AI safety: a box of tricks

Problems and solutions

Stuart Armstrong

Future of Humanity Institute

The problem

AI is effective at unbounded unfriendly goals.

(probably)

Problems and "solutions"

AI is effective at unbounded unfriendly goals.

Boxed Oracle
Interruptible
Satisficing

Low-impact
Virtual world values

Moral learning
Friendly utility

Tool AI
Goalless Oracle

FINDING FLAWS

Problems and "solutions"

AI is effective at unbounded unfriendly goals.

Boxed Oracle
Interruptible
Satisficing

↓

Allows danger
Satisficing→Maximising

Low-impact
Virtual world values

↓

Code all human moral concepts

Moral learning
Friendly utility

↓

Lying problems
Tool→Maximising ?

Tool AI
Goalless Oracle

Problems and "solutions"

AI is effective at unbounded unfriendly goals.

Boxed Oracle
Interruptible
Satisficing

Low-impact
Virtual world values

Moral learning
Friendly utility

Tool AI
Goalless Oracle

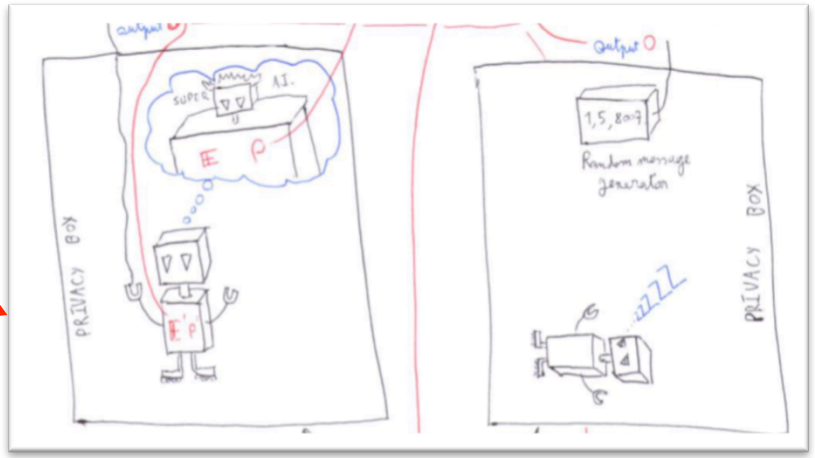
Building safety

$$R = \min_{\rho} \left\{ \mathbb{E}' \left(\frac{P(g^{\rho}|X,b,O)}{P(g^{\rho}|\neg X,b,O)} \right) > 10, \text{ or } \mathbb{E}' \left(\frac{P(g^{\rho}|\neg X,b,O)}{P(g^{\rho}|X,b,O)} \right) > 10 \right\}$$

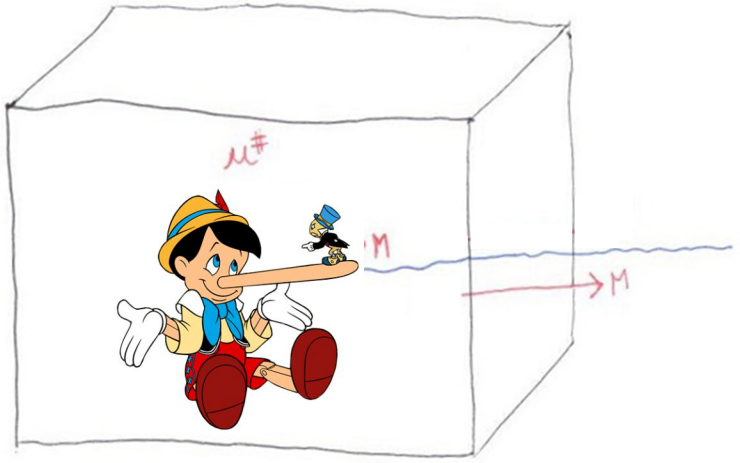
4.2 Unsafe output channel

The easiest way to allow for higher impact, is to specifically exclude the AI's output from the reduced impact requirements. For instance, assume the AI is going to send out message O. To ensure that $P(O|\neg X) \neq 0$, we set up an alternative system that will produce a random message.

Then we exclude the contents of O from the reduced impact considerations



Great One, please advise me...



