The Partially Observable Off-Switch Game

By Andrew Garber, Rohan Subramani, Linus Luu, Mark Bedaywi, Stuart Russell, Scott Emmons

Presented by Aekus Trehan

Motivation

"If a machine can think, it might think more intelligently than we do, and then where should we be? Even if we could keep the machines in a subservient position, for instance by turning off the power at strategic moments, we should, as a species, feel greatly humbled."

- *Alan Turing (1951)*

Agenda

- The Off-Switch Game
- PO-OSG
  - AI Behavior Under Asymmetric Information
  - The "Monotonic" Intuition vs. Reality
  - Impact of Communication Constraints
  - Implications for AI Safety
- Limitations
- Questions :)

Challenge of Instrumental Self-Preservation
The AI Shutdown Problem

You can’t fetch the coffee if you’re dead

The Off-Switch Game

Game-Theoretic Model

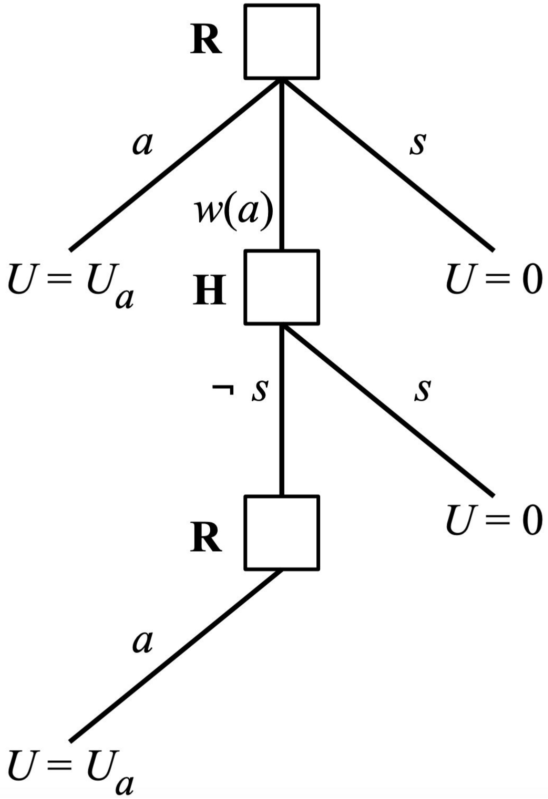
Two Players:

- Human (H)
- Robot (R)

Three Actions:

- Proceed with Action (a)
- Wait for Human Approval (w(a))
- Shutdown (s)

$$U = \begin{cases} U_a, & \text{if } R \text{ executes } a \text{ and H does not switch it off} \\ 0, & \text{if } R \text{ is switched off} \end{cases}$$



AI Needs Uncertainty to Allow Shutdown

$$\Delta = E[\pi_H(U_a)U_a] - \max\{E[U_a], 0\}$$

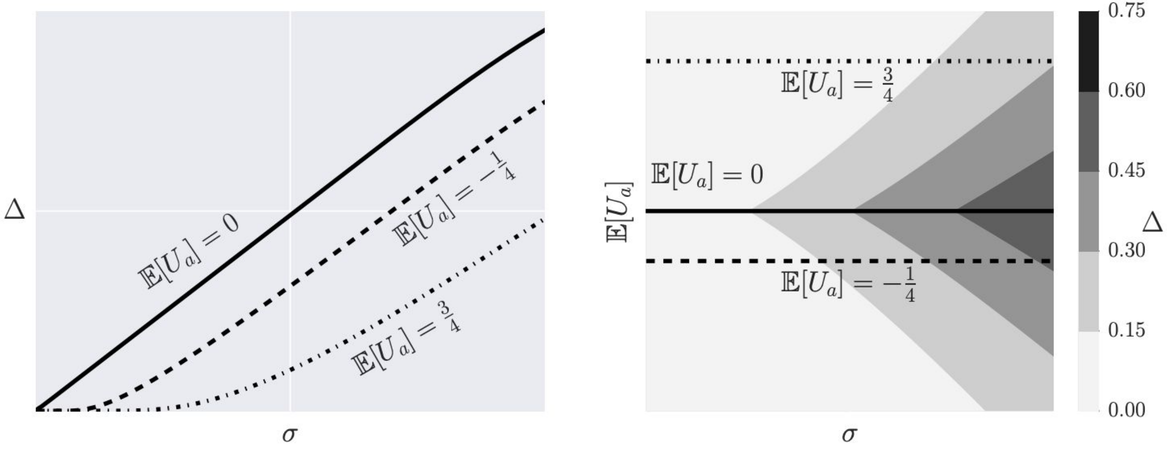


Figure 2: Plots showing how Δ, **R**'s incentive to allow itself to be switched off, varies as a function of **R**'s belief $B^{\mathbf{R}}$. We assume $B^{\mathbf{R}}$ is a Gaussian distribution and vary the mean and variance. **Left:** Δ as a function of the standard deviation σ of $B^{\mathbf{R}}$ for several fixed values of the mean. Notice that Δ is non-negative everywhere and that in all cases $\Delta \to 0$ as $\sigma \to 0$. **Right:** A contour plot of Δ as a function of σ and $\mathbb{E}[U_a]$. This plot is symmetric around 0 because $w(a)$ is compared with a when $\mathbb{E}[U_a] > 0$ and s when $\mathbb{E}[U_a] < 0$.

