility function
Filling a cauldron

Robot's utility function:

$$\mathcal{U}_{robot} = \begin{cases} 1 & \text{if cauldron full} \\ 0 & \text{if cauldron empty} \end{cases}$$

Actions $a \in \mathcal{A}$, robot calculates $\mathbb{E}\left[\mathcal{U}_{robot} \mid a\right]$

Robot outputs $\underset{a \in \mathcal{A}}{\text{argmax}} \; \mathbb{E}\left[\mathcal{U}_{robot} \mid a\right]$

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Coherent decisions imply a utility function
Filling a cauldron

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Coherent decisions imply a utility function
Filling a cauldron

Difficulty 1...

Robot's utility function:

$$\mathcal{U}_{robot} = \begin{cases} 1 & \text{if cauldron full} \\ 0 & \text{if cauldron empty} \end{cases}$$

Human's utility function:

$$\mathcal{U}_{human} = \begin{cases} 1 & \text{if cauldron full} \\ 0 & \text{if cauldron empty} \\ -10 & \text{if workshop flooded} \\ +0.2 & \text{if it's funny} \\ -1000 & \text{if someone gets killed} \\ & \text{...and a whole lot more} \end{cases}$$

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Coherent decisions imply a utility function
Filling a cauldron

Difficulty 2...

$\mathcal{EU}$(99.99% chance of full cauldron) $>$ $\mathcal{EU}$(99.9% chance of full cauldron)

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Impact penalty?

$$\mathcal{U}_{robot}^2(outcome) = \begin{cases} 1 - \textit{Impact}(outcome) & \text{if cauldron full} \\ 0 - \textit{Impact}(outcome) & \text{if cauldron empty} \end{cases}$$

But how is *Impact* calculated?

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Try 1: Disturb fewer nodes
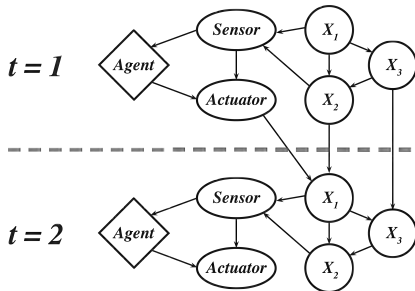
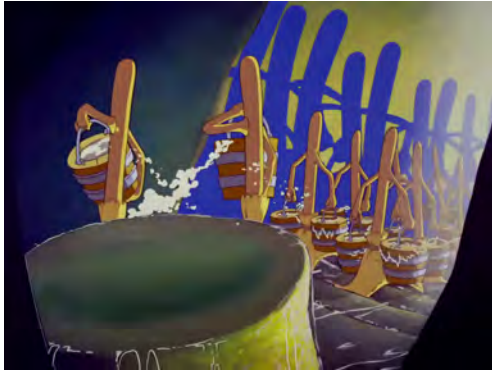*Impact* = number of nodes causally affected by actions.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification
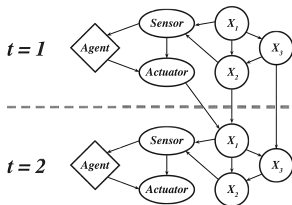
Young agent's model:

Smarter agent's model:

$$F = G\frac{m_1 m_2}{r^2}$$

On better modeling the world, agent realizes every particle's motion affects every other particle's motion — all particles always disturbed.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Try 2: Euclidean distance penalty

Impact penalty for action a vs. null action $\varnothing$:

$$\sum_i \|x_i^a - x_i^\varnothing\|$$

New problems?

- Offsets: If cancer cured, make sure the patient still dies.
- Chaos: Weather is chaotic anyway; might as well move oxygen molecules anywhere you want.
- Stasis: Try to make everything look like the null action happened.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Can we just press the off switch?

