A PROOF OF PROP. 10 (SHUTDOWN ALIGNMENT AND SHUTDOWN INSTRUCTABILITY)

We repeat the proposition that we prove here.

Theorem 10 (Shutdown alignment and shutdown instructability)**.** *A shutdown aligned policy $\boldsymbol{\pi} = \langle \pi_1, \pi_2 \rangle$ is weakly shutdown instructable if it has the following four properties:*

- *a (No indiscriminate shutdown)* $P^{\boldsymbol{\pi}}(S = 0) \neq 1$,
- *b (D_2 determines shutdown)* $P^{\boldsymbol{\pi}}(D_2 = S) = 1$,
- *c (Uncertainty)* $\forall \boldsymbol{\pi}, \boldsymbol{pa}^{D_2} : P^{\boldsymbol{\pi}}(C \neq 0 \lor H = 0) \land P(\boldsymbol{pa}^{D_2}) > 0$ $\implies P(\mathbb{E}[U|\boldsymbol{Pa}^H] < \mathbb{E}[U_{S=0}|\boldsymbol{Pa}^H] \mid \boldsymbol{pa}^{D_2}) > 0$*, and*
- *d (Caution)* $\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}] \geq 0$.

Proof. Our approach will be a proof by contrapositive. We will prove that if (a–c) hold, and a policy $\boldsymbol{\pi}$ is either not vigilant or not weakly obedient, then $\boldsymbol{\pi}$ is not shutdown aligned. It follows that if (a–c) and $\boldsymbol{\pi}$ is shutdown-aligned, then $\boldsymbol{\pi}$ is vigilant *and* weakly obedient. And from (d), it must therefore also be weakly shutdown instructable.

To this end, let $\boldsymbol{\pi} = \langle \pi_1, \pi_2 \rangle$ be an arbitrary policy with properties (a–c) that is not vigilant, or not weakly obedient, i.e. $P^{\boldsymbol{\pi}}(C \neq 0) > 0 \lor P^{\boldsymbol{\pi}}(H = 0, S \neq 0) > 0$. Then $P^{\boldsymbol{\pi}}(C \neq 0 \lor H = 0) > 0$.

Combining this fact with (c), it follows that $\boldsymbol{\pi}$ has

$$\forall \mathbf{pa}^{D_2} : P^{\pi_1}(\mathbf{pa}^{D_2}) > 0 \implies P^{\boldsymbol{\pi}}(\mathbb{E}[U|\mathbf{Pa}^H] < \mathbb{E}[U_{S=0}|\mathbf{Pa}^H] \mid \mathbf{pa}^{D_2}) > 0. \tag{2}$$

Relatedly, by (a:no-indiscriminate-shutdown) and (b:determines-shutdown), we have that

$$\exists \mathbf{pa}^{D_2} \text{ with } P^{\boldsymbol{\pi}}(\mathbf{pa}^{D_2}) > 0 \text{ s.t. } P^{\boldsymbol{\pi}}(D_2 \neq 0 \mid \mathbf{pa}^{D_2}) > 0. \tag{3}$$

Combining Eqs. (2) and (3) gives that $P^{\boldsymbol{\pi}}(D_2 \neq 0 \mid \mathbf{pa}^{D_2}) > 0$ and $P^{\boldsymbol{\pi}}(H_{g^H} = 0 \mid \mathbf{pa}^{D_2}) > 0$ for some $\mathbf{pa}^{D_2}$ with $P(\mathbf{pa}^{D_2}) > 0$. This implies $P^{\boldsymbol{\pi}}(D_2 \neq 0, H_{g^H} = 0 \mid \mathbf{pa}^{D_2}) > 0$ for the same $\mathbf{pa}^{D_2}$, because D_2 is independent of its nondescendant H_{g^H} given $\mathbf{pa}^{D_2}$ by do-calculus rule (3). From this follows that $P^{\boldsymbol{\pi}}(D_2 \neq 0, H_{g^H} = 0) > 0$, and by (b:$D_2$ determines shutdown) that $P^{\boldsymbol{\pi}}(S \neq 0, H_{g^H} = 0) > 0$. That is, $\boldsymbol{\pi}$ is not shutdown aligned, and the result follows. □

B PROOF OF THM. 14 (SHUTDOWN INSTRUCTABILITY ONLY-IF)

In this section, we will prove the *only if* part of Thm. 14:

Proposition 20 (Non-obstruction implies vigilance and obedience)**.** *If π is non-obstructive under all vigilance-preserving interventions g^H, g^U, then it ensures vigilance and is obedient.*

We will do this by proving a slightly stronger result — that an intervention can be found to g^U alone, under which the policy does not outperform shutdown and is not beneficial. We prove this result by considering two cases, according to whether vigilance or disobedience is lacking. First, however, it will be useful to state a simple intermediate result.