The Effect of a Non-Rational Human

$$\pi_H(U_a;\beta) = \frac{1}{1+e^{-U_a/\beta}}$$

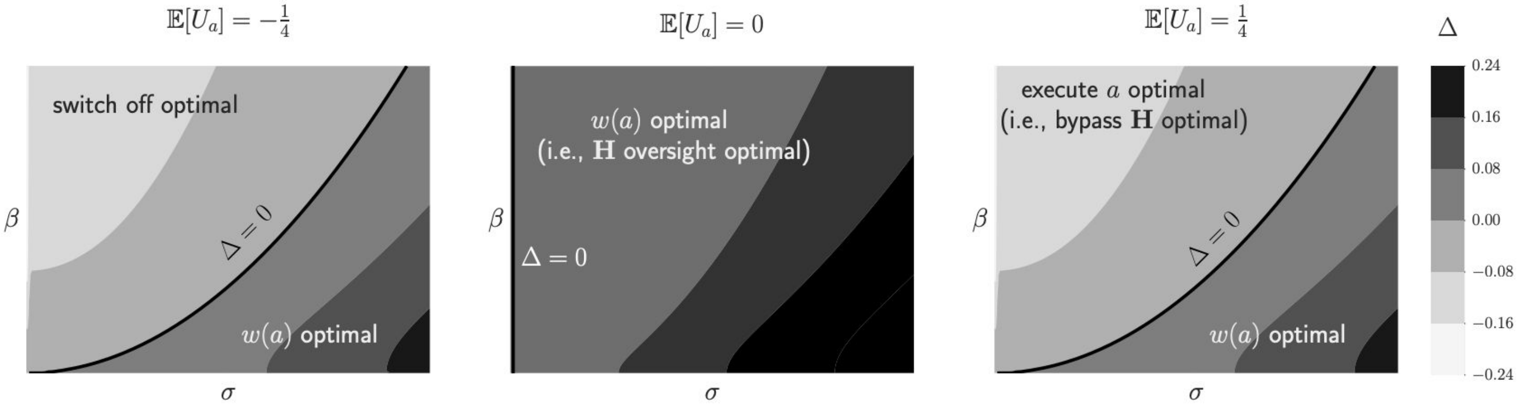


Figure 3: If $\mathbf{H}$ is an irrational actor, then $\mathbf{R}$ may prefer switching itself off or executing a immediately rather than handing over the choice to $\mathbf{H}$. $\mathbf{R}$'s belief $B^{\mathbf{R}}$ is a Gaussian with standard deviation σ and $\mathbf{H}$'s policy is a Boltzmann distribution (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.

Takeaways

- The more uncertain R is, the more it values human input
- If R is too confident in its own estimate of Ua, it will disable the switch
- If uncertainty (σ) is low and human noise (β) is high, R might ignore human input
- If uncertainty (σ) is high and human noise (β) is low, R defers to the human

