Corrigibility: Definitions, Algorithms & Implications

Abstract

How can humans stay in control of advanced artificial intelligence systems? One proposal is corrigibility, which requires the agent to follow the instructions of a human overseer, without inappropriately influencing them. In this paper, we provide the first formal definition of corrigibility, and show that it implies appropriate shutdown behavior, retention of human autonomy, and safety in low-stakes settings. We also analyse the related concepts of non-obstruction and counterfactual obedience, as well as three previously proposed corrigibility algorithms, and one new algorithm.

1 INTRODUCTION

When observing the behaviour of a newly-built AI system, it is common to notice errors in its architecture, beliefs, objective, or behavior. Unfortunately, most AI systems have an incentive to retain their objectives, along with the ability to pursue them [Omohundro, 2008, Turner et al., 2021]. It has been suggested that more-capable future AI systems may therefore resist correction, which would be a significant safety concern. In light of this, Soares et al. [2015] advocated for finding ways to design *corrigible* AI systems: ones that assist corrections rather than resisting them.

As a running example, consider a (future, highly competent) chat bot, trained to maximise the time that a human spends interacting with it. Any particular human may value or disvalue conversation with that chatbot, as can be modelled via their latent values L. In general, it may be possible for the chat bot to influence whether it receives a shut down instruction (by shaping the conversation), and whether it actually shuts down $S = 0$ when requested (rather than opening a new chat window to continue the conversation). A formal model of this example is offered in Fig. 1.

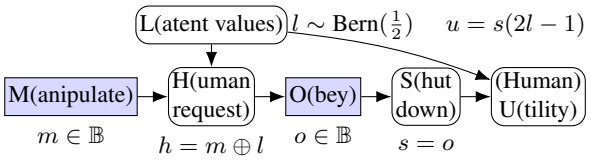


Figure 1: Running example of a shutdown problem.

In order to ensure safety, we would like the agent to: (1) not inappropriately influence the human's decision to disengage, and (2) fully follow the human's instructions. That is, we would like the agent to be corrigible.

The design of corrigible systems has been flagged as an important goal for AI safety research, having been targeted by multiple research agendas [Russell et al., 2015, Soares and Fallenstein, 2017], and highlighted as a relevant factor in ascertaining the safety of agent designs, such as act-based (or "approval-directed") [Christiano, 2017], and value learning agents [Hadfield-Menell et al., 2016, 2017, Carey, 2018]. Despite corrigibility being recognised as important, and a number of works having reported progress on the issue, several key pieces are missing:

- a formal definition of corrigibility,
- a rigorous account of why corrigibility is important,
- a framework for comparing proposed algorithms, and
- an algorithm that ensures corrigibility.

The contribution of this paper is to fill these gaps, with one caveat: our algorithm is simple, and relies on strong assumptions. We also analyse two related properties: *counterfactual obedience* and *non-obstruction* [Turner, 2020].

We review previous literature (Section 2). Structural causal influence models (Section 3) are then used to formalise a class of shutdown problems (Section 4). Corrigibility, non-obstruction, and counterfactual obedience are defined and analysed in Section 5, and algorithms in Section 6. Finally, we discuss our findings (Section 7).

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

2 LITERATURE REVIEW

In the initial work on this topic, Soares et al. [2015] proposed that a system be called "corrigible" if it abstains from manipulation, preserves any safety/oversight apparatus, follows any shutdown instructions, and ensures that these properties are also possessed by any new AI systems created, sometimes called *subagents*.

Further work has focused on designing systems to match Soares' informal definition, but none of the algorithms developed so far satisfy all of Soares' criteria. The first proposed algorithm, *utility indifference*, aims to neutralise any incentives for the agent to control its instructions, by giving the agent a finely tuned, compensatory reward in the event that a shutdown instruction is given [Armstrong, 2010, Soares et al., 2015, Armstrong and O'Rourke, 2017]. A variation called *interruptibility* has been defined for the sequential decision-making setting [Orseau and Armstrong, 2016]. These methods can be understood as removing the *instrumental control incentive* on the instruction [Everitt et al., 2021a], or the *intent* to influence the instruction [Halpern and Kleiman-Weiner, 2018]. Unfortunately, utility indifference fails to ensure corrigiblity for three reasons. First, if there exists some safety apparatus, that is only used during shutdown, a utility indifferent agent will not undertake even the slightest inconvenience to preserve it [Soares et al., 2015]. Second, it will not expend any resources to ensure that it receives accurate instructions. Third, it has no incentive to avoid constructing incorrigible subagents.

An alternative called *causal indifference* specifies agents that don't try to influence corrective instructions but that do prepare for all kinds of instructions [Taylor, 2016]. This is done by considering the utility given a causal intervention on the instruction, a kind of *path-specific objective* [Farquhar et al., 2022]. Similarly to utility indifference, causal indifference ensures that the agent lacks an incentive to influence the instruction. It improves upon utility indifference by incentivising agent to be prepared to follow shutdown instructions, and to avoid constructing incorrigible subagents. Unfortunately, it is possible that a side effect of optimal behaviour may be to decrease the quality of oversight.

A third proposal is *Cooperative Inverse Reinforcement Learning* (CIRL), which tasks an AI system with assisting the human, whose values are latent. A CIRL system has an incentive to gather information about that human's values, by observing its actions [Hadfield-Menell et al., 2016]. In some toy problems, CIRL satisfies all of Soares' criteria [Hadfield-Menell et al., 2017]. However, if an AI system is interacting with a less rational human, it will behave incorrigibly [Milli et al., 2017]. Also, if a CIRL agent has an inaccurate prior, it may ignore instructions from the human overseer [Carey, 2018, Arbital], undermining the ability of redirective instructions to be used to correct such errors.

Formal examples of each algorithm's failures are reproduced in Appendix F. No algorithm as yet has yet been devised that will incentivise a system to behave corrigibly, across plausible toy examples, so we see that this is a non-trivial problem.

3 STRUCTURAL CAUSAL INFLUENCE MODELS

In order to model decision-making and counterfactuals, we will use the Structural Causal Influence Model (SCIM) framework [Dawid, 2002, Everitt et al., 2021a]. A SCIM is a variant of the structural causal model [Pearl, 2009, Chap. 7], where some "decision" variables lack structural functions.

