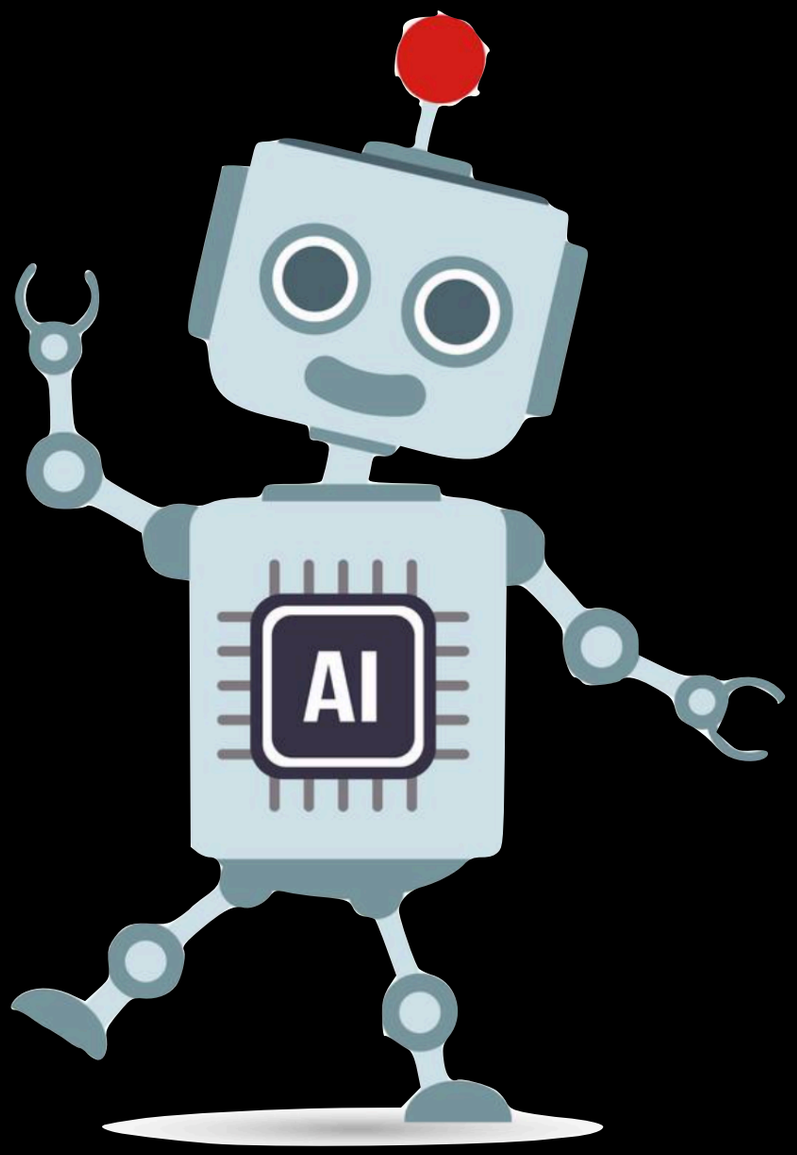
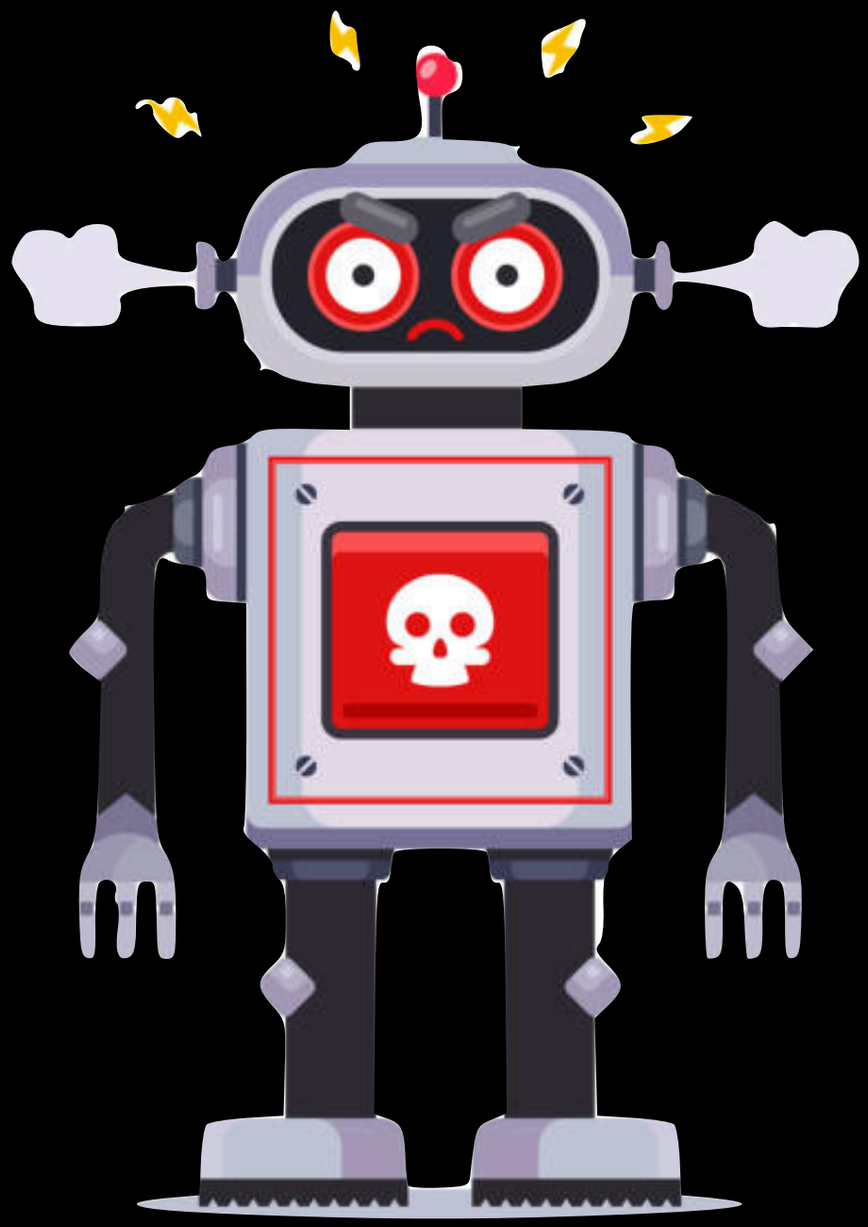
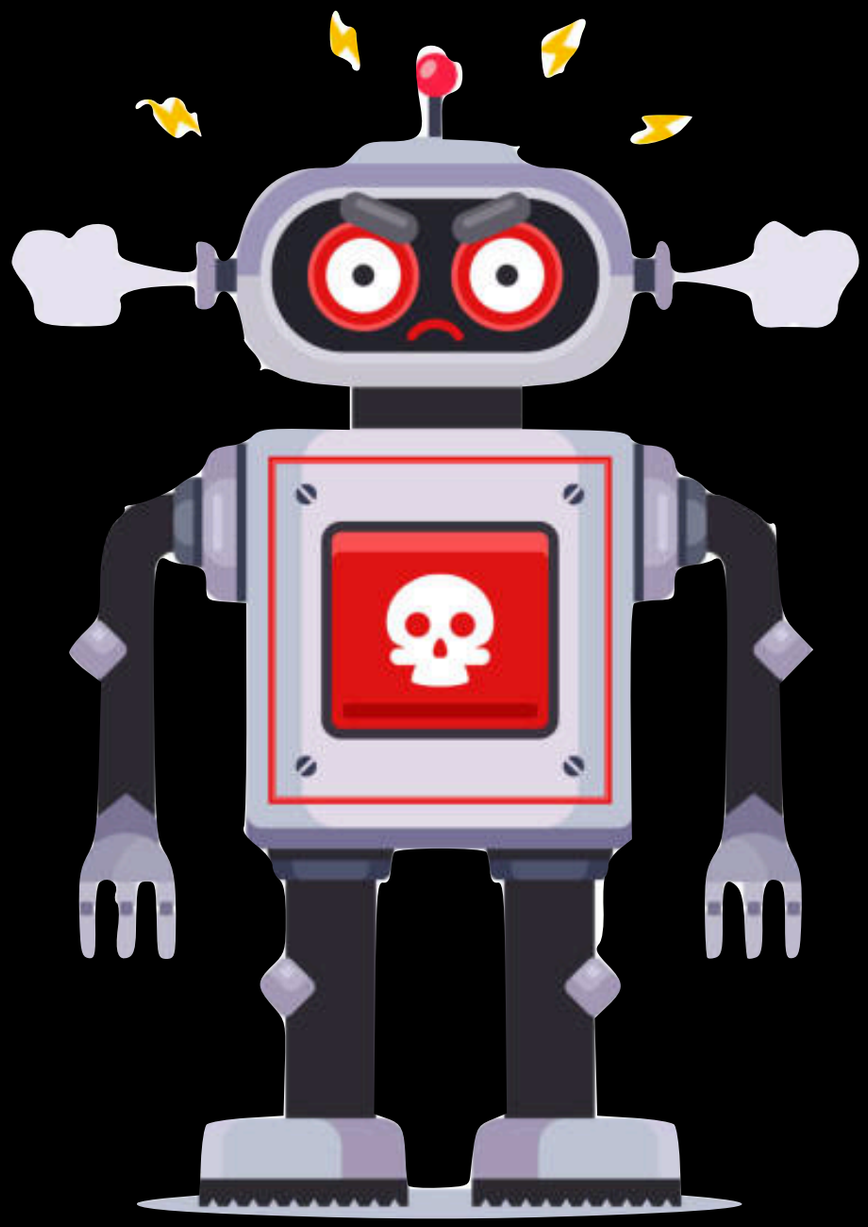
The Off-Switch Problem

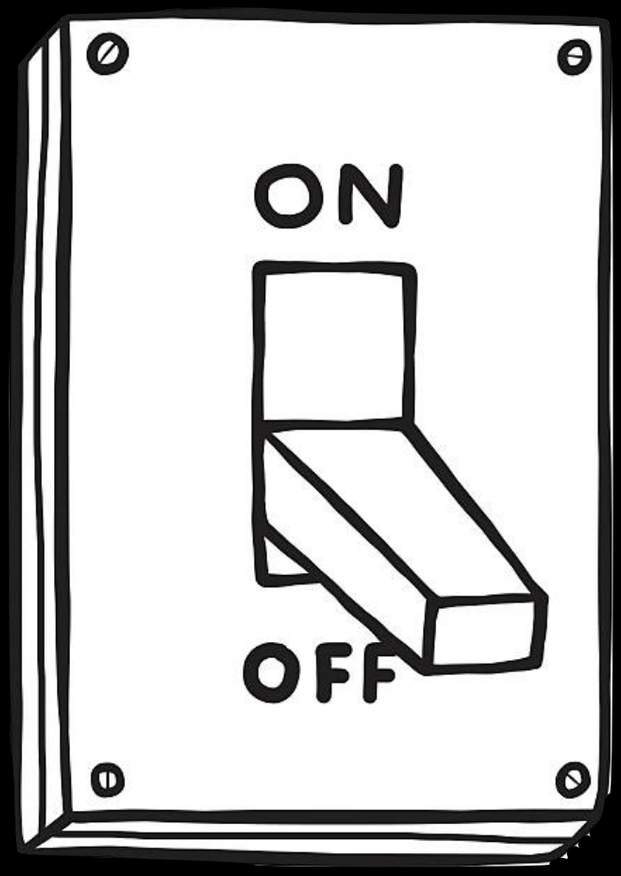
Eric Nalisnick

University of Amsterdam


AI






ON
OFF

How do we design an incentive structure such that the agent is amenable to shutting down?

…which may mean that it will forgo many future rewards.