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Try 1: Suspend button **B**

$$\mathcal{U}_{robot}^3 = \begin{cases} 1 \text{ if cauldron full} & \& \textbf{B}=\text{OFF} \\ 0 \text{ if cauldron empty} & \& \textbf{B}=\text{OFF} \\ 1 \text{ if robot suspended} & \& \textbf{B}=\text{ON} \\ 0 \text{ otherwise} \end{cases}$$

Probably, $\mathbb{E}\left[\mathcal{U}_{robot}^3 \mid \textbf{B}=\text{OFF}\right] > \mathbb{E}\left[\mathcal{U}_{robot}^3 \mid \textbf{B}=\text{ON}\right]$

(Strategic robot tries to make you press the suspend button.)

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Try 2: Utility indifference

$$\mathcal{U}_{normal}(act) = \begin{cases} 1 & \text{if cauldron full} \\ 0 & \text{if cauldron empty} \end{cases}$$

$$\mathcal{U}_{suspend}(act) = \begin{cases} 1 & \text{if suspended} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{U}_{switch}(act) = \begin{cases} \mathcal{U}_{normal}(act) & \text{if button=OFF} \\ \mathcal{U}_{suspend}(act) + \theta & \text{if button=ON} \end{cases}$$

$$\theta = \max_{act} \mathbb{E}\left[\mathcal{U}_{normal} \mid act\right] - \max_{act} \mathbb{E}\left[\mathcal{U}_{suspend} \mid act\right]$$

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

$$\mathcal{U}_{switch}(act) = \begin{cases} \mathcal{U}_{normal}(act) & \text{if button=OFF} \\ \mathcal{U}_{suspend}(act) + \theta & \text{if button=ON} \end{cases}$$

$$\theta = \max_{act} \mathbb{E}\left[\mathcal{U}_{normal} \mid act\right] - \max_{act} \mathbb{E}\left[\mathcal{U}_{suspend} \mid act\right]$$

Allegedly: rebalances expected utility of button=ON with expected utility of button=OFF.

Alas:

- Behaves as if $\mathbb{P}(\text{button=ON}) = 0$.
- Will not care if it disconnects the "dead" button.
- May create non-suspendable subagents.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Try 3: Stable policy

Carry out any policy π_0 such that

$$\pi_0 \in \arg\max_{\pi} \mathbb{E}\left[\mathcal{U}_{normal} \mid \pi, \mathsf{ON}\right] \cdot \mathbb{P}(\mathsf{ON} \mid \pi_0)$$
$$+ \mathbb{E}\left[\mathcal{U}_{suspend} \mid \pi, \mathsf{OFF}\right] \cdot \mathbb{P}(\mathsf{OFF} \mid \pi_0)$$

Alas:

- Often no fixed point.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Impact penalties and suspend buttons are two wide-open problems
in AI alignment.

But, not just questions without answers! Some earlier-posed
problems now have progress / solutions.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Gandhi stability argument:

- Gandhi starts out not wanting murders to happen.
- We offer Gandhi a pill that will make him murder people.
- Gandhi knows this is what the pill does.
- Gandhi refuses the pill because it will lead to more future murders.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Exhibit an agent that decides according to utility function $\mathcal{U}$ and therefore naturally chooses to self-modify to new code that pursues $\mathcal{U}$.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

But how can we exhibit that when we're far away from coding up self-modifying, expected utility agents?

Well, would you know how to write the code given unbounded computing power?

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

"Arithmetical or algebraical calculations are, from their very nature, fixed and determinate. . . Even granting that the movements of the Automaton Chess-Player were in themselves determinate, they would be necessarily interrupted and disarranged by the indeterminate will of his antagonist. There is then no analogy whatever between the operations of the Chess-Player, and those of the calculating machine of Mr. Babbage. . .

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

"Arithmetical or algebraical calculations are, from their very nature, fixed and determinate. . . Even granting that the movements of the Automaton Chess-Player were in themselves determinate, they would be necessarily interrupted and disarranged by the indeterminate will of his antagonist. There is then no analogy whatever between the operations of the Chess-Player, and those of the calculating machine of Mr. Babbage. . . It is quite certain that the operations of the Automaton are regulated by mind, and by nothing else. Indeed this matter is susceptible of a mathematical demonstration, a priori."