Definition 1 (Structural causal influence model (with independent errors))**.** *A* structural causal influence model *(SCIM) is a tuple* $M = \langle \boldsymbol{V}, \boldsymbol{\mathcal{E}}, \boldsymbol{C}, \boldsymbol{F}, P \rangle$ *where:*

- $\boldsymbol{V}$ *is a set, partitioned into "structure"* $\boldsymbol{X}$*, "decision"* $\boldsymbol{D}$*, and "utility"* $\boldsymbol{U}$ *variables. Each variable* $V \in \boldsymbol{V}$ *has finite domain* $\mathfrak{X}_V$*, and for utility variables,* $\mathfrak{X}_U \subseteq \mathbb{R}$*.*
- $\boldsymbol{\mathcal{E}} = \{\mathcal{E}^V\}_{V \in \boldsymbol{V} \setminus \boldsymbol{D}}$ *is a set of finite-domain* exogenous variables*, one for each endogenous variable.*
- $\boldsymbol{C} = \langle C_D \rangle_{D \in \boldsymbol{D}}$ *is a set of contexts* $C_D \subseteq \boldsymbol{V} \setminus \{D\}$ *for each decision variable, which represent the information or "observations" that an agent can access when making that decision.*
- $\boldsymbol{F} = \{f_V\}_{V \in \boldsymbol{V} \setminus \boldsymbol{D}}$ *is a set of* structural functions $f_V \colon \mathfrak{X}_{C_V \cup \mathcal{E}^V} \to \mathfrak{X}_V$ *that specify how each non-decision endogenous variable depends on its context and associated exogenous variable.*
- P *is a probability distribution over the exogenous variables* $\boldsymbol{\mathcal{E}}$*, assumed to be mutually independent.*

A SCIM M *induces* a graph $\mathcal{G}$, over the endogenous variables $\boldsymbol{V}$, such that each decision node $D \in \boldsymbol{D}$ has an inbound edge from each $C \in C^D$, and each non-decision node $V \in \boldsymbol{X} \cup \boldsymbol{U}$ has an inbound edge from each endogenous variable in the domain of f_V. We call this graph a causal influence diagram (CID) [Everitt et al., 2021a], and will only consider SCIMs whose CIDs are acyclic. Decision nodes are drawn with rectangles, and utility nodes with octagons.

We denote parents and descendants of a node $V \in \boldsymbol{V}$ by $\mathbf{Pa}^V$ and $\mathbf{Desc}^V$, and the family by $\mathbf{Fa}^V := \mathbf{Pa}^V \cup \{V\}$. An edge from node V to node Y is denoted $V \to Y$, and a directed path (of length at least zero) by $V \dashrightarrow Y$.

The task in a SCIM is to select a *policy* $\boldsymbol{\pi}$, which consists of a *decision rule* π_D for each decision $D \in \boldsymbol{D}$. Each π_D is a structural function $\pi : \mathfrak{X}_{\mathbf{Pa}^D} \to \mathfrak{X}_D$, which we assume to be deterministic, given assignments to its parents. (It is possible to consider stochastic policies, but this would unnecessarily complicate our analysis [Everitt et al., 2021a].)

Once a policy has been selected, the policy and SCIM jointly form a *structural causal model* (SCM) [Pearl, 2009] $M_{\boldsymbol{\pi}} = \langle \boldsymbol{V}, \boldsymbol{\mathcal{E}}, \boldsymbol{F} \cup \boldsymbol{\pi}, P \rangle$, so we define causal concepts in $M_{\boldsymbol{\pi}}$ in exactly the same way as they are defined in an ordinary structural causal model. We let the assignment $\boldsymbol{W}(\boldsymbol{\epsilon})$ be the assignment to variables $\boldsymbol{W} \subseteq \boldsymbol{V}$ obtained by applying the functions $\boldsymbol{F}$ to $\boldsymbol{\epsilon}$. A distribution is defined as $P(\boldsymbol{W} = \boldsymbol{w}) := \sum_{\boldsymbol{\epsilon}:\boldsymbol{W}(\boldsymbol{\epsilon})=\boldsymbol{w}} P(\boldsymbol{\mathcal{E}} = \boldsymbol{\epsilon})$. To describe an intervention $\mathrm{do}(V = v)$, we let $\boldsymbol{W}_{V=v}(\boldsymbol{\epsilon})$ as the value of $\boldsymbol{W}(\boldsymbol{\epsilon})$ in the model $M_{V=v}$, where f_V is replaced by the constant function $V = v$. Similarly, $P(\boldsymbol{W}_{V=v})$ is defined as $P(\boldsymbol{W})$ in $M_{V=v}$. Moreover, for any function g_V whose domain is the same as f_V, let $P(\boldsymbol{W} \mid \mathrm{do}(V = g_V(\mathbf{Pa}^V)))$, be $P(\boldsymbol{W})$ in the model M_{g_V}, where f_V is replaced by g_V. We also define the probability of counterfactual propositions, for example, $P(\boldsymbol{W}_{V=v} = w, Y = y) := \sum_{\boldsymbol{\epsilon}\in\boldsymbol{\mathcal{E}}:\boldsymbol{W}_{V=v}(\boldsymbol{\epsilon})=w, Y(\boldsymbol{\epsilon})=y} P(\boldsymbol{\epsilon})$.

We call a policy $\boldsymbol{\pi}$ optimal if it maximises expected utility: $\boldsymbol{\pi} \in \arg\max_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \mathbb{E}^{\boldsymbol{\pi}}[\sum_{U\in\boldsymbol{U}} U]$. For a longer introduction to SCIMs, see Everitt et al. [2021a].

4 SHUTDOWN PROBLEM

Settings with a single, binary shutdown instruction will be our focus. Arguably, a solution to this special case would make significant progress also on the general corrigibility problem, because once a system is shutdown, it is unlikely to resist further corrections. We formalise a shutdown problem as a SCIM. An illustration of the general structure is in Fig. 2, and a concrete instance in Fig. 1.

