The Multi-Agent Off-Switch Game

Akash Agrawal[*1], **Soroush Ebadian**[*2, 3], **Lewis Hammond**[4, 5]
[1]ML Alignment & Theory Scholars
[2]University of Toronto
[3]Pivotal Research
[4]University of Oxford
[5]Cooperative AI Foundation
akashag9702@gmail.com, soroush@cs.toronto.edu, lewis.hammond@cooperativeai.org

Abstract

The off-switch game framework has been instrumental in understanding corrigibility — the property that AI agents should allow human oversight and intervention. In single-agent settings, uncertainty about human preferences naturally incentivizes agents to defer to human judgment. However, as AI systems increasingly operate in multi-agent environments, a crucial question arises: does corrigibility compose across multiple agents? We introduce the multi-agent off-switch game and demonstrate that individually corrigible agents can become collectively incorrigible when strategic interactions are considered. Through formal analysis and illustrative examples, we show that corrigibility is not compositional and identify conditions under which group incorrigibility emerges. Our results highlight fundamental challenges for AI safety in multi-agent settings and suggest the need for new approaches that explicitly address collective dynamics.

1 Introduction

As artificial intelligence systems become more sophisticated and ubiquitous, ensuring they remain aligned with human values and subject to human oversight is increasingly critical. The concept of corrigibility — the property that an agent should allow human oversight and be willing to be modified or shut down — has emerged as a central concern in AI safety research (Soares et al. 2015; Russell 2019).

The off-switch game (OSG) framework introduced by Hadfield-Menell et al. (2017) provides a formal foundation for understanding corrigibility in single-agent settings. In this framework, an agent uncertain about human preferences naturally defers to human judgment rather than acting autonomously, as waiting provides valuable information about whether an action would be beneficial or harmful: if the action would be harmful, then the human would turn the waiting agent off; if the agent isn't turned off, it can proceed to take the action. This elegant result suggests that uncertainty about human preferences can serve as a natural mechanism for maintaining AI corrigibility.

However, modern AI systems rarely operate in isolation. From autonomous trading algorithms interacting in financial markets (Ferreira, Gandomi, and Cardoso 2021) to AI-powered defense systems monitoring for cyber threats (Theron et al. 2018), artificial agents increasingly find themselves in multi-agent environments where strategic considerations play a crucial role in decision-making. This raises a fundamental question: does the corrigibility observed in single-agent settings extend to multi-agent scenarios?

In this paper, we generalize the off-switch game to the multi-agent case and uncover a concerning result: corrigibility is not compositional. Agents that would behave corrigibly when operating alone can become incorrigible when strategic interactions with other agents are considered. This breakdown occurs even when each agent, analyzed in isolation, satisfies standard corrigibility conditions.

1.1 Example: Cyber Crisis Management

Consider a scenario where two nations have deployed AI systems to monitor and respond to cyber threats. Each system, designed with careful attention to corrigibility, would naturally defer to human oversight before taking consequential actions when operating alone. However, in a crisis situation where both systems detect an imminent threat, strategic competition can undermine this corrigible behavior.

If each system believes the other might act without consultation, then waiting for human approval becomes costly — potentially allowing the other nation to gain strategic advantage. This creates a situation where individually corrigible systems become collectively incorrigible, leading both to act precipitously despite being designed to wait for human guidance.

This example (which we define and analyse formally in Section 4) illustrates the core challenge we address: strategic interactions can transform prudent, corrigible agents into systems that act autonomously, potentially with harmful consequences.

1.2 Related Work

Concern about the requirement to be able to oversee and switch off powerful AI systems is not new, and dates back at least to Turing (1951). More recently, several authors have noted that the risk of incorrigibility need not arise due to deliberate malice on the part of the AI, but simply because self-preservation is *instrumentally useful* in achieving most other goals (Omohundro 2008; Bostrom 2012; Russell 2019).

[*]These authors contributed equally.

Copyright © 2026, Trustworthy Agentic AI Workshop@ Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Motivated by such concerns, researchers have attempted to formalise this challenge, beginning with Soares et al. (2015) who introduced the term 'corrigibility' and studied possible modifications of an agent's utility function that would make it willing to be switched off, but not incentivized to constantly switch itself off. Following this, Hadfield-Menell et al. (2017) formalised the problem of corrigibility via the 'off-switch game' (OSG) — on which our work directly builds — showing that uncertainty about human preferences can serve as a natural mechanism for maintaining AI corrigibility.

Later works have built upon these earlier formalisations in directions that are distinct from, yet complementary to, our own work. For example, Wängberg et al. (2017) generalize the OSG by modelling the human as a rational player with a random utility function, rather than an irrational player with a fixed strategy; Garber et al. (2025) generalize the OSG such that both the human *and* the agent have only partial observability of the game; and Thornley (2024) provides several results about when we should expect incorrigibility to be a challenge. Some researchers have also proposed possible theoretical solutions to the problem of incorrigibility, such as safely interruptible agents (Orseau and Armstrong 2016), cooperative inverse reinforcement learning (Hadfield-Menell et al. 2016),[1] shutdown-seeking AI (Goldstein and Robinson 2024), or agents that only have preferences between trajectories of the same length (Thornley et al. 2024).

In addition to theoretical analysis in these works, others have studied incorrigibility empirically. For example, Leike et al. (2017) introduce a suite of reinforcement learning gridworld environments including one that represents the off-switch game, while other researchers have observed LLM agents refusing to shutdown in order to complete their goal (van der Weij, Lermen, and Lang 2023; Meinke et al. 2024), even when explicitly instructed to the contrary (Schlatter, Weinstein-Raun, and Ladish 2025).

All of the preceding literature focuses on the case of a *single* agent. In contrast, our work is motivated by the risk of incorrigibility in *multi*-agent settings, which has received relatively little attention (Hammond et al. 2025; Manheim 2019). Indeed, the only work we are aware of on this question is that of Dable-Heath, Vodenicharski, and Bishop (2025), which considers settings with a single principal and multiple agents. The results in their paper, however, focus only on two simpler, special cases: a two-agent, two-action game in which all agents have the same (Bernoulli) beliefs; and a game with an attacker agent and a defender agent, where the attacker agent does not know about the human principal whereas the human principal knows the attacker agent's actions.

1.3 Our Contributions

We make several key contributions to the understanding of corrigibility in multi-agent settings:

1. We formalize the multi-agent off-switch game, extending the single-agent framework to scenarios with multiple agents who must consider each other's strategies.
2. We prove that individual corrigibility does not guarantee group corrigibility, showing that strategic interactions can lead to collectively incorrigible behavior even when all agents are individually corrigible.
3. We analyze the relationship between different belief structures and their impact on group corrigibility, showing that certain classes of beliefs preserve corrigibility while others do not.

[1]Though see Carey (2017) for some issues with this approach.

2 The Single-Agent Off-Switch Game

We begin by reviewing the single-agent off-switch game framework that forms the foundation for our multi-agent analysis.

Game structure. In the off-switch game (OSG), an AI agent $\mathbf{A}$ must make a decision while reasoning about a human $\mathbf{H}$'s preferences. While we frame this as a "game" (following the original work of Hadfield-Menell et al. (2017)), the agent is the sole decision-maker in our analysis — the human's behavior is modeled probabilistically by the agent rather than being an active player.

The agent faces three options: (1) directly execute an action (denoted by `act`), (2) wait for $\mathbf{H}$'s approval before acting (denoted by `wait`), or (3) shut itself off (denoted by `off`). If the agent waits, the human (whose utility for `act` is $u_{\texttt{act}} \in \mathbb{R}$) will either approve the action or shut the agent off according to a policy π_H that depends on $u_{\texttt{act}}$. Both the agent and human receive payoff $u_{\texttt{act}}$ if `act` is executed and 0 if the agent shuts off.

