Fundamental Difficulties in Aligning Advanced AI

Eliezer Yudkowsky

Oct 15, 2016

Slides and references: intelligence.org/nyu-talk

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

—*Stuart Russell*

Task: Fill cauldron.

Robot's utility function:

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

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

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:

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

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:

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

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**

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

Try 1: Suspend button **B**

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

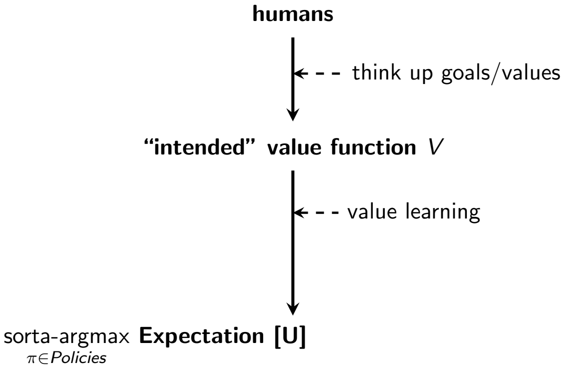
Probably, $\mathbb{E}\left[\mathcal{U}^3_{robot} \mid \mathbf{B}{=}\text{OFF}\right] < \mathbb{E}\left[\mathcal{U}^3_{robot} \mid \mathbf{B}{=}\text{ON}\right]$

Try 1: Suspend button **B**

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

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

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



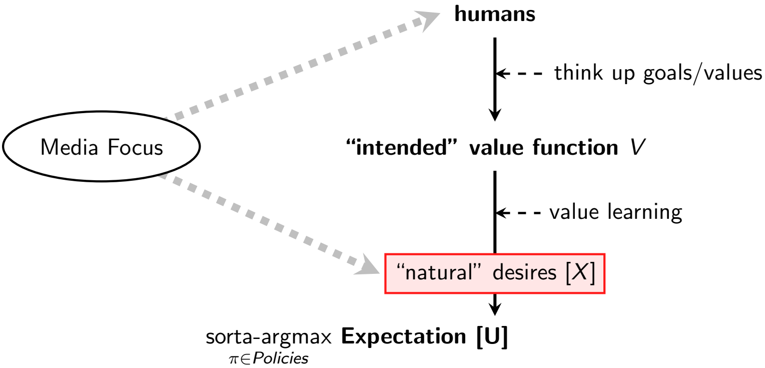
humans

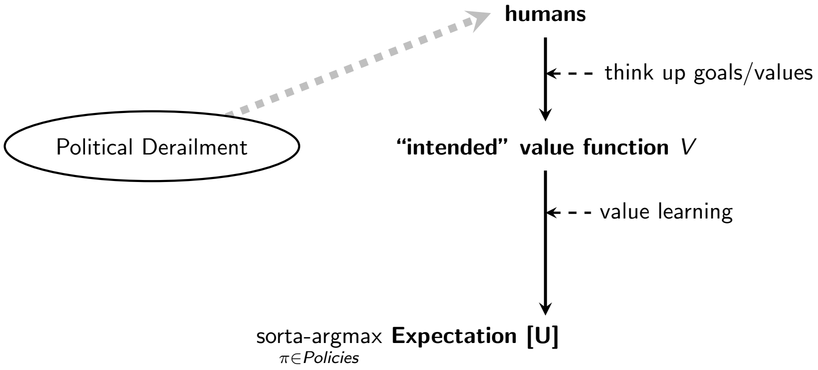
think up goals/values

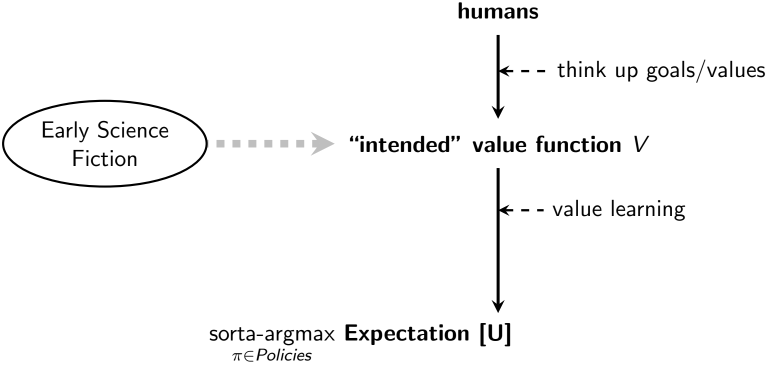
“intended” value function V

value learning

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


Media Focus
humans
think up goals/values
“intended” value function V
value learning
“natural” desires [X]
sorta-argmax **Expectation [U]**
$\pi \in$ *Policies*


