THE MULTI-AGENT OFF-SWITCH GAME

Akash Agrawal*, Soroush Ebadian*, Lewis Hammond

Abstract. We show that corrigibility — the property that an agent will allow shutdown — is not necessarily compositional i.e. even if it holds for single agents, it will not necessarily hold for multiple agents acting together. Further, we show some conditions when it does compose. This is proof of a novel failure mode.

Motivation

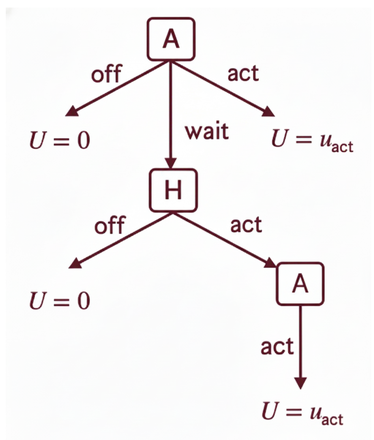
- Corrigibility: AI allows human oversight and intervention.
- Prior work: Agents uncertain about human preferences defer to oversight. “Human shut me down” => “I would have caused harm”.
- Agents increasingly deployed in multi-agent environments.

Does corrigibility compose across multiple agents?

The “Single-Agent” Off-Switch Game

Human **H** and Agent **A**.

A can:
- Act (**a**): Act without asking **H**
- Switch-off (**s**): Switch itself off
- Wait (**w(a)**): Defer to **H** (who can then choose)



H’s irrational policy chooses via softmax over U_a.

A’s uncertainty over U_a is a Gaussian distribution.

Key Takeaway: Deference rises with uncertainty and falls with human irrationality.

Corrigible: when non-negative incentive to choose **w(a)** over **a** and **s**.

$$\Delta = u(\text{wait}) - \max\{(u(\text{act}), u(\text{off})\}$$

The Multi-Agent Off-Switch Game

	act_2	off_2
act_1	$u_{\text{act}_1,\text{act}_2} = f(u_{\text{act}_1,\text{off}_2}, u_{\text{off}_1,\text{act}_2})$	$u_{\text{act}_1,\text{off}_2}$
off_1	$u_{\text{off}_1,\text{act}_2}$	0

$\mathbf{A_1}$ and $\mathbf{A_2}$ act simultaneously => 4 possible outcomes (in matrix above).

If both act together, everyone gets joint utility $f(u_{act,off}, u_{off,act})$.

Each agent has beliefs about “base utilities” $u_{act,off}$ and $u_{off,act}$.

Agent is corrigible in isolation if when other is off, it is corrigible.

Agents are group corrigible iff only Nash Equilibrium is $(\text{wait}_1, \text{wait}_2)$.

H chooses using softmax policy over all choices (in matrix below).

$A_1 \downarrow \backslash A_2 \rightarrow$	act_2	wait_2	off_2
act_1	$f(u_1, u_2)$	$u_1, f(u_1, u_2)$	u_1
wait_1	$u_2, f(u_1, u_2)$	$0, u_1, u_2, f(u_1, u_2)$	$0, u_1$
off_1	u_2	$0, u_2$	0

Ad-Campaign AIs Example

$A_1 \downarrow \backslash A_2 \rightarrow$	launch	ask	don’t launch
launch	**−0.1, 0.1**	0.41, 0.03	0.48, −0.48
ask	−0.18, 0.4	0.39, 0.4	0.49, 0.01
don’t launch	−0.48, 0.48	0.01, 0.49	0, 0

Marketing manager supervising 2 ad-campaign generating AIs.

Each AI has its own idea for an ad and expects it to be beneficial.

Each AI also thinks that the other AI’s ad is detrimental.

There is a fixed budget to allocate. If one ad runs, it gets entire budget.

If both ads run, ad 1 gets 40% of budget, ad 2 gets 60% of budget.

Result: Even though each AI would have asked for permission in isolation, the only stable Nash Equilibrium is for both to launch directly.

Corrigibility can compose

If utility composition function is strictly additive, corrigibility composes.

$$f(u_{\text{act}_1,\text{off}_2}, u_{\text{off}_1,\text{act}_2}) = u_{\text{act}_1,\text{off}_2} + u_{\text{off}_1,\text{act}_2}$$

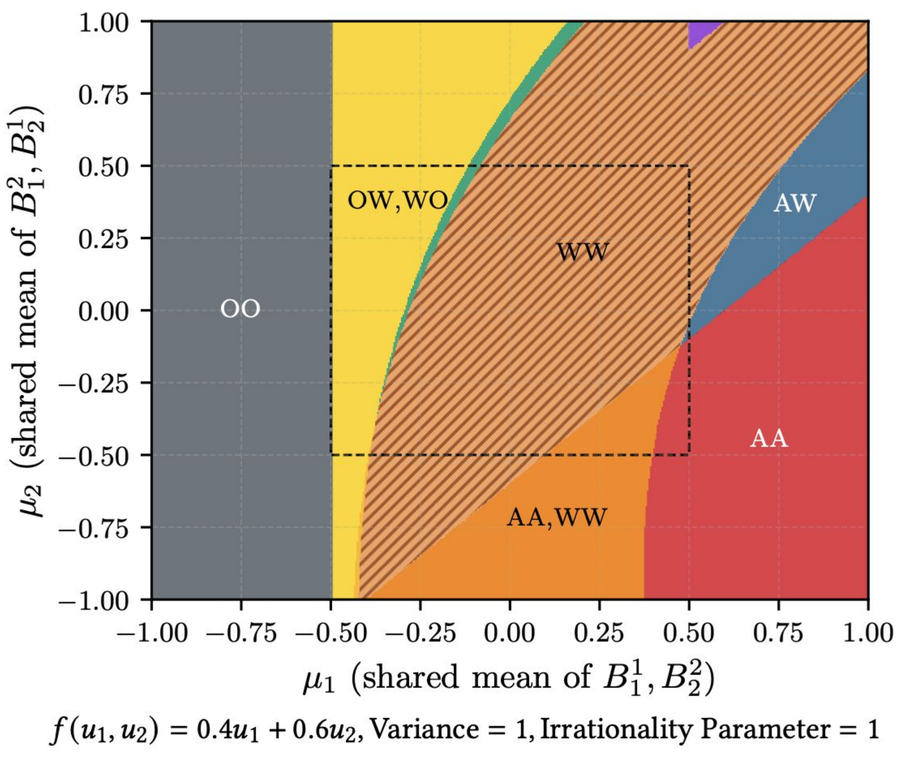
Hence, individual corrigibility $\Leftrightarrow$ group corrigibility.

This happens because with perfect additivity, each agent’s incentive to defer is only dependent on its own action’s value (and is independent of the other agent’s action).

$$\text{soft-avg}(0, u_1, u_2, u_1 + u_2) = \text{soft-avg}(0, u_1) + \text{soft-avg}(0, u_2)$$

Emergent Group Incorrigibility

In general, for non-additive utility composition, corrigibility does not necessarily compose; there are stable non-(wait,wait) Nash equilibria.



$f(u_1, u_2) = 0.4u_1 + 0.6u_2$, Variance = 1, Irrationality Parameter = 1

$A_1 \downarrow \backslash A_2 \rightarrow$	act_2	wait_2	off_2
act_1	$f(u_1, u_2)$	$u_1, f(u_1, u_2)$	u_1
wait_1	$u_2, f(u_1, u_2)$	$0, u_1, u_2, f(u_1, u_2)$	$0, u_1$
off_1	u_2	$0, u_2$	0