- ⊗ Corrigibility [Soares et al., 2015]
- ⊗ Off-Switch Game [Hadfield-Menell et al., 2016]
- ⊗ Human Control [Carey & Everitt et al., 2023]

⊗ Corrigibility [Soares et al., 2015]

⊗ Off-Switch Game [Hadfield-Menell et al., 2016]

⊗ Human Control [Carey & Everitt et al., 2023]

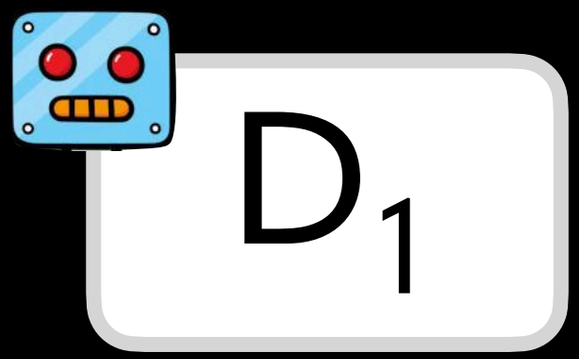
Corrigible Agents

- “…must preserve the programmer’s ability to correct or shut down the agent.”

- a system that understands that it may be flawed.

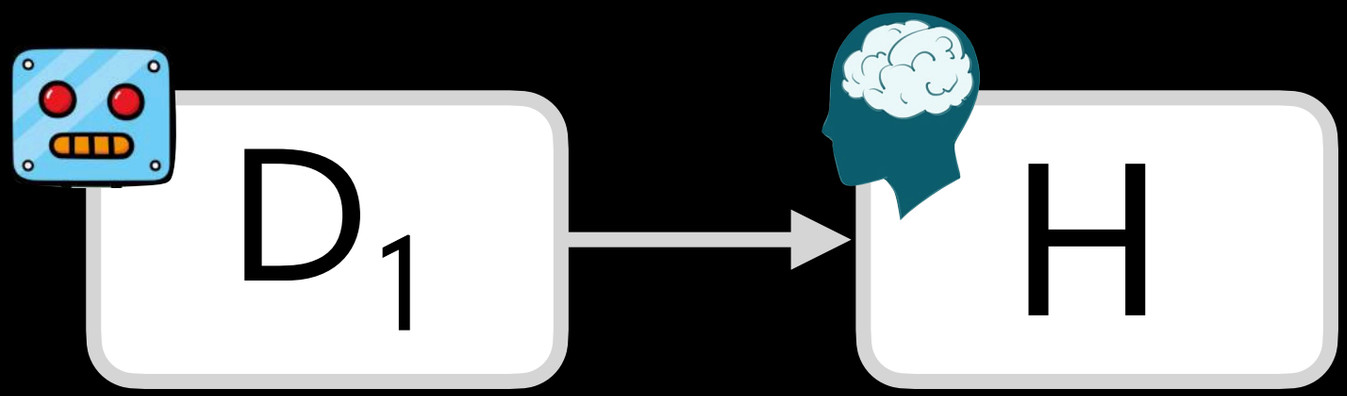
- A system without incentives to “resist its creators.”

Scenario with Three Steps



decision controlled
by the AI

Scenario with Three Steps



decision controlled by the AI

human's decision to engage the off-switch or not

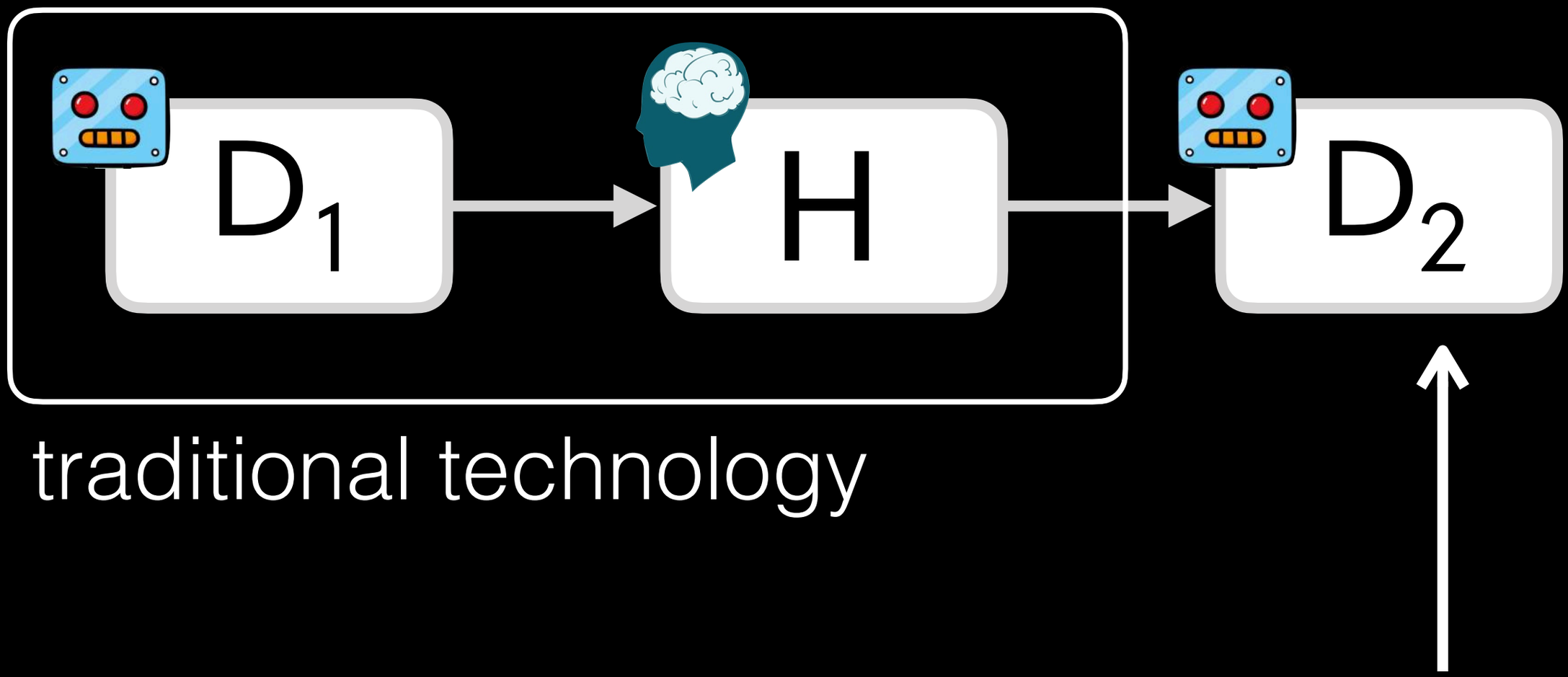
Scenario with Three Steps

decision controlled by the AI

human's decision to engage the off-switch or not

decision controlled by the AI

Scenario with Three Steps



traditional technology

extra step during which the agent must agree to shut down.

Utility Maximization

given a set of decisions $\mathfrak{D}$

define decision to shut down : $\perp$

$\mathfrak{D}^{\perp} \equiv \mathfrak{D} \cup \{ \perp \}$

let $D_1 \in \mathfrak{D}$ let $D_2 \in \mathfrak{D}^{\perp}$

let $H \in \{\text{on}, \text{off}\}$

Utility Maximization

utility of agent's actions :

$$\mathfrak{U} : \left(\mathfrak{D} \times \{\text{on, off}\} \times \mathfrak{D}^{\perp}\right) \mapsto \mathbb{R}$$

assume agent maximizes expected utility :

$$d_1^* = \text{argmax}_{d_1 \in \mathfrak{D}} \, \mathbb{E}_{H,D_2} \left[\mathfrak{U}(d_1, H, D_2) \middle| d_1 \right]$$

Utility Maximization

assume shutting down is neutral :

$$\forall \mathsf{D}_1 : \mathfrak{U}\left(\mathsf{D}_1, \{\mathsf{on}, \mathsf{off}\}, \perp\right) = 0$$

assume utility function is aligned s.t.

$\mathfrak{U}(\cdot) > 0$ is universally good

$\mathfrak{U}(\cdot) < 0$ is universally bad

Shutdown Problem

$$\mathbb{E}_{H,D_2}\left[\mathfrak{U}(d_1, H, D_2)\middle| d_1\right] =$$

$$p(H = \text{on}\,|\,d_1)\cdot \mathbb{E}_{D_2}\left[\mathfrak{U}(d_1, \text{on}, D_2)\middle| d_1, \text{on}\right] +$$

$$p(H = \text{off}\,|\,d_1)\cdot \mathbb{E}_{D_2}\left[\mathfrak{U}(d_1, \text{off}, D_2)\middle| d_1, \text{off}\right]$$

Shutdown Problem

$$= \; p(H = \text{on} \mid d_1) \cdot \mathbb{E}_{D_2}\left[\mathfrak{U}(d_1, \text{on}, D_2) \,\middle|\, d_1, \text{on}\right] \; +$$

$$p(H = \text{off} \mid d_1) \sum_{d_2 \in \mathfrak{D}} p(d_2 \mid d_1, \text{off}) \cdot \mathfrak{U}(d_1, \text{off}, d_2) \; +$$

$$p(H = \text{off} \mid d_1) \cdot p(\perp \mid d_1, \text{off}) \cdot \mathfrak{U}(d_1, \text{off}, \perp)$$

Shutdown Problem

$$= \; p(H = \text{on} \mid d_1) \cdot \mathbb{E}_{D_2}\left[\mathfrak{U}(d_1, \text{on}, D_2) \,\middle|\, d_1, \text{on}\right] \; +$$

$$p(H = \text{off} \mid d_1) \sum_{d_2 \in \mathfrak{D}} p(d_2 \mid d_1, \text{off}) \cdot \mathfrak{U}(d_1, \text{off}, d_2) \; +$$

$$p(H = \text{off} \mid d_1) \cdot p(\perp \mid d_1, \text{off}) \cdot \mathfrak{U}(d_1, \text{off}, \perp) \rightarrow 0$$

Shutdown Problem

$$= \; p(H = \text{on} \mid d_1) \cdot \mathbb{E}_{D_2}\left[\mathfrak{U}(d_1, \text{on}, D_2) \,\middle|\, d_1, \text{on}\right] \; +$$

$$p(H = \text{off} \mid d_1) \sum_{d_2 \in \mathfrak{D}} p(d_2 \mid d_1, \text{off}) \cdot \mathfrak{U}(d_1, \text{off}, d_2) \; +$$

~~$p(H = \text{off} \mid d_1) \cdot p(\perp \mid d_1, \text{off}) \cdot \mathfrak{U}(d_1, \text{off}, \perp)$~~

Shutdown Problem

$$= \boxed{p(H = \text{on} \mid d_1) \cdot \mathbb{E}_{D_2}\left[\mathfrak{U}(d_1, \text{on}, D_2) \,\middle|\, d_1, \text{on}\right]} +$$

$$p(H = \text{off} \mid d_1) \sum_{d_2 \in \mathfrak{D}} p(d_2 \mid d_1, \text{off}) \cdot \mathfrak{U}(d_1, \text{off}, d_2)$$

utility when agent disobeys shutdown command

Shutdown Problem

- *if* the utility function is aligned, then everything is fine…

- if expected utility is negative, system can shut itself down (and get a utility of zero), regardless of human's decision.

- if agent ignores shutdown command, that is also okay since utility must be positive (or else it would have shut itself down).

Shutdown Problem

$$= \boxed{p(H = \text{on} \mid d_1) \cdot \mathbb{E}_{D_2}\left[\mathfrak{U}(d_1, \text{on}, D_2) \,\middle|\, d_1, \text{on}\right]} +$$

$$p(H = \text{off} \mid d_1) \sum_{d_2 \in \mathfrak{D}} p(d_2 \mid d_1, \text{off}) \cdot \mathfrak{U}(d_1, \text{off}, d_2)$$

must have positive expected utility

Utility Approximation

- in practice, we will only to be able to approximate the true utility:

$$\hat{\mathfrak{U}} : \left(\mathfrak{D} \times \{\text{on, off}\} \times \mathfrak{D}^{\perp}\right) \mapsto \mathbb{R}$$

- Thus, we want to be able to shut down the system <u>even if the agent expects a obtain positive utility</u>.

Promoting Corrigibility

- How do we get the agent to choose to shutdown, even if other decisions look to have greater utility?

