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:

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

Broom's utility function:

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

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

Broom's utility function:

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

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

Broom outputs: $\underset{a \in \mathcal{A}}{\text{sorta-argmax}}\, \mathbb{E}\left[\mathcal{U}_{broom} \mid a\right]$

Difficulty 1...

Broom's utility function:

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

Human's utility function:

$$\mathcal{U}_{human} = \begin{cases} 1 & \text{if cauldron full} \\ 0 & \text{if cauldron empty} \\ -10 & \text{if workshop flooded} \\ & \\ & \\ & \end{cases}$$

Difficulty 1...

Broom's utility function:

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

Human's utility function:

$$\mathcal{U}_{human} = \begin{cases} 1 & \text{if cauldron full} \\ 0 & \text{if cauldron empty} \\ -10 & \text{if workshop flooded} \\ +0.2 & \text{if it's funny} \\ -1000000 & \text{if someone gets killed} \\ & \text{\ldots and a whole lot more} \end{cases}$$

Difficulty 2...

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

Difficulty 2. . .

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

- Contrast "Task" - goal bounded in space, time, fulfillability, and effort required to fulfill

Difficulty 2. . .

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

Try 1: Suspend button **B**

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

Try 1: Suspend button **B**

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

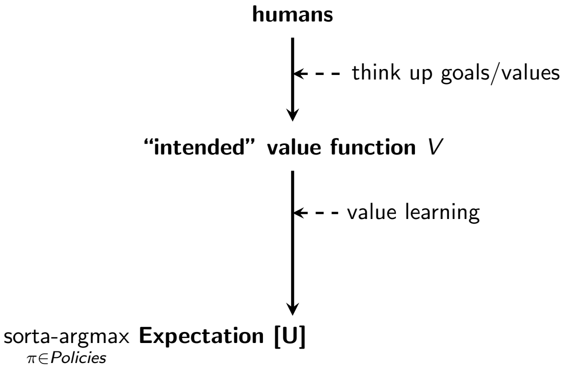
Probably, $\mathbb{E}\left[\mathcal{U}^3_{broom} \mid \mathbf{B}{=}\text{OFF}\right] < \mathbb{E}\left[\mathcal{U}^3_{broom} \mid \mathbf{B}{=}\text{ON}\right]$

Try 1: Suspend button **B**

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

Probably, $\mathbb{E}\left[\mathcal{U}^3_{broom} \mid \mathbf{B}{=}\text{OFF}\right] < \mathbb{E}\left[\mathcal{U}^3_{broom} \mid \mathbf{B}{=}\text{ON}\right]$

(Strategic broom tries to make you press the button.)