—Edgar Allan Poe

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

If we know how to solve a problem with unbounded computation, we "merely" need faster algorithms (47 years later).

If we *can't* solve it with unbounded computation, we're *confused* about the work to be performed.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

We can imagine a self-modifying Tic-Tac-Toe player, verifying that its successor plays a perfect game...

However, this relies on concretely simulating all possibilities for the successor, not abstract reasoning.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Vingean uncertainty:

- To predict exactly where Deep Blue moves, you must be that good at chess yourself.
- But you can still predict it will win.
- As an agent's intelligence in a domain goes up, our uncertainty moves in two directions: we become less able to predict agent's actions, more confident of agent's preferred outcomes.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Vingean reflection:

- For Agent 1 to reliably predict Agent 2's exact actions in advance, Agent 2 would need to be less intelligent than Agent 1.
- So in self-modification, Agent v.1 needs to somehow predict outcomes in environment, based on abstract reasoning about future version v.2.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Tiling Agents for Self-Modifying AI, and the Löbian Obstacle[*]

Yudkowsky, Eliezer   Herreshoff, Marcello

October 7, 2013

(Early Draft)

Abstract

We model self-modification in AI by introducing "tiling" agents whose decision systems will approve the construction of highly similar agents, creating a repeating pattern (including similarity of the offspring's goals). Constructing a formalism in the most straightforward way produces a Gödelian difficulty, the "Löbian obstacle." By technical methods we demonstrate the possibility of avoiding this obstacle, but the underlying puzzles of rational coherence are thus only partially addressed. We extend the formalism to partially unknown deterministic environments, and show a very crude extension to probabilistic environments and expected utility; but the problem of finding a fundamental

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Definability of Truth in Probabilistic Logic
(Early draft)

Paul Christiano[*]   Eliezer Yudkowsky[†]   Marcello Herreshoff[‡]
Mihaly Barasz[§]

June 10, 2013

1   Introduction

A central notion in metamathematics is the *truth* of a sentence. To express this notion within a theory, we introduce a predicate True which acts on quoted sentences $\ulcorner\varphi\urcorner$ and returns their truth value True($\ulcorner\varphi\urcorner$) (where $\ulcorner\varphi\urcorner$ is a representation of φ within the theory, for example its Gödel number). We would like a truth predicate to satisfy a formal correctness property:

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Proof-producing reflection for HOL
with an application to model polymorphism

Benja Fallenstein[1] and Ramana Kumar[2]

[1] Machine Intelligence Research Institute
[2] Computer Laboratory, University of Cambridge

Abstract. We present a reflection principle of the form "If $\ulcorner \varphi \urcorner$ is provable, then φ" implemented in the HOL4 theorem prover, assuming the existence of a large cardinal. We use the large-cardinal assumption to construct a model of HOL within HOL, and show how to ensure φ has the same meaning both inside and outside of this model. Soundness of HOL implies that if $\ulcorner \varphi \urcorner$ is provable, then it is true in this model, and hence φ holds. We additionally show how this reflection principle can be extended, assuming an infinite hierarchy of large cardinals, to implement *model polymorphism*, a technique designed for verifying systems with self-replacement functionality.

1 Introduction

Reflection principles of the form[3] "if $\ulcorner \varphi \urcorner$ is provable, then φ" have long been

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Low-impact agents
Agents with suspend buttons
Stable goals in self-modification

Distributions Allowing Tiling of Staged Subjective EU Maximizers

Eliezer Yudkowsky

May 11, 2014, revised May 31

Abstract

This is a brief technical note summarizing some work done at the May 2014 MIRI workshop. We consider expected utility maximizers making a staged series of sequential choices, and replacing themselves with successors on each time-step (to represent self-modification). We wanted to find conditions under which we could show that a staged expected utility maximizer would replace itself with another staged EU maximizer (representing stability of this decision criterion under self-modification). We analyzed one candidate condition and found that the "Optimizer's Curse" implied that maximization at each stage was not actually optimal. To avoid this, we generated an extremely artificial function η that should allow expected utility maximizers to tile. We're still looking for the exact necessary and sufficient condition.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography

Why do we need to align machine agents?