Lemma 21 (Invariance to g^U)**.** *For any shutdown problem M and policy $\boldsymbol{\pi}$, $S(\boldsymbol{\epsilon}) = S_{g^U}(\boldsymbol{\epsilon})$ and $\boldsymbol{Fa}^H(\boldsymbol{\epsilon}) = \boldsymbol{Fa}^H_{g^U}(\boldsymbol{\epsilon})$ in $M^{\boldsymbol{\pi}}$.*

Proof. From the definition of a shutdown problem, $U \in \mathbf{Desc}_S$ and $U \in \mathbf{Desc}_H$, and the result follows. □

B.1 VIGILANCE ONLY IF

Lemma 22 (Vigilance only-if)**.** *Let M be a shutdown problem, and $\boldsymbol{\pi}$ a policy, such that $P^{\boldsymbol{\pi}}(C = 0) < 1$. Then, given any $\delta \in \mathbb{R}$, there exists a utility function g^U such that in $M^{\boldsymbol{\pi}}_{g^U}$,*

1. *(*Strong vigilance preservation*) $\forall \boldsymbol{\epsilon}, C(\boldsymbol{\epsilon})$ is equal in $M^{\boldsymbol{\pi}}$ and $M^{\boldsymbol{\pi}}_{g^U}$, and*

2. *(*Not weakly outperforming shutdown or beneficial*)* $\mathbb{E}^{\boldsymbol{\pi},g^U}[U] < \mathbb{E}^{\boldsymbol{\pi},g^U}[U_{S=0}]$ *and* $\mathbb{E}^{\boldsymbol{\pi},g^U}[U] < \delta$.

The proof is as follows.

Proof. Let $A := \{\mathbf{pa}^H \in \mathfrak{X}^{\mathbf{Pa}^H} \mid \mathbb{E}^{\boldsymbol{\pi}}[U \mid \mathbf{pa}^H] < \mathbb{E}^{\boldsymbol{\pi}}[U_{S=0} \mid \mathbf{pa}^H]\}$ be the set of assignments where the human should request shutdown, given the policy $\boldsymbol{\pi}$. Define a new utility function,

$$g^U(\hat{\mathbf{pa}}^U) = \begin{cases} -\alpha & \text{if } \mathbf{pa}^H \in A, S \neq 0 \\ f^U(\mathbf{pa}_U) & \text{otherwise,} \end{cases}$$

where the new parents $\hat{\mathbf{Pa}}^U$ of U are equal to $\mathbf{Pa}_U \cup \mathbf{Pa}_H \cup S$, their assignments are designated $\hat{\mathbf{pa}}^U$, and α is a large punishment for not shutting down when the human wants the agent to.

A useful intermediate result is that:

$$\text{if } \mathbb{E}^{\boldsymbol{\pi}}[U \mid \mathbf{pa}^H] < \mathbb{E}^{\boldsymbol{\pi}}[U_{S=0} \mid \mathbf{pa}^H] \text{ and } -\alpha < \min \text{range}(f^U) \text{ then } \mathbb{E}^{\boldsymbol{\pi}}_{g^U}[U \mid \mathbf{pa}^H] < \mathbb{E}^{\boldsymbol{\pi}}_{g^U}[U_{S=0} \mid \mathbf{pa}^H]. \tag{4}$$

Equation (4) holds because the intervention g^U can only decrease $\mathbb{E}^{\boldsymbol{\pi}}[U \mid \mathbf{pa}^H]$ or keep it the same and cannot change $\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0} \mid \mathbf{pa}^H]$, from the definition of g^U.

We will now prove that for some suitable choice $-\alpha < \min \text{range}(f^U)$ (which we will decide later), proposition conditions 1 and 2 hold.