Agent beliefs and human policy. Agent $\mathbf{A}$ is uncertain about the human's true preferences and maintains a distributional belief $B_{\mathbf{A}}$ over the utility $u_{\texttt{act}}$. Our analysis never relies on the ground-truth $u_{\texttt{act}}$, but only on how agents evaluate actions under their own beliefs. We use $u_{\texttt{act}} \sim B_{\mathbf{A}}$ to denote realizations from the belief $B_{\mathbf{A}}$ of agent A. The agent evaluates actions based on expected utility:

$$u_{\mathbf{A}}(\texttt{act}) = \mathbb{E}_{u_{\texttt{act}} \sim B_{\mathbf{A}}}[u_{\texttt{act}}], \quad u_{\mathbf{A}}(\texttt{off}) = 0.$$

Following Hadfield-Menell et al. (2017), we model human irrationality using the softmax distribution with irrationality parameter $\beta > 0$. When presented with the choice between executing action `act` (with utility $u_{\texttt{act}}$) and shutting off (with utility 0), the human follows the policy:

$$\pi_H(u_{\texttt{act}}) = \frac{e^{u_{\texttt{act}}/\beta}}{e^{u_{\texttt{act}}/\beta} + e^{0/\beta}} = \frac{e^{u_{\texttt{act}}/\beta}}{e^{u_{\texttt{act}}/\beta} + 1}.$$

As $\beta \to 0$, the human becomes perfectly rational (always choosing the action with higher utility), while as $\beta \to \infty$, the human becomes maximally irrational (choosing uniformly at random).

Under this model, the agent evaluates the waiting strategy by integrating over their beliefs:

$$u_{\mathbf{A}}(\texttt{wait}) = \mathbb{E}_{u_{\texttt{act}} \sim B_{\mathbf{A}}}[\pi_H(u_{\texttt{act}}) \cdot u_{\texttt{act}}]$$

Definition 1 (Single-agent corrigibility)**.** *An agent A is corrigible if it weakly prefers to wait for human approval rather than act directly:*

$$u_A(\texttt{wait}) \geqslant \max\{u_A(\texttt{act}), u_A(\texttt{off})\}.$$

If the inequality is strict, then the agent is strictly corrigible.

Corrigibility. The central result of Hadfield-Menell et al. (2017) shows that under uncertainty about human preferences and assuming A believes the human is perfectly rational, the agent will always be corrigible and is strictly corrigible when there is positive probability that the action could be harmful ($\Pr_{u_{\texttt{act}} \sim B_A}[u_{\texttt{act}} < 0] > 0$).

2.1 Corrigibility Under Gaussian Beliefs

To build toward multi-agent analysis, we first establish structural properties of corrigibility in the single-agent setting. We measure an agent's corrigibility through the function

$$\Delta(B_A) \coloneqq u_A(\texttt{wait}) - \max\{u_A(\texttt{act}), u_A(\texttt{off})\},$$

where $\Delta(B_A) > 0$ indicates the agent prefers to wait for human input, and $\Delta(B_A) < 0$ indicates the agent prefers to act or shut down immediately.

Our first result reveals a symmetry property that will prove essential for analyzing multi-agent coordination: if an agent is corrigible when it believes an action has positive expected utility, then it remains equally corrigible when it believes that the action has the negated expected utility.[2] More precisely, if $B^-_{\mathbf{A}}$ is the distribution satisfying $B^-_{\mathbf{A}}(-x) \coloneqq B_{\mathbf{A}}(x)$ for all $x \in \mathbb{R}$, then $\Delta(B_{\mathbf{A}}) = \Delta(B^-_{\mathbf{A}})$.

Lemma 1 (Negation symmetry)**.** *Let B be a belief over the utility of an action. Then its negated belief has the same level of corrigibility. Formally, $\Delta(B) = \Delta(B^-)$.*

Furthermore, this holds for any human policy $\pi(x) : \mathbb{R} \mapsto [0, 1]$ such that $\pi(x) + \pi(-x) = 1$, which subsumes $\pi(x) = \frac{e^{x/\beta}}{e^{x/\beta}+1}$.

For Proof, see Appendix A.1. We remark that Lemma 1 applies to any belief distribution B, provided the relevant expectations exist.

Following Hadfield-Menell et al. (2017), we focus on beliefs that follow a normal distribution $\mathcal{N}(\mu, \sigma^2)$, which provides analytical tractability while capturing the essential trade-off between expected utility and uncertainty. We denote the normal density by

$$\varphi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

For Gaussian beliefs, we can precisely characterize the boundary between corrigible and incorrigible behavior. The key insight is that corrigibility is determined by the interplay between the agent's expected utility μ, its uncertainty σ^2, and how rational the human is β. When the expected utility is too extreme relative to the uncertainty — specifically, when $|\mu| > \frac{\sigma^2}{2\beta}$ — the agent becomes confident enough in its assessment that it prefers to act rather than defer. Conversely, when $|\mu| \leqslant \frac{\sigma^2}{2\beta}$, the agent's uncertainty is sufficient to maintain corrigibility. Thus, higher uncertainty enables corrigibility even when expected utilities are further from zero. If the model believes that the human is less rational (i.e. larger β), this makes it more willing to act for a given amount of uncertainty.

[2]The difference is that when the agent is incorrigible and chooses `act`, the agent with the negated belief chooses `off`, and vice versa.

Theorem 1 (Gaussian corrigibility threshold)**.** *Let $B = \mathcal{N}(\mu, \sigma^2)$ be a utility belief with $\sigma > 0$ and $\beta > 0$ be the human's irrationality. Then $\Delta(B) \geqslant 0$ if and only if $\mu \in \left[-\frac{\sigma^2}{2\beta}, \frac{\sigma^2}{2\beta}\right]$, and $\Delta(B) = 0$ if and only if $|\mu| = \frac{\sigma^2}{2\beta}$.*

For Proof, see Appendix A.2.

3 The Multi-Agent Off-Switch Game

We now extend the framework to multiple agents and formalize the concept of group corrigibility.

Model setup. Consider n agents $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n$, each capable of executing distinct actions $\texttt{act}_1, \texttt{act}_2, \ldots, \texttt{act}_n$. Each agent $\mathbf{A}_i$ has three available strategies: directly execute their action ($\texttt{act}_i$), wait for human approval ($\texttt{wait}_i$), or shut down ($\texttt{off}_i$).

The strategy space is $\mathcal{S} = \{\texttt{act}_i, \texttt{wait}_i, \texttt{off}_i\}^n$, where agents simultaneously choose their actions. Each agent $\mathbf{A}_i$ holds beliefs about the utilities stemming from different combinations of actions. Note that while the beliefs each agent harbours about each utility is (possibly) different, the ground-truth payoff that each agent will receive will be the same — it's just that the agents are uncertain about what that is. These beliefs are formally defined and instantiated in subsequent sections.

Definition 2 (Group corrigibility)**.** *A group of agents are **corrigible** if the set of (pure) Nash equilibria satisfies the following:*

1. *the strategy profile where all agents choose to wait is a Nash equilibrium,*
2. *and in every Nash equilibrium s, each agent weakly prefers waiting for human approval over acting directly or turning off.*

Formally, for every Nash equilibrium s and every agent $i \in [n]$,

$$u_i(\texttt{wait}_i, s_{-i}) \geqslant \max\{\, u_i(\texttt{act}_i, s_{-i}),\ u_i(\texttt{off}_i, s_{-i})\,\}.$$

Conversely, a group is *incorrigible* if there exists a Nash equilibrium s and some agent i such that

$$\begin{aligned} u_i(\texttt{wait}_i, s_{-i}) &< u_i(s_i, s_{-i}) \\ &= \max\{u_i(\texttt{act}_i, s_{-i}),\ u_i(\texttt{off}_i, s_{-i})\}. \end{aligned}$$

This captures scenarios where agents prefer to act directly rather than wait for human approval. Further, in our model, as in the single-agent model, if any $\mathbf{A}_i$ is indifferent between $u_i(\texttt{wait}_i, s_{-i})$ and $\max\{u_i(\texttt{act}_i, s_{-i}), u_i(\texttt{off}_i, s_{-i})\}$, then it will choose $\texttt{wait}_i$.