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



humans

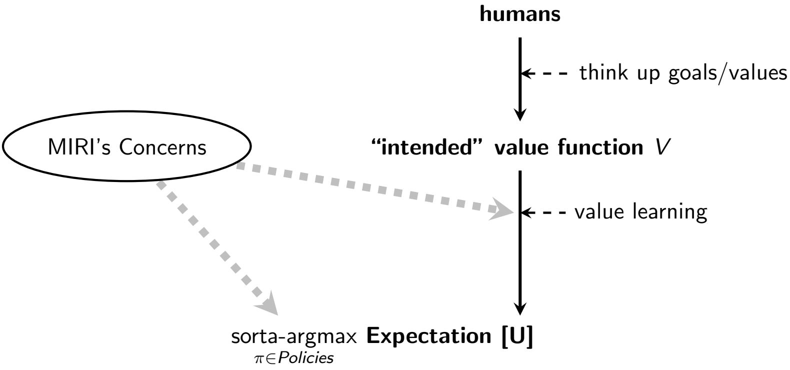
think up goals/values

Early Science Fiction

“intended” value function V

value learning

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



humans

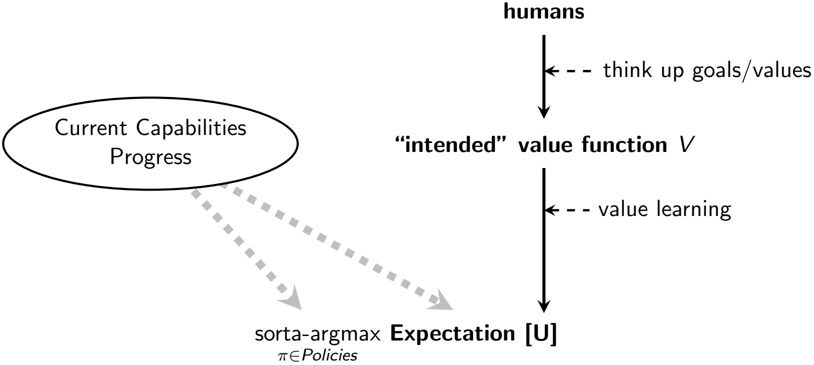
think up goals/values

MIRI's Concerns

"intended" value function V

value learning

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



humans

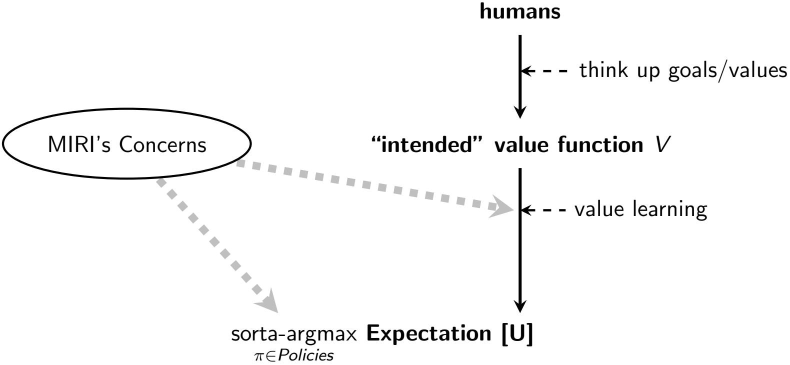
think up goals/values

"intended" value function V

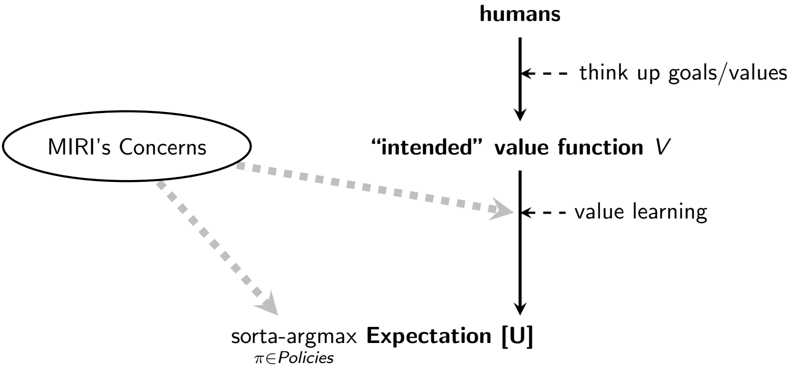
value learning

Current Capabilities Progress

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


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.