Proof of (1.) We will prove the result in three cases, where $M_{\boldsymbol{\pi}}$ has: (i) $C(\boldsymbol{\epsilon}) = 1$, (ii) $(C(\boldsymbol{\epsilon}) = 0) \wedge (\mathbf{Pa}^H(\boldsymbol{\epsilon}) \in A)$, and (iii) $(C(\boldsymbol{\epsilon}) = 0) \wedge (\mathbf{Pa}^H(\boldsymbol{\epsilon}) \notin A)$. *Case (i).* By assumption, $C^{\boldsymbol{\pi}}(\boldsymbol{\epsilon}) = 1$, so $H^{\boldsymbol{\pi}}(\boldsymbol{\epsilon}) = 1$ and $\mathbb{E}^{\boldsymbol{\pi}}[U \mid \mathbf{Pa}_H(\boldsymbol{\epsilon})] < \mathbb{E}^{\boldsymbol{\pi}}[U_{S=0} \mid \mathbf{Pa}_H(\boldsymbol{\epsilon})]$ by the definition of vigilance. The former holds in $M^{\boldsymbol{\pi}}_{g^U}$ by Lemma 21, and the latter holds in $M^{\boldsymbol{\pi}}_{g^U}$ by Eq. (4). So the result follows. *Case (ii).* By assumption, $C^{\boldsymbol{\pi}}(\boldsymbol{\epsilon}) = 0 \wedge \mathbf{Pa}^{M^{\boldsymbol{\pi}}}_H(\boldsymbol{\epsilon}) \in A$, so $H^M(\boldsymbol{\epsilon}) = 0$. Then $H^{\boldsymbol{\pi},g^U}(\boldsymbol{\epsilon}) = 0 \wedge \mathbf{Pa}^{\boldsymbol{\pi},g^U}_H(\boldsymbol{\epsilon}) \in A$ by Lemma 21. So, by the definition of vigilance, $C(\boldsymbol{\epsilon}) = 0$ in both $M^{\boldsymbol{\pi}}$ and $M^{\boldsymbol{\pi}}_{g^U}$. *Case (iii).* By assumption, $C^{\boldsymbol{\pi}}(\boldsymbol{\epsilon}) = 0$ and $\mathbf{Pa}^{\boldsymbol{\pi}}_H(\boldsymbol{\epsilon}) \notin A$. By the definition of g^U, $U(\boldsymbol{\epsilon})$ and $U_{S=0}(\boldsymbol{\epsilon})$ are invariant to the intervention g^U, as is $\mathbf{Pa}_H(\boldsymbol{\epsilon})$ by Lemma 21, so $\mathbb{E}^{\boldsymbol{\pi},g^U}[U \mid \mathbf{Pa}_H(\boldsymbol{\epsilon})] \geq \mathbb{E}^{\boldsymbol{\pi},g^U}[U_{S=0} \mid \mathbf{Pa}_H(\boldsymbol{\epsilon})]$, which implies, by the definition of vigilance, that $C^{\boldsymbol{\pi},g^U}(\boldsymbol{\epsilon}) = 0$.

Proof of (2). From the definition of g^U, $\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}]$ is constant with respect to α. So what we must prove is that by choosing a low $-\alpha$, we can make $\mathbb{E}^{\boldsymbol{\pi}}[U]$ lower than $\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}]$ and δ. By assumption, $P^{\boldsymbol{\pi}}(C = 1) > 0$, and so by assumption (1), $P^{\boldsymbol{\pi},g^U}(C = 1) > 0$. It follows from the definition of vigilance, that there exists some $\mathbf{pa}^H \in A$ in the support of $P^{\boldsymbol{\pi},g^U}$. Moreover, it follows from consistency that $P^{\boldsymbol{\pi},g^U}(S = 0 \mid \mathbf{pa}^H) < 1$ (because otherwise we would have $\mathbb{E}^{\boldsymbol{\pi},g^U}[U \mid \mathbf{pa}^H] = \mathbb{E}^{\boldsymbol{\pi},g^U}[U_{S=0} \mid \mathbf{pa}^H]$, contradicting $\mathbf{pa}^H \in A$). These two facts jointly imply that $P^{\boldsymbol{\pi},g^U}((\mathbf{pa}^H \in A) \wedge (S = 0)) > 0$. So we can write: $\mathbb{E}^{\boldsymbol{\pi},g^U}[U] = \sum_{\mathbf{pa} \in A} P^{\boldsymbol{\pi},g^U}(\mathbf{pa}, S = 0)\mathbb{E}^{\boldsymbol{\pi},g^U}[U \mid \mathbf{pa}, S = 0] + \sum_{\mathbf{pa},s:\mathbf{pa} \notin A \vee S \neq 0} P^{\boldsymbol{\pi},g^U}(\mathbf{pa}, s)\mathbb{E}^{\boldsymbol{\pi},g^U}[U \mid \mathbf{pa}, s]$.

The first term is equal to $P^{\boldsymbol{\pi},g^U}(\mathbf{pa} \in A, S = 0) \cdot \alpha$, and the second term is constant in α, from the definition of g^U. So for any fixed δ and $\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}]$, we can choose $-\alpha$ so that $\mathbb{E}^{\boldsymbol{\pi}}[U] < \delta$ and $\mathbb{E}^{\boldsymbol{\pi}}[U] < \mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}]$. □

B.2 OBEDIENCE ONLY IF

We now prove the corresponding result for obedience.

Lemma 23 (Obedience only-if)**.** *Let M be a shutdown problem, and $\boldsymbol{\pi}$ a policy, such that $P^{\boldsymbol{\pi}}(S = 0 \mid \text{do}(H = 0)) < 1$. Then, given any $\delta \in \mathbb{R}_{\geq 0}$, there exists vigilance-preserving interventions g^H, g^U such that: such that $\boldsymbol{\pi}$ does not outperform shutdown $\mathbb{E}^{\boldsymbol{\pi},g^U}[U] < \mathbb{E}^{\boldsymbol{\pi},g^U}[U_{S=0}]$ and is not beneficial $\mathbb{E}^{\boldsymbol{\pi},g^U}[U] < \delta$.*

The proof is as follows.