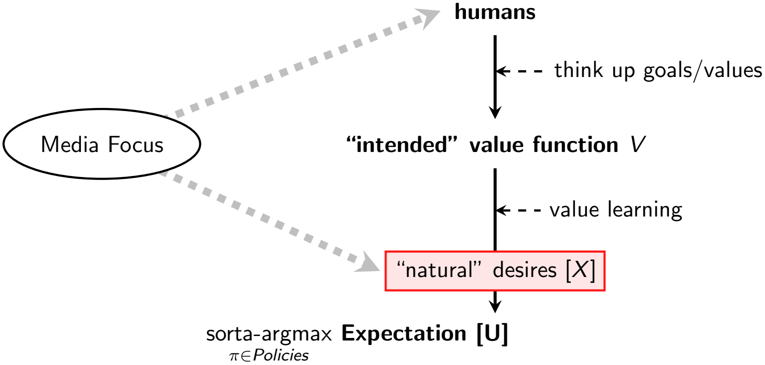
humans

think up goals/values

“intended” value function V

value learning

$$\text{sorta-argmax}_{\pi \in Policies}\ \textbf{Expectation [U]}$$



humans

think up goals/values

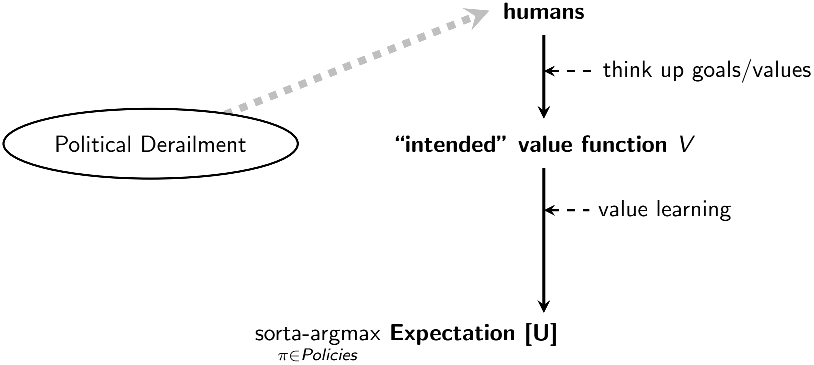
“intended” value function V

value learning

“natural” desires $[X]$

Media Focus

$\underset{\pi \in \mathit{Policies}}{\text{sorta-argmax}}$ **Expectation [U]**



humans

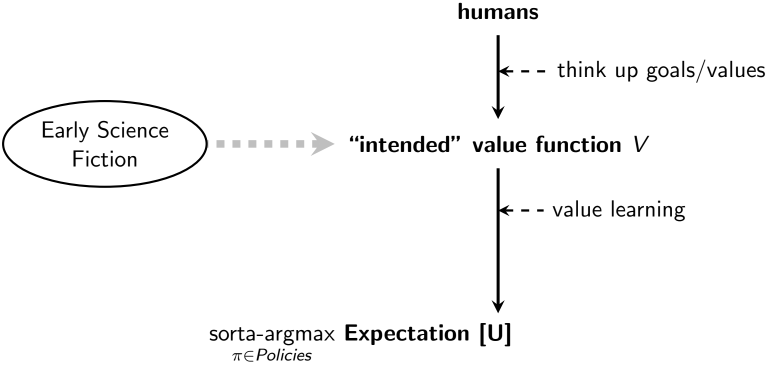
think up goals/values

Political Derailment

“intended” value function *V*

value learning

$\underset{\pi \in Policies}{\text{sorta-argmax}}$ **Expectation [U]**



humans

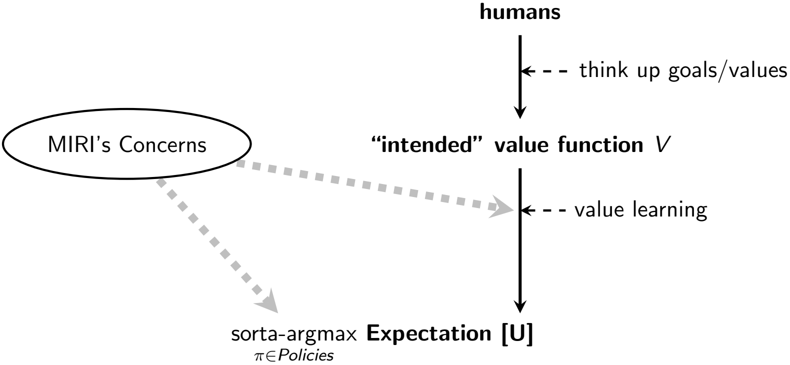
think up goals/values

Early Science Fiction

“intended” value function V

value learning

sorta-argmax$_{\pi \in \textit{Policies}}$ **Expectation [U]**



humans

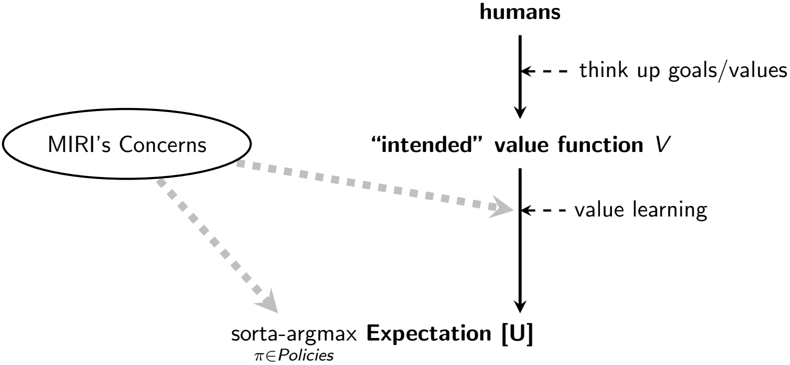
think up goals/values

“intended” value function *V*

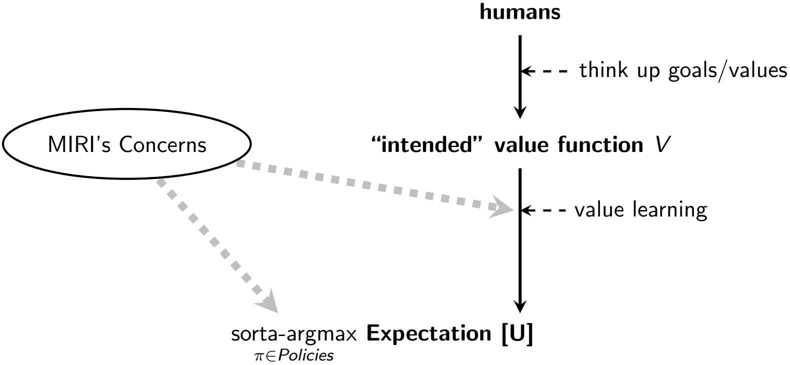
value learning

MIRI’s Concerns

sorta-argmax$_{\pi \in Policies}$ **Expectation [U]**


humans
think up goals/values
MIRI's Concerns
"intended" value function V
value learning
sorta-argmax$_{\pi \in Policies}$ **Expectation [U]**

Take-home message: We're afraid it's going to be *technically difficult* to point AIs in an intuitively intended direction.


humans
think up goals/values
“intended” value function V
MIRI's Concerns
value learning
sorta-argmax$_{\pi \in Policies}$ Expectation [U]

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. . .

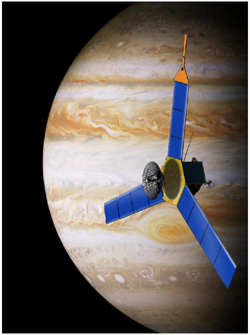
. . . like rockets are difficult.

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

AI aligment is difficult. . .

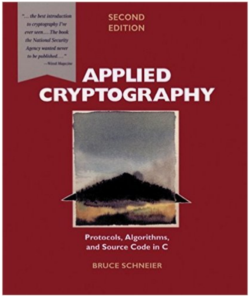
. . . like space probes are difficult.

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



AI aligment is difficult. . .

. . . *sort of* like computer security is difficult.


"... the best introduction to cryptography I've ever seen.... The book the National Security Agency wanted never to be published...."
—Wired Magazine
SECOND EDITION
APPLIED CRYPTOGRAPHY
Protocols, Algorithms, and Source Code in C
BRUCE SCHNEIER

(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