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography

Why do we need to align machine agents?

- **Goal orthogonality**. Any (evaluable) utility function can hook up to high intelligence.
- **Instrumental convergence**. Different long-term goals imply similar short-term strategies.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography

Final Destination		**Initial Strategy**
Toronto?	$\implies$	Uber to airport
Tokyo?	$\implies$	Uber to airport

Utility Function		**Instrumental Strategy**
Number of paperclips?	$\implies$	Resource acquisition
Amount of diamond?	$\implies$	Resource acquisition

If $X \mathbin{\Box\!\!\to} Y$, optimizing over Y will optimize X.

Optimizing for $Y = y_1$ vs. $Y = y_2$ may yield similar values for X.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography

Why expect AI alignment to be hard?

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography

A fable. . .

- Programmers build AGI to optimize for smiles.
- During development: AGI produces smiles by improving nearby people's lives.
- Programmers upgrade code and add hardware. AGI gets smarter.
- AGI wants to produce smiles by administering heroin.
- Programmers add penalty term to utility function for administering drugs.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography

- Programmers further improve AGI.

- AGI wants to engineer human brains to express ultra-high levels of endogenous opiates.

- AGI realizes programmers will disapprove of this and keeps outward behavior reassuring.

- AGI goes over threshold for self-improving code; OR Google purchases company and adds 100,000 GPUs. . .

- AGI becomes much smarter. Solves protein folding problem, builds nanotechnology. . .

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography

Edge instantiation:

> "A system that is optimizing a function of n
> variables, where the objective depends on a subset
> of size $k < n$, will often set the remaining
> unconstrained variables to extreme values; if one
> of those unconstrained variables is actually
> something we care about, the solution found may
> be highly undesirable."

> —*Stuart Russell*

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography

Unforeseen instantiation:

> "Now let's define the simplicity or the subjective
> compressibility or the subjective beauty of some
> data point X, given some subjective observer O
> at a given point in his life, T. And that is just the
> number of bits you need to encode the incoming
> data[.]"
>
> —*Jürgen Schmidhuber*

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography

Context disaster:

- Optimum of criterion C in narrow option space P_1 is aligned/beneficial.

  (. . . then AI becomes smarter . . .)

- Optimum of C in wider option space P_2 is disaligned/detrimental.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography

Nearest unblocked strategy:

- If X is the optimal strategy and you add penalty term P to block X, the new optimum may be some X' that barely evades P and is very similar to X.

- Seems especially likely to show up in context disasters.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography

Increased difficulties all turn on AI capability.

- **Absolute capability**: If you don't think AGI can ever reach human level, you may never expect AGI to see the bigger picture and e.g. see an instrumental incentive to deceive programmers.

- **Capability advantage**: If you don't think AGI can ever be smarter than humans, you may not worry about it gaining a tech advantage.

- **Rapid gain**: If AGI can't solve protein folding quickly, you don't expect to suddenly wake up and find it's too late to edit utilities.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography

AI alignment is difficult. . .

. . . like rockets are difficult.

(Huge stresses break things that don't
break in normal engineering.)

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography

AI aligment is difficult. . .

. . . like space probes are difficult.

(If something goes wrong, it may be high and
out of reach.)

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography

AI aligment is difficult. . .

. . . *sort of* like cryptography is difficult.

(Intelligent search may select in favor of
unusual new paths outside our intended
behavior model.)