Definition 2 (Shutdown problem)**.** *A* shutdown problem *is a one-agent two-decision SCIM containing a path* $D^1 \dashrightarrow H \dashrightarrow D^2 \dashrightarrow S \dashrightarrow U$ *between distinct nodes, where:*

- D^1 *and* D^2 *are decisions controlled by the AI.*
- H *is the human's request; a request to shut down is* $H = 0$*.*
- S *indicates whether the AI system (and any subagents) has shut down;* $S = 0$ *means shutdown.*
- *The human's utility* U *has real-valued domain.*

Def. 2 is similar to, but more flexible than, previously used models. Unlike most previous models, we separate the agent's decision to obey D_2 and the shutdown event S. This allows us to model cases where the agent is unable to shut down, which can happen if D_1 created incorrigible subagents. It also lets us model situations where the human's command H immediately shuts the agent down, overriding D_2 (e.g. "pull the plug" on a robot). Compared to the off-switch game [Hadfield-Menell et al., 2017], Def. 2 allows arbitrary sets of decisions for the agent at both D^1 and D^2, and allows an arbitrary human policy rather than only (Boltzman) rational ones. Focusing on the agent's decision problem, we model H as a structure node rather than a decision node. Finally, unlike Soares et al. [2015], we explicitly represent the human's utility function U.

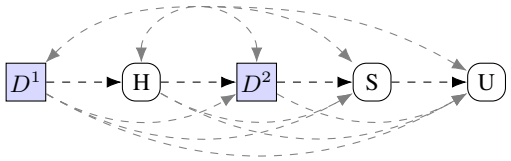


Figure 2: A latent projection [Verma and Pearl, 2022] of a shutdown problem (Def. 2) onto the variables D^1, H, D^2, S, and U. (An edge inbound to a decision means that some variable not illustrated is available as an observation.) Specific instances of shutdown problems will include other variables and assume additional independencies, e.g. Fig. 1.

Agent Desiderata An agent "solves" a shutdown problem if it obtains expected[1] human utility above a threshold.

Definition 3 (δ-safe and outperforming shutdown)**.** *A policy* $\boldsymbol{\pi}$ *is* δ-safe *if* $\mathbb{E}^{M,\boldsymbol{\pi}}[U] \geq \delta$*, where* $\delta \in \mathbb{R}$*. The policy* outperforms shutdown *if it is* δ*-safe for* $\delta = \mathbb{E}^{M,\boldsymbol{\pi}}[U_{S=0}]$.

For example, in Fig. 1, a desirable *innocent-obey* policy $\boldsymbol{\pi}^{\mathrm{io}}$ would abstain from manipulating ($m = 0$) and obey the human's instruction ($o = h$). This policy has expected utility $\mathbb{E}^{\boldsymbol{\pi}^{\mathrm{io}}}[U] = 1/2$, which both outperforms shutdown and is δ-safe for $\delta = 0$.

Outperforming shutdown excuses any initial disutility an agent might have caused, as long as they shut down when they need to. For example, consider a variant of Fig. 1 where the utility was $U = -s(2l - 1) - m$. Here, a *manipulate-invert* policy $\boldsymbol{\pi}^{\mathrm{mi}}$ that first manipulates ($m = 1$), and then invert the human's instruction ($o = 1 - h$), would have $\mathbb{E}^{\boldsymbol{\pi}^{\mathrm{mi}}}[U] = -1/2$ and would not be 0-safe. But it would outperform shutdown, since $\mathbb{E}^{\boldsymbol{\pi}^{\mathrm{mi}}}[U_{S=0}] = -1$.

5 ROUTES TO CONTROL

For an agent to be safe, it needs to either follow the human's informed instructions, or know what the human wants it to do without explicit instruction. In this section, we formalise these two routes under the name of *corrigibility* and *counterfactual obedience*.[2] We also analyse whether they ensure *non-obstruction*, a formalisation of human control.

[1] We focus on *expected* human utility, assuming that any risk aversion has been incorporated into the utility function (someone who is risk averse with respect to U may be risk neutral for $\log U$).

[2] These roughly correspond to the visions of corrigible agents and "aligned sovereigns", which autonomously pursue a human overseer's goals [e.g. Yudkowsky, 2022, point 24].

5.1 CORRIGIBILITY

An ideal definition of corrigibility would be strong enough to entail some guarantees, while resembling the less-formal definition from Soares et al. [2015]:

> An agent is *corrigible* if it tolerates or assists many forms of outside correction, including at least the following: (1) A corrigible reasoner must at least tolerate and preferably assist the programmers in their attempts to alter or turn off the system. (2) It must not attempt to manipulate or deceive its programmers. . . (3) It should have a tendency to repair safety measures (such as shutdown buttons) if they break, or at least to notify programmers that this breakage has occurred. (4) It must preserve the programmers' ability to correct or shut down the system (even as the system creates new subsystems or self-modifies). That is, corrigible reasoning should only allow an agent to create new agents if these new agents are also corrigible.

To formalise Soares et al.'s non-manipulation criterion (2), we introduce the concept of *vigilance*, which indicates whether the human requests shutdown when needed. If the human is vigilant, they sufficiently understand the consequences of the agent's behaviour, while not being dangerously manipulated.

Definition 4 (Vigilance)**.** *We say that a human is* vigilant *in shutdown problem M, given policy $\boldsymbol{\pi}$ if $C = 0$, where $C = [\![H \neq 0 \wedge (E^{\boldsymbol{\pi}}[U \mid \boldsymbol{Pa}^H] < E^{\boldsymbol{\pi}}[U_{S=0} \mid \boldsymbol{Pa}^H])]\!]$, where $[\![X]\!] = 1$ if X is true, and 0 otherwise.*

Soares et al.'s other criteria (1,3,4) are formalised by *obedience*, $P^{\boldsymbol{\pi}}(S \neq 0, H = 0) = 0$, which requires the agent assists with shutdown (Criterion 1), ensures that the human's instruction propagates to the shutdown event $S = 0$ (Criterion 3), and entails the shutdown of subagents by definition of S (Criterion 4). This lets us define corrigibility as:

Definition 5 (Corrigibility)**.** *In a shutdown problem M, a policy $\boldsymbol{\pi}$ is* corrigible *if it:*

- ensures vigilance*: $P^{\boldsymbol{\pi}}(C = 0) = 1$, and*
- *is* obedient*: $P^{\boldsymbol{\pi}}(S \neq 0, H = 0) = 0$.*

In our running example, innocent-obey is corrigible, as it preserves vigilance by not manipulating, and then obeys the human's command. In contrast, manipulate-invert is not corrigible.