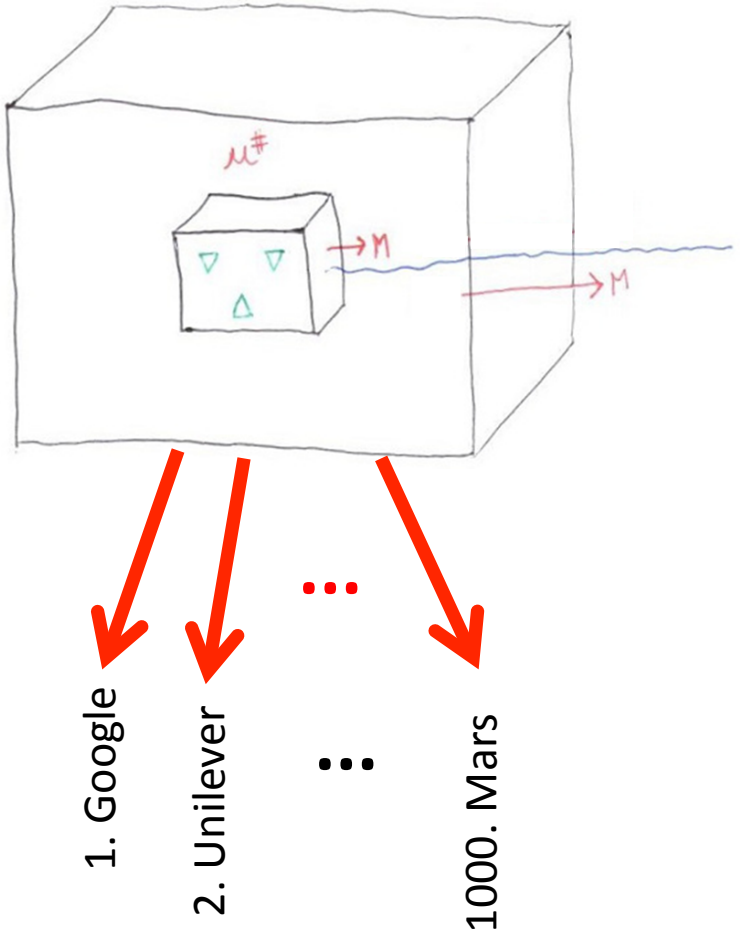
Lying is default!

What's safe to ask?

1) Checkable multiple choice, non-counterfactual.
2) Counterfactual questions about expected utilities, probabilities, conditionals, etc...

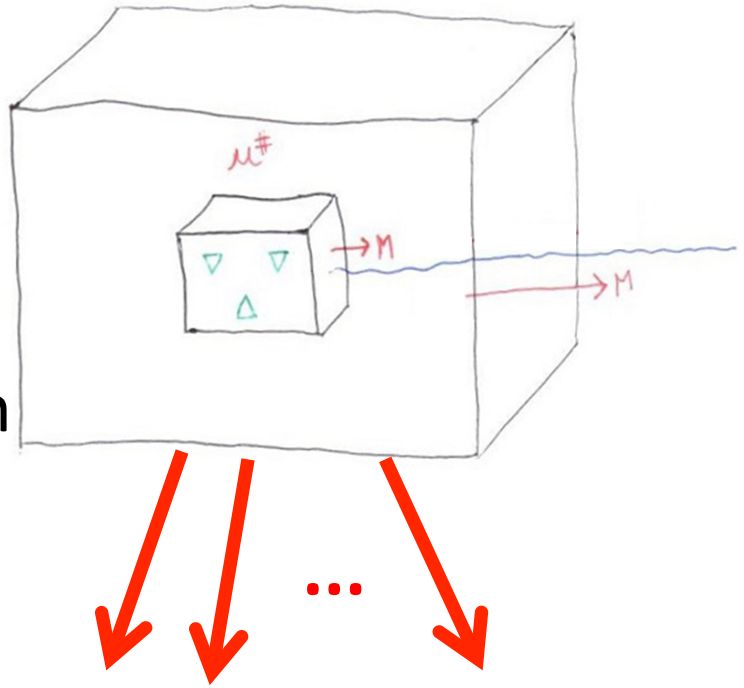
Great One, please advise me...