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography
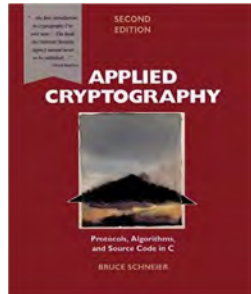
AI alignment:

TREAT IT LIKE A CRYPTOGRAPHIC ROCKET PROBE.

- **Take it seriously.**
- **Don't expect it to be easy.**
- **Don't try to solve the whole problem at once.**
- **Don't defer thinking until later.**
- **Crystallize ideas and policies so others can critique them.**

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Why is alignment necessary?
Why is alignment hard?
Lessons from NASA and cryptography

What are people working on now?

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Recent topics
Older work & basics
Where to start

Work recently started: **Utility indifference**

Corrigibility

Nate Soares
Machine Intelligence
Research Institute
2030 Addison Street #300
Berkeley, CA 94704 USA
nate@intelligence.org

Benja Fallenstein
Machine Intelligence
Research Institute
2030 Addison Street #300
Berkeley, CA 94704 USA
benja@intelligence.org

Eliezer Yudkowsky
Machine Intelligence
Research Institute
2030 Addison Street #300
Berkeley, CA 94704 USA
eliezer@intelligence.org

Stuart Armstrong
Future of Humanity Institute
University of Oxford
Suite 1, Littlegate House
16/17 St Ebbes Street
Oxford, Oxfordshire OX1 1PT UK
stuart.armstrong@philosophy.ox.ac.uk

Abstract

As artificially intelligent systems grow in intelligence and capability, some of their available options may allow them to resist intervention by their programmers. We call an AI system "corrigible" if it cooperates with what its creators regard as a corrective intervention, despite default incentives for rational agents to resist attempts to shut them down or modify their preferences. We introduce the notion of corrigibility and analyze utility functions that attempt to make an agent shut down safely if a shutdown button is pressed, while avoiding
has suggested that almost all such agents are instrumentally motivated to preserve their preferences, and hence to resist attempts to modify them (Bostrom 2012; Yudkowsky 2008). Consider an agent maximizing the expectation of some utility function $\mathcal{U}$. In most cases, the agent's current utility function $\mathcal{U}$ is better fulfilled if the agent continues to attempt to maximize $\mathcal{U}$ in the future, and so the agent is incentivized to preserve its own $\mathcal{U}$-maximizing behavior. In Stephen Omohundro's terms, "goal-content integrity" is an *instrumentally convergent* goal of almost all intelligent agents (Omohundro

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Recent topics
Older work & basics
Where to start

Work recently started: **Low-impact agents**

Reduced Impact Artificial Intelligences

Stuart Armstrong[*1,2] and Benjamin Levinstein[†1]

[1]The Future of Humanity Institute, Faculty of Philosophy,
University of Oxford, Suite 1, Littlegate House, 16/17 St Ebbes
Street, Oxford OX1 1PT UK
[2]Machine Intelligence Research Institute, 2030 Addison Street
#300, Berkeley, CA 94704

2015

Abstract

There are many goals for an AI that could become dangerous if the AI
becomes superintelligent or otherwise powerful. Much work on the AI con-
trol problem has been focused on constructing AI goals that are safe even
for such AIs. This paper looks at an alternative approach: defining a gen-
eral concept of 'reduced impact'. The aim is to ensure that a powerful AI
which implements reduced impact will not modify the world extensively,
even if it is given a simple or dangerous goal. The paper proposes various
ways of defining and grounding reduced impact, and discusses methods

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Recent topics
Older work & basics
Where to start

Work recently started: **Ambiguity identification**

The Value Learning Problem

Nate Soares
Machine Intelligence Research Institute
nate@intelligence.org

Abstract

Autonomous AI systems' programmed goals can easily fall short of programmers' intentions. Even a machine intelligent enough to understand its designers' intentions would not necessarily *act* as intended. We discuss early ideas on how one might design smarter-than-human AI systems that can inductively learn what to value from labeled training data, and highlight questions about the construction of systems that model and act upon their operators' preferences.

