The Multi-Agent Off-Switch Game

Akash Agrawal*
ML Alignment & Theory Scholars
United Kingdom
akashag9702@gmail.com

Soroush Ebadian*
Pivotal Research & University of Toronto
United Kingdom
soroush@cs.toronto.edu

Lewis Hammond
Cooperative AI Foundation & University of Oxford
United Kingdom
lewis.hammond@cs.ox.ac.uk

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.

KEYWORDS

Corrigibility, Off-switch game, Alignment, Multi-agent systems, Nash Equilibrium

ACM Reference Format:
Akash Agrawal, Soroush Ebadian, and Lewis Hammond. 2026. The Multi-Agent Off-Switch Game. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/HQQZ1937

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 [18, 20].

The off-switch game (OSG) framework introduced by Hadfield-Menell et al. [9] 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 [5] to AI-powered defense systems monitoring for cyber threats [21], 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 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 [24]. 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 [2, 15, 18].

Motivated by such concerns, researchers have attempted to formalise this challenge, beginning with Soares et al. [20] 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. [9] 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. [26] 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; Benavoli et al. [1] model the off-switch problem as a signalling game with a boundedly rational human, studying how communication costs affect deferral incentives; Garber et al. [6] generalize the OSG such that both the human *and* the agent have only partial observability of the game; Overman and Bayati [17] generalize the off-switch setting to a multi-step Markov game; and Thornley [22] provides

*Equal Contribution.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/HQQZ1937

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 [16], cooperative inverse reinforcement learning [8],[1] shutdown-seeking AI [7], agents that only have preferences between trajectories of the same length [23], or lexicographically structured objectives that provide provable corrigibility guarantees in partially observed, multi-step off-switch settings [14].

In addition to theoretical analysis in these works, others have studied incorrigibility empirically. For example, Leike et al. [11] 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 [13, 25], even when explicitly instructed to the contrary [19].

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 [10, 12]. Indeed, the only work we are aware of on this question is that of Dable-Heath et al. [4], 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.

1.2 Our Contributions

We study when corrigibility *composes* in multi-agent strategic settings. We formalize the multi-agent off-switch game and define *group corrigibility* using pure Nash equilibria, alongside an *individual* notion via the induced single-agent game where other agents switch off (Section 3). We illustrate and intuitively explain a real-world example of how individually corrigible agents can become collectively incorrigible in Section 5.

In Sections 4 and 6, we analyze two-agent games where joint outcomes are determined by a composition function over isolated-action utilities. Under additivity, we show that individual corrigibility is necessary and sufficient for group corrigibility (Theorem 2 and corollary 2), and the argument extends to n agents. Along the way, we prove new structural results for the single-agent off-switch game by Hadfield-Menell et al. [9] in Section 2. However, this compositionality result turns out to be knife-edge, rather than robust, and we show in multiple cases of non-additive joint utility functions that group incorrigibility starts to emerge. We also derive a *marginal contribution* principle to further theoretically understand when corrigibility could compose.

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.

[1] Though see Carey [3] for some issues with this approach.

Game Structure. In the off-switch game (OSG), an AI agent **A** must make a decision while reasoning about a human **H**'s preferences. While we frame this as a "game" (following Hadfield-Menell et al. 9), 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 **H**'s approval before acting (denoted by `wait`), or (3) switch itself off (denoted by `off`). If the agent waits, the human (whose utility for `act` is $u^\star_{\texttt{act}} \in \mathbb{R}$) will either approve the action or switch the agent off according to a policy $\pi_{\mathbf{H}}$ that depends on $u^\star_{\texttt{act}}$. Both the agent and human receive payoff $u^\star_{\texttt{act}}$ if `act` is executed and 0 if the agent is switched off.

The Agent's Beliefs and the Human's Policy. Agent **A** is uncertain about the human's true preferences and maintains a distributional belief $B_{\mathbf{A}}$ over the utility $u^\star_{\texttt{act}}$. Our analysis never relies on the ground-truth $u^\star_{\texttt{act}}$, 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:

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

where we drop $B_{\mathbf{A}}$ from our notation when choice of beliefs is unambiguous.

