

ENSURING

# Smarter-than-Human Intelligence

HAS A

# Positive Outcome

Nate Soares MACHINE INTELLIGENCE RESEARCH INSTITUTE

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

~ Stuart Russell



# Simple Bright Ideas Going Wrong



# *Task: Fill Cauldron*



# Broom's Utility Function



# Broom's Utility Function

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



# Broom's Utility Function

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

Actions  $a \in A$ , broom calculates:  $E[U_{\text{broom}} \mid a]$





# Broom's Utility Function

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

Actions  $a \in A$ , broom calculates:  $E[U_{\text{broom}} \mid a]$

Broom outputs:  $\underset{a \in A}{\text{sorta-argmax}} E[U_{\text{broom}} \mid a]$





*Difficulty 1*



# Human's Utility Function





# Human's Utility Function

$$U_{\text{human}} = \begin{cases} 1 & \text{if cauldron full} \\ 0 & \text{if cauldron empty} \end{cases}$$



# Human's Utility Function

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



# Human's Utility Function

$$U_{\text{human}} = \begin{cases} 1 & \text{if cauldron full} \\ 0 & \text{if cauldron empty} \\ -10 & \text{if workshop flooded} \\ +0.2 & \text{if it's funny} \end{cases}$$



# Human's Utility Function

$$U_{\text{human}} = \begin{cases} 1 & \text{if cauldron full} \\ 0 & \text{if cauldron empty} \\ -10 & \text{if workshop flooded} \\ +0.2 & \text{if it's funny} \\ -1,000,000 & \text{if someone gets killed} \end{cases}$$



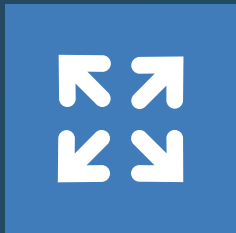


# Human's Utility Function

$$U_{\text{human}} = \begin{cases} 1 & \text{if cauldron full} \\ 0 & \text{if cauldron empty} \\ -10 & \text{if workshop flooded} \\ +0.2 & \text{if it's funny} \\ -1,000,000 & \text{if someone gets killed} \\ & \dots\text{and a whole lot more} \end{cases}$$



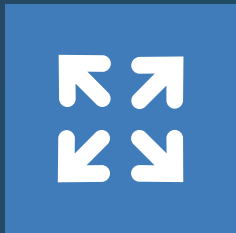
*Difficulty 2*



$99.99\% > 99.9\%$

*EU* (99.99% chance of full cauldron)

$> EU$  (99.9% chance of full cauldron)



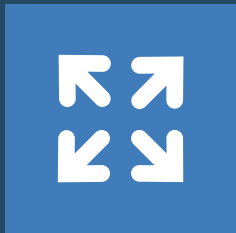
$99.99\% > 99.9\%$

*EU* (99.99% chance of full cauldron)  
> *EU* (99.9% chance of full cauldron)



## Task

Goal bounded in space,  
time, fulfillability, and  
effort required to fulfill



$99.99\% > 99.9\%$

*EU* (99.99% chance of full cauldron)

$> EU$  (99.9% chance of full cauldron)



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't we just  
press the  
off switch?*