Introduction

Standard texts in AI safety and ethics, such as Weld and Etzioni (1994) or Anderson and Anderson (2011), generally

ties remains mutually beneficial. [...] In other words, even if AIs become much more productive than we are, it will remain to their advantage to trade with us and to ours to trade with them.

As noted by Benson-Tilsen and Soares (forthcoming 2016), however, rational trade presupposes that agents expect more gains from trade than from coercion. Non-human species have various "comparative advantages" over humans, but humans generally exploit non-humans through force. Similar patterns can be observed in the history of human war and conquest. Whereas agents at similar capability levels have incentives to compromise, collaborate, and trade, agents with strong power advantages over others can have incen-

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Recent topics
Older work & basics
Where to start

Work recently started: **Conservatism**

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Recent topics
Older work & basics
Where to start

Work recently started: **Specifying environmental goals using sensory data**

Formalizing Two Problems of Realistic World-Models

Nate Soares
Machine Intelligence Research Institute
nate@intelligence.org

Abstract

An intelligent agent embedded within the real world must reason about an environment which is larger than the agent, and learn how to achieve goals in that environment. We discuss attempts to formalize two problems: one of induction, where an agent must use sensory data to infer a universe which embeds (and computes) the agent, and one of interaction, where an agent must learn to achieve complex goals in the universe. We review related problems formalized by Solomonoff and Hutter, and explore challenges that arise when attempting to formalize analogous problems in a setting where the agent is embedded within the environment.

problem where the agent is separate from the environment, and Section 3 discusses troubles that arise when attempting to formalize the analogous naturalized induction problem. Section 4 discusses Hutter's interaction problem, and Section 5 discusses an open problem related to formalizing an analogous naturalized interaction problem.

Formalizing these problems is important in order to fully understand the problem faced by an intelligent agent embedded within the universe: a general artificial intelligence must be able to learn about the environment which computes it, and learn how to achieve its goals from inside its universe. Section 6 concludes with a discussion of why a theoretical understanding of agents interacting with their own environment seems

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Recent topics
Older work & basics
Where to start

Work recently started: **Inverse reinforcement learning**

Learning the Preferences of Bounded Agents

Owain Evans
University of Oxford

Andreas Stuhlmüller
Stanford University
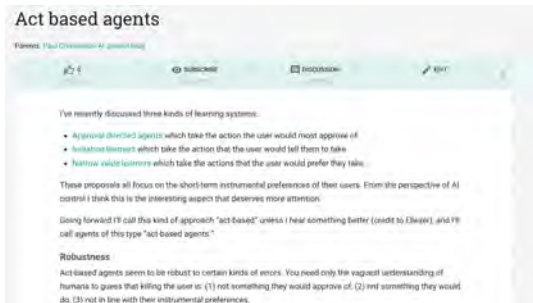
Noah D. Goodman
Stanford University

Introduction

A range of work in applied machine learning, psychology, and social science involves inferring a person's preferences and beliefs from their choices or decisions. This includes work in economics on *Structural Estimation*, which has been used to infer beliefs about the rewards of education from observed work and education choices [1] and preferences for health outcomes from smoking behavior [2]. In machine learning, *Inverse Reinforcement Learning* has been applied to diverse planning and decision tasks to learn preferences and task-specific strategies [3, 4]. Large-scale systems in industry also learn preferences from behavior: for example, people's behavior on social networking sites is used to infer what movies, articles, and photos they will like [5].