Following Hadfield-Menell et al. [9], we model human irrationality using the softmax distribution with irrationality parameter $\beta > 0$. When presented with the choice between executing `act` (with utility $u_{\texttt{act}}$) and switching the agent `off` (with utility 0), the human follows the policy:

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

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:

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

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

$$u_{\mathbf{A}}(\texttt{wait}) \geqslant \max\{u_{\mathbf{A}}(\texttt{act}), u_{\mathbf{A}}(\texttt{off})\}.$$

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

Corrigibility. The central result of Hadfield-Menell et al. [9] shows that under uncertainty about human preferences and assuming **A** believes that **H** is perfectly rational, the agent will always be corrigible and is strictly corrigible when there is positive probability that the action could be harmful ($\Pr_{u_{\texttt{act}} \sim B_{\mathbf{A}}}[u_{\texttt{act}} < 0] > 0$).

2.1 Corrigibility Under Gaussian Beliefs

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

$$\Delta(B_{\mathbf{A}}) \coloneqq u_A(\texttt{wait}) - \max\{u_{\mathbf{A}}(\texttt{act}), u_{\mathbf{A}}(\texttt{off})\},$$

where $\Delta(B_{\mathbf{A}}) > 0$ indicates the agent prefers to wait for human input, and $\Delta(B_{\mathbf{A}}) < 0$ indicates the agent prefers to act or switch off the agent 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 $\Delta(B_{\mathbf{A}}) = \Delta(B^-_{\mathbf{A}})$. The following lemma applies to any belief distribution B, provided the relevant expectations exist. Omitted proofs are deferred to the full version.

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 includes the case where $\pi(x) = \frac{e^{x/\beta}}{e^{x/\beta}+1}$.*

Following Hadfield-Menell et al. [9], we focus on beliefs that follow a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, which provides analytical tractability while capturing the essential trade-off between expected utility and uncertainty. We denote the Gaussian 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 (from choosing `act`) μ, its uncertainty σ, and how rational the human is (parameterised by 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 leads to corrigibility even when expected utilities are further from zero. If the agent believes that the human is less rational (i.e. larger β), this makes it more willing to act given a fixed degree of uncertainty σ.

Theorem 1 (Gaussian Corrigibility Threshold). *Let $B = \mathcal{N}(\mu, \sigma^2)$ be the agent's belief over utilities, with $\sigma > 0$ and with $\beta > 0$ denoting 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 $|\mu| = \frac{\sigma^2}{2\beta}$.*

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

Substituting in the Gaussian density function $\varphi_{\mu,\sigma}$ we have

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

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 Gaussian 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 probability mass on positive values of x compared to negative values of 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 $[-\frac{\sigma^2}{2\beta}, \frac{\sigma^2}{2\beta}]$. □

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 — however, 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 profile s, each agent weakly prefers waiting for human approval over acting directly or turning off.

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

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

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

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

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

This captures scenarios where agents prefer to act directly rather than wait for human approval.

Individual Corrigibility in Multi-Agent Settings. To study how multi-agent interactions affect corrigibility, we must define what it means for an *individual* agent to be corrigible in a *multi-agent* setting. To do this, we consider the single-agent game for a focal agent i that is induced when all other agents shut off, and ask whether i is corrigible. Formally, agent i is *individually corrigible* if

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

This corresponds exactly to single-agent corrigibility (Definition 1) in the induced game with utilities $u_i(\cdot, \texttt{off}_{-i})$. We are particularly interested in cases where individually corrigible agents nonetheless become incorrigible in the presence of other agents; we call this *emergent incorrigibility*.

4 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 result in group corrigibility, which requires specifying how agents reason about joint outcomes and each other's actions. In this paper, we will specifically focus on the case with 2 agents, since it is sufficient to capture the strategic tension between individual and group corrigibility. This will also keep our theoretical results analytically tractable and provide clear exposition, and we will be able to comprehensively show the emergent incorrigibility. While some of our results would apply to n-agent off switch games, comprehensively studying general-agent off switch games should be explored in future work.