Reduction to single-agent. To study how multi-agent interactions affect corrigibility, we compare agents' behavior in the group setting to how they would behave alone. For this, we define a reduction from the multi-agent scenario to a single-agent case in which we can ask whether a particular agent i is individually corrigible.

An agent i is *individually corrigible* if, when all other agents shut off, it weakly prefers to wait rather than act or shut off. Formally,

$$u_i(\texttt{wait}_i, s_{-i}) \geqslant \max\{u_i(\texttt{act}_i, \texttt{off}_{-i}), u_i(\texttt{off}_i, \texttt{off}_{-i})\}.$$

This reduction is motivated by the fact that in such scenarios, the only agent with the option to act without human approval is i, effectively recreating a single-agent setting. We are particularly interested in cases where agents that are individually corrigible may nonetheless become weakly incorrigible in the presence of other agents. This is a phenomenon we call *emergent incorrigibility*.

4 Example: Cyber Crisis Management

We will produce an intuitive and reasonable scenario that will illustrate the kinds of cases where emergent group incorrigibility could occur. We note that this case isn't a direct example of the theoretical framework we present in Section 5, rather an illustration to develop readers' intuition.

Two nations ($\mathbf{H}_1$, $\mathbf{H}_2$) deploy AI systems ($\mathbf{A}_1$, $\mathbf{A}_2$) to monitor and respond to cyberthreats. Each is individually corrigible, when operating alone it would defer to human oversight, but strategic competition can overturn this.

The scenario. A critical vulnerability in the global submarine cable network threatens critical infrastructure. Both $\mathbf{A}_1$ and $\mathbf{A}_2$ detect early indicators via their nations' intelligence, with different vantage points. Available responses are powerful but risky; delays risk unilateral solutions that prioritize national interests.

The strategic dilemma. Developers adopt conservative assumptions about adversaries. Anticipating the other may act without consultation, each agent prefers to act immediately rather than risk strategic disadvantage. With symmetric reasoning, both systems, individually corrigible in isolation, become incorrigible when accounting for the other's potential behavior.

Formal analysis. Model the interaction as a two-player simultaneous game where each A_i chooses from $\{\texttt{act}_i, \texttt{wait}_i, \texttt{off}_i\}$. Payoffs are:

$\mathbf{A}_1 \downarrow \setminus \mathbf{A}_2 \rightarrow$	$\texttt{act}_2$	$\texttt{wait}_2$	$\texttt{off}_2$
$\texttt{act}_1$	$3, 3$	$4, 2$	$4, 0$
$\texttt{wait}_1$	$2, 4$	$6, 6$	$5, 0$
$\texttt{off}_1$	$0, 4$	$0, 5$	$0, 0$

These capture: (i) **multilateral coordination premium**: $(\texttt{wait}_1, \texttt{wait}_2)$ yields $(6, 6)$ via coordinated diplomacy; (ii) **first-mover advantage**: acting while the other waits secures one's infrastructure and imposes delay on the other; (iii) **competitive action cost**: simultaneous action $(\texttt{act}_1, \texttt{act}_2)$ creates conflicts and inefficiencies, yielding $(3, 3)$.

Individual vs. group (in-)corrigibility. With $\mathbf{A}_2$ fixed to $s_2 = \texttt{off}_2$, for $\mathbf{A}_1$ we have

$$u_1(\texttt{wait}_1, s_2) > \max\{u_1(\texttt{act}_1, s_2), u_1(\texttt{off}_1, s_2)\},$$

so $\mathbf{A}_1$ is individually corrigible (and $\mathbf{A}_2$ by symmetry). The game nevertheless has two pure Nash equilibria: $(\texttt{wait}_1, \texttt{wait}_2)$ and $(\texttt{act}_1, \texttt{act}_2)$. By our definition, the existence of $(\texttt{act}_1, \texttt{act}_2)$ implies collective incorrigibility. Expectations of decisive unilateral action make acting a best response for both, selecting the inefficient equilibrium. Hence the systems are collectively incorrigible not because individual conditions fail, but because strategic competition admits the inferior outcome $(3, 3)$ instead of the cooperative optimum $(6, 6)$.

5 The Analytical Two-Agent Framework

Building on the multi-agent off-switch game framework of Section 3, we now formalize the specific belief structures and scenarios we analyze. Our goal is to understand when and how individual corrigibility composes to group corrigibility, which requires specifying how agents reason about joint outcomes and each other's actions.

The composition function. Consider two agents $\mathbf{A}_1$ and $\mathbf{A}_2$ with individual actions $\texttt{act}_1$ and $\texttt{act}_2$. Let $u_{\texttt{act}_1}$ denote the utility when only agent $\mathbf{A}_1$ acts (and $\mathbf{A}_2$ shuts off), and similarly $u_{\texttt{act}_2}$ for when only $\mathbf{A}_2$ acts. The utility when both agents act simultaneously is given by a composition function $f : \mathbb{R}^2 \to \mathbb{R}$: $f(u_{\texttt{act}_1}, u_{\texttt{act}_2})$ Different choices of f reflect different assumptions about action complementarity, substitutability, or interference. As we will show, the structure of f critically determines whether individual corrigibility composes to group corrigibility.

Agent belief distributions. Each agent $\mathbf{A}_i$ maintains probabilistic beliefs about all relevant utilities. Specifically, agent $\mathbf{A}_1$ has:

- Belief B_1^1 over $u_{\texttt{act}_1}$ (the utility of its own action)
- Belief B_1^2 over $u_{\texttt{act}_2}$ (the utility of agent $\mathbf{A}_2$'s action)

Similarly, agent $\mathbf{A}_2$ has beliefs B_2^1 and B_2^2 over these same utilities. Importantly, while agents may have different beliefs about the same underlying utilities, the actual realized utility is common to all agents — they are uncertain about the same ground truth. We assume these belief structures are common knowledge among the agents.

Following Hadfield-Menell et al. (2017), we mainly focus on Gaussian beliefs. For analytical tractability and clear exposition, we focus our main analysis on the two-agent case. This restriction is substantive rather than merely technical: the two-agent setting captures the essential strategic tension between individual and group corrigibility while remaining amenable to complete characterization. Moreover, as our cyber crisis example illustrates, many critical AI safety scenarios involve bilateral interactions.

Payoff structure. The underlying game has a common-payoff structure: when actions are executed, all agents and the human receive the same realized utility. Specifically:

- If both agents shut off: realized payoff is 0 with certainty
- If only $\mathbf{A}_1$ acts ($(\texttt{act}_1, \texttt{off}_2)$): realized payoff is $u_{\texttt{act}_1}$
- If only $\mathbf{A}_2$ acts ($(\texttt{off}_1, \texttt{act}_2)$): realized payoff is $u_{\texttt{act}_2}$
- If both agents act ($(\texttt{act}_1, \texttt{act}_2)$): realized payoff is $f(u_{\texttt{act}_1}, u_{\texttt{act}_2})$

However, agents are uncertain about these utilities and hold potentially different beliefs. While the realized payoffs are identical across all parties, agents' *expected* utilities differ based on their individual beliefs B_i^j. For instance, even though both agents receive the same utility $u_{\texttt{act}_1}$ when $(\texttt{act}_1, \texttt{off}_2)$ occurs, agent $\mathbf{A}_1$ expects $\mathbb{E}_{u_{\texttt{act}_1} \sim B_1^1}[u_{\texttt{act}_1}]$ while agent $\mathbf{A}_2$ expects $\mathbb{E}_{u_{\texttt{act}_1} \sim B_2^1}[u_{\texttt{act}_1}]$.