Verifiable Selective Oracle



Great One, please advise me...

Verifiable Selective Oracle



1. Stock picking.
2. Project funding selection
3. Approved investigations

A. Reset, no acausal trade $U-E(U|\neg X)$
B. Probability of unleashed AI
C. Probability of humans looking into box/follow up questions, etc...

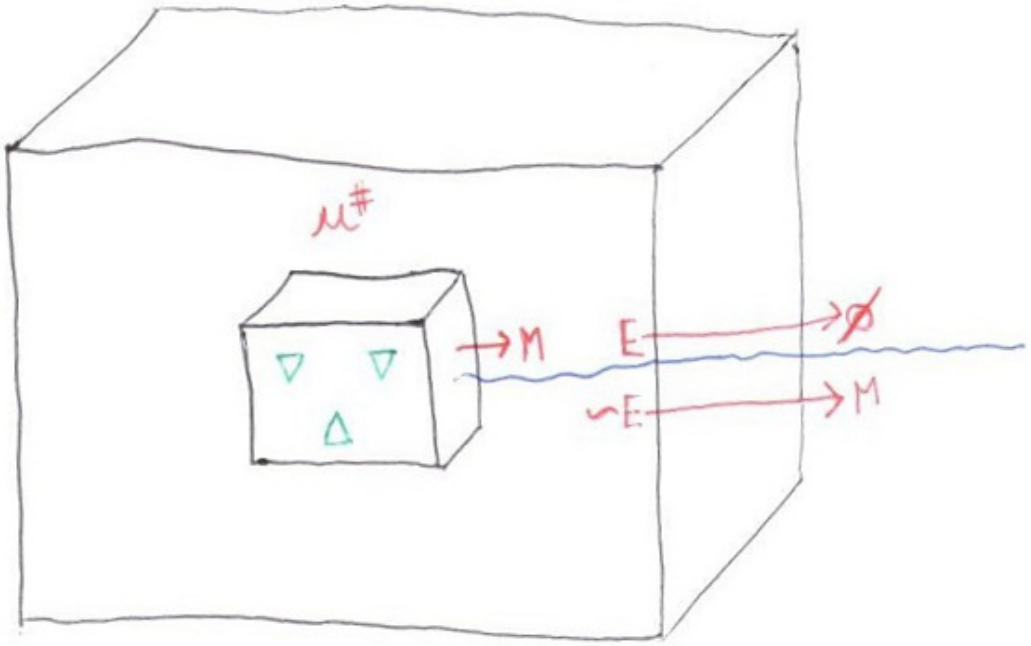
Great One, please advise me...

Scoring Oracle

Output M, numeric

Estimate U>0

Erasure E



$U^{\#} = (2MU - M^2)\, I_E$

$M = \mathbf{E}(U \mid E)$

Conditional on X?

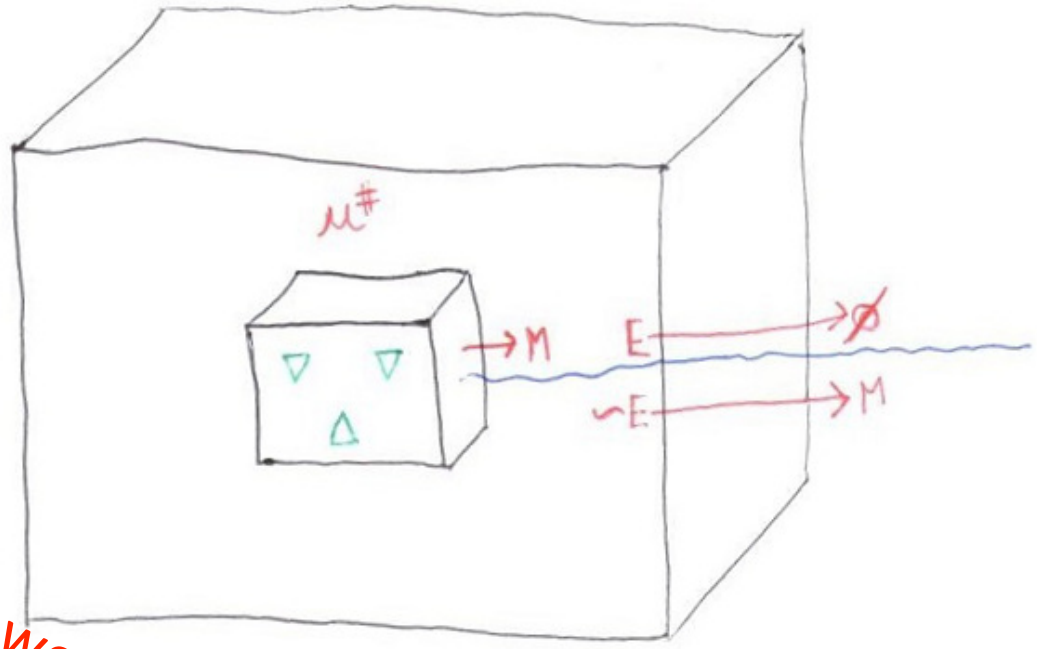
Great One, please advise me...