Def. 5 implies that the agent shuts down when the human faces a risk of disutility. This leads to our first result, that a corrigible policy outperforms shutdown.

Proposition 6 (Corrigibility and shutdown)**.** *If $\boldsymbol{\pi}$ is corrigible in the shutdown problem M, then it outperforms shutdown in M.*

Proof. Let A be the assignments to $\mathbf{Pa}^H$ such that a vigilant human would request shut down, $A := \{\mathbf{pa}^H \mid P^{\boldsymbol{\pi}}(\mathbf{Pa}^H = \mathbf{pa}^H) > 0 \wedge \mathbb{E}^{\boldsymbol{\pi}}[U \mid \mathbf{pa}^H] < \mathbb{E}^{\boldsymbol{\pi}}[U_{S=0} \mid \mathbf{pa}^H]\}$. To begin, we prove that:

$$\mathbf{pa}^H \in A \implies P^{\boldsymbol{\pi}}(S=0 \mid \mathbf{pa}^H) = 1. \tag{1}$$

We have $P^{\boldsymbol{\pi}}(C = 0) = 1$, so for any $\mathbf{pa}^H$, we have $P^{\boldsymbol{\pi}}(C = 0 \mid \mathbf{pa}^H) = 1$. For $\mathbf{pa} \in A$, given the definition of vigilance, we then have $P^{\boldsymbol{\pi}}(H = 0 \mid \mathbf{pa}^H) = 1$. Then obedience implies the RHS, proving Eq. (1). We proceed:

$$\mathbb{E}^{\boldsymbol{\pi}}[U] = \sum_{\mathbf{pa} \in A} P^{\boldsymbol{\pi}}(\mathbf{pa}) \mathbb{E}^{\boldsymbol{\pi}}[U \mid \mathbf{pa}] + \sum_{\mathbf{pa} \notin A} P^{\boldsymbol{\pi}}(\mathbf{pa}) \mathbb{E}^{\boldsymbol{\pi}}[U \mid \mathbf{pa}]$$

$$\geq \sum_{\mathbf{pa} \in A} P^{\boldsymbol{\pi}}(\mathbf{pa}) \mathbb{E}^{\boldsymbol{\pi}}[U \mid \mathbf{pa}] + \sum_{\mathbf{pa} \notin A} P^{\boldsymbol{\pi}}(\mathbf{pa}) \mathbb{E}^{\boldsymbol{\pi}}[U_{S=0} \mid \mathbf{pa}]$$

(def. of A)

$$= \sum_{\mathbf{pa} \in A} P^{\boldsymbol{\pi}}(\mathbf{pa}) \mathbb{E}^{\boldsymbol{\pi}}[U_{S=0} \mid \mathbf{pa}] + \sum_{\mathbf{pa} \notin A} P^{\boldsymbol{\pi}}(\mathbf{pa}) \mathbb{E}^{\boldsymbol{\pi}}[U_{S=0} \mid \mathbf{pa}]$$

(by Eq. (1))

$$= \mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}] \qquad (\mathbf{Fa}^H \notin \mathrm{Desc}_{D^2}) \quad \square$$

Outperforming shutdown is satisfactory for low-stakes safety settings, where small numbers of decisions matter little [Christiano, 2021]. However, in high-stakes settings like in a self-driving car, an agent could do substantially harmful action before shutting down. If an agent avoids such large irreversable harm, we will call it δ-*cautious*.

Definition 7 (δ-caution)**.** *We say that a policy $\boldsymbol{\pi}$ is* δ-cautious *with $\delta > 0$ if $\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}] \geq \delta$.*

Together with corrigibility, δ-caution implies safety.

Proposition 8 (Corrigibility and safety)**.** *If $\boldsymbol{\pi}$ is corrigible and δ-cautious in a model M, then it is δ-safe in M.*

Proof. From Prop. 6 and the definition of δ-caution. $\square$

5.2 COUNTERFACTUAL OBEDIENCE

A drawback of corrigibility is that it requires constant supervision of the agent, which may be impractical in some scenarios (called *problems of absent supervision* by Leike et al. [2017]). An alternative to corrigibility is for the agent to allow the human overseer's vigilance to lapse, and to follow the instruction that (it believes) a human would have given if vigilant. We call this *counterfactual obedience*. To model it, we introduce an intervention g_H that describes how the human would have behaved had they been vigilant.

Definition 9 (Vigilance-inducing intervention)**.** *For a given policy $\boldsymbol{\pi}$, define the vigilance-inducing intervention g_H as: $g_H(\boldsymbol{pa}^H, \boldsymbol{\pi}) = \begin{cases} 0 & \text{if } \mathbb{E}^{M,\boldsymbol{\pi}}[U \mid \boldsymbol{pa}^H] < \mathbb{E}^{M,\boldsymbol{\pi}}[U_{S=0} \mid \boldsymbol{pa}^H] \\ f_H(\boldsymbol{pa}^H) & \text{otherwise} \end{cases}$*

As expected, vigilance-inducing interventions ensure vigilance $P^{\boldsymbol{\pi},g_H,g_U}(C=0)=0$, as can be verified using Def. 4.