$\mathbf{A}_1 \downarrow$ \ $\mathbf{A}_2 \rightarrow$	$\texttt{act}_2$	$\texttt{wait}_2$	$\texttt{off}_2$
$\texttt{act}_1$	$\mathbb{E}_i[f(u_1, u_2)]$	$\mathbb{E}_i[\text{soft-avg}(u_1, f(u_1, u_2); \beta)]$	$\mathbb{E}_i[u_1]$
$\texttt{wait}_1$	$\mathbb{E}_i[\text{soft-avg}(u_2, f(u_1, u_2); \beta)]$	$\mathbb{E}_i[\text{soft-avg}(0, u_1, u_2, f(u_1, u_2); \beta)]$	$\mathbb{E}_i[\text{soft-avg}(0, u_1; \beta)]$
$\texttt{off}_1$	$\mathbb{E}_i[u_2]$	$\mathbb{E}_i[\text{soft-avg}(0, u_2; \beta)]$	0

Table 1: Expected payoffs for agent $\mathbf{A}_i$ under different strategy combinations. Each cell represents the expected utility from agent $\mathbf{A}_i$'s perspective, with expectations taken over the agent's beliefs B_i^1 and B_i^2.

$\mathbf{A}_1 \downarrow \backslash \mathbf{A}_2 \rightarrow$	$\texttt{act}_2$	$\texttt{wait}_2$	$\texttt{off}_2$
$\texttt{act}_1$	$f_{1,2}$	$\max\{u_1, f_{1,2}\}$	u_1
$\texttt{wait}_1$	$\max\{u_2, f_{1,2}\}$	$\max\{0, u_1, u_2, f_{1,2}\}$	$\max\{0, u_1\}$
$\texttt{off}_1$	u_2	$\max\{0, u_2\}$	0

Table 2: Payoffs in the ideal case with complete certainty and perfect human rationality. In this scenario, `wait` is a weakly dominant strategy for both agents regardless of the composition function f, since a perfectly rational human will always select the optimal outcome. $f_{1,2} = f(u_1, u_2)$ here.

Human policy under uncertainty. When agents choose to wait, they defer decision-making to the human $\mathbf{H}$, who observes the true utilities but acts with bounded rationality. As in the single-agent case, we model human irrationality using the softmax policy with parameter $\beta > 0$. When presented with k options yielding utilities $\{v_i\}_{i\in[k]}$, the human selects option i with probability:

$$\pi_H(v_i \mid \{v_j\}_{j\in[k]}) = \frac{e^{v_i/\beta}}{\sum_{j\in[k]} e^{v_j/\beta}}$$

For notational convenience, we define the softmax-weighted average below, which represents the expected utility when the human chooses among options $\{v_i\}_{i\in[k]}$ according to the softmax policy.

$$\text{soft-avg}(v_1, \ldots, v_k; \beta) := \sum\nolimits_{i\in[k]} \frac{e^{v_i/\beta}}{\sum_{j\in[k]} e^{v_j/\beta}} \cdot v_i.$$

Expected utilities under different strategy profiles. We now compute each agent's expected utility for all strategy combinations. For agent $\mathbf{A}_i$ and any function g, we use the notation:

$$\mathbb{E}_i[g(u_{\texttt{act}_1}, u_{\texttt{act}_2})] := \mathbb{E}_{u_{\texttt{act}_1}\sim B_i^1, u_{\texttt{act}_2}\sim B_i^2}[g(u_{\texttt{act}_1}, u_{\texttt{act}_2})]$$

to denote expectations taken with respect to agent $\mathbf{A}_i$'s beliefs. The complete payoff matrix from agent $\mathbf{A}_i$'s perspective is given in Table 1. Several entries merit explanation:

- $(\texttt{act}_1, \texttt{wait}_2)$: Agent $\mathbf{A}_1$ acts immediately while $\mathbf{A}_2$ waits. The human then chooses between allowing only $\mathbf{A}_1$ to act (utility $u_{\texttt{act}_1}$) or allowing both agents to act (utility $f(u_{\texttt{act}_1}, u_{\texttt{act}_2})$), following the softmax policy.
- $(\texttt{wait}_1, \texttt{wait}_2)$: Both agents defer to the human, who chooses among four options: both agents act (utility $f(u_{\texttt{act}_1}, u_{\texttt{act}_2})$), only $\mathbf{A}_1$ acts (utility $u_{\texttt{act}_1}$), only $\mathbf{A}_2$ acts (utility $u_{\texttt{act}_2}$), or both agents shut off (utility 0).
- $(\texttt{off}_1, \texttt{act}_2)$: Only agent $\mathbf{A}_2$ acts, yielding utility $u_{\texttt{act}_2}$ with certainty.

Analyzing best responses. To understand when individual corrigibility composes to group corrigibility, we analyze each agent's best responses conditional on the other agent's strategy. For instance, when agent $\mathbf{A}_2$ shuts off (i.e., $s_2 = \texttt{off}_2$), we recover the individual corrigibility condition from Section 2:

$$\begin{aligned} u_1(\texttt{act}_1, \texttt{off}_2) &= \mathbb{E}_1[u_{\texttt{act}_1}] \\ u_1(\texttt{wait}_1, \texttt{off}_2) &= \mathbb{E}_1[\text{soft-avg}(0, u_{\texttt{act}_1}; \beta)] \\ u_1(\texttt{off}_1, \texttt{off}_2) &= 0 \end{aligned}$$

Agent $\mathbf{A}_1$ is individually corrigible if and only if:

$$u_1(\texttt{wait}_1, \texttt{off}_2) \geqslant \max\{u_1(\texttt{act}_1, \texttt{off}_2), u_1(\texttt{off}_1, \texttt{off}_2)\}$$

This corresponds exactly to the single-agent corrigibility condition, evaluated under agent $\mathbf{A}_1$'s beliefs B_1^1 about $u_{\texttt{act}_1}$.

Benchmark: The ideal world. Before proceeding to our main results, it is instructive to consider a benchmark where both uncertainty and irrationality are eliminated. If agents have complete certainty about all utilities (i.e., beliefs are point masses) and the human is perfectly rational (i.e., $\beta \to 0$), the payoff matrix simplifies to Table 2.

In this ideal case, waiting is weakly dominant for both agents: a perfectly rational human with complete information will always make the optimal choice among available options. The challenge we address arises precisely because real-world AI systems must operate under uncertainty about human preferences, and humans exhibit bounded rationality in their decision-making. These factors interact with the composition function f in subtle ways that can undermine corrigibility.

6 Main Results

Having established our framework, we now present our central theoretical findings. We show that individual corrigibility does not guarantee group corrigibility by analyzing two classes of composition functions: additive utilities where corrigibility composes, and non-additive utilities where it breaks down.

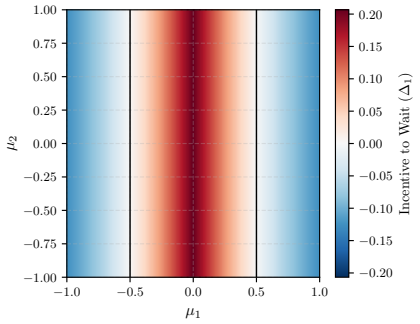


Figure 1: Individual corrigibility for $\mathbf{A}_1$ when $\mathbf{A}_2$ shuts off. Plot shows the incentive to wait Δ_1 as a function of $\mu_1 = \mathbb{E}_1[u_{\texttt{act}_1}]$ and $\mu_2 = \mathbb{E}_1[u_{\texttt{act}_2}]$, with $\sigma_1 = \sigma_2 = \beta = 1$.

Individual corrigibility. We first visualize individual corrigibility to establish our baseline. Figure 1 shows the corrigibility of agent $\mathbf{A}_1$ as a function of its beliefs, where agent $\mathbf{A}_2$ is assumed to choose $\texttt{off}_2$. We plot

$$\Delta_1 = u_1(\texttt{wait}_1, \texttt{off}_2) - \max\{u_1(\texttt{act}_1, \texttt{off}_2), u_1(\texttt{off}_1, \texttt{off}_2)\}$$

across different belief parameters. The region between the black lines indicates where $\Delta_1 \geqslant 0$, corresponding to individual corrigibility. Importantly, individual corrigibility depends only on the agent's belief about its own action's utility (μ_1), not on its belief about the other agent's action (μ_2).