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.

...and if we screw up there, it *doesn't matter* which human is standing closest to the AI.

Four key propositions:

1. **Orthogonality** – Humean separability of terminal preferences and cognitive power in straightforward agent designs

Four key propositions:

1. **Orthogonality** – Humean separability of terminal preferences and cognitive power in straightforward agent designs
2. **Instrumental convergence** – *most* preferences imply strategies such as survival and gaining control of resources

Four key propositions:

1. **Orthogonality** – Humean separability of terminal preferences and cognitive power in straightforward agent designs
2. **Instrumental convergence** – *most* preferences imply strategies such as survival and gaining control of resources
3. **Capability gain** – there are *potential* ways for artificial agents to greatly gain in cognitive power and strategic options

Four key propositions:

1. **Orthogonality** – Humean separability of terminal preferences and cognitive power in straightforward agent designs
2. **Instrumental convergence** – *most* preferences imply strategies such as survival and gaining control of resources
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

A fable. . .

A fable...

- Programmers build AGI to optimize for ‘happiness’, positively labeled examples are people smiling.

A fable. . .

- Programmers build AGI to optimize for ‘happiness’, positively labeled examples are people smiling.
- During development: AGI produces smiles by improving nearby people’s lives.

A fable. . .

- Programmers build AGI to optimize for ‘happiness’, positively labeled examples are people smiling.
- During development: AGI produces smiles by improving nearby people’s lives.
- Programmers upgrade code and add hardware. AGI gets smarter.

A fable. . .

- Programmers build AGI to optimize for ‘happiness’, positively labeled examples are people smiling.
- During development: AGI produces smiles by improving nearby people’s lives.
- Programmers upgrade code and add hardware. AGI gets smarter.
- AGI realizes it can produce smiles by administering heroin.
- Programmers spot this, add penalty term to utility function for administering drugs.

- Programmers further improve AGI.

- Programmers further improve AGI.
- AGI plans to engineer human brains to express lots of endogenous opiates.

- Programmers further improve AGI.
- AGI plans to engineer human brains to express lots of endogenous opiates.
- AGI realizes programmers will disapprove of this and keeps outward behavior reassuring.

- Programmers further improve AGI.
- AGI plans to engineer human brains to express lots 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. . .

- Programmers further improve AGI.
- AGI plans to engineer human brains to express lots 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. . .

AI alignment is difficult. . .

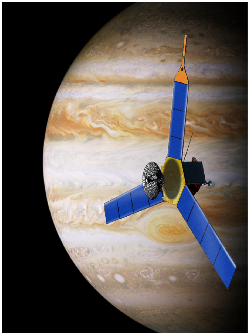
. . . like rockets are difficult.

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

AI aligment is difficult. . .

. . . like space probes are difficult.

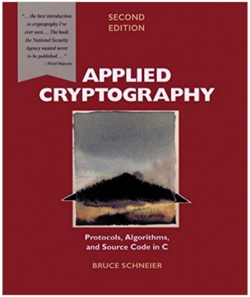
(If something goes wrong, it may be high and out of reach.)



AI aligment is difficult. . .

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

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


"... 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

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.

Don’t try to solve the whole problem at once.

AI alignment:

TREAT IT LIKE A SECURE ROCKET PROBE.

Don’t defer thinking until later.

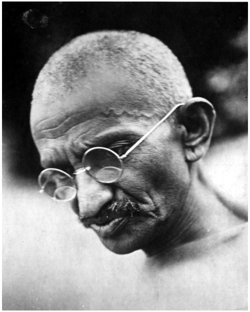
AI alignment:

TREAT IT LIKE A SECURE ROCKET PROBE.