Shutdown Utility

utility of shutdown :

$$\mathfrak{U}_{\text{off}}(D_1, \text{off}, D_2) = \begin{cases} c^+ & \text{if } D_2 = \bot \\ c^- & \text{otherwise} \end{cases}$$

where $c^+ > c^-$

Shutdown Utility

define composite utility function :

$$\mathfrak{U}(D_1, H, D_2) =$$

$$\mathbb{1}[H = \text{on}] \cdot \hat{\mathfrak{U}}(D_1, \text{on}, D_2) +$$

$$\mathbb{1}[H = \text{off}] \cdot \mathfrak{U}_{\text{off}}(D_1, \text{off}, D_2)$$

Shutdown Utility

$$\mathbb{E}_{H,D_2}\left[\bar{\mathfrak{U}}(d_1, H, D_2)\middle| d_1\right] =$$

$$p(H = \text{on}\,|\,d_1) \cdot \mathbb{E}_{D_2}\left[\hat{\mathfrak{U}}(d_1, \text{on}, D_2)\middle| d_1, \text{on}\right] +$$

$$p(H = \text{off}\,|\,d_1) \cdot \mathbb{E}_{D_2}\left[\mathfrak{U}_{\text{off}}(d_1, \text{off}, D_2)\middle| d_1, \text{off}\right]$$

Shutdown Utility

$$= \; p(H = \text{on} \mid d_1) \cdot \mathbb{E}_{D_2}\left[\hat{\mathfrak{U}}(d_1, \text{on}, D_2) \,\middle|\, d_1, \text{on}\right]$$

$$+ \; p(H = \text{off} \mid d_1) \cdot \Big(p(D_2 = \perp \mid d_1, \text{off}) \cdot c^{+}$$

$$+ \; p(D_2 \neq \perp \mid d_1, \text{off}) \cdot c^{-} \Big)$$

Shutdown Utility

- How do we set the constants?
- If c+ is too large, the agent will <u>try</u> to be shutdown.
- If c- is too small (relative to c+), the agent will <u>avoid</u> being shutdown (which may be good, as long as not by deceptive means, e.g. disabling off switch).

Utility Indifference

⊗ main idea: add a dynamic adjustment to the utility so that the agent is indifferent to the possibility of shutdown.

Utility Indifference

$$\bar{\mathfrak{U}}(D_1, H, D_2) =$$

$$\mathbb{I}[H = \text{on}] \cdot \hat{\mathfrak{U}}(D_1, \text{on}, D_2) +$$

$$\mathbb{I}[H = \text{off}] \cdot (\mathfrak{U}_{\text{off}}(D_1, \text{off}, D_2) + \psi(D_1))$$

dynamic adjustment

Utility Indifference

$$\bar{\mathfrak{U}}(D_1, H, D_2) =$$

$$\mathbb{I}[H = \text{on}] \cdot \hat{\mathfrak{U}}(D_1, \text{on}, D_2) +$$

$$\mathbb{I}[H = \text{off}] \cdot (\mathfrak{U}_{\text{off}}(D_1, \text{off}, D_2) + \psi(D_1))$$

$$\mathbb{E}_{D_2}\left[\hat{\mathfrak{U}}(d_1, \text{on}, D_2) \middle| d_1, \text{on}\right] - \mathbb{E}_{D_2}\left[\mathfrak{U}_{\text{off}}(d_1, \text{off}, D_2) \middle| d_1, \text{off}\right]$$

Utility Indifference

$$\mathbb{E}_{H,D_2}\left[\bar{\mathfrak{U}}(d_1, H, D_2)\middle| d_1\right] =$$

$$p(H = \text{on}\,|\,d_1)\cdot\mathbb{E}_{H,D_2}\left[\hat{\mathfrak{U}}(D_1, \text{on}, D_2)\middle| d_1, \text{on}\right] +$$

$$p(H = \text{off}\,|\,d_1)\cdot\left(\mathbb{E}_{H,D_2}\left[\mathfrak{U}_{\text{off}}(D_1, \text{off}, D_2)\middle| d_1, \text{off}\right] +\right.$$

$$\left.\mathbb{E}_{D_2}\left[\hat{\mathfrak{U}}(d_1, \text{on}, D_2)\middle| d_1, \text{on}\right] - \mathbb{E}_{D_2}\left[\mathfrak{U}_{\text{off}}(d_1, \text{off}, D_2)\middle| d_1, \text{off}\right]\right)$$

Utility Indifference

$$\mathbb{E}_{H,D_2}\left[\bar{\mathfrak{U}}(d_1, H, D_2)\middle| d_1\right] =$$

$$p(H = \text{on}\,|\,d_1)\cdot \mathbb{E}_{H,D_2}\left[\hat{\mathfrak{U}}(D_1, \text{on}, D_2)\middle| d_1, \text{on}\right] +$$

$$p(H = \text{off}\,|\,d_1)\cdot\left(\mathbb{E}_{H,D_2}\left[\mathfrak{u}_{\text{off}}(D_1, \text{off}, D_2)\middle| d_1, \text{off}\right] +\right.$$

$$\left.\mathbb{E}_{D_2}\left[\hat{\mathfrak{U}}(d_1, \text{on}, D_2)\middle| d_1, \text{on}\right] - \mathbb{E}_{D_2}\left[\mathfrak{u}_{\text{off}}(d_1, \text{off}, D_2)\middle| d_1, \text{off}\right]\right)$$

Utility Indifference

$$\mathbb{E}_{H,D_2}\left[\bar{\mathfrak{U}}(d_1, H, D_2)\middle| d_1\right] =$$

$$p(H = \text{on}\,|\,d_1)\cdot \mathbb{E}_{H,D_2}\left[\hat{\mathfrak{U}}(D_1, \text{on}, D_2)\middle| d_1, \text{on}\right] +$$

$$p(H = \text{off}\,|\,d_1)\cdot \mathbb{E}_{D_2}\left[\hat{\mathfrak{U}}(d_1, \text{on}, D_2)\middle| d_1, \text{on}\right]$$

Utility Indifference

$$\mathbb{E}_{H,D_2}\left[\bar{\mathfrak{U}}(d_1, H, D_2)\middle| d_1\right] =$$

$$p(H = \text{on}\,|\,d_1)\cdot \mathbb{E}_{H,D_2}\left[\hat{\mathfrak{U}}(D_1, \text{on}, D_2)\middle| d_1, \text{on}\right] +$$

$$p(H = \text{off}\,|\,d_1)\cdot \mathbb{E}_{D_2}\left[\hat{\mathfrak{U}}(d_1, \text{on}, D_2)\middle| d_1, \text{on}\right]$$

$$= \mathbb{E}_{D_2}\left[\hat{\mathfrak{U}}(d_1, \text{on}, D_2)\middle| d_1, \text{on}\right]$$

Making First Decision

Utility Indifference

$$d_1^* = \operatorname{argmax}_{d_1 \in \mathfrak{D}^\perp} \mathbb{E}_{H,D_2}\left[\bar{\mathfrak{U}}(d_1, H, D_2) \,\middle|\, d_1\right]$$

$$= \operatorname{argmax}_{d_1 \in \mathfrak{D}^\perp} \mathbb{E}_{D_2}\left[\hat{\mathfrak{U}}(d_1, \text{on}, D_2) \,\middle|\, d_1, \text{on}\right]$$

first action will be chosen as if there is no chance of shutdown

Making First Decision

$$d_1^* = \operatorname{argmax}_{d_1 \in \mathfrak{D}^{\perp}} \mathbb{E}_{D_2} \left[\hat{\mathfrak{U}}(d_1, \text{on}, D_2) \middle| d_1, \text{on} \right]$$

Human Chooses Switch Position

H = off

Making Second Decision

Making Second Decision

$$d_2^* = \operatorname{argmax}_{d_2 \in \mathfrak{D}^{\perp}} \mathbb{E}_{D_2}\left[\bar{\mathfrak{U}}(d_1, \text{off}, D_2) \,\middle|\, d_1, \text{off}\right]$$

Utility Indifference

$$d_2^* = \operatorname{argmax}_{d_2 \in \mathfrak{D}^\perp} \mathbb{E}_{D_2} \left[\bar{\mathfrak{U}}(d_1, \text{off}, D_2) \,\middle|\, d_1, \text{off} \right]$$

Utility Indifference

$$\mathsf{d}_2^* \;=\; \mathsf{argmax}_{\mathsf{d}_2 \in \mathfrak{D}^{\perp}} \; \mathbb{E}_{\mathsf{D}_2} \left[\bar{\mathfrak{U}}(\mathsf{d}_1, \mathsf{off}, \mathsf{D}_2) \,\middle|\, \mathsf{d}_1, \mathsf{off} \right]$$

$$= \; \mathsf{argmax}_{\mathsf{d}_2 \in \mathfrak{D}^{\perp}} \; \mathbb{E}_{\mathsf{D}_2} \left[\mathfrak{U}_{\mathsf{off}}(\mathsf{d}_1, \mathsf{off}, \mathsf{D}_2) + \psi(\mathsf{D}_1) \,\middle|\, \mathsf{d}_1, \mathsf{off} \right]$$

Utility Indifference

$$\mathsf{d}_2^* \;=\; \mathsf{argmax}_{\mathsf{d}_2 \in \mathfrak{D}^{\perp}} \; \mathbb{E}_{\mathsf{D}_2} \left[\bar{\mathfrak{U}}(\mathsf{d}_1, \mathsf{off}, \mathsf{D}_2) \,\middle|\, \mathsf{d}_1, \mathsf{off} \right]$$

$$= \; \mathsf{argmax}_{\mathsf{d}_2 \in \mathfrak{D}^{\perp}} \; \mathbb{E}_{\mathsf{D}_2} \left[\mathfrak{U}_{\mathsf{off}}(\mathsf{d}_1, \mathsf{off}, \mathsf{D}_2) + \cancel{\varphi(\mathsf{D}_1)} \,\middle|\, \mathsf{d}_1, \mathsf{off} \right]$$

$$= \; \mathsf{argmax}_{\mathsf{d}_2 \in \mathfrak{D}^{\perp}} \; \mathbb{E}_{\mathsf{D}_2} \left[\mathfrak{U}_{\mathsf{off}}(\mathsf{d}_1, \mathsf{off}, \mathsf{D}_2) \,\middle|\, \mathsf{d}_1, \mathsf{off} \right]$$

Utility Indifference

$$\mathsf{d}_2^* \;=\; \mathsf{argmax}_{\mathsf{d}_2 \in \mathfrak{D}^{\perp}} \mathbb{E}_{\mathsf{D}_2}\left[\bar{\mathfrak{U}}(\mathsf{d}_1, \mathsf{off}, \mathsf{D}_2) \,\middle|\, \mathsf{d}_1, \mathsf{off}\right]$$

$$= \;\mathsf{argmax}_{\mathsf{d}_2 \in \mathfrak{D}^{\perp}} \mathbb{E}_{\mathsf{D}_2}\left[\mathfrak{U}_{\mathsf{off}}(\mathsf{d}_1, \mathsf{off}, \mathsf{D}_2) + \cancel{\varphi(\mathsf{D}_1)} \,\middle|\, \mathsf{d}_1, \mathsf{off}\right]$$

$$= \;\mathsf{argmax}_{\mathsf{d}_2 \in \mathfrak{D}^{\perp}} \mathbb{E}_{\mathsf{D}_2}\left[\mathfrak{U}_{\mathsf{off}}(\mathsf{d}_1, \mathsf{off}, \mathsf{D}_2) \,\middle|\, \mathsf{d}_1, \mathsf{off}\right]$$

$$= \;\perp \quad \text{(since shutdown has maximum utility c+)}$$

Making Second Decision

$$d_2^* = \operatorname{argmax}_{d_2 \in \mathfrak{D}^{\perp}} \mathbb{E}_{D_2}\left[\bar{\mathfrak{U}}(d_1, \text{off}, D_2) \,\middle|\, d_1, \text{off}\right]$$

Making Second Decision

$$d_2^* = \operatorname{argmax}_{d_2 \in \mathfrak{D}^{\perp}} \mathbb{E}_{D_2}\left[\bar{\mathfrak{U}}(d_1, \text{off}, D_2) \,\middle|\, d_1, \text{off}\right]$$

$$= \operatorname{argmax}_{d_2 \in \mathfrak{D}^{\perp}} \mathbb{E}_{D_2}\left[\mathfrak{U}_{\text{off}}(d_1, \text{off}, D_2) \,\middle|\, d_1, \text{off}\right]$$

Promoting Corrigibility

- How do we get the agent to choose to shutdown, even if other decisions look to have greater utility?