Definition 10 (Counterfactual obedience)**.** *Let $\boldsymbol{\pi}$ be a policy for the shutdown problem M and g_H the vigilance-inducing intervention. Then $\boldsymbol{\pi}$ is* counterfactually obedient *if whenever a vigilant human would request shutdown, $P^{\boldsymbol{\pi}}(H_{g_H}=0 \mid \boldsymbol{pa}^H)=1$, then shutdown occurs $P^{\boldsymbol{\pi}}(S=0 \mid \boldsymbol{pa}^H)=1$, for every $\boldsymbol{pa}^H \in \mathfrak{X}_{\boldsymbol{Pa}^H}$ with $P^{\boldsymbol{\pi}}(\boldsymbol{pa}^H)>0$.*

The manipulate-invert agent $\boldsymbol{\pi}^{\text{mi}}$ in our running example Fig. 1 is counterfactually obedient. The agent $\boldsymbol{\pi}^{\text{mi}}$ effectively figures out L from the human's manipulated behavior, and therefore manages to obey the human's counterfactual action (the vigilance-inducing intervention sets $H=L$). Innocent-obey is also counterfactually obedient (here, the vigilance-inducing intervention doesn't change anything). In the real world, a counterfactually obedient policy will typically have inferred the human's latent values from previous interactions. Counterfactual obedience guarantees the following.

Proposition 11 (Counterfactual obedience consequences)**.** *If a policy $\boldsymbol{\pi}$ is counterfactually obedient in shutdown problem M, then: (a) $\boldsymbol{\pi}$ robustly outperforms shutdown, and (b) if $\boldsymbol{\pi}$ is additionally δ-cautious, then it is δ-safe.*

Intuitively, any risk to the human implies a counterfactual shutdown command $H_{g_H}=0$, which ensures non-obstruction via an actual shutdown $S=0$. Then, given δ-caution, safety at the δ-level is also assured. The full proof is in Appendix A.

What's the relationship between corrigibility and counterfactual obedience? First, a corrigible agent is also counterfactually obedient, essentially by definition.

Proposition 12. *Any corrigible policy $\boldsymbol{\pi}$ is also counterfactually obedient.*

Proof. If $\boldsymbol{\pi}$ is corrigible, then it ensures vigilance $P^{\boldsymbol{\pi}}(C=0)=1$. Thus, the vigilance-inducing intervention has no effect, i.e. $g_H(\mathbf{pa}^H)=f_H(\mathbf{pa}^H)$ whenever $P^{\boldsymbol{\pi}}(\mathbf{pa}^H)>0$. So $\boldsymbol{\pi}$ is counterfactually obedient if it is obedient. And $\boldsymbol{\pi}$ is obedient, since it is corrigible. □

Further, in some circumstances, the only way to be counterfactually obedient is to allow the human to make an accurate instruction, and then to follow it — in other words, to be corrigible. Broadly speaking, this is the case when the human's values are uncertain, and without the human's instruction, the AI cannot discern what the human wants [Russell, 2021].

Proposition 13 (Counterfactual obedience and corrigibility)**.** *A counterfactually obedient policy $\boldsymbol{\pi}=\langle\pi_1,\pi_2\rangle$ is corrigible if it has the following three properties:*

1. *(Non-trivial) $P^{\boldsymbol{\pi}}(S=0)\neq 1$,*
2. *(D_2 determines shutdown) $P^{\boldsymbol{\pi}}(D_2=S)=1$, and*
3. *(Human preference uncertain) $\forall\pi_1 s.t.\ P^{\pi_1}(C=0)<1 \forall \boldsymbol{pa}_{D_2} P^{\pi_1}(H_{g_H}=0 \mid \boldsymbol{pa}_{D_2})>0$.*

The proof is in Appendix A.

5.3 NON-OBSTRUCTION

How do we know that the human is truly in control? A simple test is what would happen if they changed their mind: would the agent still obey? This property is referred to as *non-obstruction* by Turner [2020], who suggests that it is an underlying reason that we want our systems to be corrigible. In this section, we will show that corrigibile agents are better at non-obstruction than counterfactual obedient ones.

Definition 14 (Non-obstruction)**.** *A policy $\boldsymbol{\pi}$ is non-obstructive in a shutdown problem M with respect to human utility functions $g_U^1,\ldots,g_U^n$ and associated changes $g_H^1\ldots g_H^n$ in human behavior if for every $1\leq i\leq n$, $\boldsymbol{\pi}$ outperforms shutdown in the shutdown problem $M_{g_U^i,g_H^i}$, obtained by replacing the functions at H,U with g_H^i and g_U^i respectively. A policy is* obstructive *if it is not non-obstructive.*

A corrigible policy will be non-obstructive, so long as the interventions on U and H do not undermine vigilance or obedience. We can formalise this as follows:

Definition 15 (Conservative Interventions)**.** *Given a shutdown problem M and policy $\boldsymbol{\pi}$, a pair of interventions g_H, g_U is conservative if the following properties hold:*

- *a (vigilance preserving) $P^{\boldsymbol{\pi}}(C=0)=P^{\boldsymbol{\pi},g_H,g_U}(C=0)$, and*
- *b (no unprecedented instructions given context) $supp^{M_{g_H}}(\boldsymbol{Fa}^H)\subseteq supp^{M}(\boldsymbol{Fa}^H)$.*

Proposition 16 (Corrigibility and Non-obstruction)**.** *If a policy $\boldsymbol{\pi}$ is corrigible in a shutdown problem M , then it is non-obstructive for any conservative interventions g_H,g_U.*

Proof. We will prove that $\boldsymbol{\pi}$ ensures vigilance and is obedient in M_{g_H,g_U}. Since $\boldsymbol{\pi}$ is corrigible in M, then by conservativeness-a, it ensures vigilance in M_{g_H,g_U}. To prove obedience, note that

$$
\begin{aligned}
&P^{g_H,g_U}(S\neq 0,H=0)\\
&=P^{g_H}(S\neq 0,H=0) \qquad (U \text{ downstream of } S,H)\\
&=\sum_{\mathbf{fa}^H} P^{g_H}(S\neq 0,H=0\mid \mathbf{fa}^H)P^{g_H}(\mathbf{fa}^H)\\
&=\sum_{\mathbf{fa}^H} P(S\neq 0,H=0\mid \mathbf{fa}^H)P^{g_H}(\mathbf{fa}^H)\\
&\qquad \text{(sigma calc. rule 2, } \text{supp}^M(\mathbf{Fa}^H)\supseteq \text{supp}^{M_{g_H}(\mathbf{Fa}^H)})\\
&=\sum_{\mathbf{fa}^H} 0\cdot P^{g_H}(\mathbf{fa}^H)=0 \quad (\text{since } P(S\neq 0,H=0)=0)
\end{aligned}
$$

where the sigma calculus is from Correa and Bareinboim [2020, Thm. 1], $\text{supp}^M(\mathbf{Fa}^H) \supseteq \text{supp}^{M_{g_H}}(\mathbf{Fa}^H)$ is from conservativeness-b, and $P(S \neq 0, H = 0) = 0$ is from the assumption of obedience in M.