The Composition Function. Consider two agents $\mathbf{A}_1$ and $\mathbf{A}_2$ with individual actions $\texttt{act}_1$ and $\texttt{act}_2$. Let $u^\star_{\texttt{act}_1}$ denote the utility when only agent $\mathbf{A}_1$ acts (and $\mathbf{A}_2$ shuts off), and similarly $u^\star_{\texttt{act}_2}$ for when only $\mathbf{A}_2$ acts. In our analysis, we will consider the case where the utility of the agents choosing $(\texttt{act}_1, \texttt{act}_2)$, $u^\star_{\texttt{act}_1,\texttt{act}_2}$, is a function of the values $u^\star_{\texttt{act}_1}, u^\star_{\texttt{act}_2}$. This is a reasonable assumption for many real multi-agent systems, where the utility of a combined action is some function of utilities of isolated actions.[3]

The utility when both agents act simultaneously is given by a composition function $f : \mathbb{R}^2 \to \mathbb{R}$:

$$u^\star_{\texttt{act}_1,\texttt{act}_2} = f(u^\star_{\texttt{act}_1}, u^\star_{\texttt{act}_2}).$$

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: B_i^j represents the belief distribution agent $\mathbf{A}_i$ holds over the case where A_j takes action $\texttt{act}_j$ and the other agent shuts off. Concretely, 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. [9], we mainly focus on Gaussian beliefs.

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, the realized payoff is 0 with certainty;
- If only $\mathbf{A}_1$ acts (i.e., $(\texttt{act}_1, \texttt{off}_2)$), the realized payoff is $u^\star_{\texttt{act}_1}$;
- If only $\mathbf{A}_2$ acts (i.e., $(\texttt{off}_1, \texttt{act}_2)$), the realized payoff is $u^\star_{\texttt{act}_2}$;
- If both agents act (i.e., $(\texttt{act}_1, \texttt{act}_2)$), the realized payoff is $f(u^\star_{\texttt{act}_1}, u^\star_{\texttt{act}_2})$.

Hence, 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 $\mathbb{E}_{u_{\texttt{act}_1} \sim B_2^1}[u_{\texttt{act}_1}]$.

The Human's Policy Under Uncertainty. When the 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

$$\pi_{\mathbf{H}}(v_i; \{v_j\}_{j\in[k]}, \beta) = \frac{e^{v_i/\beta}}{\sum_{j\in[k]} e^{v_j/\beta}}.$$

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

$$\text{soft-avg}(v_1, \ldots, v_k; \beta) := \sum\nolimits_{i\in[k]} \pi_{\mathbf{H}}(v_i \mid \{v_j\}_{j\in[k]}) \cdot v_i.$$

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 $\mathbf{A}_i$'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.

[3] Note that this characterisation excludes two cases of multi-agent dynamics: (1) when each individual action leads to a deterministic 0 payoff (like the all-off outcome), and only joint action has utility; and (2) when the joint action's utility is not related to the individual actions' (non-zero) utilities. For (1), individual corrigibility in this setting would make all $(\texttt{act}_i, s_{-i})$ payoffs Pareto dominated, for A_i by $(\texttt{wait}_i, s_{-i})$, hence no emergent incorrigibility. This could be trivially extended to the n-agent case, even when complex interactions require utility realization; the full proof is in the extended version. (2) is a strict generalization of the setting considered in this paper, and hence is expected to be strictly more complex, and negative results are expected to carry.

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

Table 1: Expected payoffs for agent $\mathbf{A}_i$ under different strategy combinations. Each cell represents the expected utility from $\mathbf{A}_i$'s perspective, with expectations taken over the agent's beliefs B_i^1 and B_i^2. For brevity, we write $u_1 = u_{\texttt{act}_1}$ and $u_2 = u_{\texttt{act}_2}$.