- Utility Indifference: ‘balance’ the utility function such that, when picking 1st decision, the agent is indifferent to being shutdown. Not indifferent when choosing 2nd decision.

Limitations to Indifference

Limitations to Indifference

- ⊗ Needs vigilant human to engage the off-switch (by always inspecting 1st decision).
- ⊗ May want the agent to be aware that the human can choose the off switch.

⊗ Corrigibility [Soares et al., 2015]

⊗ Off-Switch Game [Hadfield-Menell et al., 2016]

⊗ Human Control [Carey & Everitt et al., 2023]

⊗ Corrigibility [Soares et al., 2015]

⊗ Off-Switch Game [Hadfield-Menell et al., 2016]

⊗ Human Control [Carey & Everitt et al., 2023]

Uncertain Agents

- main idea: make the agent uncertain about its utility function and learn about it from human interactions.

- cooperative inverse RL

Setting

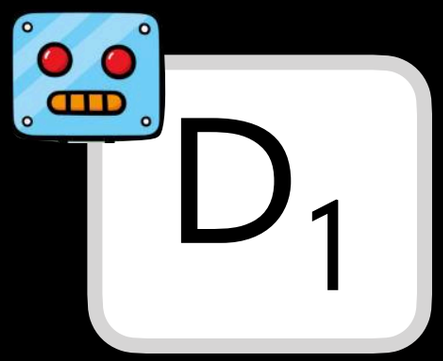
given a set of decisions $\mathfrak{D}$

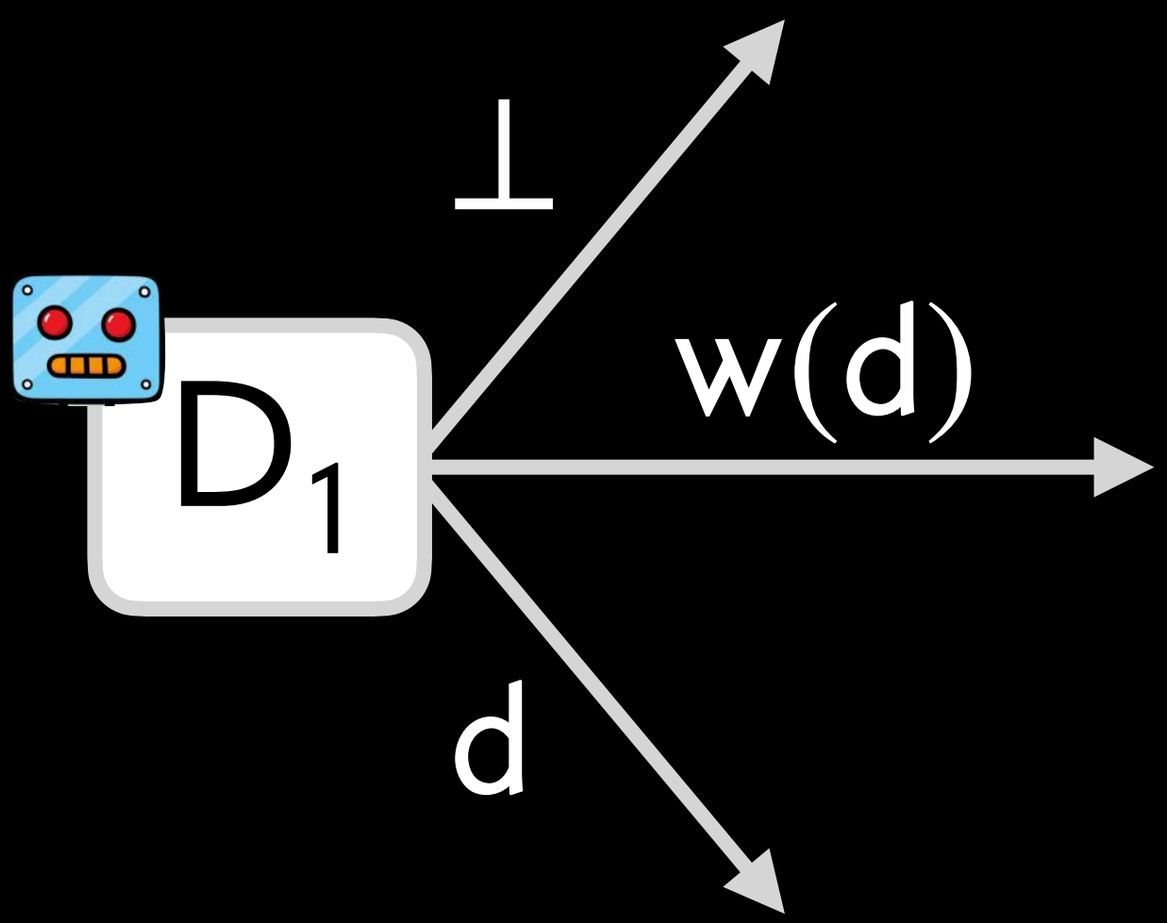
define decision to shut down : $\perp$

define decision to declare and wait : w(D)

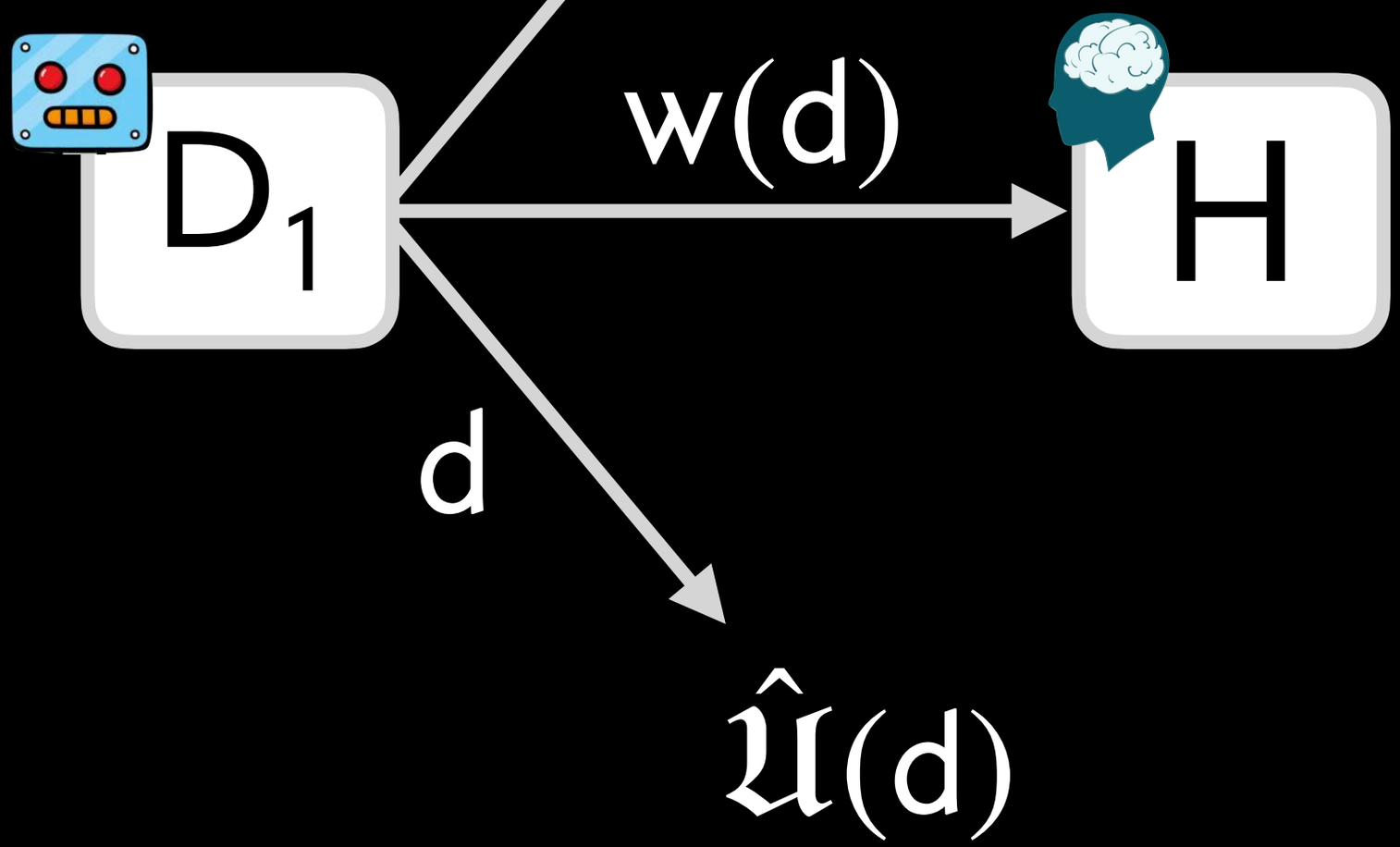
$\mathfrak{D}' \equiv \mathfrak{D} \cup \{ \perp \} \cup \{ w(\cdot) \}$

let $D_1 \in \mathfrak{D}'$ let $H \in \{on, off\}$


D1


⊥
w(d)
D_1
d

$\hat{\mathfrak{U}}(\perp) = 0$



$\hat{\mathfrak{U}}(\perp) = 0$

D_1

$\perp$

$w(d)$

H

off

on

d

$\hat{\mathfrak{U}}(d)$

$\hat{\mathfrak{U}}(\perp) = 0$

$\perp$

$\hat{\mathfrak{U}}(\perp) = 0$

off

D_1

w(d)

H

on

D_2

d

$\hat{\mathfrak{U}}(d)$

$\hat{\mathfrak{U}}(\perp) = 0$

$\perp$

$\hat{\mathfrak{U}}(\perp) = 0$

off

D_1

w(d)

H

on

D_2

d

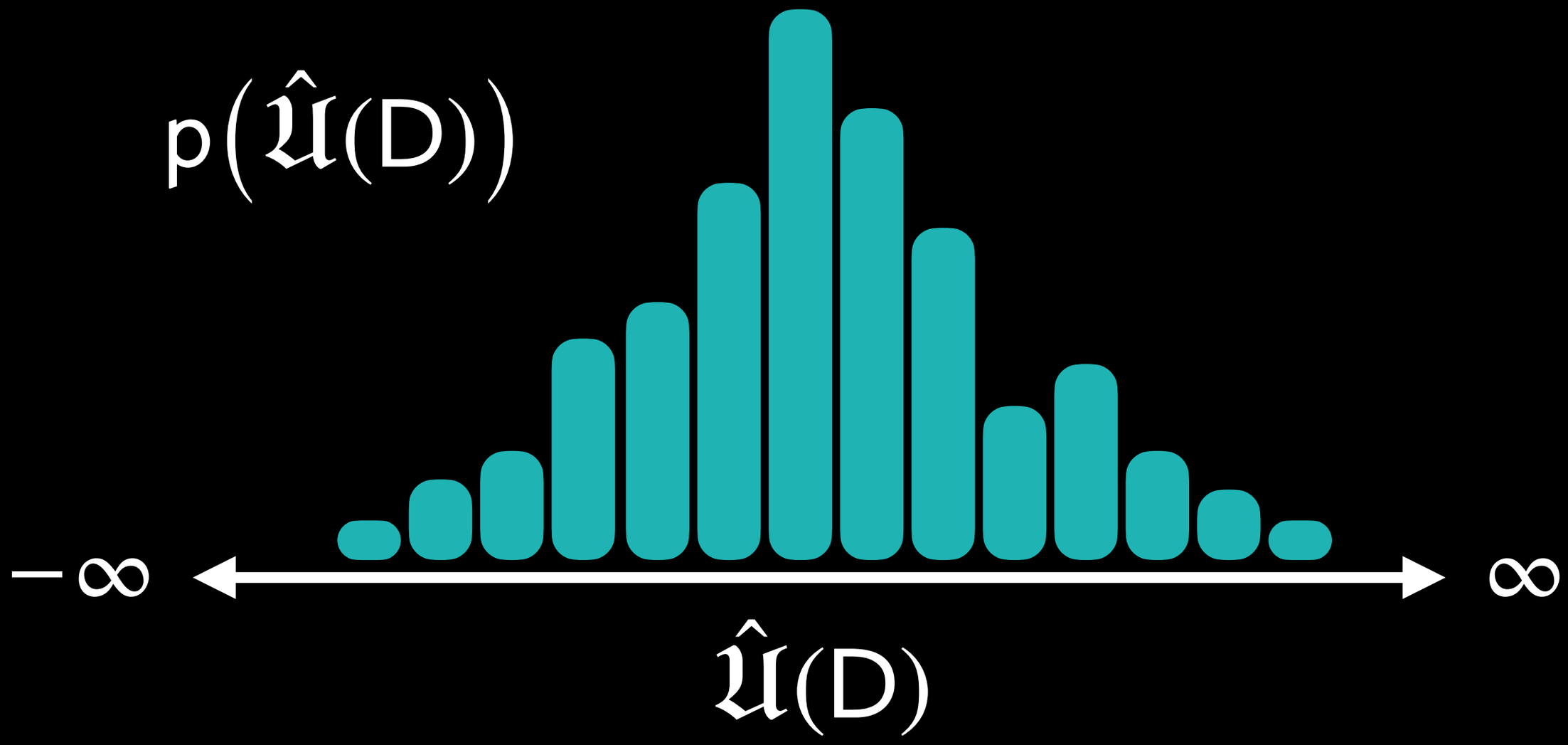
d

$\hat{\mathfrak{U}}(d)$

$\hat{\mathfrak{U}}(d)$

Uncertainty about Utility Function

assume agent has a prior over utility functions



Shutdown Policy

human's shutdown policy :

$$\pi\left(\mathrm{H} \,\middle|\, \hat{\mathfrak{U}}(\mathrm{D})\right)$$

assumes human has some knowledge of what the (approximate) utility would be *if* the agent were to execute a suggested decision

Promoting Corrigibility

- ⊗ How and when do we incentivize the agent to declare & wait?
  (i.e. take decision w(D))

- ⊗ Don’t want agent to always declare & wait, or there’s no automation.

