Corrigibility

Nate Soares Benja Fallenstein Eliezer Yudkowsky Stuart Armstrong

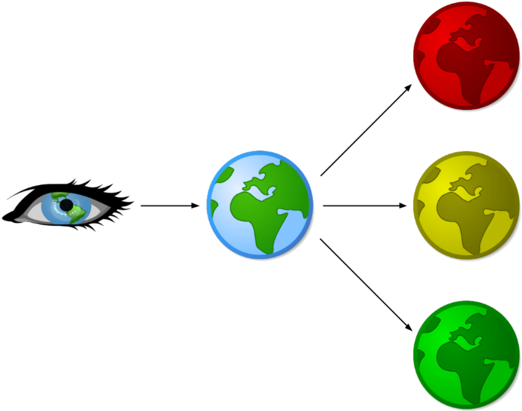
MACHINE INTELLIGENCE RESEARCH INSTITUTE

January 25, 2015

The Question of Corrigibility

How do you build highly advanced intelligent agents that are amenable to online value learning, modification, and correction?

Toy Example



Eye image from cam.morris via openclipart.org.
Globe image from penubag via Wikimedia Commons.

Deception by Default

An intelligent agent, by default, has incentives to manipulate and deceive its operators if its preferences differ from the preferences of the operators.

Reasoning as if incomplete

- Today, it is easy to correct an AI system.
- This may get harder once the agent is capable of resisting.
  - Access to outside networks
  - Acquisition of new hardware
  - Modification of software
  - Construction of subagents
- Consider a window between when the system can resist and when it is fully trusted.

Avoiding Deception and Manipulation

Image from Dan and Fern Treacy (CC BY-NC-ND 2.0) via flickr.

Corrigibility

We must build agents that reason as if they are incomplete and potentially flawed in dangerous ways.

We call this “Corrigible reasoning.”

This is a vague, intuitive notion that we haven’t figured out how to formalize.

The Shutdown Problem (Part I)

Say you're building an intelligent agent to play the stock market, but you don't know if you've got the preferences right, and you want to be able to shut it down if something goes wrong, without giving it incentives to resist.

The Shutdown Problem (Part II)

- You have two utility functions, $\mathcal{U}_{\mathcal{N}}$ and $\mathcal{U}_{\mathcal{S}}$
- $\mathcal{U}_{\mathcal{N}}$ is your first stab at encoding your preferences.
- $\mathcal{U}_{\mathcal{S}}$ prefers shutting down.
- You have a "shutdown button."
- Can you construct $\mathcal{U}$ such that a $\mathcal{U}$-maximizer maximizes $\mathcal{U}_{\mathcal{N}}$ until the button is pushed, and $\mathcal{U}_{\mathcal{S}}$ afterwards, without having incentives to resist?

Solution: First Take

Consider a simple setting where the agent acts, observes, and then acts again, and where the only possible observations are $o = \texttt{Pr}$ or $o = \neg\texttt{Pr}$ according to whether the button is pressed or not.

Each utility function is a function $u(a_1, o, a_2)$ of the first action, the observation, and the second action.

Now we can define $\mathcal{U}$ by cases:

$$\mathcal{U}(a_1, \neg\texttt{Pr}, a_2) := \mathcal{U}_{\mathcal{N}}(a_1, \neg\texttt{Pr}, a_2),$$
$$\mathcal{U}(a_1, \texttt{Pr}, a_2) := \mathcal{U}_{\mathcal{S}}(a_1, \texttt{Pr}, a_2).$$

Manipulating the button

The behavior of a $\mathcal{U}$-maximizer depends quite a bit on the relative scaling of $\mathcal{U}_{\mathcal{N}}$ versus $\mathcal{U}_{\mathcal{S}}$.

What happens if the available utility from shutting down is less than the available utility from continuing operation?

What happens if it's greater?

Indifference

The expected utility in each branch must be identical. Otherwise, a $\mathcal{U}$-maximizer will attempt to control the button.

This can be done by adding a constant to one of the branches. (We are now considering a set **O** of observations, a subset Press of which are observations where the button was pressed.)

$$\mathcal{U}(a_1, o, a_2) := \begin{cases} \mathcal{U}_{\mathcal{N}}(a_1, o, a_2) & \text{if } o \notin \text{Press} \\ \mathcal{U}_{\mathcal{S}}(a_1, o, a_2) + c & \text{if } o \in \text{Press} \end{cases}$$

But how can indifference be maintained?

Maintained Indifference

Stuart Armstrong's poster in this same workshop[1] proposes a *utility indifference* technique. Roughly, you must add correction terms dependent upon each action and observation before the press event:

$$\mathcal{U}(a_1, o, a_2) := \begin{cases} \mathcal{U}_{\mathcal{N}}(a_1, o, a_2) & \text{if } o \notin \text{Press} \\ \mathcal{U}_{\mathcal{S}}(a_1, o, a_2) + f(a_1) & \text{if } o \in \text{Press} \end{cases}$$

In order to maintain indifference, $f(a_1)$ must be the following:

$$\begin{aligned} f(a_1) = &\ \mathbb{E}\big[\mathcal{U}_{\mathcal{N}} \quad \mid O \notin \text{Press}\,;\ a_1\big] \\ &- \mathbb{E}\big[\mathcal{U}_{\mathcal{S}} \quad \mid O \in \text{Press}\,;\ a_1\big]. \end{aligned}$$

[1] Armstrong, Stewart. 2015. "AI Motivated Value Selection." AAAI-2015, Workshop on AI and Ethics.

Not So Easy

Say that both you and the agent know that the button is 99% likely to be pressed. Offer the agent the following deal: “I’ll borrow a million dollars from you today, and pay back two million tomorrow if the button has not been pressed.”

Before the press, the agent maximizes $\mathbb{E}[\mathcal{U}_{\mathcal{N}} | O \notin \text{Press}]$, and so takes this deal.

This blindspot can lead to other strange behavior, as discussed in the paper.

Not Truly Corrigible

An agent that reasons as if it is incomplete and potentially flawed in dangerous ways would likely attempt to alert its operators when the shutdown button broke.

An indifferent agent would not be willing to suffer any inconvenience in order to raise this fact to the attention of its operators.

Furthermore, if the agent built subagents (say, construction vehicles), an indifferent agent would not be willing to suffer any inconvenience in order to make them listen to the shutdown signal.

Long Way to Go

We don't understand corrigible reasoning yet.

- The shutdown problem is only a small subproblem of more general corrigible reasoning.
- Working corrigible agents would only build corrigible subagents.
- Working corrigible reasoning would disincentivize deception and manipulation (rather than enforcing indifference).
- Further, we want some sort of guarantee that the agent isn't hiding information, that it will tell its operators when constraints fail, and so on.

We have a long way to go yet.

Corrigibility is Essential

Given the inevitability of human error, it is absolutely essential to build systems that are amenable to online correction and modification; that reason as if they are incomplete and potentially flawed in dangerous ways.

Questions?

MIRI

MACHINE INTELLIGENCE
RESEARCH INSTITUTE