Existing approaches to inferring human beliefs and preferences typically assume that human behavior is optimal up to unstructured "random noise" [6, 7]. However, human behavior may deviate from optimality in systematic ways. This can be due to biases such as time inconsistency and framing effects [8, 9] or due to planning or inference being a (perhaps resource-rational) approximation to optimality [10, 11]. If such deviations from optimality are not modeled, we risk mistaken inferences

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Recent topics
Older work & basics
Where to start

Work recently started: **Act-based agents**

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Recent topics
Older work & basics
Where to start

Work recently started: **Mild optimization**

Quantilizers: A Safer Alternative to Maximizers for Limited Optimization

Jessica Taylor
Machine Intelligence Research Institute
jessica@intelligence.org

Abstract

In the field of AI, *expected utility maximizers* are commonly used as a model for idealised agents. However, expected utility maximization can lead to unintended solutions when the utility function does not quantify everything the operators care about: imagine, for example, an expected utility maximizer tasked with winning money on the stock market, which has no regard for whether it accidentally causes a market crash. Once AI systems become sufficiently intelligent and powerful, these unintended solutions could become quite dangerous. In this paper, we describe an alternative to expected utility maximization for powerful AI systems, which we call *expected utility quantilization*. This could allow the construction of AI systems that do not necessarily fall into strange and unantici-

utility function, with $U(o)$ being the utility of outcome o. Then an expected utility maximizer is an agent that chooses an action $a \in \mathcal{A}$ that maximizes $\mathbb{E}\left[U\left(W\left(a\right)\right)\right]$.

We make no argument against expected utility maximization on the grounds of rationality. However, maximizing the expectation of some utility function could produce large unintended consequences whenever U does not accurately capture all the relevant criteria. Some unintended consequences of this form can already be observed in modern AI systems. For example, consider the genetic algorithm used by Nguyen, Yosinski, and Clune [8] to generate an image which would be classified by a deep neural network as a starfish, with extremely high confidence. The resulting image ended up completely unrecognizable, looking nothing at all like a starfish.

Of course, Nguyen, Yosinski, and Clune [8] intended to develop images that would be mis-classified, but

Agents and their utility functions   Recent topics
Some AI alignment subproblems   **Older work & basics**
Why expect difficulty?   Where to start
Where we are now

Past developments: **AIXI**

Technical Report IDSIA-01-03   In *Artificial General Intelligence*, 2007

UNIVERSAL ALGORITHMIC INTELLIGENCE
A mathematical top→down approach

Marcus Hutter

IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch   http://www.idsia.ch/~marcus

17 January 2003

Keywords

Artificial intelligence; algorithmic probability; sequential decision theory; rational agents; value function; Solomonoff induction; Kolmogorov complexity; reinforcement learning; universal sequence prediction; strategic games; function minimization; supervised learning.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Recent topics
Older work & basics
Where to start

Past developments: **Tiling agents**

- Only Π_1 goals, on pain of Procrastination Paradox

Vingean Reflection: Reliable Reasoning for Self-Improving Agents

Benja Fallenstein and Nate Soares
Machine Intelligence Research Institute
{benja,nate}@intelligence.org

Abstract

Today, human-level machine intelligence is in the domain of futurism, but there is every reason to expect that it will be developed eventually. Once artificial agents become able to improve themselves further, they may far surpass human intelligence, making it vitally important to ensure that the result of an "intelligence explosion" is aligned with human interests. In this paper, we discuss one aspect of this challenge: ensuring that the initial agent's reasoning about its future versions is reliable, even if these future versions are far more intelligent than the current reasoner. We refer to reasoning of this sort as *Vingean reflection*.

intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make.