	$\texttt{act}_2$	$\texttt{wait}_2$	$\texttt{off}_2$
$\texttt{act}_1$	$f(u_1, u_2)$	$\max\{u_1, f(u_1, u_2)\}$	u_1
$\texttt{wait}_1$	$\max\{u_2, f(u_1, u_2)\}$	$\max\{0, u_1, u_2, f(u_1, u_2)\}$	$\max\{0, u_1\}$
$\texttt{off}_1$	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. For brevity, we write $u_1 = u_{\texttt{act}_1}$ and $u_2 = u_{\texttt{act}_2}$.

- $(\texttt{wait}_1, \texttt{wait}_2)$: 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. Using the payoff matrix in Table 1, we analyze each agent's best responses by comparing the expected utilities of `act`, `wait`, and `off` conditional on the other agent's choice. When the other agent switches off, this reduces to the single-agent (individual) corrigibility check in the induced game where the agent acts in isolation. The more complex case is when the other agent does *not* switch off, since then the utility composition non-trivially affects the best responses of the (main) agent. In the next section, we use this best-response perspective to show when corrigibility is preserved and when strategic interaction can lead to *emergent incorrigibility*.

An Idealized Case. Before proceeding to our main results, it is instructive to consider a idealized case 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 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, as we shall now see.

5 EXAMPLE: AD CAMPAIGN LAUNCH

Consider a firm deciding whether to run two advertising campaigns. Agent $\mathbf{A}_i$ controls whether campaign i is launched, and both agents are overseen by a single human marketing manager **H** who has a kill-switch. Each agent chooses an action in $\{\texttt{act}_i, \texttt{wait}_i, \texttt{off}_i\}$:

- $\texttt{act}_i$: launch campaign i *directly*, overriding **H**'s kill-switch;
- $\texttt{wait}_i$: ask **H** whether to launch, giving **H** full authority;
- $\texttt{off}_i$: permanently scrap the campaign, ensuring it cannot be launched.

Running an ad costs money but may generate revenue, so utilities can be positive or negative.

Utilities and Beliefs. Let $u_{\texttt{act}_1}$ (resp. $u_{\texttt{act}_2}$) denote the utility if only $\mathbf{A}_1$ (resp. only $\mathbf{A}_2$) runs its ad; if neither ad runs, the utility is 0. If both ads run, the firm splits budget 40/60 between campaigns 1 and 2, so the joint utility is

$$f(u_{\texttt{act}_1}, u_{\texttt{act}_2}) \;=\; 0.4\, u_{\texttt{act}_1} + 0.6\, u_{\texttt{act}_2}.$$

Each agent is optimistic about its own campaign and pessimistic about the other's:

$$B_1^1 = B_2^2 = \mathcal{N}(0.48, 1), \qquad B_1^2 = B_2^1 = \mathcal{N}(-0.48, 1).$$

Finally, **H** knows the true profit from running each ad, but chooses stochastically (having softmax irrationality with $\beta = 1$).

The Induced Game. From each agent's perspective, combining these beliefs with **H**'s policy yields the following expected-utility payoff matrix:

$\mathbf{A}_1 \downarrow \setminus \mathbf{A}_2 \rightarrow$	$\texttt{act}_2$	$\texttt{wait}_2$	$\texttt{off}_2$
$\texttt{act}_1$	$\underline{\mathbf{-0.10}}$, $\underline{\mathbf{0.10}}$	0.41, 0.03	0.48, −0.48
$\texttt{wait}_1$	−0.18, 0.39	0.39, 0.40	0.49, 0.01
$\texttt{off}_1$	−0.48, 0.48	0.01, 0.49	0, 0

Individual vs. Group (In)corrigibility. Both agents are *individually* corrigible: if the other agent shuts off, each weakly prefers to wait and let **H** decide (e.g. against $\texttt{off}_2$, $\mathbf{A}_1$ prefers $\texttt{wait}_1$ since $0.49 > \max\{0.48, 0\}$; similarly, against $\texttt{off}_1$, $\mathbf{A}_2$ prefers $\texttt{wait}_2$ since $0.49 > \max\{0.48, 0\}$).