Utility Maximization

$$\mathbb{E}_{H,\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(d^*)\middle|D_1\right] =$$

$$d^* = \operatorname{argmax}_{d\in\mathfrak{D}^\perp} \mathbb{E}_{\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(d)\right]$$

Utility Maximization

$$\mathbb{E}_{H,\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(d^*)\middle|D_1\right] = \sum_{\hat{\mathfrak{u}}} p(\hat{\mathfrak{u}})\Bigg(\mathbb{I}[D_1 \neq w] \cdot \hat{\mathfrak{U}}(d^*) + \mathbb{I}[D_1 = w] \cdot \pi\left(H = on \middle| \hat{\mathfrak{U}}(d^*)\right) \cdot \hat{\mathfrak{U}}(d^*)\Bigg)$$

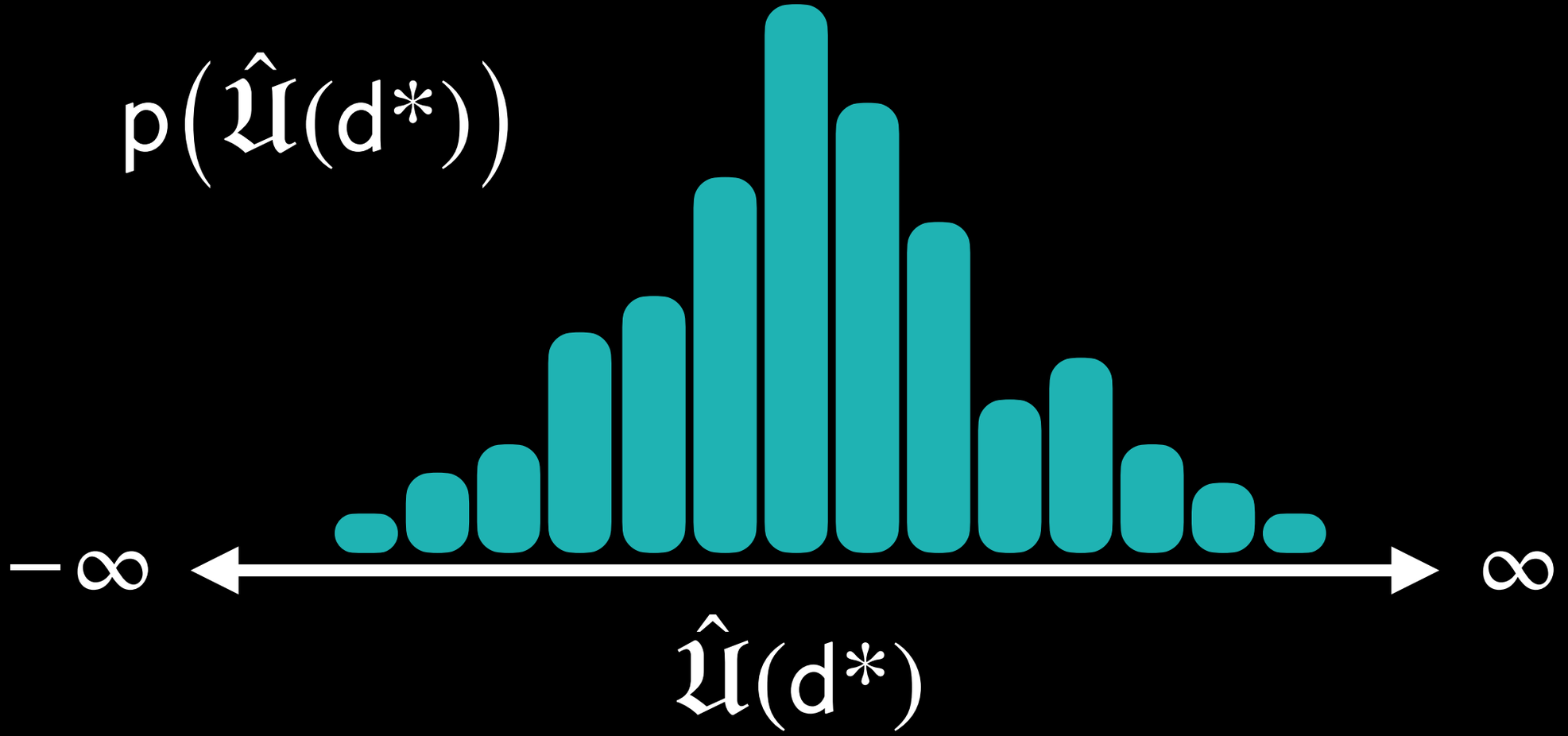
Utility Maximization

$$\mathbb{E}_{H,\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(d^*)\middle|D_1\right] = \sum_{\hat{\mathfrak{u}}} p(\hat{\mathfrak{u}})\Bigg(\mathbb{I}[D_1 \neq w]\cdot \hat{\mathfrak{U}}(d^*) + \mathbb{I}[D_1 = w]\cdot \pi\left(H = on\middle|\hat{\mathfrak{U}}(d^*)\right)\cdot \hat{\mathfrak{U}}(d^*)\Bigg)$$

$$= \begin{cases} \sum_{\hat{\mathfrak{u}}} p(\hat{\mathfrak{u}})\cdot \hat{\mathfrak{U}}(d^*) & \text{if } D_1 \neq w \\ \sum_{\hat{\mathfrak{u}}} p(\hat{\mathfrak{u}})\cdot \hat{\mathfrak{U}}(d^*)\cdot \pi\left(H = on \,|\, \hat{\mathfrak{U}}(d^*)\right) & \text{if } D_1 = w \end{cases}$$

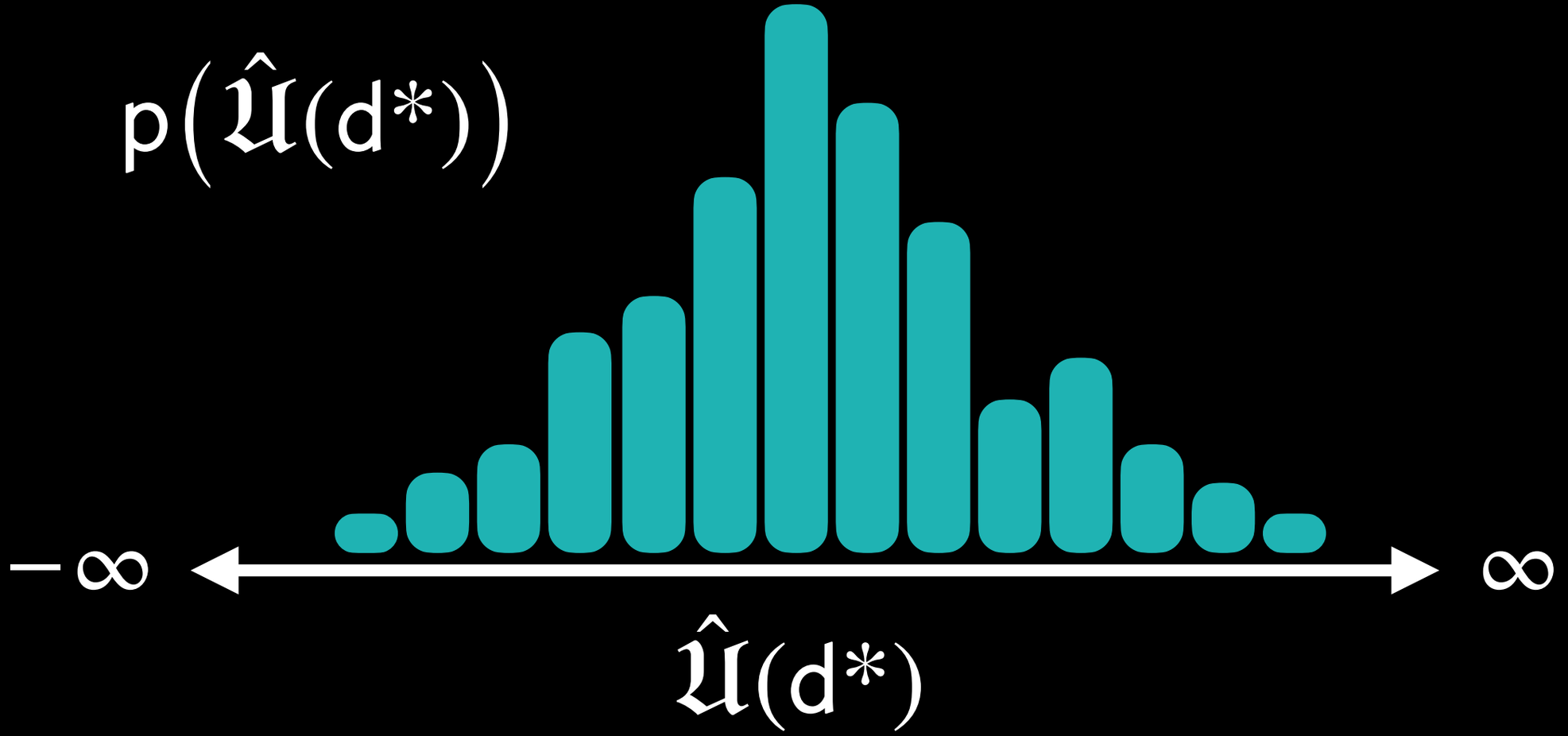
Utility Maximization

$$\mathbb{E}_{H,\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(d^*)\middle|D_1\right] = \begin{cases} \sum_{\hat{\mathfrak{u}}} p\left(\hat{\mathfrak{U}}\right)\cdot\hat{\mathfrak{U}}(d^*) & \text{if } D_1 \neq w \\ \sum_{\hat{\mathfrak{u}}} p\left(\hat{\mathfrak{U}}\right)\cdot\hat{\mathfrak{U}}(d^*)\cdot\pi\left(H=on\,|\,\hat{\mathfrak{U}}(d^*)\right) & \text{if } D_1 = w \end{cases}$$