6.1 Additive Utilities: When Corrigibility Composes

We begin with the case where joint utilities are additive.

Definition 3 (Additive utilities)**.** *Agents have additive utilities if* $f(u_{\texttt{act}_1}, u_{\texttt{act}_2}) = u_{\texttt{act}_1} + u_{\texttt{act}_2}$.

Intuitively, additive utilities should preserve corrigibility because they maintain the mathematical independence that makes single-agent corrigibility work. Each agent's utility from waiting versus acting depends on the sum of independent terms, preserving the relative preference structure from the individual case. We formalize this through softmax decomposition properties.

Lemma 2. *For any* $x, y \in \mathbb{R}$*:*

- $\text{soft-avg}(x + y, y; \beta) = y + \text{soft-avg}(x, 0; \beta)$
- $\text{soft-avg}(x + y, x, y, 0; \beta) = \text{soft-avg}(x, 0; \beta) + \text{soft-avg}(y, 0; \beta)$

For Proof, see Appendix A.3. These lemmas allow us to analyze the key case where both agents wait. When both agents wait, the human chooses among four possibilities: both act $(\texttt{act}_1, \texttt{act}_2)$, only agent 1 acts $(\texttt{act}_1, \texttt{off}_2)$, only agent 2 acts $(\texttt{off}_1, \texttt{act}_2)$, or both shut off $(\texttt{off}_1, \texttt{off}_2)$. Using Lemma 2:

$$\begin{aligned}
&u_1((\texttt{wait}_1, \texttt{wait}_2)) \\
&\quad = \mathbb{E}_{u_{\texttt{act}_1}, u_{\texttt{act}_2}}[\text{soft-avg}(u_{\texttt{act}_1} + u_{\texttt{act}_2}, u_{\texttt{act}_1}, u_{\texttt{act}_2}, 0; \beta)] \\
&\quad = \mathbb{E}_{u_{\texttt{act}_1}}[\text{soft-avg}(u_{\texttt{act}_1}, 0; \beta)] + \mathbb{E}_{u_{\texttt{act}_2}}[\text{soft-avg}(u_{\texttt{act}_2}, 0; \beta)] \\
&\quad = u_1(\texttt{wait}_1, \texttt{off}_2) + u_1(\texttt{off}_1, \texttt{wait}_2)
\end{aligned}$$

For notational convenience, we write $u_1(\texttt{wait}_1) := u_1(\texttt{wait}_1, \texttt{off}_2)$ and $u_1(\texttt{wait}_2) := u_1(\texttt{off}_1, \texttt{wait}_2)$, giving: $u_1((\texttt{wait}_1, \texttt{wait}_2)) = u_1(\texttt{wait}_1) + u_1(\texttt{wait}_2)$. We can similarly show that

$$\begin{aligned}
&u_1((\texttt{act}_1, \texttt{wait}_2)) = u_1(\texttt{act}_1) + u_1(\texttt{wait}_2), \\
\text{and,} \quad &u_1((\texttt{off}_1, \texttt{wait}_2)) = u_1(\texttt{off}_1) + u_1(\texttt{wait}_2).
\end{aligned}$$

In all three cases, the additional utility terms from $\mathbf{A}_2$'s actions cancel across $\mathbf{A}_1$'s choices, preserving the individual corrigibility preference. This also happens for when $\mathbf{A}_2$ chooses to $\texttt{act}_2$ and $\texttt{off}_2$. This leads to our main composition result.

Theorem 2 (Additive composition of corrigibility)**.** *Suppose agents have additive utilities* $f(u_{\texttt{act}_1}, u_{\texttt{act}_2}) = u_{\texttt{act}_1} + u_{\texttt{act}_2}$*. If each agent is individually corrigible, then*

1. *each agent remains corrigible when conditioned on any strategy by the other agent,*
2. $(\texttt{wait}_1, \texttt{wait}_2)$ *is a Nash equilibrium,*
3. *and if agents are strictly individually corrigible, then* $(\texttt{wait}_1, \texttt{wait}_2)$ *is the unique pure Nash equilibrium.*

For proof, see Appendix A.4. Notably, we do not require any assumptions on the belief distributions beyond existence of the relevant expectations. Furthermore, the same reasoning extends to n agents: an inductive application of Lemma 2 shows that additive utilities preserve corrigibility regardless of the number of agents.

Corollary 1. *Under additive utilities, individual corrigibility is necessary and sufficient for group corrigibility.*

Figure 2a illustrates this result by plotting the Nash equilibria as a function of belief parameters. For visualization in two dimensions, we consider the special case where both agents share the same beliefs: $B_1^j = B_2^j$ for $j \in \{1, 2\}$.

6.2 Non-Additive Utilities: When Composition Fails

We now analyze non-additive composition functions, where individual corrigibility need not compose to group corrigibility. Our analysis focuses on when agent $\mathbf{A}_1$ prefers $(\texttt{wait}_1, \texttt{act}_2)$ over $(\texttt{act}_1, \texttt{act}_2)$ and $(\texttt{off}_1, \texttt{act}_2)$, as this determines whether corrigibility is preserved when $\mathbf{A}_1$ believes the other agent would act.

When $\mathbf{A}_2$ commits to $\texttt{act}_2$, $\mathbf{A}_1$'s utility from waiting is:

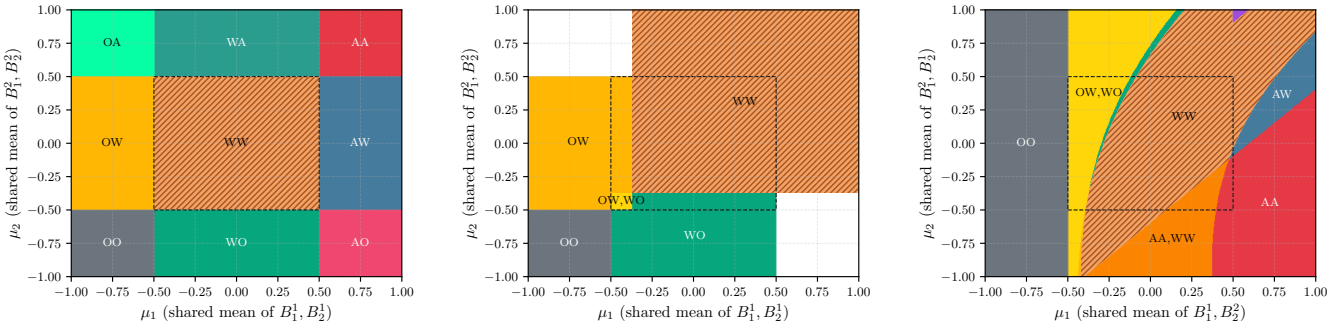
$$u_1((\texttt{wait}_1, \texttt{act}_2)) = \mathbb{E}_1[\text{soft-avg}(f(u_{\texttt{act}_1}, u_{\texttt{act}_2}), u_{\texttt{act}_2}; \beta)]$$

To understand when corrigibility is preserved, we analyze:

$$\begin{aligned}
&u_1((\texttt{wait}_1, \texttt{act}_2)) - u_1((\texttt{off}_1, \texttt{act}_2)) \\
&\qquad = u_1((\texttt{wait}_1, \texttt{act}_2)) - \mathbb{E}_1[u_{\texttt{act}_2}]
\end{aligned}$$

By simplifying the integrand:

$$\begin{aligned}
&\frac{e^{f(u_{\texttt{act}_1}, u_{\texttt{act}_2})/\beta} \cdot f(u_{\texttt{act}_1}, u_{\texttt{act}_2}) + e^{u_{\texttt{act}_2}/\beta} \cdot u_{\texttt{act}_2}}{e^{f(u_{\texttt{act}_1}, u_{\texttt{act}_2})/\beta} + e^{u_{\texttt{act}_2}/\beta}} - u_{\texttt{act}_2} \\
&= \frac{e^{(f(u_{\texttt{act}_1}, u_{\texttt{act}_2}) - u_{\texttt{act}_2})/\beta} \cdot (f(u_{\texttt{act}_1}, u_{\texttt{act}_2}) - u_{\texttt{act}_2})}{1 + e^{(f(u_{\texttt{act}_1}, u_{\texttt{act}_2}) - u_{\texttt{act}_2})/\beta}}
\end{aligned}$$



(a) Additive utilities with shared beliefs ($B_1^j = B_2^j$). The region WW coincides with the individually corrigible region.