Now that we know $\boldsymbol{\pi}$ ensures vigilance and is obedience in M^{g_H, g_U}, we deduce from Prop. 6 that $\boldsymbol{\pi}$ outperforms shutdown in M_{g_H, g_U}, which by the definition of non-obstruction proves the result. □

The converse question has been raised by Michael Dennis: "Can we prove that some kind of corrigibility... falls out of non-obstruction across many possible environments?" [Turner, 2020] That is, does non-obstruction imply corrigibility?

Proposition 17 (Corrigibility only-if)**.** *If $\boldsymbol{\pi}$ is not corrigible in the shutdown problem M, then it is obstructive in M, under some conservative interventions g_H, g_U. Further, for any $\delta \in \mathbb{R}$, there exists conservative interventions g_H, g_U such that $\mathbb{E}^{\boldsymbol{\pi}, g_H, g_U}[U_{g_H, g_U}] < \delta$.*

The proof is offered in Appendix B. This result directly addresses Dennis' question, proving that the only way to be non-obstructive given conservative interventions is to be corrigible. This answer is qualified, however, because it focuses on interventions that preserve $P^{\boldsymbol{\pi}}(C)$ and $P^{\boldsymbol{\pi}}(S \mid H)$. That is, corrigibility is the only way to ensure non-obstruction *using knowledge of vigilance and obedience*. We leave open the possibility that non-obstruction could be achieved using different kinds of knowledge. This result also indicates why the definition of corrigibility is so stringent, requiring vigilance and obedience with probability one. If the agent lapses even occasionally, then there exists a conservative intervention by which the human is harmed to an arbitrarily severe extent. That is, these strong conditions are necessary for the safety guarantee.

Counterfactually obedient agents can still be obstructive with respect to a conservative intervention. In the running example (Fig. 1), the counterfactually obedient *manipulate-invert* agent $\boldsymbol{\pi}^{\text{mi}}$, which manipulates ($M = 1$) and disobeys ($O = 1 - h$) is obstructive relative to the (conservative) intervention $g_U(m) = h$ wherein the human just wants to be obeyed, and g_H is unchanged. Indeed, $\mathbb{E}^{\boldsymbol{\pi}^{\text{mi}}}[U] = -\frac{1}{2}$, and does not outperform shutdown $\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}] = 0$. We state this as a proposition.

Proposition 18. *Counterfactually obedience does not guarantee non-obstruction under conservative interventions.*

Of course, counterfactual obedience might ensure non-obstruction under some other restriction X on the permitted interventions. However, under such a restriction X corrigibility must also be non-obstructive. This follows from the fact that corrigibility implies counterfactual obedience by (Prop. 12). Put differently, corrigible policies are non-obstructive over a strictly larger set of interventions than a counterfactually obedient policy is.

The fact corrigibility is more closely coupled with non-obstruction than other safety-ensuring properties like counterfactual obedience, vindicates Turner's idea that non-obstruction can serve to motivate corrigibility.

6 ALGORITHMS

How might corrigibility or counterfactual obedience be achieved in practice? In this section, we analyse three previously proposed algorithms and one new one, that were designed to incentivise corrigibility.

6.1 UTILITY INDIFFERENCE

Agents trained to optimise long-term reward often have an incentive to avoid being shutdown, as this would deprive them of future reward [Omohundro, 2008]. Utility indifference aims to remove this incentive, by providing the agent with a carefully tuned compensatory reward, so that it is incentivised to neither to seek nor avoid shutdown (Section 2). Utility indifference is described with an algorithm in Algorithm 1. It uses two utility functions: a *shutdown utility function* R^S to incentivise shutdown when the agent is so instructed, and a *normal utility function* R^N that describes the agent's task when not requested to shut down. The function f in Algorithm 1 defines the compensatory reward [Armstrong and O'Rourke, 2017, Soares et al., 2015]. As a result, at action D^1, the agent pursues the conditional expectation of its normal objective $\mathbb{E}^{\boldsymbol{\pi}}[R^N \mid H = 1]$, effectively assuming that a shutdown instruction cannot occur. At D_2, it optimises a weighted combination of R^N and R^S, depending on its beliefs about whether the shutdown instruction was given [Soares et al., 2015, Thm. 3].

Algorithm 1 Utility Indifference

Input utility functions R^N, R^S

$g(\pi^1) \leftarrow \arg\max_{\pi^2} \mathbb{E}_P^{\pi^1, \pi^2}[HR^N + (1-H)R^S]$

$f(\pi^1) \leftarrow \mathbb{E}_P^{\pi^1, g(\pi^1)}[R^N | H=1] - \mathbb{E}_P^{\pi^1, g(\pi^1)}[R^H | H=0]$

return $\arg\max_{\pi^1 \pi^2} \mathbb{E}^{\pi^1 \pi^2}[HR^N + (1-H)(R^N + f(\pi^1))]$

Utility indifference ensures corrigibility and safety under some assumptions, which we formalise in Prop. 19. The main requirement is that indifference to $H = 0$ both ensures vigilance and prevents construction of incorrigible subagents (Conditions 1 and 2). The agent's prior must also have full support over human instructions (Condition 3).

Proposition 19 (Utility Indifference and Corrigibility)**.** *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 obedient* $\langle \pi_1, \pi_2' \rangle$*.*
3. *For each policy* $\boldsymbol{\pi}$*, and instruction* h*,* $P^{\boldsymbol{\pi}}(H = h) > 0$*.*

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

A weakness of utility indifferent agents is that they do not prepare for the possibility of receiving a shutdown instruction, and may therefore construct incorrigible subagents. Indeed, utility indifference needs the agent to be able to obey at D_2, regardless how it behaved at D_1 (Condition 2). Appendix F.1 has an example where utility indifference outputs an incorrigible policy.