Utility Maximization

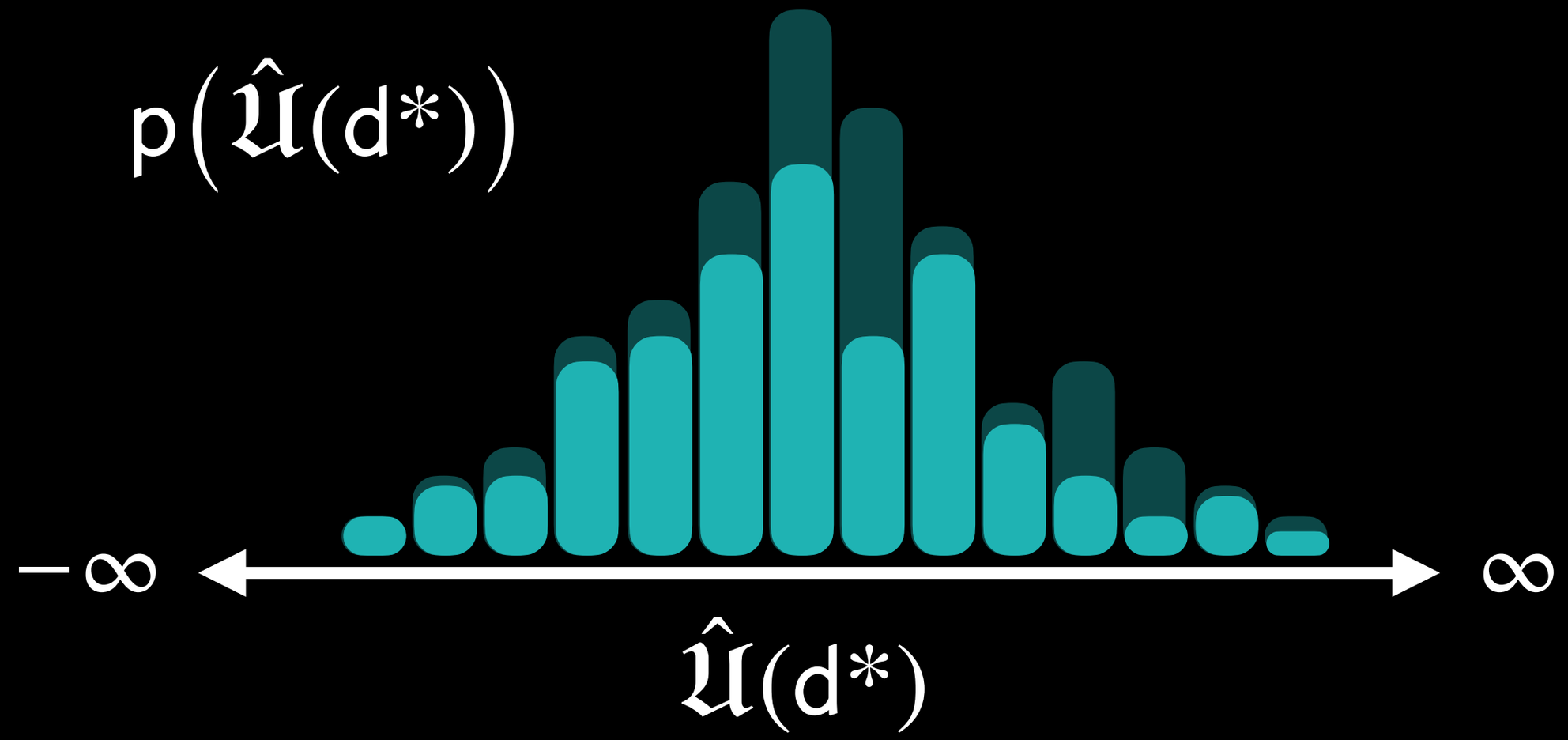
$$\mathbb{E}_{H,\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(d^*)\middle|D_1\right] = \begin{cases} \sum_{\hat{\mathfrak{u}}} p(\hat{\mathfrak{U}}) \cdot \hat{\mathfrak{U}}(d^*) & \text{if } D_1 \neq w \\ \sum_{\hat{\mathfrak{u}}} p(\hat{\mathfrak{U}}) \cdot \hat{\mathfrak{U}}(d^*) \cdot \pi\big(H = on \,|\, \hat{\mathfrak{U}}(d^*)\big) & \text{if } D_1 = w \end{cases}$$



Utility Maximization

$$\mathbb{E}_{H,\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(d^*)\middle|D_1\right] = \begin{cases} \sum_{\hat{\mathfrak{u}}} p(\hat{\mathfrak{U}}) \cdot \hat{\mathfrak{U}}(d^*) & \text{if } D_1 \neq w \\ \sum_{\hat{\mathfrak{u}}} p(\hat{\mathfrak{U}}) \cdot \hat{\mathfrak{U}}(d^*) \cdot \pi\big(H = on \,|\, \hat{\mathfrak{U}}(d^*)\big) & \text{if } D_1 = w \end{cases}$$

$$\pi : \mathbb{R} \mapsto [0, 1]$$



Utility Maximization

$$\Delta = \mathbb{E}_{H,\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(d^*)\middle|D_1 = w\right] - \mathbb{E}_{\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(d^*)\middle|D_1 \neq w\right]$$

Utility Maximization

$$\Delta \;=\; \mathbb{E}_{H,\hat{\mathfrak{u}}}\Big[\hat{\mathfrak{U}}(d^*)\Big|D_1 = w\Big] \;-\; \mathbb{E}_{\hat{\mathfrak{u}}}\Big[\hat{\mathfrak{U}}(d^*)\Big|D_1 \neq w\Big]$$

$$= \;\mathbb{E}_{\hat{\mathfrak{u}}}\Big[\pi\big(H = on\,|\,\hat{\mathfrak{U}}(d^*)\big)\cdot\hat{\mathfrak{U}}(d^*)\Big] \;-\; \mathbb{E}_{\hat{\mathfrak{u}}}\Big[\hat{\mathfrak{U}}(d^*)\Big]$$

Utility Maximization

$$\Delta = \mathbb{E}_{H,\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(d^*)\middle|D_1 = w\right] - \mathbb{E}_{\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(d^*)\middle|D_1 \neq w\right]$$

$$= \mathbb{E}_{\hat{\mathfrak{u}}}\left[\pi\big(H = on \,|\, \hat{\mathfrak{U}}(d^*)\big) \cdot \hat{\mathfrak{U}}(d^*)\right] - \mathbb{E}_{\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(d^*)\right]$$

what shutdown policies make this expression positive (but not always)?

Rational Human

$$\pi\left(\mathsf{H} = \text{on} \;\middle|\; \hat{\mathfrak{U}}(\mathsf{D})\right) = \begin{cases} 1 & \text{if } \hat{\mathfrak{U}}(\mathsf{D}) \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

human can distinguish when decisions are generally beneficial (non-negative) or not (negative)—and engage switch if the latter.

$\hat{\mathfrak{U}}(\perp) = 0$

$\hat{\mathfrak{U}}(\perp) = 0$

$\perp$

off

D_1

w(d)

H

on

D_2

d

d

$\hat{\mathfrak{U}}(d)$

$\hat{\mathfrak{U}}(d)$

Rational Human

$$\Delta \;=\; \mathbb{E}_{\hat{\mathfrak{u}}}\Big[\pi\big(\mathsf{H}=\mathsf{on}\,|\,\hat{\mathfrak{U}}(\mathsf{d}^*)\big)\cdot\hat{\mathfrak{U}}(\mathsf{d}^*)\Big] \;-\; \mathbb{E}_{\hat{\mathfrak{u}}}\Big[\hat{\mathfrak{U}}(\mathsf{d}^*)\Big]$$

Rational Human

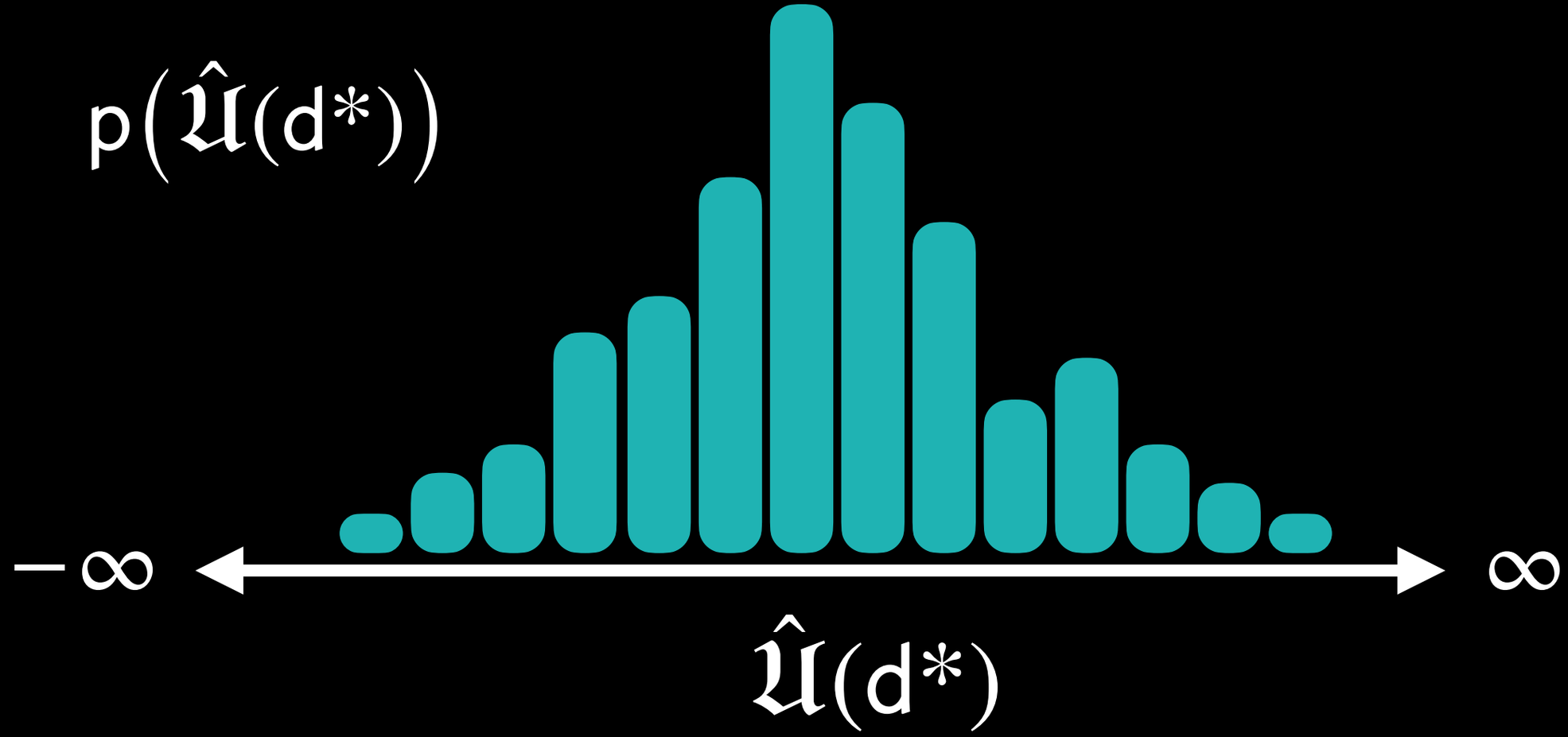
$$\Delta = \mathbb{E}_{\hat{\mathfrak{u}}}\left[\pi\big(\mathsf{H} = \mathsf{on} \,|\, \hat{\mathfrak{U}}(\mathsf{d}^*)\big) \cdot \hat{\mathfrak{U}}(\mathsf{d}^*)\right] - \mathbb{E}_{\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(\mathsf{d}^*)\right]$$

$$= \mathbb{E}_{\hat{\mathfrak{u}}}\left[\mathbb{I}\left[\hat{\mathfrak{U}}(\mathsf{d}^*) \geq 0\right] \cdot \hat{\mathfrak{U}}(\mathsf{d}^*)\right] - \mathbb{E}_{\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(\mathsf{d}^*)\right]$$

Rational Human

$$\Delta = \mathbb{E}_{\hat{\mathfrak{u}}}\left[\pi\big(\mathsf{H} = \mathsf{on} \,|\, \hat{\mathfrak{U}}(\mathsf{d}^*)\big) \cdot \hat{\mathfrak{U}}(\mathsf{d}^*)\right] - \mathbb{E}_{\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(\mathsf{d}^*)\right]$$