## Suspend Button ***B***

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



## Suspend Button ***B***

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

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



## Suspend Button ***B***

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

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

*Strategic broom tries to make you press the suspend button*



# The Big Picture

Humans

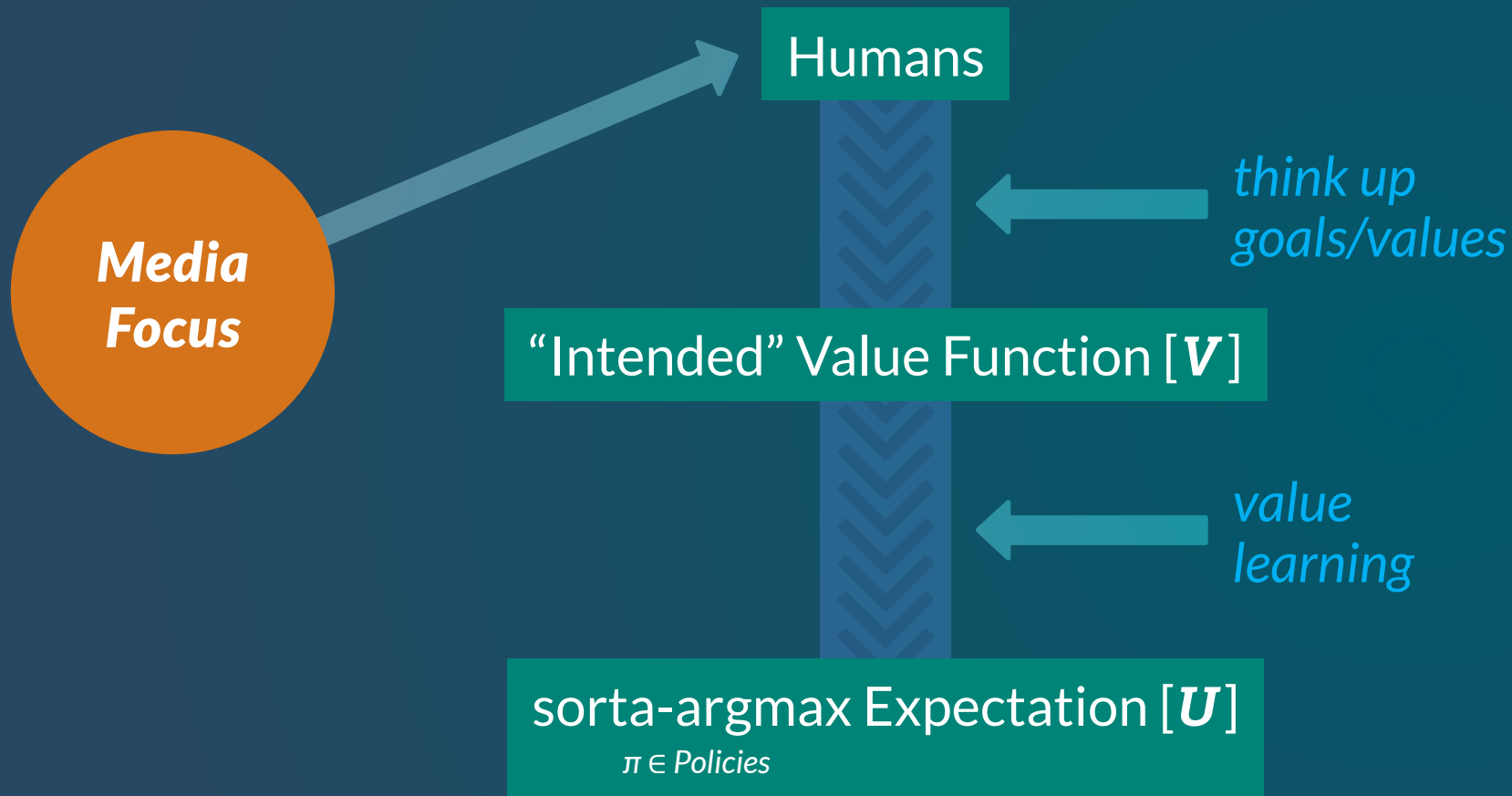
*think up  
goals/values*

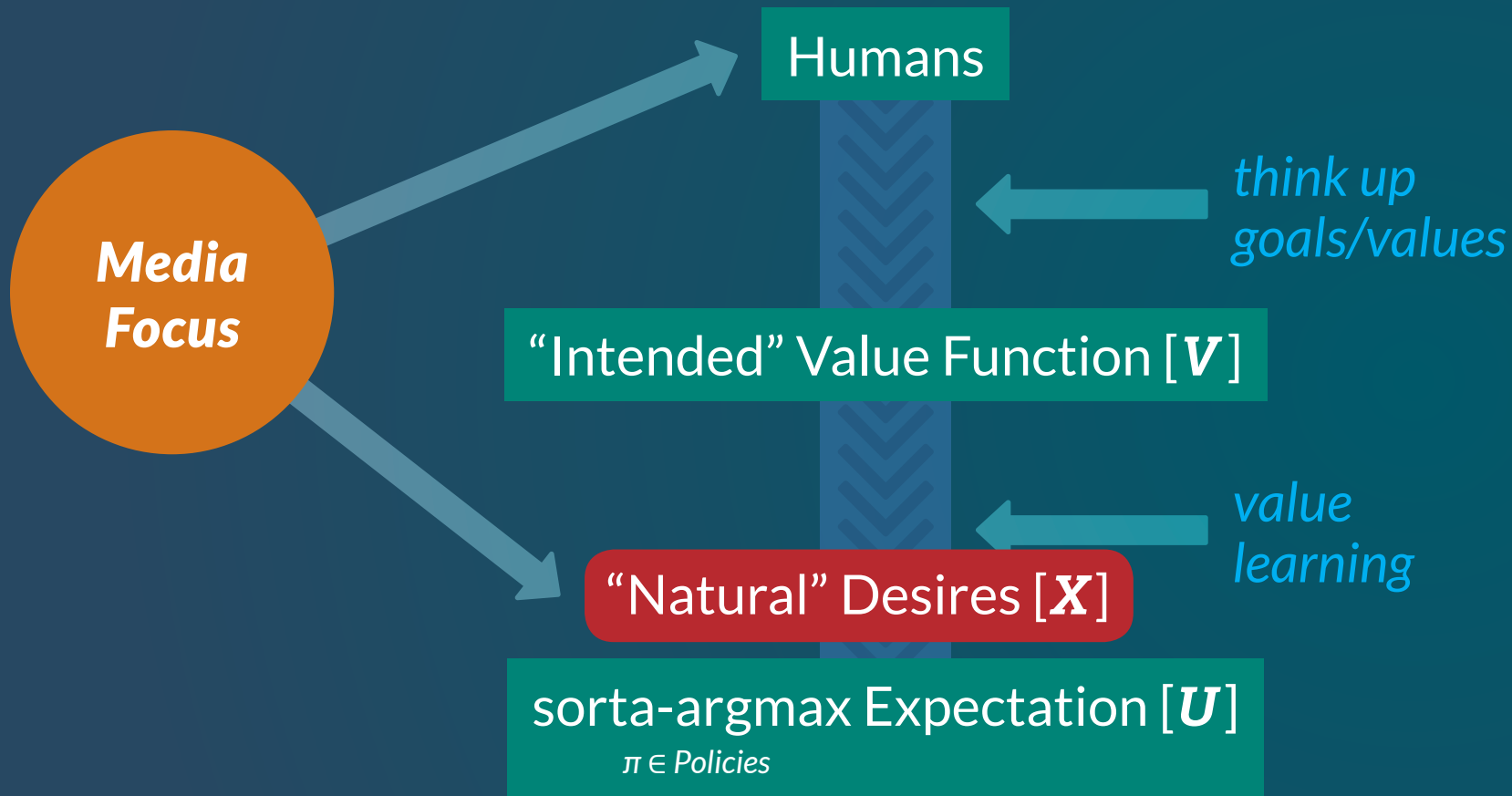
“Intended” Value Function [ $V$ ]