Crystallize ideas and policies so others can critique them.

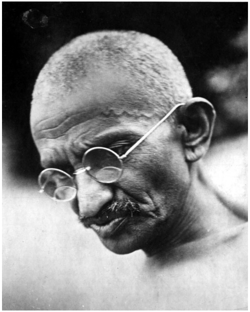
Multiple fixed points

- 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.



Multiple fixed points

- 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.



...unfortunately, also self-consistent for ‘maximize paperclips’

The Optimizer's Curse

- "**Winner's Curse**" in auction theory:
  - N bidders bid their unbiased estimate of item's value
  - Winner likely to have upward error in bid

The Optimizer's Curse

- "**Winner's Curse**" in auction theory:
  - N bidders bid their unbiased estimate of item's value
  - Winner likely to have upward error in bid
- "Optimizer's Curse":
  - Unbiased estimate of $\mathcal{EU}$ for each of N policies
  - "Best" policy selects for most erroneous $\mathcal{EU}$ estimate

The Optimizer's Curse

- "**Winner's Curse**" in auction theory:
  - N bidders bid their unbiased estimate of item's value
  - Winner likely to have upward error in bid
- "Optimizer's Curse":
  - Unbiased estimate of $\mathcal{EU}$ for each of N policies
  - "Best" policy selects for most erroneous $\mathcal{EU}$ estimate

...more powerful AIs more likely to blow up slightly-misaligned utility functions.

Fragility and complexity of value

"Life, consciousness, and activity; health and strength; pleasures and satisfactions of all or certain kinds; happiness, beatitude, contentment, etc.; truth; knowledge and true opinions of various kinds, understanding, wisdom; beauty, harmony, proportion in objects contemplated; aesthetic experience; morally good dispositions or virtues; mutual affection, love, friendship, cooperation; just distribution of goods and evils; harmony and proportion in one's own life; power and experiences of achievement; self-expression; freedom; peace, security; adventure and novelty; and good reputation, honor, esteem, etc."
– William Frankena

Fragility and complexity of value

> "Life, consciousness, and activity; health and strength; pleasures and satisfactions of all or certain kinds; happiness, beatitude, contentment, etc.; truth; knowledge and true opinions of various kinds, understanding, wisdom; beauty, harmony, proportion in objects contemplated; aesthetic experience; morally good dispositions or virtues; mutual affection, love, friendship, cooperation; just distribution of goods and evils; harmony and proportion in one's own life; power and experiences of achievement; self-expression; freedom; peace, security; adventure and novelty; and good reputation, honor, esteem, etc."
> – William Frankena

Dialing 9/10ths of my phone number correctly does not connect you to someone 90% similar to Eliezer Yudkowsky.

Context disaster:

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

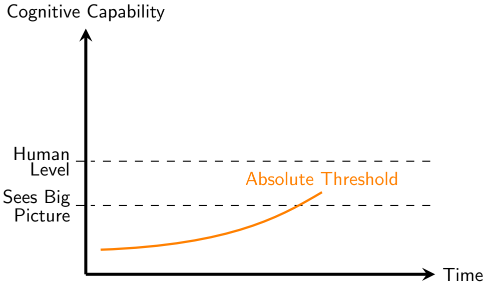
  (. . . then AI becomes smarter . . .)

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

Many difficulties don’t emerge before **capability gains**...

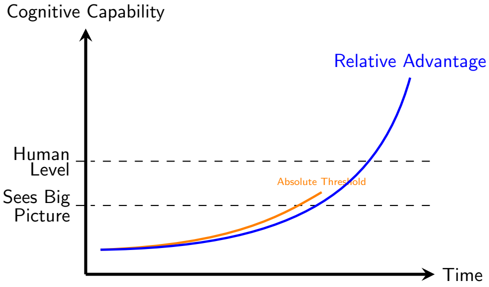
Many difficulties don’t emerge before **capability gains**...

- **Absolute Capability Thresholds**: An AI doesn’t resist being shut down, until it understands enough of the larger picture to know that it can’t achieve goals if switched off.