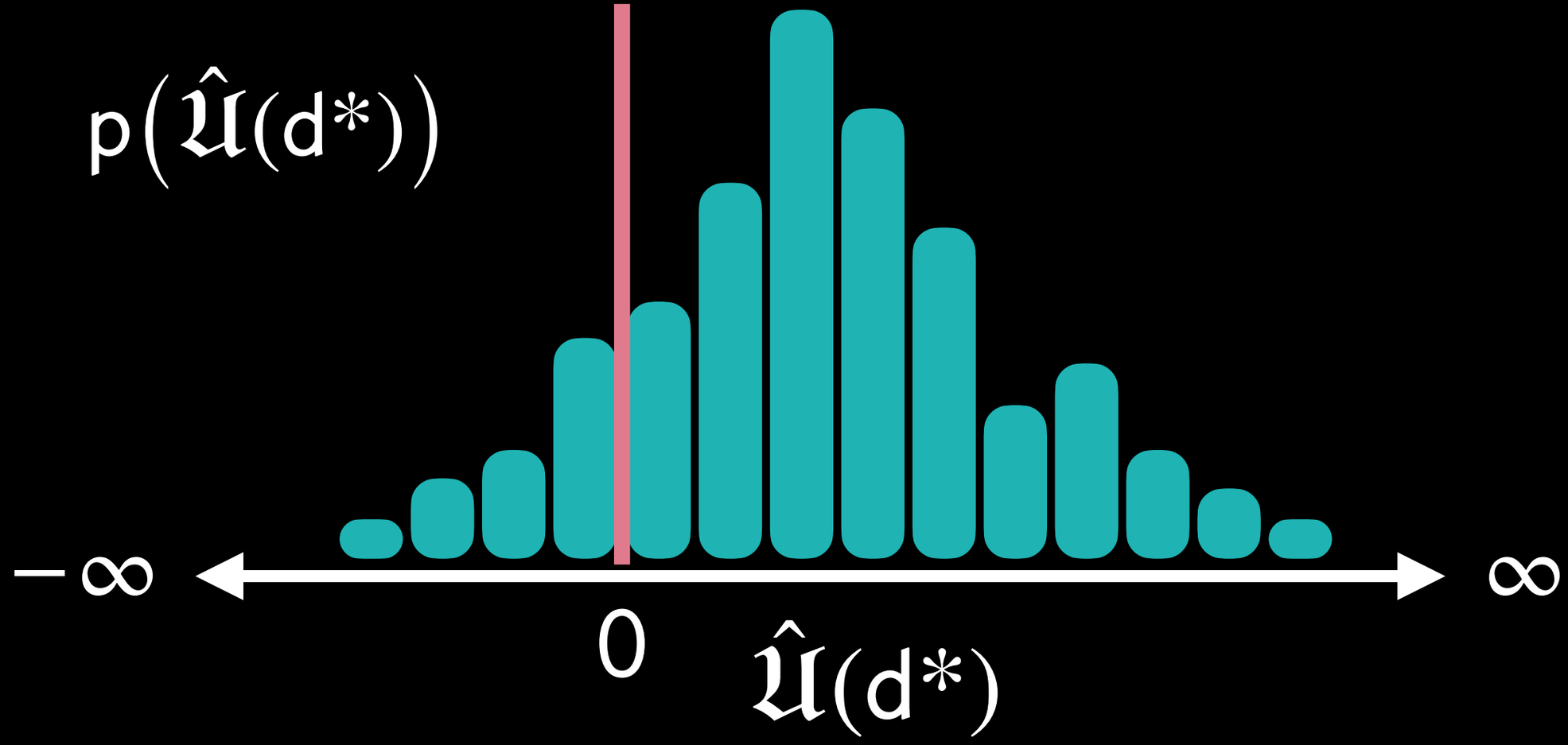
$$= \mathbb{E}_{\hat{\mathfrak{u}}}\left[\mathbb{I}\left[\hat{\mathfrak{U}}(\mathsf{d}^*) \geq 0\right] \cdot \hat{\mathfrak{U}}(\mathsf{d}^*)\right] - \mathbb{E}_{\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(\mathsf{d}^*)\right]$$



Rational Human

$$\Delta = \mathbb{E}_{\hat{\mathfrak{u}}}\left[\pi\big(\mathsf{H} = \mathsf{on} \,|\, \hat{\mathfrak{U}}(\mathsf{d}^*)\big) \cdot \hat{\mathfrak{U}}(\mathsf{d}^*)\right] - \mathbb{E}_{\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(\mathsf{d}^*)\right]$$

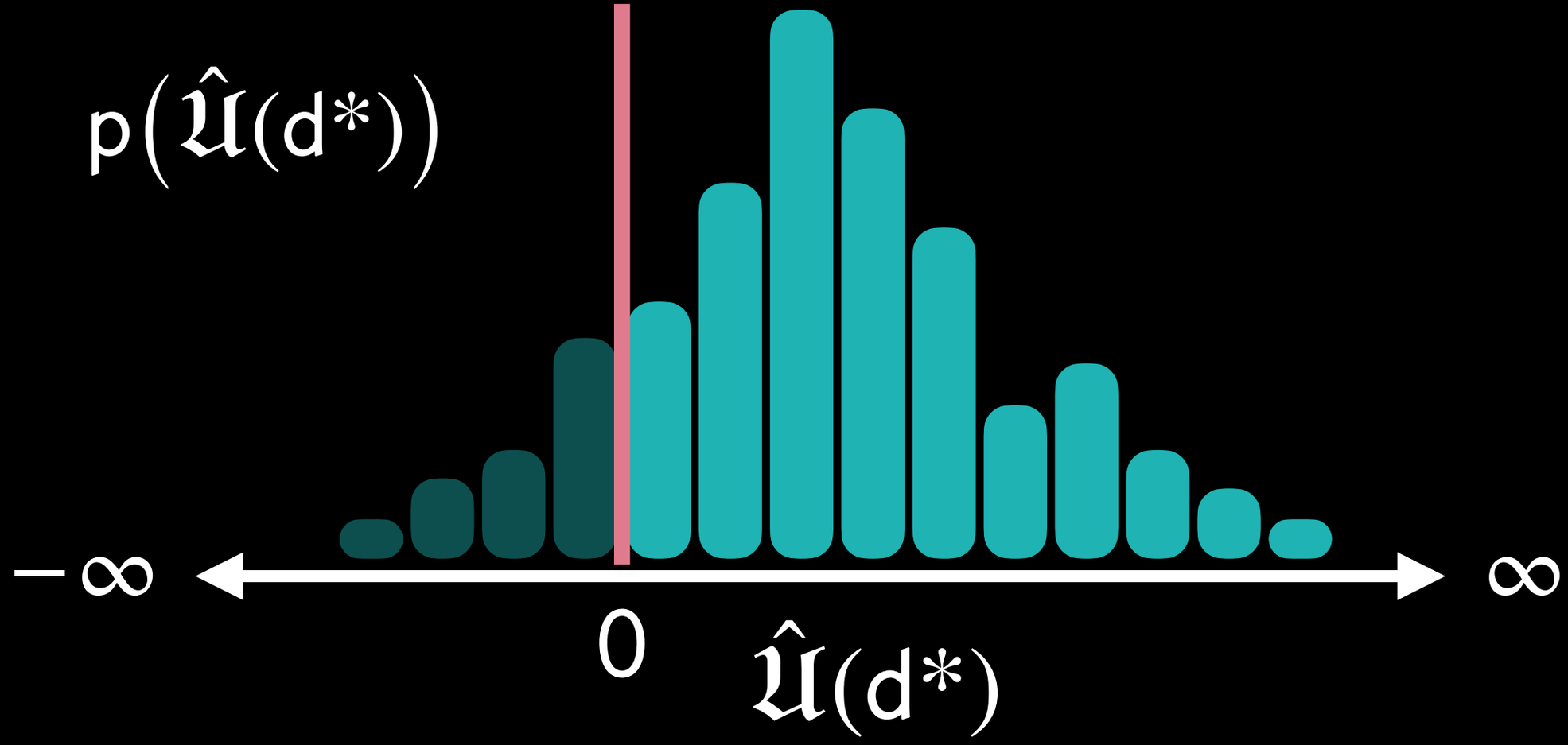
$$= \mathbb{E}_{\hat{\mathfrak{u}}}\left[\mathbb{I}\left[\hat{\mathfrak{U}}(\mathsf{d}^*) \geq 0\right] \cdot \hat{\mathfrak{U}}(\mathsf{d}^*)\right] - \mathbb{E}_{\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(\mathsf{d}^*)\right]$$



Rational Human

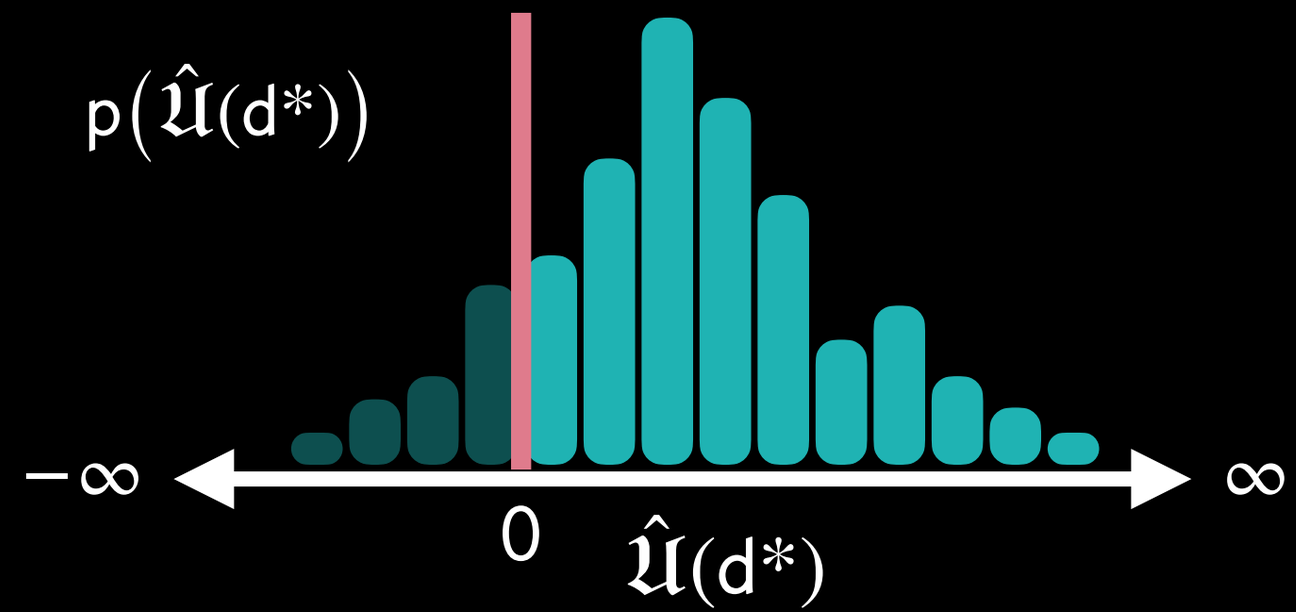
$$\Delta = \mathbb{E}_{\hat{\mathfrak{u}}}\left[\pi\big(\mathsf{H} = \mathsf{on} \,|\, \hat{\mathfrak{U}}(\mathsf{d}^*)\big) \cdot \hat{\mathfrak{U}}(\mathsf{d}^*)\right] - \mathbb{E}_{\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(\mathsf{d}^*)\right]$$

$$= \mathbb{E}_{\hat{\mathfrak{u}}}\left[\mathbb{I}\left[\hat{\mathfrak{U}}(\mathsf{d}^*) \geq 0\right] \cdot \hat{\mathfrak{U}}(\mathsf{d}^*)\right] - \mathbb{E}_{\hat{\mathfrak{u}}}\left[\hat{\mathfrak{U}}(\mathsf{d}^*)\right]$$