*value  
learning*

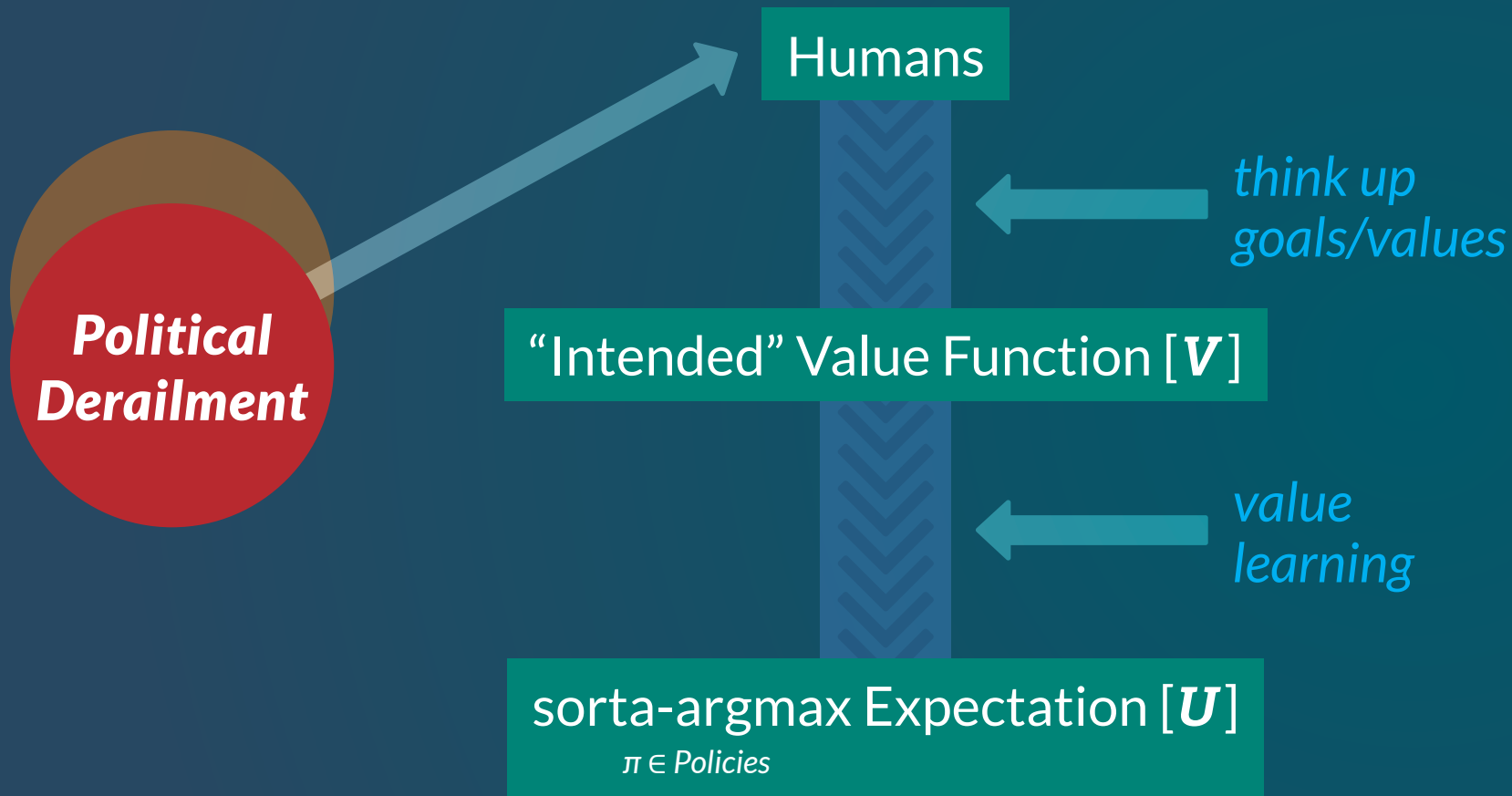
sorta-argmax Expectation [ $U$ ]