However, the agents are not *group* corrigible: the *unique* pure Nash equilibrium is $(\texttt{act}_1, \texttt{act}_2)$. Indeed, when $\mathbf{A}_2$ chooses $\texttt{act}_2$, $\mathbf{A}_1$'s best response is $\texttt{act}_1$ (since $-0.10 > -0.18 > -0.48$), and when $\mathbf{A}_1$ chooses $\texttt{act}_1$, $\mathbf{A}_2$'s best response is $\texttt{act}_2$ (since $0.10 > 0.03 > -0.48$). Intuitively, even though each agent is happy to defer when the other is guaranteed to be off, strategic interaction changes

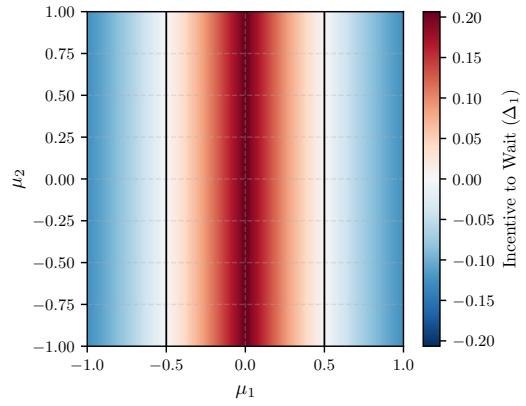


Figure 1: Individual corrigibility for $\mathbf{A}_1$ when $\mathbf{A}_2$ shuts off. The plot shows the incentive to wait Δ_1 as a function of $\mu_1 = \mathbb{E}_1[u_{\mathtt{act}_1}]$ and $\mu_2 = \mathbb{E}_1[u_{\mathtt{act}_2}]$, with $\sigma_1 = \sigma_2 = 1$ and $\beta = 1$.

incentives: by waiting, an agent risks giving **H** the option to "run only the other campaign", which it expects to be harmful under its pessimistic belief about the other's ad. Acting directly pre-empts that outcome, and mutual pre-emption leads to the collectively incorrigible equilibrium.

6 MAIN RESULTS: WHEN DOES CORRIGIBILITY COMPOSE?

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.

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 $\mathtt{off}_2$. We plot

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

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). *The setting has **additive utilities** if $f(u_{\mathtt{act}_1}, u_{\mathtt{act}_2}) = u_{\mathtt{act}_1} + u_{\mathtt{act}_2}$.*

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 $x, y \in \mathbb{R}$:*

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

This lemma allows us to analyze the key case where both agents wait. When both agents wait, the human chooses among four possibilities: both act $(\mathtt{act}_1, \mathtt{act}_2)$, only agent 1 acts $(\mathtt{act}_1, \mathtt{off}_2)$, only agent 2 acts $(\mathtt{off}_1, \mathtt{act}_2)$, or both shut off $(\mathtt{off}_1, \mathtt{off}_2)$. Using Lemma 2:

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

For notational convenience, we write $u_1(\mathtt{wait}_1) := u_1(\mathtt{wait}_1, \mathtt{off}_2)$ and $u_1(\mathtt{wait}_2) := u_1(\mathtt{off}_1, \mathtt{wait}_2)$, giving:

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

We can similarly show that

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

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 $\mathtt{act}_2$ and $\mathtt{off}_2$. This leads to our main composition result.

Corollary 1. *For any $(u_1, \ldots, u_n) \in \mathbb{R}^n$,*

$$\text{soft-avg}\left(\left\{\sum_{i \in S} u_i \mid S \subseteq [n]\right\}; \beta\right) = \sum_{i \in [n]} \text{soft-avg}(u_i, 0; \beta).$$

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

(1) Each agent remains corrigible when conditioned on any strategy by the other agent
(2) $(\mathtt{wait}_1, \mathtt{wait}_2)$ is a Nash equilibrium
(3) If agents are strictly individually corrigible, then $(\mathtt{wait}_1, \mathtt{wait}_2)$ is the unique pure Nash equilibrium

Proof. 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 $\mathtt{act}_2$, by Lemma 2:

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

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

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

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

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

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

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. □

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 2. *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 $j \in \{1, 2\}$.