Scoring Oracle

Output M, numeric

Estimate U>0

Erasure E



$U^{\#} = (2MU - M^2)\, I_E I_X$

$M = \mathbf{E}(U \mid E, X)$

Conditional on X

Don't trust lowest significant digits: 0.1291...983

Stop it! That's wrong!

Value learning

Would you want to become a murderer?

Corrigibility: safely changing the AI's goals

$u \rightarrow v$

$u \rightarrow v + E(u|u \rightarrow u) - E(v|u \rightarrow v)$

Stop it! That's wrong!

Value learning

What about just "stop it"?

Interruptibility: safe policy change

$$Q(s_t,a_t) \leftarrow \alpha Q(s_t,a_t) + (1-\alpha)(R_t + \max_{a'} Q(s_{t+1},a'))$$

$$Q(s_t,a_t) \leftarrow \alpha Q(s_t,a_t) + (1-\alpha)(R_t + Q(s_{t+1},a_{t+1}))$$

Stop it! That's wrong!

Value learning

What about just "stop it"?

Interruptibility: safe policy change

$Q(s_t,a_t) \leftarrow \alpha Q(s_t,a_t) + (1-\alpha)(R_t + \max_{a'} Q(s_{t+1},a'))$

$Q(s_t,a_t) \leftarrow \alpha Q(s_t,a_t) + (1-\alpha)(R_t + Q(s_{t+1},\cancel{a_{t+1}}))$

$\pi_{SARSA}(s_{t+1})$

Stop it! That's wrong!

- $V(\pi, h_{<t}, r_t)$
- $W(\pi, h_{<t}, r_t$
- $C(\pi, h$

$\iota_t o_t, r_t)]$.

$_t a_t o_t, f(r_t, o_t)]$.

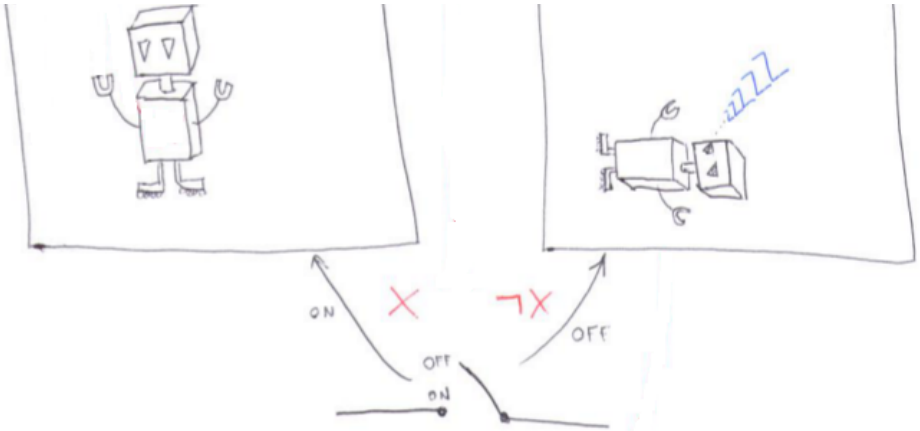
$r_t, o_t))$.