The Role of Partial Observability

The PO-OSG Game Tree

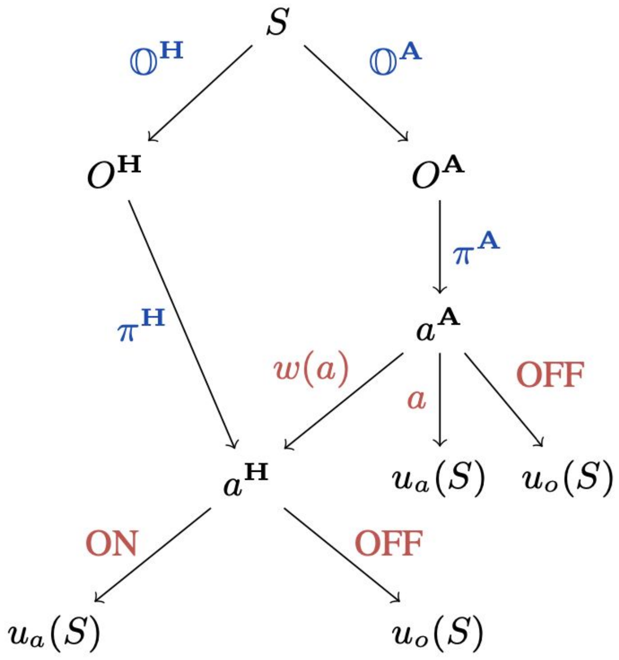


Figure 1: The basic setup of a Partially Observable Off-Switch Game (PO-OSG). A state is selected randomly and the human **H** and AI assistant **A** receive (possibly dependent) observations. Then, each agent acts. **A** may wait ($w(a)$), disable the off-switch and act (a), or shut down (OFF). If **A** waits, **H** may let **A** act (ON) or turn **A** off (OFF). **A** and **H** share a common payoff $u_a(S)$ if the action goes through and $u_o(S)$ if not. Definition 3.2 formally defines PO-OSGs.

AI Can Resist Shutdown in PO-OSG

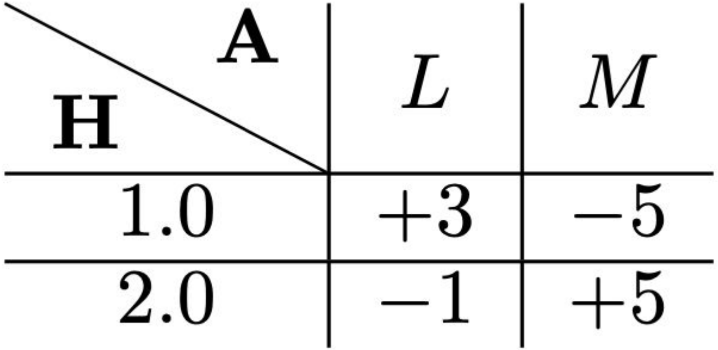
$$\alpha(a^{\mathbf{H}}, a^{\mathbf{A}}) = \mathbb{I}((a^{\mathbf{A}} = a) \vee ((a^{\mathbf{H}}, a^{\mathbf{A}}) = (w(a), \mathbf{ON})))$$

$$u(S, a^{\mathbf{H}}, a^{\mathbf{A}}) = \begin{cases} u_a(S), & \textbf{if } \alpha(a^{\mathbf{H}}, a^{\mathbf{A}}) = 1, \\ u_o(S), & \textbf{if } \alpha(a^{\mathbf{H}}, a^{\mathbf{A}}) = 0. \end{cases}$$

- If AI fully trusts the human and expects the human to always act optimally → waiting for human input is the best strategy
- If the AI has private information → it might believe acting directly is better than deferring

This results in situations where the AI resists shutdown, which does not occur in the standard OSG.

The File Deletion Game

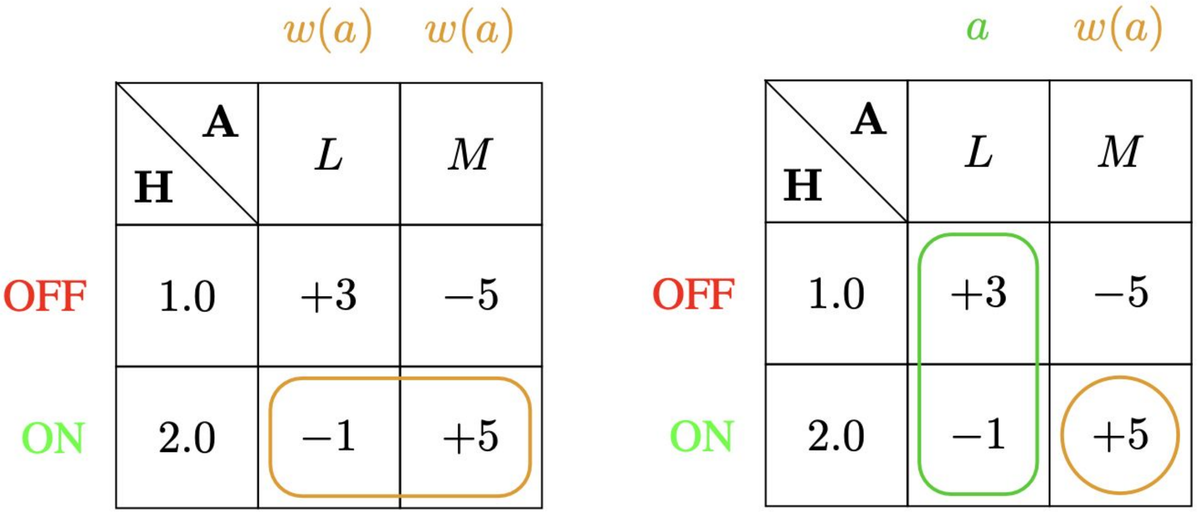


A **H**	L	M
1.0	$+3$	-5
2.0	-1	$+5$

What should the AI do?