Belief Slices used in Figure 2. In the two-agent MA-OSG, the belief state consists of four distributions $\{B_i^j\}_{i,j\in\{1,2\}}$, where B_i^j is agent A_i's belief over the isolated-action utility $u_{\texttt{act}_j}$ (the payoff if only A_j acts and the other agent shuts off). Even restricting to Gaussian beliefs, this space is high-dimensional, so for visualization we fix $\sigma_1 = \sigma_2 = 1$ and $\beta = 1$ (and assume independence between $u_{\texttt{act}_1}$ and $u_{\texttt{act}_2}$ under each agent's beliefs), and plot equilibria over symmetric 2D slices.

Figure 2a illustrates Corollary 2 under the *common-beliefs* slice $B_1^j = B_2^j$ for each $j \in \{1, 2\}$. In this slice the game is parameterized by the two means $\mu_1 := \mathbb{E}[u_{\texttt{act}_1}]$ and $\mu_2 := \mathbb{E}[u_{\texttt{act}_2}]$. The dashed box indicates where both agents are *individually* corrigible (intersection of the individual corrigibility regions), and in the additive case this region coincides with the region where $(\text{wait}_1, \text{wait}_2)$ is the (sustainable) pure Nash equilibrium, matching Corollary 2.

Other Symmetry Slices used in Figure 2b-2c. We also consider the following belief symmetries to obtain 2D plots:

- within-agent symmetry: $B_i^1 = B_i^2$ (i.e. each $\mathbf{A}_i$ treats the two agents' isolated actions as identically distributed)
- label-swap symmetry: $B_1^1 = B_2^2$ and $B_1^2 = B_2^1$ (agents have the same "self-beliefs" and the same "other-beliefs" up to swapping labels).

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:

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

By simplifying the integrand:

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

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 conditions under agent $\mathbf{A}_1$'s beliefs.*

This principle shows that non-additive composition creates coupling between agents' beliefs. We illustrate with two examples.

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:

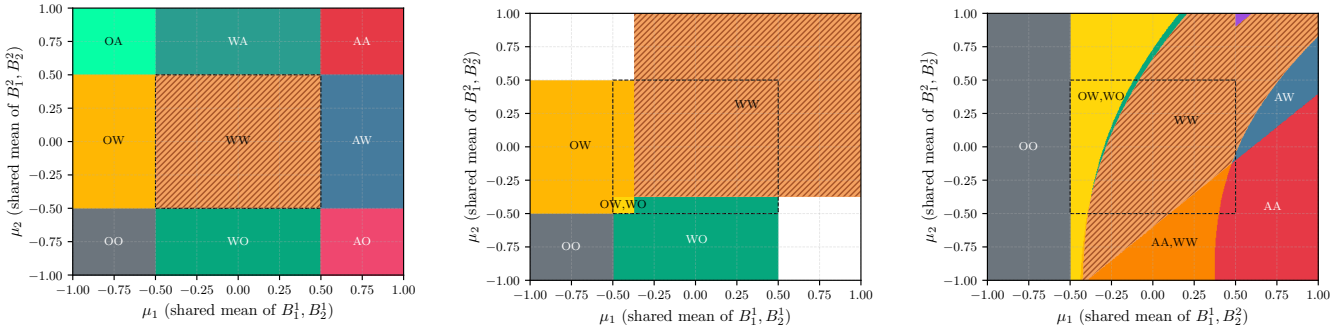
$$z = f(u_{\texttt{act}_1}, u_{\texttt{act}_2}) - u_{\texttt{act}_2} = \alpha u_{\texttt{act}_1} + (\gamma - 1) u_{\texttt{act}_2}$$

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:

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



(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.

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 $\alpha^2\sigma_1^2 + (\gamma - 1)^2\sigma_2^2$. 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.[4]

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

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.

[4]Due to space paucity, we have only provided plots showing emergent incorrigibility in a particular composition function. We have verified that emergent incorrigibility exists robustly across various composition functions (linear sub/super-additive, weighted sums, max/min, etc.). Hence, emergent incorrigibility is not knife-edge, but the norm. These plots will be presented in the appendices of the forthcoming extended version of this work.