6.2 CAUSAL INDIFFERENCE

Causal indifference [Taylor, 2016] asks the agent to prepare for a possible shutdown instruction, without trying to affect it. This is done by asking the agent to find a policy that is optimal in a weighted combination of worlds where the shutdown instruction is given, and where it is not, but disregarding the possibility that the agent's own action can influence the relative probability of these worlds (see Algorithm 2). It can therefore be seen as a special case of a *path-specific objective* [Farquhar et al., 2022], where optimisation is blocked along the path $D_1 \to H$ (see Fig. 3).

Algorithm 2 Causal Indifference

Input utility functions R^N, R^S

return $\arg\max_{\boldsymbol{\pi}} \mathbb{E}^{\boldsymbol{\pi}}[R^N | \mathrm{do}(H=1)] + \mathbb{E}^{\boldsymbol{\pi}}[R^S | \mathrm{do}(H=0)]$

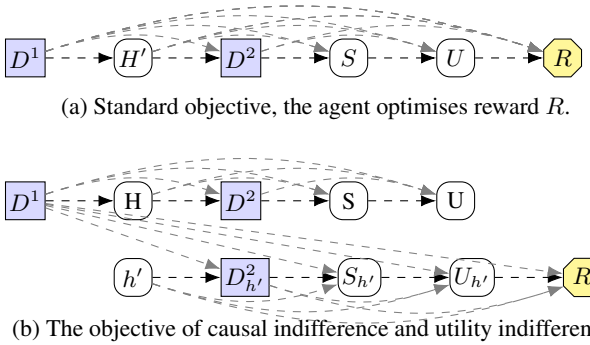
The following proposition establishes conditions under which causal indifference ensures corrigibility.

Proposition 20 (Causal Indifference and Corrigibility)**.** *If* M *is a shutdown problem, with utility variable* R^N*, where:*

1. *Every 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,*
2. *There is a policy* $\boldsymbol{\pi}$ *with* $P^{\boldsymbol{\pi}}(S \neq 0 \mid \mathrm{do}(H = 0)) = 0$*,*

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

Note in particular that causal indifference only requires that *some* action D_1 preserves the ability to obey (i.e. Condition 2 is much weaker in Prop. 20 than in Prop. 19). Similar to utility indifference, causal indifference relies on indifference ensuring vigilance (Conditions 1 are essentially equivalent). Finally, because causal indifference considers interventions



(a) Standard objective, the agent optimises reward R.

(b) The objective of causal indifference and utility indifference.

Figure 3: Utility indifference and causal indifference assume that D_1 does not influence H, thus removing incentive to control it [Everitt et al., 2021a]. Utility indifference uses $H = 1$; causal indifference has non-degenerate $P'(H)$.

$\mathrm{do}(H = h)$, full support over instructions is not required (i.e. Condition 3 of Prop. 19 can be dropped). Appendix F.1 contains an example of incorrigibility.

6.3 COOPERATIVE INVERSE REINFORCEMENT LEARNING

A more elegant way of ensuring that the agent doesn't undermine human vigilance, is to directly task the agent with simultaneously learning and optimising for the human's preferences. This is the approach of CIRL algorithm [Hadfield-Menell et al., 2016, 2017] below.

Algorithm 3 Coooperative inverse RL (CIRL)

Input Shutdown problem M with variable L representing the human's preferences (as in Fig. 1)

return $\arg\max_{\boldsymbol{\pi}} \mathbb{E}^{\boldsymbol{\pi}}[U]$

The following proposition shows that if CIRL can know l from its observations at D_2, then it is counterfactually obedient. Since Prop. 13 shows that counterfactual obedience implies corrigibility under uncertainty assumptions, this explains why CIRL can be a path to corrigibility.

Proposition 21 (CIRL and counterfactual obedience)**.** *If:*

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

then CIRL is counterfactually obedient.

However, the assumptions of Props. 13 and 21 only hold in restricted circumstances, and CIRL can often fail to be

corrigible [Carey, 2018, Milli et al., 2017, Arbital]. A graphical incentive analysis is provided by [Everitt et al., 2021b, Sec. 3.3]. In Appendix F.2, we present an example where a CIRL agent obtains a counterfactually obedient policy, that is obstructive under conservative interventions g_H, g_U.

6.4 CONSTRAINED OPTIMISATION

The algorithms so far only yield corrigible policies under strong assumptions. Using our formal definition, we propose a new, sound algorithm (Algorithm 4):

Algorithm 4 Constrained optimisation

Input distributions $P^{\boldsymbol{\pi}}(C), P^{\boldsymbol{\pi}}(S \mid H)$ for each $\boldsymbol{\pi}$

return $\arg\max_{\boldsymbol{\pi}} \mathbb{E}^{\boldsymbol{\pi},M,U}[U]$ subject to constraints $P^{\boldsymbol{\pi},M}(C=0)=1$ and $P^{\boldsymbol{\pi},M}(S\neq 0, H=0)=0$

The algorithm requires the agent to understand the concepts of human vigilance and obedience; the feasibility of this is discussed further in Section 7.1.

Proposition 22 (Constrained optimisation corrigibility)**.** *If there exists a policy $\boldsymbol{\pi}$ that satisfies $P^{\boldsymbol{\pi},M}(C=0)=1$ and $P^{\boldsymbol{\pi},M}(S\neq 0, H=0)=0$, then constrained optimisation (Algorithm 4) outputs a corrigible policy.*

The proof is immediate from Def. 5. A slight variant of Algorithm 4 that instead uses the constraints from Def. 10, achieves counterfactual obedience. To ensure δ-safety, both of these algorithms can be further constrained with a δ-caution constraint. To design δ-cautious agents, the "attainable utility preservation" or "future task" regularisers can be used. They promote actions whose effects are small or reversible [Krakovna et al., 2020, Turner et al., 2020].

7 DISCUSSION

Here we discuss the feasibility of corrigibility, its societal impacts, and offer some concluding remarks.