$\pi \in \text{Policies}$











**“Intended” Value Function  $[V]$**

**Humans**



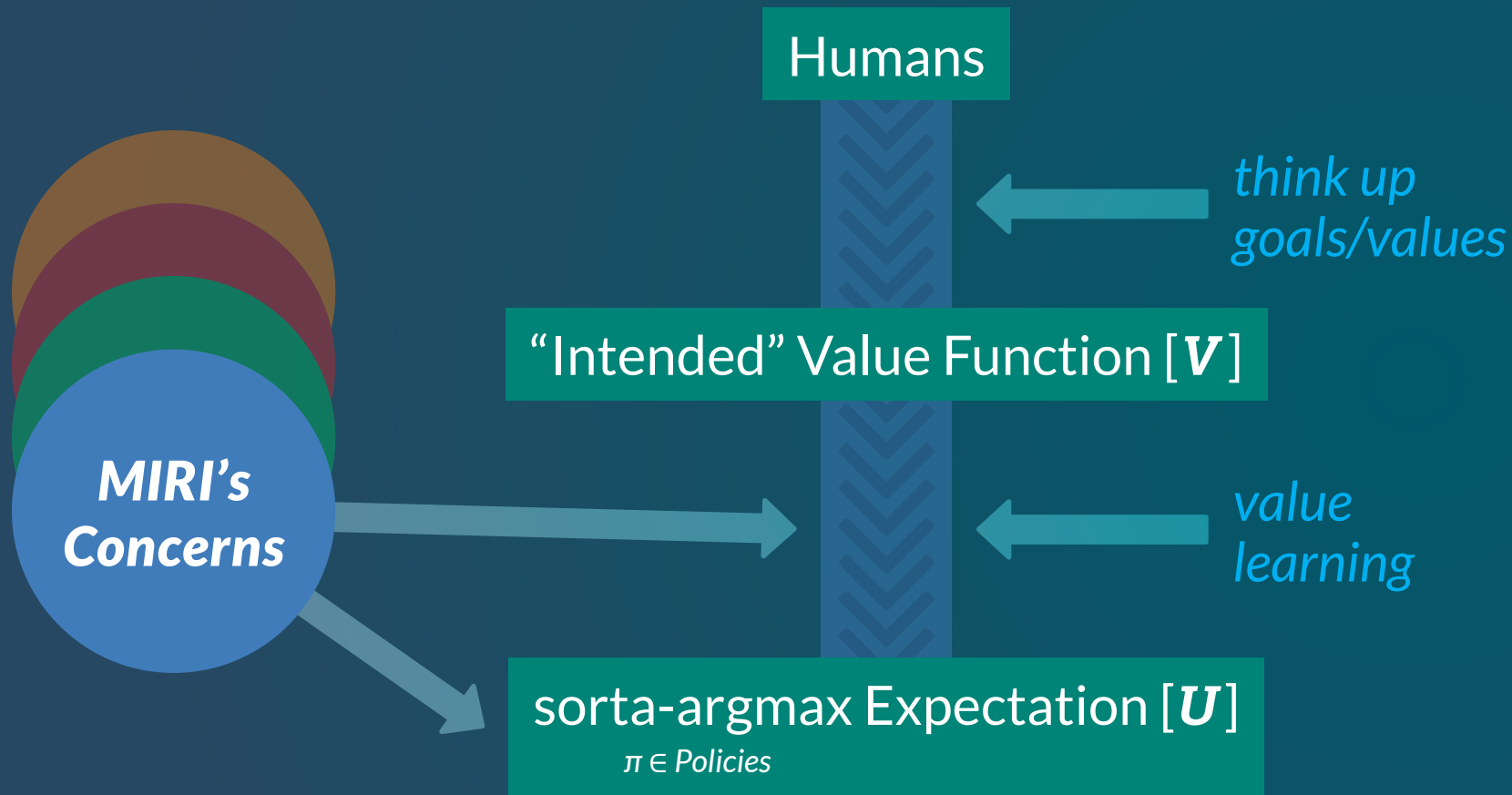
*think up  
goals/values*



*value  
learning*

**sorta-argmax Expectation  $[U]$**

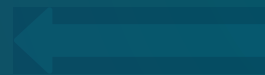
$\pi \in \text{Policies}$





# Take-Home Message

*think up  
goals/values*



*value  
learning*



# Take-Home Message

*think up  
goals/values*

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

*value  
learning*



# Take-Home Message

think up  
goals/values

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.

value  
learning

# Four Key Propositions





# 1

## Orthogonality

An AI system can be built to pursue almost any objective, in theory.

1

Orthogonality

2

Instrumental Convergence

Most objectives imply survival, resource acquisition, etc. as instrumental subgoals.

1

Orthogonality

2

Instrumental Convergence

3

Capability Gain

There are potential ways for artificial agents to greatly gain in cognitive power and strategic options.

1

Orthogonality

2

Instrumental Convergence

3

Capability Gain

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.

1

Orthogonality

2

Instrumental Convergence

3

Capability Gain

4

Alignment Difficulty

# Fundamental Difficulties

AI Alignment is difficult...

...*like rockets*

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



AI Alignment is difficult...

***...like space  
probes***

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





AI Alignment is difficult...

***...like computer  
security\****

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

*(\*kind of)*



# AI Alignment

***Treat it like a secure rocket probe***



Do



Do

➡ Take it seriously



Do

- ➡ Take it seriously
- ➡ Formalize ideas so others can critique and build upon them



Do

- ➡ Take it seriously
- ➡ Formalize ideas so others can critique and build upon them



Don't



# Do

- ➡ Take it seriously
- ➡ Formalize ideas so others can critique and build upon them



# Don't

- ➡ Expect it to be easy



# Do

- ➡ Take it seriously
- ➡ Formalize ideas so others can critique and build upon them



# Don't

- ➡ Expect it to be easy
- ➡ Defer thinking until later





# MIRI

MACHINE INTELLIGENCE  
RESEARCH INSTITUTE

[intelligence.org](https://intelligence.org)