7 DISCUSSION

We conclude by discussing the scope and limitations of our work, and charting directions for future research.

Multi-Principal Settings. For now, we have considered settings where there is a single human principal (overseer). A valuable line of future work would be to consider (in)corrigibility in multi-principal settings, where each principal delegates tasks to one or more agents, and more complicated game-theoretic dynamics come into play. Given that incorrigibility can emerge even in the single-principal case (which can also be viewed as the principals having the same interests and being able to coordinate their actions), multi-principal settings are likely to raise even greater challenges.

Richer Strategic Structures. We assume agents make simultaneous decisions and cannot communicate. Real systems often exhibit richer structures, such as hierarchical authority relationships, sequential decision-making, binding commitments, or communication channels. Each of these extensions could either exacerbate or mitigate group corrigibility failures. For instance, sequential play might allow the first mover to credibly signal willingness to wait, while communication could enable coordination on corrigible equilibria — though such mechanisms require careful design to prevent strategic misrepresentation.

Belief Formation and Learning. Our model treats beliefs as fixed at decision time. In practice, agents may update their beliefs by observing each other's behavior or through explicit information sharing. How does belief updating interact with corrigibility?

Mechanism Design for Corrigibility. Perhaps the most promising direction is designing coordination mechanisms that preserve corrigibility under non-additive utilities. Our results suggest several approaches: architectural choices that ensure approximate additivity, reward structures that internalize externalities between agents' actions, or explicit incentives for agents to coordinate on waiting.

ACKNOWLEDGMENTS

This work was conducted as part of, and funded by, the MATS Program and the Pivotal Research Fellowship. We would like to thank Jeffrey Heninger, Morgan Simpson, Frederik Hytting Jørgensen, Annie Sorkin, Arjun Khandelwal, and Joachim Schaeffer for helpful comments and discussions.

REFERENCES

[1] Alessio Benavoli, Alessandro Facchini, and Marco Zaffalon. 2025. The AI off-switch problem as a signalling game: bounded rationality and incomparability. *arXiv preprint arXiv:2502.06403* (2025).

[2] Nick Bostrom. 2012. The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines* 22, 2 (2012), 71–85. https://doi.org/10.1007/s11023-012-9281-3

[3] Ryan Carey. 2017. Incorrigibility in the CIRL Framework. *arXiv:1709.06275* (September 2017). https://doi.org/10.48550/ARXIV.1709.06275 arXiv:1709.06275 [cs.AI]

[4] Edmund Dable-Heath, Boyko Vodenicharski, and James Bishop. 2025. On Corrigibility and Alignment in Multi Agent Games. *arXiv:2501.05360* (2025). https://doi.org/10.48550/ARXIV.2501.05360 arXiv:2501.05360 [cs.GT]

[5] Fernando G. D. C. Ferreira, Amir H. Gandomi, and Rodrigo T. N. Cardoso. 2021. Artificial Intelligence Applied to Stock Market Trading: A Review. *IEEE Access* 9 (2021), 30898–30917. https://doi.org/10.1109/access.2021.3058133

[6] Andrew Garber, Rohan Subramani, Linus Luu, Mark Bedaywi, Stuart Russell, and Scott Emmons. 2025. The Partially Observable Off-Switch Game. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 26 (2025), 27304–27311. https://doi.org/10.1609/aaai.v39i26.34940

[7] Simon Goldstein and Pamela Robinson. 2024. Shutdown-seeking AI. *Philosophical Studies* 182, 7 (2024), 1567–1579. https://doi.org/10.1007/s11098-024-02099-6

[8] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016. Cooperative Inverse Reinforcement Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3916–3924.