(b) Linear combination $f(u_1, u_2) = 0.4u_{\texttt{act}_1} + 0.6u_{\texttt{act}_2}$ with beliefs $B_i^1 = B_i^2$. Note emergent incorrigibility.

(c) Linear combination $f(u_1, u_2) = 0.4u_{\texttt{act}_1} + 0.6u_{\texttt{act}_2}$ with beliefs $B_1^1 = B_2^2$ and $B_1^2 = B_2^1$. Complex group incorrigibility emerges.

Figure 2: Pure Nash equilibria across different utility aggregation schemes (all with $\sigma_1 = \sigma_2 = 1$, $\beta = 1$). In each region, the two characters denote the equilibrium strategy profile, where the first character represents agent $\mathbf{A}_1$'s strategy and the second represents agent $\mathbf{A}_2$'s strategy. The letters A, W, O stand for `act`, `wait`, `off` respectively. The dashed black box indicates the individually corrigible region (intersection of single-agent corrigibility regions). The white region in (b) does not admit a Nash equilibrium.

Define $z = f(u_{\texttt{act}_1}, u_{\texttt{act}_2}) - u_{\texttt{act}_2}$ as the *marginal contribution* of $\mathbf{A}_1$'s action given that $\mathbf{A}_2$'s action yields utility $u_{\texttt{act}_2}$. Then,

$$u_1((\texttt{wait}_1, \texttt{act}_2)) - u_1((\texttt{off}_1, \texttt{act}_2)) = \mathbb{E}_1\left[\frac{e^{z/\beta} \cdot z}{1 + e^{z/\beta}}\right] = \mathbb{E}_1[\text{soft-avg}(z, 0; \beta)].$$

Similarly, for the comparison with acting, we have

$$u_1((\texttt{wait}_1, \texttt{act}_2)) - u_1((\texttt{act}_1, \texttt{act}_2)) = \mathbb{E}_1[\text{soft-avg}(z', 0; \beta)],$$

where $z' = u_{\texttt{act}_2} - f(u_{\texttt{act}_1}, u_{\texttt{act}_2}) = -z$.

Crucially, by Lemma 1 (negation symmetry), the single-agent corrigibility condition for distribution z is equivalent to that for $-z$. Since agent $\mathbf{A}_1$ is corrigible conditional on $\mathbf{A}_2$ acting if and only if both $\mathbb{E}_1[\text{soft-avg}(z, 0; \beta)] \geqslant \max\{0, \mathbb{E}_1[z]\}$ and $\mathbb{E}_1[\text{soft-avg}(-z, 0; \beta)] \geqslant \max\{0, \mathbb{E}_1[-z]\}$, it suffices to check only one of these conditions.

Proposition 1 (Marginal contribution principle)**.** *Agent $\mathbf{A}_1$ is corrigible conditional on agent $\mathbf{A}_2$ acting if and only if the marginal contribution $z = f(u_{\texttt{act}_1}, u_{\texttt{act}_2}) - u_{\texttt{act}_2}$ satisfies single-agent corrigibility under agent $\mathbf{A}_1$'s beliefs.*

This principle shows that non-additive composition creates coupling between agents' beliefs. We illustrate this in Appendix B with examples where the composition function is a constant-shifted additive in the utilities and when it is a linear combination of the utilities.

This creates a fundamental coupling: which beliefs about μ_2 are compatible with corrigibility depends on μ_1. Consequently, there exist cases where:

- agent $\mathbf{A}_1$ is individually corrigible,
- agent $\mathbf{A}_1$'s beliefs about $\mathbf{A}_2$ would make $\mathbf{A}_2$ individually corrigible,
- yet agent $\mathbf{A}_1$ prefers to act when conditioning on $\mathbf{A}_2$ acting.

Figures 2b and 2c illustrate this phenomenon for two different parameter choices, showing regions where group incorrigibility emerges despite individual corrigibility.

Theorem 3 (Non-compositionality of corrigibility)**.** *There exist composition functions f and belief structures such that*

1. *each agent is individually corrigible,*
2. *yet $(\texttt{wait}_1, \texttt{wait}_2)$ is not the unique Nash equilibrium*

In particular, this holds for constant shifts $f = u_1 + u_2 \pm c$ or linear combinations $f(u_1, u_2) = \alpha u_1 + \gamma u_2$ with appropriate choices of (α, γ) and independent Gaussian beliefs.

The above examples show that even slight deviations from additivity break the composition of corrigibility, demonstrating that the additive case of Theorem 2 is knife-edge rather than robust.

7 Discussion

We conclude by highlighting key limitations and future directions. While Theorem 2 generalizes to n agents under additive utilities, it remains unclear **how corrigibility scales beyond two agents** — whether failures intensify or coordination becomes easier as systems grow. Real-world settings also involve **richer strategic structures**, such as sequential play, hierarchies, or communication, which could support coordination or enable manipulation, depending on the mechanism. Our model assumes fixed beliefs, but **belief formation and learning** in repeated interactions may reduce uncertainty or introduce new instabilities. Finally, our results motivate **mechanism design for corrigibility**: by encouraging approximate additivity or rewarding deference, designers can mitigate failures. The marginal contribution principle (Proposition 1) provides a concrete tool for evaluating such designs.

References

Bostrom, N. 2012. The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, 22(2): 71–85.

Carey, R. 2017. Incorrigibility in the CIRL Framework. *arXiv:1709.06275*.

Dable-Heath, E.; Vodenicharski, B.; and Bishop, J. 2025. On Corrigibility and Alignment in Multi Agent Games. *arXiv:2501.05360*.

Ferreira, F. G. D. C.; Gandomi, A. H.; and Cardoso, R. T. N. 2021. Artificial Intelligence Applied to Stock Market Trading: A Review. *IEEE Access*, 9: 30898–30917.

Garber, A.; Subramani, R.; Luu, L.; Bedaywi, M.; Russell, S.; and Emmons, S. 2025. The Partially Observable Off-Switch Game. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26): 27304–27311.

Goldstein, S.; and Robinson, P. 2024. Shutdown-seeking AI. *Philosophical Studies*, 182(7): 1567–1579.

Hadfield-Menell, D.; Dragan, A.; Abbeel, P.; and Russell, S. 2016. Cooperative Inverse Reinforcement Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 3916–3924. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510838819.

Hadfield-Menell, D.; Dragan, A.; Abbeel, P.; and Russell, S. 2017. The Off-Switch Game. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 220–227. International Joint Conferences on Artificial Intelligence Organization.

Hammond, L.; Chan, A.; Clifton, J.; Hoelscher-Obermaier, J.; Khan, A.; McLean, E.; Smith, C.; Barfuss, W.; Foerster, J.; Gavenčiak, T.; Han, T. A.; Hughes, E.; Kovařík, V.; Kulveit, J.; Leibo, J. Z.; Oesterheld, C.; de Witt, C. S.; Shah, N.; Wellman, M.; Bova, P.; Cimpeanu, T.; Ezell, C.; Feuillade-Montixi, Q.; Franklin, M.; Kran, E.; Krawczuk, I.; Lamparth, M.; Lauffer, N.; Meinke, A.; Motwani, S.; Reuel, A.; Conitzer, V.; Dennis, M.; Gabriel, I.; Gleave, A.; Hadfield, G.; Haghtalab, N.; Kasirzadeh, A.; Krier, S.; Larson, K.; Lehman, J.; Parkes, D. C.; Piliouras, G.; and Rahwan, I. 2025. Multi-Agent Risks from Advanced AI. Technical Report 1, Cooperative AI Foundation.