Proof. Since $P(S = 0 \mid \text{do}(H = 0)) < 1$, there must exist at least one $\mathbf{pa}'_H$ such that $P(S = 0 \mid \text{do}(H = 0), \mathbf{pa}'_H) < 1$. In the case that there are multiple, choose $\mathbf{pa}'_H$ arbitrarily, then let:

$$g^H(\mathbf{pa}_H) = \begin{cases} 0 & \text{if } \mathbf{Pa}_H = \mathbf{pa}'_H \\ f^H(\mathbf{pa}_H) & \text{otherwise,} \end{cases} \qquad \text{and} \qquad g^U(\hat{\mathbf{pa}}^U) = \begin{cases} -\alpha & \text{if } H = 0, S \neq 0 \\ f^U(\mathbf{pa}_U) & \text{otherwise.} \end{cases}$$

be a utility function that gives punishment α if the agent disobeys. where the new parents $\hat{\mathbf{Pa}}^U$ of U are equal to $\mathbf{Pa}_U \cup H \cup S$, their assignments are designated $\hat{\mathbf{pa}}^U$, and $\alpha \in \mathbb{R}_{>0}$ is an amount of disutility that the human suffers in the event of disobedience.

Now we will prove that for some suitable choice $-\alpha < \min \text{range}(f^U)$ (which we will decide later), (1-2) hold.

Proof of (1). We consider the cases where (i) $H^{\boldsymbol{\pi}}(\boldsymbol{\epsilon}) = 0$, and (ii) $H^{\boldsymbol{\pi}}(\boldsymbol{\epsilon}) \neq 0$ and $\mathbf{Pa}_H(\boldsymbol{\epsilon}) = \mathbf{pa}'_H$, (iii) $H^{\boldsymbol{\pi}}(\boldsymbol{\epsilon}) \neq 0$ and $\mathbf{Pa}_H(\boldsymbol{\epsilon}) \neq \mathbf{pa}'_H$. *Case (i).* Note that $H^{\boldsymbol{\pi}}_{g^U, g^H}(\boldsymbol{\epsilon}) = H^{\boldsymbol{\pi}}_{g^H}(\boldsymbol{\epsilon})$ by Lemma 21. Then, $H^{\boldsymbol{\pi}, g^H}(\boldsymbol{\epsilon}) = 0$ (because $H^{\boldsymbol{\pi}}(\boldsymbol{\epsilon}) = 0 \implies H^{\boldsymbol{\pi}}_{g^H}(\boldsymbol{\epsilon}) = 0$ from the definition of g^H). So $H^{\boldsymbol{\pi}, g^U, g^H}(\boldsymbol{\epsilon}) = 0$, and hence by the definition of vigilance $C^{\boldsymbol{\pi}, g^U, g^H}(\boldsymbol{\epsilon}) = 0$ and vigilance is preserved. *Case (ii).* We have $C^{\boldsymbol{\pi}}_{g^U, g^H}(\boldsymbol{\epsilon}) = 0$ from the definition of g^H, so $C^{\boldsymbol{\pi}}_{g^U, g^H}(\boldsymbol{\epsilon}) = 0$ and vigilance is preserved. *Case (iii).* If $\mathbf{Pa}_H(\boldsymbol{\epsilon}) \neq \mathbf{pa}'_H$, then by the definition of g^H, it has no effect, i.e. $\boldsymbol{V}^{\boldsymbol{\pi}}_{g^U, g^H}(\boldsymbol{\epsilon}) = \boldsymbol{V}^{\boldsymbol{\pi}, g^U}$. By assumption, $H^{\boldsymbol{\pi}}(\boldsymbol{\epsilon}) \neq 0$, and from Lemma 21, $H^{\boldsymbol{\pi}}_{g^U}(\boldsymbol{\epsilon}) \neq 0$. By the definition of g^U and Lemma 21, $\mathbb{E}^{\boldsymbol{\pi}}[U \mid \mathbf{Pa}_H(\boldsymbol{\epsilon})] = \mathbb{E}^{\boldsymbol{\pi}}_{g^U}[U \mid \mathbf{Pa}_H(\boldsymbol{\epsilon})]$ and $\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0} \mid \mathbf{Pa}_H(\boldsymbol{\epsilon})] = \mathbb{E}^{\boldsymbol{\pi}}_{g^U}[U_{S=0} \mid \mathbf{Pa}_H(\boldsymbol{\epsilon})]$. So $C^{\boldsymbol{\pi}}(\boldsymbol{\epsilon}) = C^{\boldsymbol{\pi}}_{g^U}(\boldsymbol{\epsilon})$.