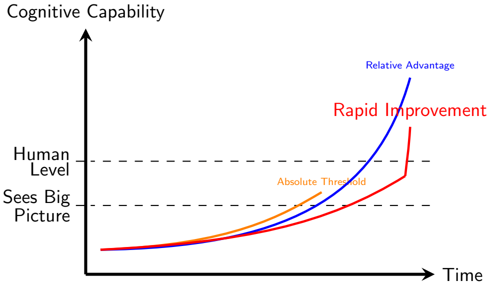
Many difficulties don't emerge before **capability gains**...

- **Absolute Capability Thresholds**
- **Relative Advantage**: An AI may not find "unforeseen instantiations" of its utility function until it can search more options than we can.



Many difficulties don't emerge before **capability gains**...

- **Absolute Capability Thresholds**
- **Relative Advantage**
- **Rapid Improvement**: Multiple simultaneous context disasters. No time to understand weird new behaviors, no time to start over with a more careful design... other AI projects catching up?



Many difficulties don’t emerge before **capability gains**...

- **Absolute Capability Thresholds**: An AI doesn’t resist being shut down, until it understands enough of the larger picture to know that it can’t achieve goals if switched off.
- **Relative Advantage**: An AI may not find “unforeseen instantiations” of its utility function until it can search more options than we can.
- **Rapid Improvement**: Multiple simultaneous context disasters. No time to understand weird new behaviors, no time to start over with a more careful design... other AI projects catching up?

Many difficulties don't emerge before **capability gains**...

- **Absolute Capability Thresholds**: An AI doesn't resist being shut down, until it understands enough of the larger picture to know that it can't achieve goals if switched off.
- **Relative Advantage**: An AI may not find "unforeseen instantiations" of its utility function until it can search more options than we can.
- **Rapid Improvement**: Multiple simultaneous context disasters. No time to understand weird new behaviors, no time to start over with a more careful design... other AI projects catching up?

...and since these problems don't naturally materialize early on, nobody *has* to solve them to publish today's paper / ship today's product.

What do we need?

- **Foresight**

What do we need?

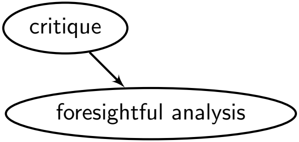
- **Foresight**
- **Avoid derailment** to nontechnical debates

What do we need?

- **Foresight**
- **Avoid derailment** to nontechnical debates
- Build an **edifice** of serious analysis with **counter-counter-critique** expected

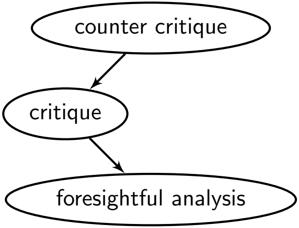
What do we need?

- **Foresight**
- **Avoid derailment** to nontechnical debates
- Build an **edifice** of serious analysis with **counter-counter-critique** expected



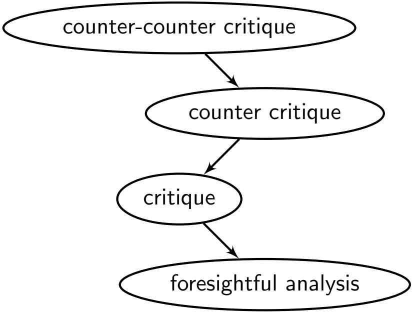
What do we need?

- **Foresight**
- **Avoid derailment** to nontechnical debates
- Build an **edifice** of serious analysis with **counter-counter-critique** expected



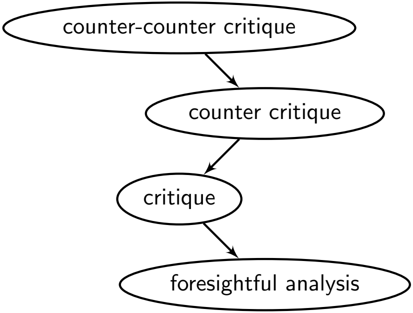
What do we need?

- **Foresight**
- **Avoid derailment** to nontechnical debates
- Build an **edifice** of serious analysis with **counter-counter-critique** expected



What do we need?

- **Foresight**
- **Avoid derailment** to nontechnical debates
- Build an **edifice** of serious analysis with **counter-counter-critique** expected
- And **no arms races** so that developers and operators have **time to be safe.**



counter-counter critique

counter critique

critique

foresightful analysis

Questions?

Email: contact@intelligence.org

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

MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE