Fundamental Difficulties in Aligning Advanced AI

Nate Soares

Adapted from a talk by Eliezer Yudkowsky
“The primary concern is not spooky emergent consciousness but simply the ability to make high-quality decisions.”

—Stuart Russell
Task: Fill cauldron.
Broom’s utility function:

\[ U_{broom} = \begin{cases} 
1 & \text{if cauldron full} \\
0 & \text{if cauldron empty} 
\end{cases} \]
Broom’s utility function:

\[ U_{broom} = \begin{cases} 
1 & \text{if cauldron full} \\
0 & \text{if cauldron empty} 
\end{cases} \]

Actions \( a \in \mathcal{A} \), broom calculates: \( \mathbb{E} [U_{broom} \mid a] \)
Broom’s utility function:

\[
U_{\text{broom}} = \begin{cases} 
1 & \text{if cauldron full} \\
0 & \text{if cauldron empty} 
\end{cases}
\]

Actions \( a \in \mathcal{A} \), broom calculates: \( \mathbb{E}[U_{\text{broom}} | a] \)

Broom outputs: sorta-argmax \( \mathbb{E}_{a \in \mathcal{A}}[U_{\text{broom}} | a] \)
**Difficulty 1...**

Broom’s utility function:

\[
U_{\text{broom}} = \begin{cases} 
1 & \text{if cauldron full} \\
0 & \text{if cauldron empty} 
\end{cases}
\]

Human’s utility function:

\[
U_{\text{human}} = \begin{cases} 
1 & \text{if cauldron full} \\
0 & \text{if cauldron empty} \\
10 & \text{if workshop flooded} 
\end{cases}
\]
**Difficulty 1…**

Broom’s utility function:

\[ U_{broom} = \begin{cases} 
1 & \text{if cauldron full} \\
0 & \text{if cauldron empty} 
\end{cases} \]

Human’s utility function:

\[ U_{human} = \begin{cases} 
1 & \text{if cauldron full} \\
0 & \text{if cauldron empty} \\
-10 & \text{if workshop flooded} \\
+0.2 & \text{if it’s funny} \\
-1000000 & \text{if someone gets killed} \\
\ldots & \text{and a whole lot more} 
\end{cases} \]
Difficulty 2...

\[ \mathcal{EU}(99.99\% \text{ chance of full cauldron}) > \mathcal{EU}(99.9\% \text{ chance of full cauldron}) \]
Difficulty 2...

\[ \mathcal{EU}(99.99\% \text{ chance of full cauldron}) > \mathcal{EU}(99.9\% \text{ chance of full cauldron}) \]

- Contrast “Task” - goal bounded in space, time, fulfillability, and effort required to fulfill
Difficulty 2. . .

\[ \mathcal{EU}(99.99\% \text{ chance of full cauldron}) > \mathcal{EU}(99.9\% \text{ chance of full cauldron}) \]

- Contrast “Task” - goal bounded in space, time, fulfillability, and effort required to fulfill
- “Task AGI” - not just top goal, but optimization subroutines are Tasks: nothing open-ended anywhere
Can we just press the off switch?
Simple bright ideas going wrong
The big picture
Fundamental difficulties

Nate Soares  Fundamental Difficulties in Aligning Advanced AI
Simple bright ideas going wrong
The big picture
Fundamental difficulties
Simple bright ideas going wrong
The big picture
Fundamental difficulties

Nate Soares
Fundamental Difficulties in Aligning Advanced AI
Try 1: Suspend button $B$

\[ U_{broom}^3 = \begin{cases} 
1 & \text{if cauldron full} & \text{and } B=\text{OFF} \\
0 & \text{if cauldron empty} & \text{and } B=\text{OFF} \\
1 & \text{if broom suspended} & \text{and } B=\text{ON} \\
0 & \text{otherwise} & \end{cases} \]
Try 1: Suspend button $B$

$U_{\text{broom}}^3 =$

\[
\begin{cases} 
1 \text{ if cauldron full} & \text{ & } B=\text{OFF} \\
0 \text{ if cauldron empty} & \text{ & } B=\text{OFF} \\
1 \text{ if broom suspended} & \text{ & } B=\text{ON} \\
0 \text{ otherwise}
\end{cases}
\]

Probably, $\mathbb{E} [U_{\text{broom}}^3 \mid B=\text{OFF}] < \mathbb{E} [U_{\text{broom}}^3 \mid B=\text{ON}]$
Try 1: Suspend button $B$

$$U_{broom}^3 = \begin{cases} 
1 & \text{if cauldron full} \\
0 & \text{if cauldron empty} \\
1 & \text{if broom suspended} \\
0 & \text{otherwise}
\end{cases}$$

$$\mathbb{E} [U_{broom}^3 | B=OFF] < \mathbb{E} [U_{broom}^3 | B=ON]$$

(Strategy broom tries to make you press the button.)
Simple bright ideas going wrong
The big picture
Fundamental difficulties
Fundamental Difficulties in Aligning Advanced AI

Nate Soares

Simple bright ideas going wrong
The big picture
Fundamental difficulties

humans

← -- think up goals/values

“intended” value function $V$

← -- value learning

sorta-argmax $\pi \in \text{Policies}$

Expectation $[U]$
Simple bright ideas going wrong
The big picture
Fundamental difficulties

Media Focus

humans

“intended” value function $V$

→ -- think up goals/values
→ -- value learning

“natural” desires $[X]$

sorta-argmax $\pi \in Policies$

Expectation $[U]$
Simple bright ideas going wrong

The big picture

Fundamental difficulties

Political Derailment

humans

← -- think up goals/values

“intended” value function $V$

← -- value learning

$\text{sorta-argmax} \quad \text{Expectation}[U]$

$\pi \in \text{Policies}$
Simple bright ideas going wrong
The big picture
Fundamental difficulties

Early Science Fiction

humans

\[ \text{think up goals/values} \]

\[ \text{value learning} \]

"intended" value function \( V \)

sorta-argmax \[ \pi \in \text{Policies} \]

Expectation \[ [U] \]

Nate Soares
Fundamental Difficulties in Aligning Advanced AI
Simple bright ideas going wrong
The big picture
Fundamental difficulties

MIRI’s Concerns

humans

"intended" value function \( V \)

\[ \text{sorta-argmax} \quad \pi \in \text{Policies} \]

Expectation [U]

\[ \leftarrow\leftarrow \text{think up goals/values} \]

\[ \leftarrow\leftarrow \text{value learning} \]
Take-home message: We’re afraid it’s going to be technically difficult to point AIs in an intuitively intended direction.
Take-home message: We’re afraid it’s going to be technically difficult to point AIs in an intuitively intended direction.

...and if we screw up there, it doesn’t matter which human is standing closest to the AI.
Four key propositions:

1. **Orthogonality** – An AI system can be built to pursue almost any objective, in theory
Four key propositions:

1. **Orthogonality** – An AI system can be built to pursue almost any objective, in theory

2. **Instrumental convergence** – most objectives imply survival, resource acquisition, etc. as instrumental subgoals
Four key propositions:

1. **Orthogonality** – An AI system can be built to pursue almost any objective, in theory

2. **Instrumental convergence** – most objectives imply survival, resource acquisition, etc. as instrumental subgoals

3. **Capability gain** – there are potential ways for artificial agents to greatly gain in cognitive power and strategic options
Four key propositions:

1. **Orthogonality** – An AI system can be built to pursue almost any objective, in theory
2. **Instrumental convergence** – most objectives imply survival, resource acquisition, etc. as instrumental subgoals
3. **Capability gain** – there are potential ways for artificial agents to greatly gain in cognitive power and strategic options
4. **Alignment difficulty** – there’s at least one part of “build an AI that does a big right thing” which is a deep, technical, hard AI problem
AI alignment is difficult. . .

. . .like rockets are difficult.

(Huge stresses break things that don’t break in normal engineering.)
AI alignment is difficult...

...like space probes are difficult.

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

...sort of like computer security is difficult.

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

Treat it like a secure rocket probe.
AI alignment:

Treat it like a secure rocket probe.

Take it seriously.
AI alignment:

Treat it like a secure rocket probe.

Don’t expect it to be easy.
AI alignment:

Treat it like a secure rocket probe.
AI alignment:

Treat it like a secure rocket probe.

Don’t defer thinking until later.
AI alignment:

Treat it like a secure rocket probe.

Formalize ideas so others can critique and build upon them.
Questions?

Email: contact@intelligence.org

MIRI
MACHINE INTELLIGENCE RESEARCH INSTITUTE