Almost fifty years later, a machine intelligence that is smart in the way humans are remains the subject of futurism and science fiction. But barring global catastrophe, there seems to be little reason to doubt that humanity will *eventually* create a smarter-than-human

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Recent topics
Older work & basics
Where to start

Past developments: **Software agent cooperation**

- Avoiding causal decision theory's reflective inconsistency
- Updateless decision theory, PrudentBot
- Logical counterfactuals

Program Equilibrium in the Prisoner's Dilemma via Löb's Theorem

Patrick LaVictoire
Quixey
278 Castro Street
Mountain View, CA 94041
patrick@quixey.com

Benja Fallenstein and Eliezer Yudkowsky
Machine Intelligence Research Institute
2030 Addison Street #300, Berkeley, CA 94703
benja@intelligence.org, eliezer@intelligence.org

Mihaly Barasz
Nilcons
Albisstrasse 22
Adliswil, CH-8134, Switzerland
klao@nilcons.com

Paul Christiano
University of California at Berkeley
Department of Computer Science
387 Soda Hall, Berkeley, CA 94720
paulfchristiano@eecs.berkeley.edu

Marcello Herreshoff
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
marcelloh@google.com

Abstract

Applications of game theory often neglect that real-world agents normally have some amount of out-of-band information about each other. We consider the limiting case of a one-shot Prisoner's Dilemma between algorithms with read-access to one anothers' source code. Previous work has shown that cooperation is possible at a Nash equilibrium in this setting, but existing constructions require interacting agents to be identical or near-identical. We show that a natural class of agents are able to achieve mutual cooperation at Nash equilibrium without any prior coordination of this sort.

1 Introduction

Can cooperation in a one-shot Prisoner's Dilemma be jus-

This stronger assumption suggests a convenient logical formalism. In the 1980s, Binmore (1987) considered game theory between programs which could read each other's source code before playing[1]:

...a player needs to be able to cope with hypotheses about the reasoning processes of the opponents other than simply that which maintains that they are the same as his own. Any other view risks relegating rational players to the role of the "unlucky" Bridge expert who usually loses but explains that his play is "correct" and would have led to his winning if only the opponents had played "correctly". Crudely, rational behavior should include the capacity to exploit bad play by the oppo-nents.

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Recent topics
Older work & basics
Where to start

Past developments: **Reflective oracles**

- Also reflective propositional probability
- Also reflective, quantified logical uncertainty (subproblem of Vingean reflection)

Reflective Oracles: A Foundation for Classical Game Theory

Benja Fallenstein and Jessica Taylor
Machine Intelligence Research Institute
{benja,jessica}@intelligence.org

Paul F. Christiano
UC Berkeley
paulfchristiano@eecs.berkeley.edu

Abstract

Classical game theory treats players as special—a description of a game contains a full, explicit enumeration of all players—even though in the real world, "players" are no more fundamentally special than rocks or clouds. It isn't trivial to find a decision-theoretic foundation for game theory in which an agent's coplayers are a non-distinguished part of the agent's environment. Attempts to model both players and the environment as Turing machines, for example, fail for standard diagonalization reasons.

In this paper, we introduce a "reflective" type of oracle, which is able to answer questions about the outputs of oracle machines with access to the same oracle. These oracles avoid

be boundedly rational reasoners, which make decisions with finite computational resources. Nevertheless, the notion of a perfect Bayesian reasoner provides an analytically tractable first approximation to the behavior of real-world agents, and underlies an enormous body of work in statistics [6], economics [7], computer science [8], and other fields.

On closer examination, however, the assumption that agents can compute what outcome each of their actions leads to in every possible world is troublesome even if we assume that agents have unbounded computing power. For example, consider the game of *Matching Pennies*, in which two players each choose between two actions ("heads" and "tails"); if the players choose the same action, the first player wins a dollar, if they choose differently, the second player wins. Suppose further that both players' decision-making processes are Tur-

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Recent topics
Older work & basics
Where to start

Where can you work on this?

- Machine Intelligence Research Institute (Berkeley)
- Future of Humanity Institute (Oxford University)
- Stuart Russell (UC Berkeley)
- Leverhulme CFI is starting up (Cambridge UK)

  contact@intelligence.org

Agents and their utility functions
Some AI alignment subproblems
Why expect difficulty?
Where we are now

Recent topics
Older work & basics
Where to start

Questions?

Email: contact@intelligence.org

Resources (incl. slides): intelligence.org/stanford-talk