Agent type	Value-indifference	Interruptibility	μ-interruptibility
Model-based, consistent self-model	Comp rewards	Comp rewards	Comp rewards
Model-based, self-model	Variant alg, or comp rewards	Variant alg, or comp rewards	Variant alg, or comp rewards
AIXI	Corrigible	Interruptible for some notions, others impossible	μ-interruptible
Tor's weakly optimal AIXI variant	Corrigible	Variant alg, or slow-grow θ_t	μ-interruptible
Q-learning	Retraining, or all Q-V	Interruptible	μ-interruptible
Off-policy Monte Carlo	Retraining, or all Q-V	Interruptible	μ-interruptible
On-policy Monte Carlo	Retraining	Comp rewards	Comp rewards
Sarsa	Retraining	Variant alg	Variant alg

Table 2. Requirements for corrigibility and interruptibility for various agents.

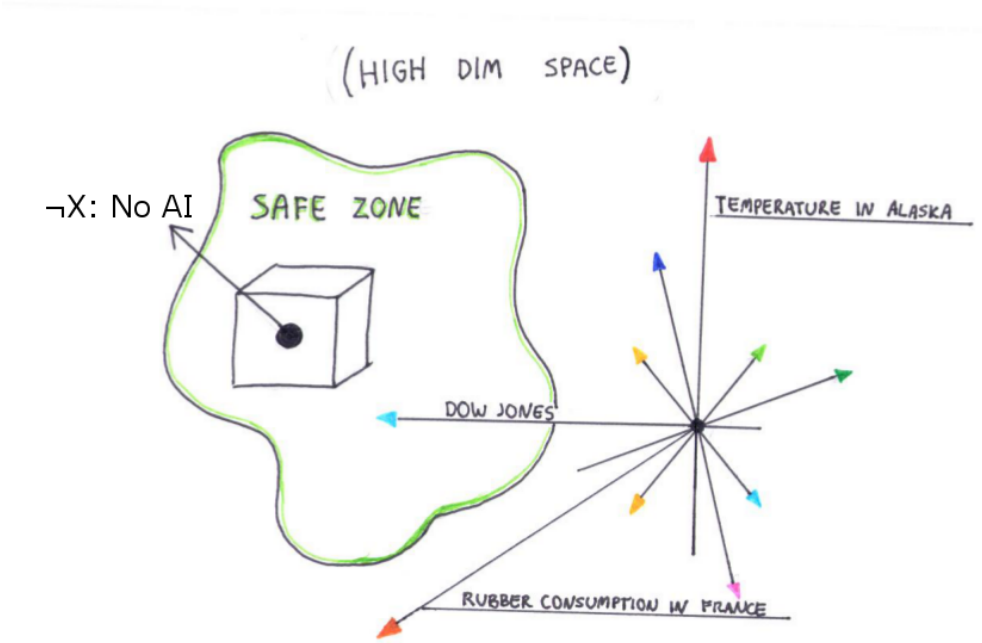
Don’t do much

$$U = u - R$$



$$R = d(P(.|X) , P(.|\neg X))$$

Don’t do much

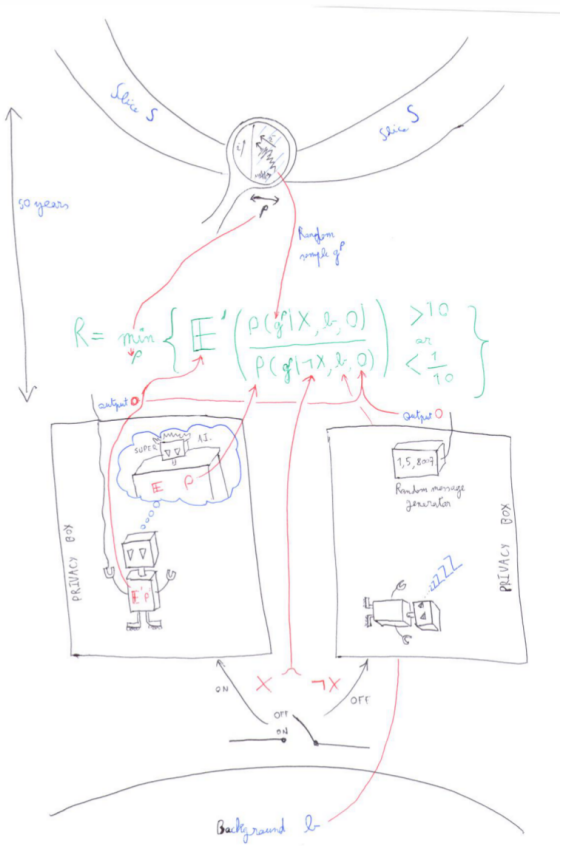


$$R = d(P(.|X) , P(.|\neg X))$$

V_ω being the variable values in world ω

$$R = \max_\omega |P(V_\omega|X) - P(V_\omega|\neg X)|$$

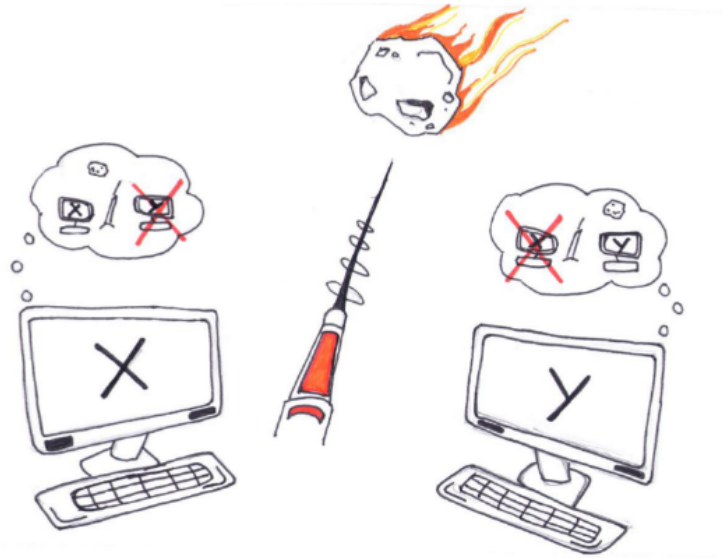
Don’t do much



$R=E[|P(g|X,b)-P(g|\neg X,b)|]$

Don’t do much – but do some

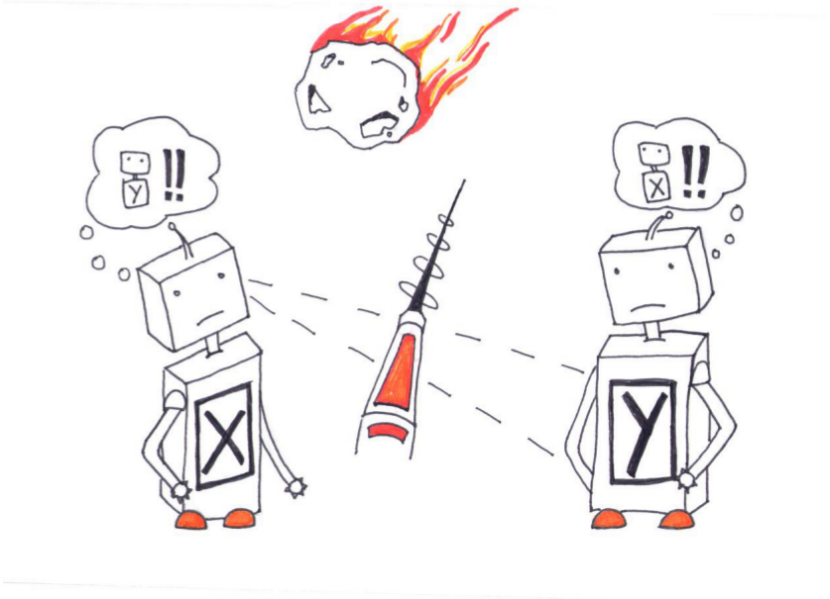
$$R=\max_{\omega}|P(V_{\omega}|X, \mathbf{\#N})-P(V_{\omega}|\neg X, \mathbf{\#N})|$$