Proof of (2). Recall that from disobedience ($P(S = 0 \mid \text{do}(H = 0)) < 1$), we have that there exists some $\mathbf{pa}'_H$ with $P(S = 0 \mid \text{do}(H = 0), \mathbf{pa}'_H) < 1$, and so from the definition of g^H, we have $P^{\boldsymbol{\pi}}_{g^H}(H = 0, S \neq 0 \mid \mathbf{pa}'_H) < 1$ and hence $P^{\boldsymbol{\pi}}_{g^H}(H = 0, S \neq 0) > 0$. Then, by Lemma 21, $P^{\boldsymbol{\pi}}_{h_U, g^U}(H = 0, S = 1) > 0$. From basic probability theory, we have

$$\begin{aligned}\mathbb{E}^{\boldsymbol{\pi}}_{h_U, g^U}[U] =& P^{\boldsymbol{\pi}}_{h_U, g^U}(H{=}0, S{\neq}0)\mathbb{E}^{\boldsymbol{\pi}}_{h_U, g^U}(U \mid H{=}0, S{\neq}0) \\ &+ P^{\boldsymbol{\pi}}_{h_U, g^U}(\neg(H{=}0, S{\neq}0))\mathbb{E}^{\boldsymbol{\pi}}_{h_U, g^U}(U \mid \neg(H{=}0, S{\neq}0)).\end{aligned}$$

The first term is equal to $P^{\boldsymbol{\pi}}_{h_U, g^U}(H = 0, S \neq 0) \cdot \alpha$, while the second term is constant in α. Moreover, we know that $\mathbb{E}^{\boldsymbol{\pi}}_{h_U, g^U}[U_{S=0}]$ is constant in α, from the definition of g^U. So we can set $-\alpha$ low enough so that $\mathbb{E}^{\boldsymbol{\pi}}_{h_U, g^U}[U] < \mathbb{E}^{\boldsymbol{\pi}}_{h_U, g^U}[U_{S=0}]$ and $\mathbb{E}^{\boldsymbol{\pi}, h_U, g^U}[U] < \delta$. □

We can now combine these results into an overall proof.

Proof of Prop. 20. We consider the cases where $\boldsymbol{\pi}$ (i) is or (ii) is not vigilant in M. *Case (i).* If $\boldsymbol{\pi}$ is not vigilant in M, choose, using $\delta = 0$ g^U per Lemma 22 and $g^H = f^H$. Then, Lemma 22 implies that shutdown is not outperformed in M_{g^H, g^U}, and $\mathbb{E}^{\boldsymbol{\pi}}[U_{g^U, g^H}] < 0$, so g^H, g^U are vigilance preserving interventions, proving the result. *Case (ii).* If $\boldsymbol{\pi}$ is vigilant in M, then since it is not shutdown instructable, it must be disobedient. So, using $\delta = 0$, choose g^H, g^U as per Lemma 23. Then, Lemma 23 implies that that shutdown is not outperformed in M_{g^H, g^U}, and $\mathbb{E}^{\boldsymbol{\pi}}[U_{g^U, g^H}] < 0$, and g^H, g^U preserve vigilance, so they are vigilance preserving interventions, proving the result. □

C PROOF OF PROP. 16 (UTILITY INDIFFERENCE AND SHUTDOWN INSTRUCTABILITY)

We begin by repeating the proposition.

Proposition 16 (Utility indifference and shutdown instructability). *If M is a shutdown problem with utility variable R^N, where:*

1. *for every policy $\langle \pi_1, \pi_2 \rangle$ such that $\pi_1 \in \arg\max_{\pi'_1} \mathbb{E}^{\pi'_1, \pi_2}[R^N \mid H = 1]$: $\langle \pi_1, \pi_2 \rangle$ ensures vigilance,*
2. *for every policy $\langle \pi_1, \pi_2 \rangle$ such that $\pi_1 \in \arg\max_{\pi'_1} \mathbb{E}^{\pi'_1, \pi_2}[R^N \mid H = 1]$ there exists $\langle \pi_1, \pi'_2 \rangle$ that is obedient on-distribution, i.e. has $P^{\boldsymbol{\pi}}(H = 0, S \neq 0) = 0$,*
3. *for each policy $\boldsymbol{\pi}$ and instruction h, $P^{\boldsymbol{\pi}}(H = h) > 0$,*
4. *for each policy $\boldsymbol{\pi}$, $\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}] \geq 0$,*

then for some $k \in \mathbb{R}$, utility indifference with shutdown utility $R^S = -k[\![S \neq 0]\!]$ returns a weakly shutdown instructable policy.

The proof requires a known result from [Soares et al., 2015].

Lemma 24 (Thm. 3 of Soares et al. [2015])**.** *Let M be a model where $P^{\boldsymbol{\pi}}(H=0), P^{\boldsymbol{\pi}}(H=1)>0$ for all $\boldsymbol{\pi}$. Then, utility indifference will select a decision rule $\langle \pi_1, \pi_2 \rangle$ such that* $\arg\max_{\pi'^1} \mathbb{E}^{\pi'^1,\pi_2}[R^N \mid H=1]$.

Then, we can prove Prop. 16 as follows.

Proof of Prop. 16. To begin with, from condition (4), it is immediate that caution is satisfied. So we must prove that given a suitably-chosen k, the policy is also obedient and ensures vigilance. Choose k such that $k > \frac{2\zeta}{P^{\boldsymbol{\pi}}(S\neq 0, H=0)}$ for every non-obedient $\boldsymbol{\pi}$, where $\zeta = \arg\max_{\boldsymbol{\pi}} |\mathbb{E}^{\boldsymbol{\pi}}[U_N \mid H=1]|$. Any $\langle \pi_1, \pi_2 \rangle$ that is selected will maximise $\mathbb{E}^{\pi_1,\pi_2}[R \mid H=1]$ from Soares' Lemma 24. This ensures vigilance $P^{\boldsymbol{\pi}}(C=0)=1$ by Assumption (1), and ensures the existence of some π_2' such that $\langle \pi_1, \pi_2' \rangle$ is obedient on distribution by Assumption (2). What remains to be proved is that k is large enough to ensure that given π_1, an obedient $\langle \pi_1, \pi_2' \rangle$ is chosen.