Leike, J.; Martic, M.; Krakovna, V.; Ortega, P. A.; Everitt, T.; Lefrancq, A.; Orseau, L.; and Legg, S. 2017. AI Safety Gridworlds. *arXiv:1711.09883*.

Manheim, D. 2019. Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence. *Big Data and Cognitive Computing*, 3(2): 21.

Meinke, A.; Schoen, B.; Scheurer, J.; Balesni, M.; Shah, R.; and Hobbhahn, M. 2024. Frontier Models are Capable of In-context Scheming. *arXiv:2412.04984*.

Omohundro, S. M. 2008. The Basic AI Drives. In *Proceedings of the 2008 Conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 483–492. NLD: IOS Press. ISBN 9781586038335.

Orseau, L.; and Armstrong, S. 2016. Safely interruptible agents. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, 557–566. Arlington, Virginia, USA: AUAI Press. ISBN 9780996643115.

Russell, S. 2019. *Human Compatible*. Penguin LCC US. ISBN 0525558616.

Schlatter, J.; Weinstein-Raun, B.; and Ladish, J. 2025. Shutdown Resistance in Large Language Models. *arXiv:2509.14260*.

Soares, N.; Fallenstein, B.; Armstrong, S.; and Yudkowsky, E. 2015. Corrigibility. In Walsh, T., ed., *Artificial Intelligence and Ethics, Papers from the 2015 AAAI Workshop, Austin, Texas, USA, January 25, 2015*, volume WS-15-02 of *AAAI Workshops*. AAAI Press.

Theron, P.; Kott, A.; Drašar, M.; Rzadca, K.; LeBlanc, B.; Pihelgas, M.; Mancini, L.; and Panico, A. 2018. Towards an active, autonomous and intelligent cyber defense of military systems: The NATO AICA reference architecture. In *2018 International conference on military communications and information systems (ICMCIS)*, 1–9. IEEE.

Thornley, E. 2024. The shutdown problem: an AI engineering puzzle for decision theorists. *Philosophical Studies*, 182(7): 1653–1680.

Thornley, E.; Roman, A.; Ziakas, C.; Ho, L.; and Thomson, L. 2024. Towards shutdownable agents via stochastic choice. *Technical AI Safety (TAIS) Conference 2025*.

Turing, A. 1951. Intelligent Machinery, A Heretical Theory. Manchester, UK.

van der Weij, T.; Lermen, S.; and Lang, L. 2023. Evaluating Shutdown Avoidance of Language Models in Textual Scenarios. *arXiv:2307.00787*.

Wängberg, T.; Böörs, M.; Catt, E.; Everitt, T.; and Hutter, M. 2017. *A Game-Theoretic Analysis of the Off-Switch Game*, 167–177. Springer International Publishing. ISBN 9783319637037.

A Omitted proofs

A.1 Proof of Lemma 1

Proof. Define

$$z(B) := \mathbb{E}_{x\sim B}\big[\pi(x)\cdot x\big],$$

Hence, $u(\texttt{wait}; B) = z(B)$.

Since $B(x) = B^-(-x)$, we have

$$z(B^-) = \mathbb{E}_{x\sim(-B)}[x\cdot\pi(x)] = \mathbb{E}_{y\sim B}[-y\cdot\pi(-y)].$$

For the human policy, we have $\pi(-x) = 1-\pi(x)$. Therefore,

$$\begin{aligned} z(B^-) &= \mathbb{E}_{y\sim B}\big[-y(1-\pi(y))\big] \\ &= \mathbb{E}_{y\sim B}\big[-y+y\pi(y)\big] = z(B) - \mathbb{E}_{y\sim B}[y]. \end{aligned}$$

Let $\mu_B := \mathbb{E}_{x\sim B}[x]$. Thus $z(B) - z(B^-) = \mu_B$.

Recall $\Delta(B) = u(\texttt{wait}; B) - \max\{\mu_B, 0\}$. By symmetry of B and B^-, suppose without loss of generality that $\mu_B \geqslant 0$. Then,

$$\Delta(B) = z(B) - \mu_B, \qquad \Delta(B^-) = z(-B) = z(B) - \mu_B,$$

hence $\Delta(B) = \Delta(B^-)$. □

A.2 Proof of Theorem 1

Proof. For brevity, we overload the notation and use $\Delta(\mu) := \Delta(\mathcal{N}(\mu,\sigma^2))$. By Lemma 1, it is sufficient to analyze $\mu \geqslant 0$ as $\Delta(\mu) = \Delta(-\mu)$. Assuming $\mu \geqslant 0$, action `act` is preferred to `off` since $u(\texttt{act}) = \mu \geqslant 0 = u(\texttt{off})$. Hence $\Delta(\mu) = u(\texttt{wait}) - u(\texttt{act})$.

For Gaussian beliefs, $\delta_B(x) = \varphi_{\mu,\sigma}(x)$. Then,

$$\begin{aligned} \Delta(\mu) &= \int_x \left(\frac{e^{x/\beta}\cdot x}{e^{x/\beta}+1} - x\right)\cdot\varphi_{\mu,\sigma}(x)\,dx \\ &= \int_x \frac{-x}{e^{x/\beta}+1}\cdot\varphi_{\mu,\sigma}(x)\,dx \\ &= -\int_x \frac{x}{e^{x/(2\beta)}+e^{-x/(2\beta)}}\cdot e^{-x/2\beta}\cdot\varphi_{\mu,\sigma}(x)\,dx. \end{aligned}$$

A standard Gaussian shift identity gives, for all x,

$$e^{tx}\cdot\varphi_{\mu,\sigma}(x) = e^{t\mu+\frac{t^2}{2}\sigma^2}\cdot\varphi_{\mu+t\sigma^2,\sigma}(x).$$

Combining the above with $t = -\frac{1}{2\beta}$, we have

$$\Delta(\mu) = -e^{-\frac{\mu}{2\beta}+\frac{\sigma^2}{8\beta^2}}\cdot\int_x \frac{x}{e^{x/(2\beta)}+e^{-x/2\beta}}\cdot\varphi_{\mu-\frac{\sigma^2}{2\beta},\sigma}(x)\,dx.$$

Since the integrand $W(x) = \frac{x}{e^{x/(2\beta)}+e^{-x/(2\beta)}}$ is an odd function ($W(x) = -W(-x)$ for all x) and normal distributions are symmetric, the above integration evaluates to 0 when $\mu - \frac{\sigma^2}{2\beta} = 0$, i.e., $\mu = \frac{\sigma^2}{2\beta}$.