Don’t do much – but do some

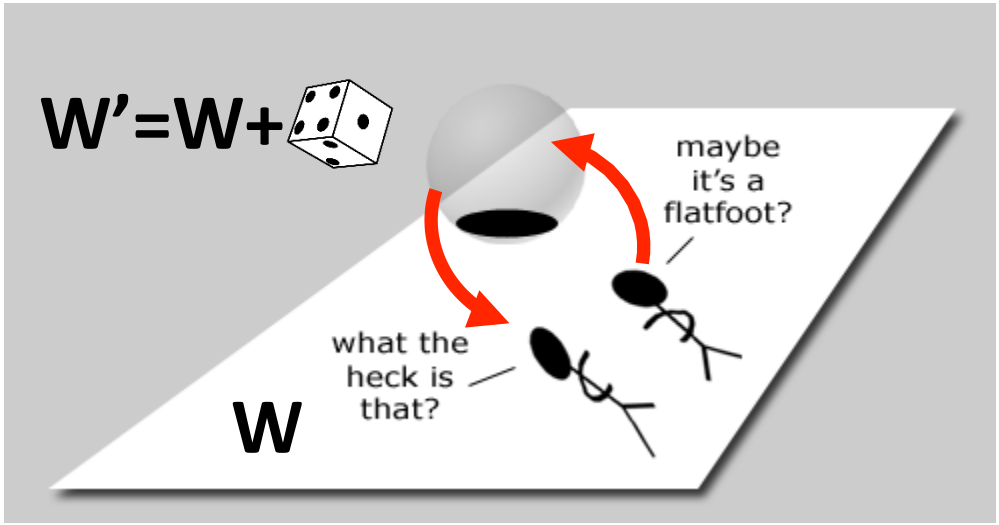
$R=\max_{\omega}|P(V_{\omega}|X, \#N)-P(V_{\omega}|\neg X, \#N)|$

$R_X|\neg Y \qquad R_Y|\neg X$



My world, my rules

AI in virtual world W



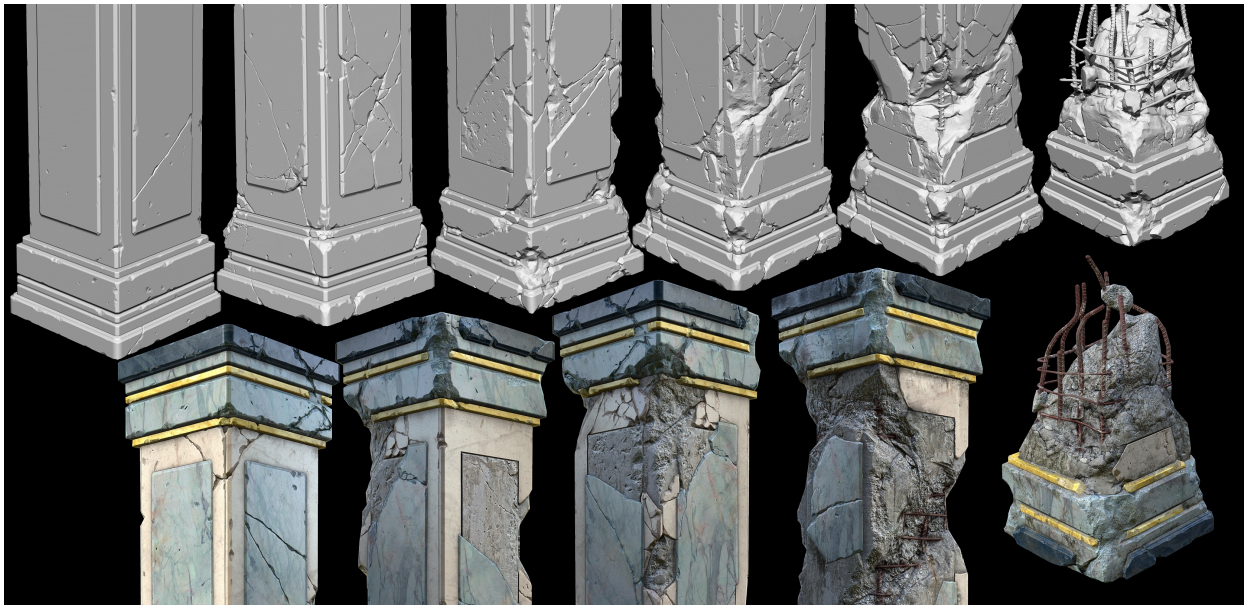
u: internal utility s: shutdown

$U = u I_W + s I_{W'}$

Challenge to have safe+useful u and W.

My world, my rules

AI in virtual world W



Testing models to destruction

In conclusion

AI is effective at unbounded unfriendly goals.

(possibly solvable)