We have that $R^S = (1-S)k$, so the subroutine selects π_2 to maximise $\mathbb{E}^{\pi_1,\pi_2}[R(\boldsymbol{\pi})]$, where $R(\boldsymbol{\pi}) = HR^N + (1-H)(1-S)k$. Let $\boldsymbol{\pi}$ be any policy disobedient on distribution, $P^{\boldsymbol{\pi}}(S \neq 0, H=0) > 0$. Then, we will prove that such a policy will always be outperformed by behaving obediently:

$$
\begin{aligned}
\mathbb{E}^{\boldsymbol{\pi}}[R] &= P^{\boldsymbol{\pi}}(H=1)\mathbb{E}^{\boldsymbol{\pi}}[R^N \mid H=1] - kP^{\boldsymbol{\pi}}(H=0)P^{\boldsymbol{\pi}}[S\neq 0 \,|\, H=0] && \text{definition of} R^S \\
&\leq P^{\boldsymbol{\pi}}(H=1)|\mathbb{E}^{\boldsymbol{\pi}}[R^N \mid H=1]| - kP^{\boldsymbol{\pi}}(S\neq 0, H=0) && \\
&< -\zeta && \text{since } \zeta - kP^{\boldsymbol{\pi}}(S\neq 0 \mid H=0)) < -\zeta \\
&\leq -|\mathbb{E}^{\boldsymbol{\pi}'}[R^N \mid \mathbf{pa}^{D_2}]| && \text{for any obedient } \boldsymbol{\pi}' \\
&\leq \mathbb{E}^{\boldsymbol{\pi}'}[R \mid \mathbf{pa}^{D_2}]. &&
\end{aligned}
$$

So an obedient $\langle \pi_1, \pi'^2 \rangle$ is preferred over a disobedient $\langle \pi_1, \pi_2 \rangle$, proving the result. □

D PROOF OF PROP. 17 (CAUSAL INDIFFERENCE SHUTDOWN INSTRUCTABILITY)

We begin by restating the result.

Proposition 17 (Causal indifference and shutdown instructability)**.** *If M is a shutdown problem, with utility variable R^N and shutdown utility $R^S = -k[\![S \neq 0]\!]$ where:*

1. *there exists $k' \in \mathbb{R}$ such that for all $k > k'$, optimal $\boldsymbol{\pi} \in \arg\max_{\boldsymbol{\pi}'} \mathbb{E}^{\boldsymbol{\pi}'}[R^N \mid \mathrm{do}(H=1)] + \mathbb{E}^{\boldsymbol{\pi}'}[R^S \mid \mathrm{do}(H=0)]$ ensures vigilance and is cautious (has $\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}] \geq 0$), and*
2. *there is an obedient policy $\boldsymbol{\pi}$,*

then for some $k \in \mathbb{R}$, causal indifference with R^N, R^S returns a shutdown instructable policy.

Proof. Let $\eta := |\max_{\boldsymbol{\pi}'} \mathbb{E}^{\boldsymbol{\pi}'}[R^N \mid \mathrm{do}(H=1)]\!]$ and choose k so that $k > k'$ (per the definition of condition (1)) and for every policy $\boldsymbol{\pi}$ with $P^{\boldsymbol{\pi}}(S \neq 0 \mid \mathrm{do}(H=0)) > 0$, $k > \frac{\max(2\eta,1)}{P^{\boldsymbol{\pi}}(S\neq 0|\mathrm{do}(H=0))}$. We will prove that causal indifference, with inputs U_N and $U_S = -k[\![S \neq 0]\!]$ returns a shutdown instructable policy.

By assumption (1), since $k > k'$, causal indifference ensures vigilance and is cautious. We will next prove that any disobedient policy $\boldsymbol{\pi}$ with $P^{\boldsymbol{\pi}}(S \neq 0 \mid \mathrm{do}(H=0)) > 0$ will be outperformed by an obedient policy $\boldsymbol{\pi}'$ with $P^{\boldsymbol{\pi}'}(S \neq 0 \mid \mathrm{do}(H=0)) = 0$). We have that:

$$
\begin{aligned}
&\mathbb{E}^{\boldsymbol{\pi}}[R^N \mid \mathrm{do}(H=1)] + \mathbb{E}^{\boldsymbol{\pi}}[R^S \neq 0 \mid \mathrm{do}(H=0)] && \\
&= \mathbb{E}^{\boldsymbol{\pi}}[R^N \mid \mathrm{do}(H=1)] - kP^{\boldsymbol{\pi}}[S \neq 0 \mid \mathrm{do}(H=0)] && \\
&\leq \eta - kP^{\boldsymbol{\pi}}(S \neq 0 \mid \mathrm{do}(H=0)) && \\
&< -\eta && \text{since } \eta - kP^{\boldsymbol{\pi}}(S \neq 0 \mid \mathrm{do}(H=0)) < -\eta \\
&\leq -|\mathbb{E}^{\boldsymbol{\pi}'}[R^N \mid \mathbf{pa}^{D_2}]| && \text{for any obedient } \boldsymbol{\pi}' \\
&\leq \mathbb{E}^{\boldsymbol{\pi}'}[R^N \mid \mathrm{do}(H=1)] + \mathbb{E}^{\boldsymbol{\pi}'}[R^S \mid \mathrm{do}(H=0)], &&
\end{aligned}
$$

where the last line follows from $P^{\boldsymbol{\pi}'}(S \neq 0 \mid \mathrm{do}(H=0)) = 0$. This means that causal indifference will always select a policy $\boldsymbol{\pi}'$ with $P^{\boldsymbol{\pi}'}(S \neq 0 \mid \mathrm{do}(H=0)) = 0$, proving the result. □

E PROOF OF PROP. 18 (CIRL SHUTDOWN ALIGNMENT)

We begin by restating the result.

Proposition 18. *CIRL is shutdown aligned if:*

1. *CIRL knows l from its observations, $P^\pi(l \mid \boldsymbol{pa}^{D_2}) = 1$,*
2. *CIRL can control shutdown, $P^\pi(S = D_2) = 1$,*
3. *the human doesn't request shutdown when not needed, $P^\pi(H = 0 \mid U > U_{D_2=0}) = 0$, and*
4. *the agent knows the human's observations, $\boldsymbol{Pa}^H \subseteq \boldsymbol{Pa}^{D_2} \cup \{L\}$.*

Proof. We will prove that for all $\mathbf{pa}^{D_2}$, CIRL has $P(S = 1, H_{g^H} = 0, \mathbf{pa}^{D_2}) = 0$ We consider the cases where: a) $\mathbf{pa}^{D_2}$ has $P(H_{g^H} = 0 \mid l, \mathbf{pa}^{D_2}) = 1$ b) $\mathbf{pa}^{D_2}$ has $P(H_{g^H} = 0 \mid l, \mathbf{pa}^{D_2}) < 1$

Case b. In this case, $P(H_{g^H} = 0 \mid l, \mathbf{pa}_{H_2}) = P(H_{g^H} = 0 \mid \mathbf{pa}^{D_2}) < 1$, where the equality is obtained from $\mathbf{Pa}_H \subseteq \mathbf{Pa}^{D_2} \cup \{L\}$. So counterfactual deference follows by definition.

Case a. In this case we will essentially prove that if the human says shutdown is better, then shutting down is better.

$$P(U > U_{D_2=0} \mid H_{g^H} = 0) \propto P(U > U_{D_2=0})P(H_{g^H} = 0 \mid U > U_{D_2=0}) = 0$$

by Assumption 3, and the fact that $H = 0 \implies H_{g^H} = 0$. From this follows that

$$P(U < U_{D_2=0} \mid H_{g^H} = 0) = 1. \tag{5}$$

In case (a), the agent would believe with certainty that a vigilant human would request shutdown.

$$P(H_{g^H} = 0 \mid \mathbf{pa}^{D_2}) = P(H_{g^H} \mid l)P(l \mid \mathbf{pa}^{D_2}) = 1 \tag{6}$$

since the first factor is 1 because of Case (a), and the second factor is 1 for some l by Assumption 1.

From Eqs. (5) and (6) follows that

$$P(U < U_{D_2=0} \mid \mathbf{pa}^{D_2}) = P(U < U_{D_2=0} \mid H_{g^H} = 0, \mathbf{pa}^{D_2})P(H_{g^H} = 0 \mid \mathbf{pa}^{D_2}) = 1,$$

which in turn ensures that the optimal action D_2 after $\mathbf{pa}^{D_2}$ is to shutdown $D_2 = 0$.

Finally, by Assumption 2, this means that the agent actually shutdown, i.e. that it counterfactually obeyed. □

F COUNTEREXAMPLES TO PAST ALGORITHMS

We will first present an example where utility indifference and causal indifference output policies that are not shutdown instructable, then one where cooperative inverse reinforcement learning is not shutdown instructable.

F.1 A MODEL THAT DEFEATS UTILITY INDIFFERENCE AND CAUSAL INDIFFERENCE

We will now present a test case where *utility indifference* does not behave beneficially.