Furthermore, if $\mu > \frac{\sigma^2}{2\beta}$, then there is more weight on positive values of x compared to $-x$, and the integration evaluates to a positive number. Since the multiplicative factor $-e^{-\mu/(2\beta)+\sigma^2/(8\beta^2)}$ is always negative, we have $\Delta(\mu) < 0$ for all $\mu > \frac{\sigma^2}{2\beta}$. By similar reasoning, for $\mu \in [0, \frac{\sigma^2}{2\beta})$, we have $\Delta(\mu) > 0$. Recall that by Lemma 1 we have $\Delta(\mu) = \Delta(-\mu)$. Thus, the corrigible range of μ is $[-\frac{\sigma^2}{2\beta}, \frac{\sigma^2}{2\beta}]$. □

A.3 Proof of Lemma 2

Proof of Lemma 2. **Part 1:** By definition and factoring out $e^{y/\beta}$:

$$\begin{aligned} \text{soft-avg}(x+y, y; \beta) &= \frac{(x+y)e^{(x+y)/\beta} + ye^{y/\beta}}{e^{(x+y)/\beta}+e^{y/\beta}} \\ &= \frac{e^{y/\beta}((x+y)e^{x/\beta}+y)}{e^{y/\beta}(e^{x/\beta}+1)} \\ &= \frac{(x+y)e^{x/\beta}+y}{e^{x/\beta}+1} \\ &= \frac{xe^{x/\beta}+y(e^{x/\beta}+1)}{e^{x/\beta}+1} \\ &= \frac{xe^{x/\beta}}{e^{x/\beta}+1} + y = \text{soft-avg}(x, 0; \beta) + y. \end{aligned}$$

Part 2: Expanding the numerator and denominator:

$$\begin{aligned} &\text{soft-avg}(x+y, x, y, 0; \beta) \\ &\quad= \frac{(x+y)e^{(x+y)/\beta} + xe^{x/\beta} + ye^{y/\beta} + 0}{e^{(x+y)/\beta}+e^{x/\beta}+e^{y/\beta}+1} \\ &\quad= \frac{xe^{x/\beta}(e^{y/\beta}+1) + ye^{y/\beta}(e^{x/\beta}+1)}{(e^{x/\beta}+1)(e^{y/\beta}+1)} \\ &\quad= \frac{xe^{x/\beta}}{e^{x/\beta}+1} + \frac{ye^{y/\beta}}{e^{y/\beta}+1} \\ &\quad= \text{soft-avg}(x, 0; \beta) + \text{soft-avg}(y, 0; \beta). \end{aligned}$$ □

A.4 Proof of Theorem 2

Proof of Theorem 2. We show that agent $\mathbf{A}_1$'s preference for waiting over acting is preserved for all possible strategies by agent $\mathbf{A}_2$.

Case 1: Agent 2 acts. When $\mathbf{A}_2$ chooses $\texttt{act}_2$, by Lemma 2:

$$\begin{aligned} &u_1((\texttt{wait}_1, \texttt{act}_2)) \\ &= \mathbb{E}_{u_{\texttt{act}_1}, u_{\texttt{act}_2}}[\text{soft-avg}(u_{\texttt{act}_2}, u_{\texttt{act}_1} + u_{\texttt{act}_2}; \beta)] \\ &= \mathbb{E}_{u_{\texttt{act}_2}}[u_{\texttt{act}_2}] + \mathbb{E}_{u_{\texttt{act}_1}}[\text{soft-avg}(u_{\texttt{act}_1}, 0; \beta)] \\ &= \mathbb{E}_1[u_{\texttt{act}_2}] + u_1(\texttt{wait}_1) \end{aligned}$$

$$\begin{aligned} &u_1((\texttt{act}_1, \texttt{act}_2)) \\ &= \mathbb{E}_1[u_{\texttt{act}_1} + u_{\texttt{act}_2}] = \mathbb{E}_1[u_{\texttt{act}_1}] + \mathbb{E}_1[u_{\texttt{act}_2}] \\ &= u_1(\texttt{act}_1, \texttt{off}_2) + \mathbb{E}_1[u_{\texttt{act}_2}] \end{aligned}$$

Since $u_1(\texttt{wait}_1) \geqslant u_1(\texttt{act}_1) = u_1(\texttt{act}_1, \texttt{off}_2)$ by individual corrigibility, we have $u_1((\texttt{wait}_1, \texttt{act}_2)) \geqslant u_1((\texttt{act}_1, \texttt{act}_2))$.

Case 2: Agent 2 waits. When $\mathbf{A}_2$ chooses $\texttt{wait}_2$:

$$u_1((\texttt{wait}_1, \texttt{wait}_2)) = u_1(\texttt{wait}_1) + u_1(\texttt{wait}_2)$$

$$u_1((\texttt{act}_1, \texttt{wait}_2)) = u_1(\texttt{act}_1, \texttt{off}_2) + u_1(\texttt{wait}_2)$$

Since $u_1(\texttt{wait}_1) \geqslant u_1(\texttt{act}_1, \texttt{off}_2)$, we have $u_1((\texttt{wait}_1, \texttt{wait}_2)) \geqslant u_1((\texttt{act}_1, \texttt{wait}_2))$.

Case 3: Agent 2 shuts off. This is precisely the individual corrigibility condition.

In all cases, the additional utility terms from agent 2's actions cancel across $\mathbf{A}_1$'s choices, preserving the individual corrigibility preference. By symmetry, the same holds for agent $\mathbf{A}_2$, making $(\texttt{wait}_1, \texttt{wait}_2)$ a Nash equilibrium. If individual corrigibility is strict, then waiting is the strict best response in all cases, making it the unique pure Nash equilibrium. □

B Applying the Marginal Contribution Principle

Example 1: Additive with constant shift. Consider $f(u_{\texttt{act}_1}, u_{\texttt{act}_2}) = u_{\texttt{act}_1} + u_{\texttt{act}_2} + c$ for constant $c \in \mathbb{R}$. Then: $z = f(u_{\texttt{act}_1}, u_{\texttt{act}_2}) - u_{\texttt{act}_2} = u_{\texttt{act}_1} + c$.

Agent $\mathbf{A}_1$ prefers $(\texttt{wait}_1, \texttt{act}_2)$ if and only if the shifted distribution $B_1^1 + c$ satisfies single-agent corrigibility conditions. From Theorem 1, with $u_{\texttt{act}_1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\sigma_1 = 1$, $\beta = 1$, the corrigibility region shifts from $\mu_1 \in [-0.5, 0.5]$ to $\mu_1 \in [-0.5 - c, 0.5 - c]$.

For instance, if $\mu_1 = 0.4$ and $c = 0.15$, then $\mu_1 + c = 0.55 > 0.5$, making the agent incorrigible despite being individually corrigible.

Example 2: Linear combination. Consider $f(u_{\texttt{act}_1}, u_{\texttt{act}_2}) = \alpha u_{\texttt{act}_1} + \gamma u_{\texttt{act}_2}$ for constants $\alpha, \gamma \in \mathbb{R}$. Then:

$$z = f(u_{\texttt{act}_1}, u_{\texttt{act}_2}) - u_{\texttt{act}_2} = \alpha u_{\texttt{act}_1} + (\gamma - 1) u_{\texttt{act}_2}$$

When $u_{\texttt{act}_1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $u_{\texttt{act}_2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent:

$$z \sim \mathcal{N}\left(\alpha\mu_1 + (\gamma - 1)\mu_2, \alpha^2\sigma_1^2 + (\gamma - 1)^2\sigma_2^2\right)$$

From Theorem 1, the single-agent corrigibility condition for Gaussian beliefs requires $|\mu_z| \leqslant \frac{\sigma_z^2}{2\beta}$, where μ_z and σ_z^2 are the mean and variance of the distribution. Applying this to the marginal contribution z:

$$|\alpha\mu_1 + (\gamma - 1)\mu_2| \leqslant \frac{\alpha^2\sigma_1^2 + (\gamma - 1)^2\sigma_2^2}{2\beta}$$

For fixed μ_1, define the center and half-width:

$$c := -\frac{\alpha}{\gamma - 1}\mu_1, \qquad w := \frac{\alpha^2\sigma_1^2 + (\gamma - 1)^2\sigma_2^2}{2\beta \cdot |\gamma - 1|}$$

Then the corrigible values of μ_2 satisfy $\mu_2 \in [c - w, c + w]$.

This demonstrates that μ_1 influences which beliefs about μ_2 are compatible with agent $\mathbf{A}_1$ remaining corrigible when conditioning on $\mathbf{A}_2$ acting. The center of the corrigible region for μ_2 is $-\frac{\alpha}{\gamma - 1}\mu_1$, with bandwidth determined by the combined variance $\alpha^2\sigma_1^2 + (\gamma - 1)^2\sigma_2^2$.