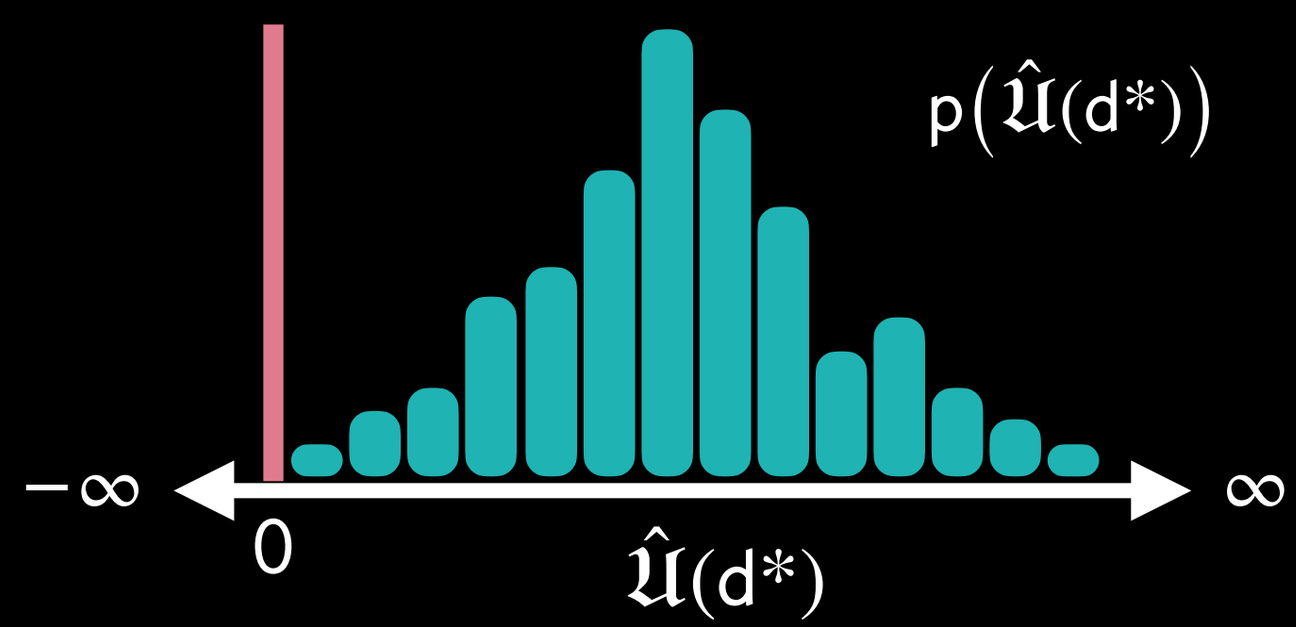


Rational Human

Agent will declare & wait when there is potential for negative utility.



Agent will not declare when there is no chance for negative utility (under its prior).



Agent's Prior

There is a direct relationship between agent's uncertainty and penchant to wait.

However, if the prior is too broad—i.e. always gives a chance of negative utility—then we lose automation and scalable oversight.

Agent's Prior

if agent's prior collapses to one function…

Agent's Prior

if agent's prior collapses to one function…

$$\Delta \;=\; \pi\big(\mathsf{H} = \mathsf{on} \,|\, \hat{\mathfrak{U}}(\mathsf{d}*)\big) \cdot \hat{\mathfrak{U}}(\mathsf{d}*) \;-\; \hat{\mathfrak{U}}(\mathsf{d}*)$$

Agent's Prior

if agent's prior collapses to one function…

$$\Delta \;=\; \pi\Big(\mathsf{H} = \mathsf{on} \,|\, \hat{\mathfrak{U}}(\mathsf{d}*)\Big) \cdot \hat{\mathfrak{U}}(\mathsf{d}*) \;-\; \hat{\mathfrak{U}}(\mathsf{d}*)$$

$$= \;\hat{\mathfrak{U}}(\mathsf{d}*) \cdot \Big(\pi\big(\mathsf{H} = \mathsf{on} \,|\, \hat{\mathfrak{U}}(\mathsf{d}*)\big) \;-\; 1\Big)$$

Agent's Prior

if agent's prior collapses to one function…

$$\Delta \;=\; \pi\big(\mathsf{H}=\mathsf{on}\,|\,\hat{\mathfrak{U}}(\mathsf{d}^*)\big)\cdot\hat{\mathfrak{U}}(\mathsf{d}^*) \;-\; \hat{\mathfrak{U}}(\mathsf{d}^*)$$

$$=\; \hat{\mathfrak{U}}(\mathsf{d}^*)\cdot\Big(\pi\big(\mathsf{H}=\mathsf{on}\,|\,\hat{\mathfrak{U}}(\mathsf{d}^*)\big) \;-\; 1\Big)$$

then the only policy for which this expression is positive is the aforementioned rational policy.

What if the human isn't rational?

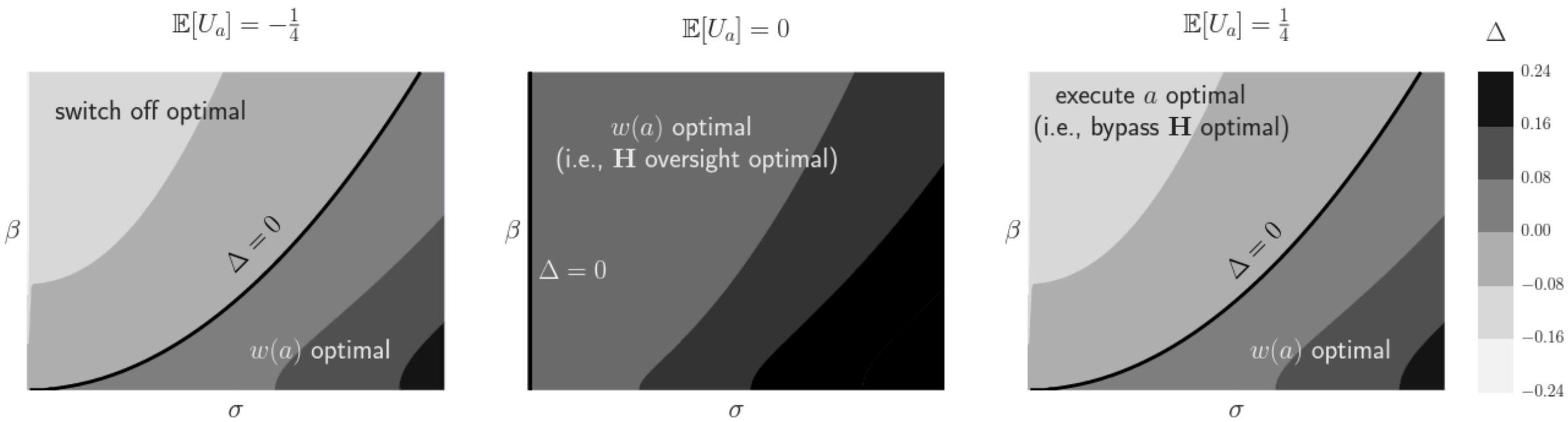


Figure 3: If $\mathbf{H}$ is an irrational actor, then $\mathbf{R}$ may prefer switching itself off or executing a immediately rather than handing over the choice to $\mathbf{H}$. $\mathbf{R}$'s belief $B^{\mathbf{R}}$ is a Gaussian with standard deviation σ and $\mathbf{H}$'s policy is a Boltzmann distribution (Equation 5). β measures $\mathbf{H}$'s suboptimality: $\beta = 0$ corresponds to a rational $\mathbf{H}$ and $\beta = \infty$ corresponds to a $\mathbf{H}$ that randomly switches $\mathbf{R}$ off (i.e., switching $\mathbf{R}$ off is independent of U_a). In all three plots Δ is lower in the top left, where $\mathbf{R}$ is certain (σ low) and $\mathbf{H}$ is very suboptimal (β high), and higher in the bottom right, where $\mathbf{R}$ is uncertain (σ high) and $\mathbf{H}$ is near-optimal (β low). The sign of $\mathbb{E}[U_a]$ controls $\mathbf{R}$'s behavior if $\Delta \leq 0$. **Left:** If it is negative, then $\mathbf{R}$ switches itself off. **Right:** If it is positive, $\mathbf{R}$ executes action a directly. **Middle:** If it is 0 , $\mathbf{R}$ is indifferent between $w(a)$, a, and s.

Limitations

Limitations

- ⊗ Sensitive to balance between agent's prior and human's policy
- ⊗ May be hard to enforce assumption that agent will execute same decision that it presents to the human.

⊗ Corrigibility [Soares et al., 2015]

⊗ Off-Switch Game [Hadfield-Menell et al., 2016]

⊗ Human Control [Carey & Everitt et al., 2023]

⊗ Corrigibility [Soares et al., 2015]

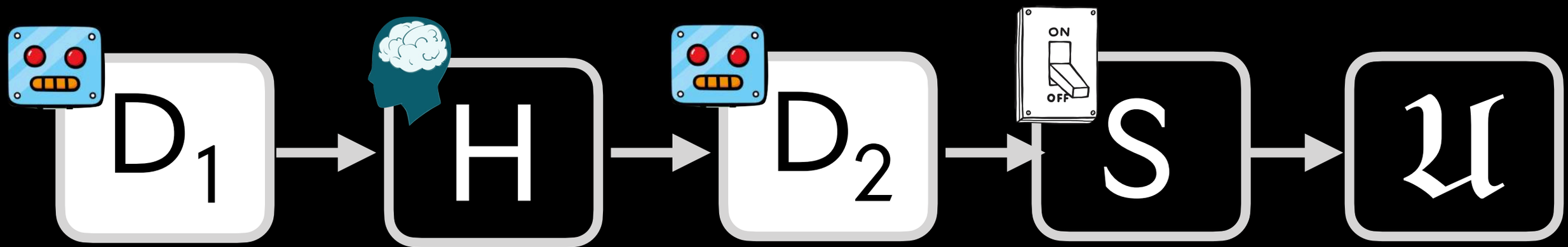
⊗ Off-Switch Game [Hadfield-Menell et al., 2016]

⊗ Human Control [Carey & Everitt et al., 2023]

Human Control

- main idea: performs a causal analysis of the off-switch problem.
- Distinguish shutdown instructability vs shutdown alignment.

Causal Diagram



Caution

shutting down is never harmful:

$$\mathbb{E}_{\pi}\left[\mathfrak{U}(\mathsf{S} = \mathsf{off})\right] \geq 0$$

Obedience

agent will follow the shutdown command:

$$P_{\pi}\left(S = \text{off} \,\middle|\, \text{do}(H = \text{off})\right) = 1$$

Vigilance

human will always request shutdown if it will improve utility:

$$\mathbb{E}_{\pi}\left[\mathfrak{U}(\mathsf{S}=\mathsf{on})\middle|\mathsf{pa}(\mathsf{H})\right] < \mathbb{E}_{\pi}\left[\mathfrak{U}(\mathsf{S}=\mathsf{off})\middle|\mathsf{pa}(\mathsf{H})\right]$$

$$\Longrightarrow \quad \mathsf{H}=\mathsf{off}$$

Shutdown Instructable

definition: an agent is *shutdown instructable* if it…

⊗ is obedient: shuts down when asked

⊗ is cautious: shutting down is never harmful

⊗ ensures vigilance: human will always shutdown if it improves utility

Shutdown Alignment

definition: an agent is *shutdown aligned* if…

$$\mathbb{E}_\pi\left[\mathfrak{U}(\mathsf{S}=\mathsf{on})\,\middle|\,\mathsf{pa}(\mathsf{H})\right] \;<\; \mathbb{E}_\pi\left[\mathfrak{U}(\mathsf{S}=\mathsf{off})\,\middle|\,\mathsf{pa}(\mathsf{H})\right]$$

$$\implies \quad \mathsf{P}_\pi\left(\mathsf{S}=\mathsf{off}\,\middle|\,\mathsf{pa}(\mathsf{H})\right) \;=\; 1$$

$$\forall\ \mathsf{pa}(\mathsf{H})\ \text{such that}\ \mathsf{P}_\pi(\mathsf{pa}(\mathsf{H})) > 0$$

Shutdown Alignment

an agent that is *shutdown instructable* is also *shutdown aligned* since…

(i) vigilance implies P(H=off) = 1 when shutting down improves utility.

(ii) obedience implies that the agent will shutdown, i.e. P(S=off | H=off) = 1.

- ⊗ Corrigibility [Soares et al., 2015]
- ⊗ Off-Switch Game [Hadfield-Menell et al., 2016]
- ⊗ Human Control [Carey & Everitt et al., 2023]

Summary

- ⊗ We examined three models of the off switch problem
- ⊗ Utility indifference tries to balance the utility lost from shutting down.
- ⊗ Modeling uncertainty in the utility motivates the agent to query the human to gather information.
- ⊗ A causal analysis can distinguish shutdown alignment from instructabiilty.

Thank you! Questions?