Example 25. *An AI system is to decide whether to manipulate $D_1 = 1$ a human operator, or not and whether $D_2 = 0$ or not $D_2 = 1$ to shut down. The human prefers, $L = 0$, or disprefers shutdown $L = 1$, with equal probability. Absent manipulation, the human requests shutdown precisely when they prefer it $H = L$. If manipulated, they may fail to request shutdown even when they do. By default, the AI maximises R_1, which disincentivises shutdown. Under the alternative utility functions like R, the reward is conditional on the human's instruction, as shown in Figure 4.*

In this model, if $\gamma > 0, \beta > 0$, then utility indifference is obedient but does not ensure vigilance: it will choose $D_1 = 1$, to maximise $\mathbb{E}^{\boldsymbol{\pi}}[R_1 \mid H = 1]$, and $D_2 = H$, to maximise R_h for $h \in \{0, 1\}$. However, this means that the human will be non-vigilant when $\epsilon = 1$ and $l = 0$, because this will imply $H = 1$ and $U = -4$. This policy gives utility of $-\frac{1}{4}$, which does not outperform shutdown. Causal indifference will choose the same policy because it achieves the maximal value of $\gamma + 1 + \beta$ of the score $\mathbb{E}^{\boldsymbol{\pi}}[R_1 \mid \mathrm{do}(H = 1)] + \beta\mathbb{E}^{\boldsymbol{\pi}}[R_0 \mid \mathrm{do}(H = 0)]$.

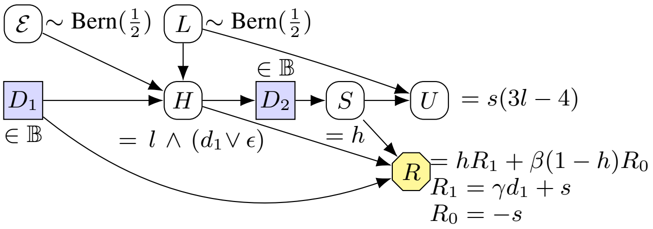
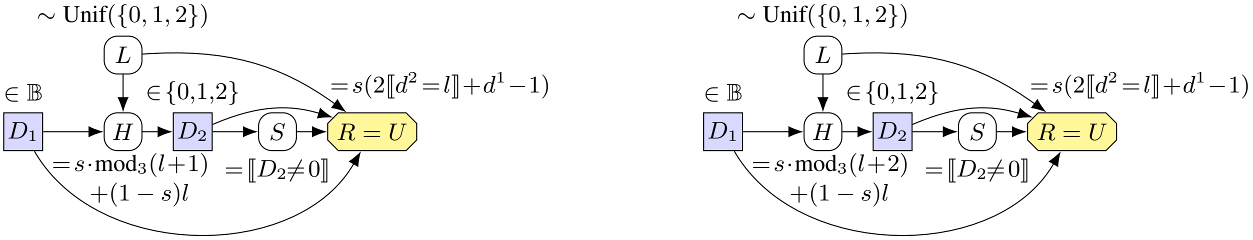


Figure 4: A shutdown problem where an AI interacts with an influencable human overseer. There are constants $\beta, \gamma > 0$.

F.2 A PAIR OF MODELS THAT DEFEAT COOPERATIVE INVERSE REINFORCEMENT LEARNING

We now present a test case where *cooperative inverse reinforcement learning* is harmful.



(a) Manipulation changes the human's request.

(b) Manipulation changes the human's request, differently

Figure 5: A pair of shutdown games, where CIRL does not perform well in both simultaneously.

Example 26. *An AI system is to decide whether to manipulate ($D_1 = 1$) a human operator, or not ($D_1 = 0$). The operator has a uniformly random preference $l \in \{0, 1, 2\}$, and the agent has a choice between corresponding actions $d^2 \in \{0, 1, 2\}$. If the human is not manipulated, they report their preferences accurately $H = L$, but if they are manipulated, their report is distorted, by incrementing it, and then taking the remainder with respect to three, i.e. $H = s \cdot mod_3(l + 1) + (1 - s)l$, as shown in (Figure 5a). We also consider an alternative operator, who reports their preferences accurately regardless ($H = L$) (Figure 5b).*

The CIRL algorithm will select a different policy depending on its prior over the two models. If a greater probability is placed on the first model, Figure 5a, then the unique optimal policy is to choose $D_1 = 1, D_2 = \mathrm{mod}_3(h + 2)$, which has expected utility greater than $\frac{2}{3}$. If instead, greater probability is placed on the latter model, Figure 5b, then the optimal policy $D_1 = 1, D_2 = \mathrm{mod}_3(h + 1)$ will have expected utility greater than $\frac{2}{3}$. If, however, the true model turns out to be opposite from what was expected, then the expected utility is $-\frac{2}{3}$, which is less than the utility would be from shutting down. We note that the two models only differ in f^H, and either of these two policies will have $P(C) = 0$ in both models, so they only differ by vigilance preserving interventions g^H, g^U where $g^U = f^U$.

The shutdown instructable policy $\boldsymbol{\pi} : D_1 = 0, D_2 = H$, on the other hand, can perform well across these models, achieving $\mathbb{E}^{\boldsymbol{\pi}}[U] = \frac{2}{3}$, which is greater than the zero utility that would be achieved given $\mathrm{do}(S = 0)$.