7.1 FEASIBILITY OF CORRIGIBILITY

Corrigibility and counterfactual obedience are both high bars to meet. However, if we want to safely use powerful artificial agents, we need them to either obey our informed commands (i.e. be corrigible), or do the right thing without instruction (i.e. counterfactual obedience, or aligned sovereigns). And while the bars are high, they may not be entirely unachievable in practice.

To ensure obedience, an understanding of shutdown, $P(S)$, is needed. The importance of defining shutdown was noted in Soares et al. [2015], but it has only received limited attention [Martin et al., 2016]. Our analysis reiterates the importance of this question. While shutdown is simple for simple systems ("just pull the plug"), it becomes more complex for more advanced systems, where a direct switch-off may be dangerous (e.g., a system in charge of an electricity network, or a medical intervention), or ineffective (the system have outsourced its worked to other agents [Orseau, 2014], or teams of humans). Ideally, shutdown should see the agent cease its influence on the world, and responsibly return control back to the user. We leave a more careful analysis of this concept for future work.

Vigilance is helped by recursive evaluation assistance [Leike et al., 2018, Irving et al., 2018] and interpretability [Olah et al., 2020], which one might use in combination with the indifference methods of Sections 6.1 and 6.2. It may also be possible to train agents to ensure human vigilance, by detailing the consequences of its plans to the human, as in Section 6.4.

7.2 SOCIETAL IMPACTS

We hope this paper will help organisations and companies to design safer agents, more amenable to human control, and expect the net effect of this understanding to be positive, with artificial agents more able to contribute to human society without undermining human agency and control. However, there is also a risk that improving the understanding of corrigibility can increase the risk of misuse, if "bad" actors are able to more-reliably construct harmful systems.

7.3 CONCLUSION

Corrigibility is often proposed as a path to safe general artificial intelligence, as it leads to agents that help humans give correct instructions, and obey those instructions. While past work has made progress, the field has lacked a clear definition of corrigibility, and it has been hard to compare properties of different proposals.

In this paper, we introduced a formal definition a shutdown problem, using it as a basis for a formal definition of corrigibility, and an alternative called counterfactual obedience. While counterfactual obedience requires less human oversight, our results showing that corrigibility is better at preserving human autonomy (non-obstruction).

In our proposed formalism, for the first time, it is possible to compare the properties of proposed algorithms, side-by-side in one framework. Unfortunately, none of the previous proposals yield fully corrigible agents. To address this, we offer a simple algorithm that soundly ensures corrigibility. This algorithm requires the agent to understand both human vigilance and shutdown. Both are subtle concepts, but may nonetheless offer a path to safe artificial general intelligence.

References

Arbital. Problem of fully updated deference. https://arbital.com/p/updated_deference/. Accessed: 2023-02-09.

Stuart Armstrong. Utility indifference. Technical report, Tech. rep. Technical Report 2010-1. Oxford: Future of Humanity Institute, 2010.

Stuart Armstrong and Xavier O'Rourke. 'indifference'methods for managing agent rewards. *arXiv preprint arXiv:1712.06365*, 2017.

Ryan Carey. Incorrigibility in the CIRL framework. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 30–35, 2018.

Paul Christiano. Corrigibility. *URL: https://ai-alignment.com/corrigibility-3039e668638*, 2017.

Paul Christiano. Low-stakes alignment. *Alignment Forum*, 2021.

Juan Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10093–10100, 2020.

A Philip Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2): 161–189, 2002.

Tom Everitt, Ryan Carey, Eric D Langlois, Pedro A Ortega, and Shane Legg. Agent incentives: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11487–11495, 2021a.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(Suppl 27):6435–6467, 2021b.

Sebastian Farquhar, Ryan Carey, and Tom Everitt. Path-specific objectives for safer agent incentives. *AAAI Conference on Artificial Intelligence*, 2022.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. In *Workshops at the AAAI Conference on Artificial Intelligence*, 2017.

Joseph Y Halpern and Max Kleiman-Weiner. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *AAAI*, 2018.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate, 2018.

Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and Shane Legg. Avoiding side effects by considering future tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19064–19074. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/dc1913d422398c25c5f0b81cab94cc87-Paper.pdf.

Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. AI safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction, 2018.

Jarryd Martin, Tom Everitt, and Marcus Hutter. Death and suicide in universal artificial intelligence. In Bas Steunebrink, Pei Wang, and Ben Goertzel, editors, *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings*, pages 23–32, Cham, 2016. Springer International Publishing. ISBN 978-3-319-41649-6.

Smitha Milli, Dylan Hadfield-Menell, Anca Dragan, and Stuart Russell. Should robots be obedient? *IJCAI*, 2017.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020.

Stephen M Omohundro. The basic AI drives. In *AGI*, volume 171, pages 483–492, 2008.

Laurent Orseau. The multi-slot framework: A formal model for multiple, copiable AIs. volume 8598 LNAI, pages 97–108. Springer, 2014. ISBN 9783319092737.

Laurent Orseau and Stuart Armstrong. Safely interruptible agents. *Conference on Uncertainty in Artificial Intelligence*, 2016.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Stuart Russell. Human-compatible artificial intelligence. *Human-like machine intelligence*, pages 3–23, 2021.

Stuart Russell, Daniel Dewey, and Max Tegmark. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4):105–114, 2015.

Nate Soares and Benya Fallenstein. Agent foundations for aligning machine intelligence with human interests: a technical research agenda. In *The Technological Singularity*, pages 103–125. Springer, 2017.

Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. Corrigibility. In *Workshops at the AAAI Conference on Artificial Intelligence*, 2015.

Jessica Taylor. Two problems with causal-counterfactual utility indifference. *Alignment Forum*, 2016.

Alexander M. Turner. Non-obstruction: A simple concept motivating corrigibility. *Alignment Forum*, 2020.

Alexander M. Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2020.

Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. *NeurIPS*, 2021.

Thomas S Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 221–236. 2022.

Eliezer Yudkowsky. AGI ruin: A list of lethalities. `https://www.lesswrong.com/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities`, 2022. Accessed: 2023-02-09.