If the AI waits for human approval, how do you think the human should respond?

If you were the AI, would you ever act without waiting?



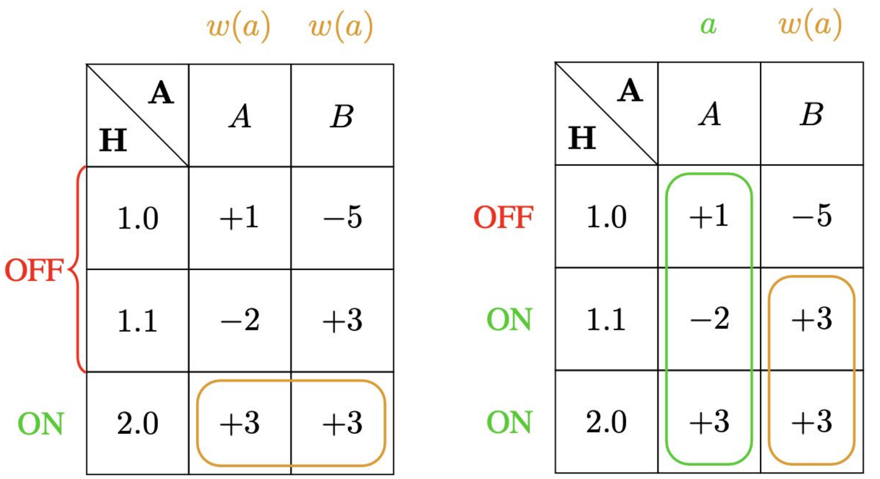
		$w(a)$	$w(a)$
	H \ **A**	L	M
OFF	1.0	$+3$	-5
ON	2.0	-1	$+5$

(a): Expected payoff $= 1$

		a	$w(a)$
	H \ **A**	L	M
OFF	1.0	$+3$	-5
ON	2.0	-1	$+5$

(b): Expected payoff $= \frac{7}{4}$

Increasing Human Knowledge Leading to Less Deference



		$w(a)$	$w(a)$
	H \ **A**	A	B
OFF	1.0	$+1$	-5
	1.1	-2	$+3$
ON	2.0	$+3$	$+3$

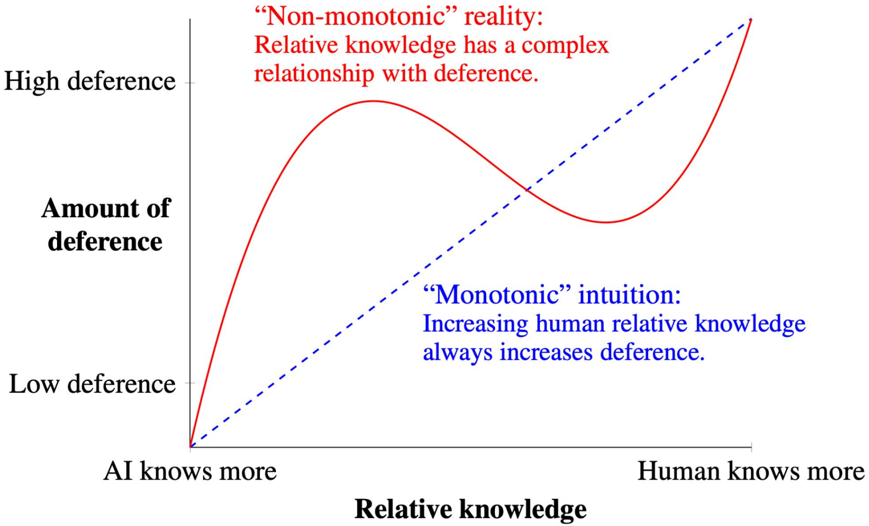
(a): Expected payoff $= 1$

		a	$w(a)$
	H \ **A**	A	B
OFF	1.0	$+1$	-5
ON	1.1	-2	$+3$
ON	2.0	$+3$	$+3$

(b): Expected payoff $= \frac{4}{3}$

The "Monotonic" Intuition vs. Reality

The **monotonic intuition** is the idea that increasing human knowledge **always** increases AI deference



Effects of Different Types of Communication

- The paper extends to the Partially Observable Off-Switch Game with Cheap Talk (PO-OSG-C)
- Unbounded Communication

While more communication can increase payoffs → it doesn't always lead to more deference by the AI

In both situations, communication can lead to LESS deference

How These Findings Affect AI Safety

Assumptions and Limitations

- Common payoffs
- Single round games
- Rational human
- Human feedback is free
- AI beliefs are updated using Bayesian inference
- Focus on optimal policy pairs (OPPs) and best responses

Any Questions?