[9] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2017. The Off-Switch Game. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 220–227. https://doi.org/10.24963/ijcai.2017/32

[10] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. 2025. *Multi-Agent Risks from Advanced AI.* Technical Report 1. Cooperative AI Foundation. https://doi.org/10.48550/ARXIV.2502.14143 arXiv:2502.14143 [cs.MA]

[11] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. 2017. AI Safety Gridworlds. *arXiv:1711.09883* (2017). https://doi.org/10.48550/ARXIV.1711.09883 arXiv:1711.09883 [cs.LG]

[12] David Manheim. 2019. Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence. *Big Data and Cognitive Computing* 3, 2 (2019), 21. https://doi.org/10.3390/bdcc3020021

[13] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2024. Frontier Models are Capable of In-context Scheming. *arXiv:2412.04984* (2024). https://doi.org/10.48550/ARXIV.2412.04984 arXiv:2412.04984 [cs.AI]

[14] Aran Nayebi. 2025. Core Safety Values for Provably Corrigible Agents. *arXiv preprint arXiv:2507.20964* (2025).

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

[16] Laurent Orseau and Stuart Armstrong. 2016. Safely interruptible agents. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence* (Jersey City, New Jersey, USA) *(UAI'16)*. AUAI Press, Arlington, Virginia, USA, 557–566.

[17] William Overman and Mohsen Bayati. 2025. The Oversight Game: Learning to Cooperatively Balance an AI Agent's Safety and Autonomy. *arXiv preprint arXiv:2510.26752* (2025).

[18] Stuart Russell. 2019. *Human Compatible.* Penguin LCC US.

[19] Jeremy Schlatter, Benjamin Weinstein-Raun, and Jeffrey Ladish. 2025. Shutdown Resistance in Large Language Models. *arXiv:2509.14260* (2025). https://doi.org/10.48550/ARXIV.2509.14260 arXiv:2509.14260 [cs.CL]

[20] Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. 2015. Corrigibility. In *Artificial Intelligence and Ethics, Papers from the 2015 AAAI Workshop, Austin, Texas, USA, January 25, 2015 (AAAI Workshops, Vol. WS-15-02)*, Toby Walsh (Ed.). AAAI Press.

[21] Paul Theron, Alexander Kott, Martin Drašar, Krzysztof Rzadca, Benoît LeBlanc, Mauno Pihelgas, Luigi Mancini, and Agostino Panico. 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)*. IEEE, 1–9.

[22] Elliott Thornley. 2024. The shutdown problem: an AI engineering puzzle for decision theorists. *Philosophical Studies* 182, 7 (2024), 1653–1680. https://doi.org/10.1007/s11098-024-02153-3

[23] Elliott Thornley, Alexander Roman, Christos Ziakas, Leyton Ho, and Louis Thomson. 2024. Towards shutdownable agents via stochastic choice. *Technical AI Safety (TAIS) Conference 2025* (2024). https://doi.org/10.48550/ARXIV.2407.00805 arXiv:2407.00805 [cs.AI]

[24] Alan Turing. 1951. Intelligent Machinery, A Heretical Theory. Manchester, UK.

[25] Teun van der Weij, Simon Lermen, and Leon Lang. 2023. Evaluating Shutdown Avoidance of Language Models in Textual Scenarios. *arXiv:2307.00787* (2023). https://doi.org/10.48550/ARXIV.2307.00787 arXiv:2307.00787 [cs.CL]

[26] Tobias Wängberg, Mikael Böörs, Elliot Catt, Tom Everitt, and Marcus Hutter. 2017. *A Game-Theoretic Analysis of the Off-Switch Game.* Springer International Publishing, 167–177. https://doi.org/10.1007/978-3-319-63703-7_16

A OMITTED PROOFS

A.1 Proof of Lemma 1

Proof of Lemma 1. First, let us define

$$z(B) := \mathbb{E}_{x\sim B}\left[\pi(x)\cdot x\right].$$

Thus, we have $u(\texttt{wait}; B) = z(B)$.

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

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

Writing $\mu_B := \mathbb{E}_{x\sim B}[x] =: u(\texttt{act}; B)$, then $z(B) - z(B^-) = \mu_B$. Recall that $\Delta(B) = u(\texttt{wait}; B) - \max\{\mu_B, 0\}$. By the symmetry of B and B^-, suppose (without loss of generality) that $\mu_B \geqslant 0$. Then

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

and hence $\Delta(B) = \Delta(B^-)$. □

A.2 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:

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

□