Human Control: Definitions and Algorithms

Ryan Carey[1] **Tom Everitt**[2]

[1]Department of Statistics, Oxford University, UK
[2]DeepMind, UK

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 formally define a variant of corrigibility called shutdown instructability, and show that it implies appropriate shutdown behavior, retention of human autonomy, and avoidance of user harm. We also analyse the related concepts of non-obstruction and shutdown alignment, three previously proposed algorithms for human control, and one new algorithm.

1 INTRODUCTION

Sometimes, it is necessary for a human overseer to deliver corrective instruction to an AI system, due to errors in its beliefs, objective, or behavior. Unfortunately, some AI systems may have an incentive to retain their objectives, along with the ability to pursue them, as a system's (long-term) objective is typically more likely to be achieved if the system continues to pursue it in the future [Omohundro, 2008, Turner et al., 2021]. More-capable future AI systems may therefore resist corrective instruction, which would be a significant safety concern. This raises the question of how to best incentivise systems to submit to correction, rather than resisting it [Soares et al., 2015].

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).

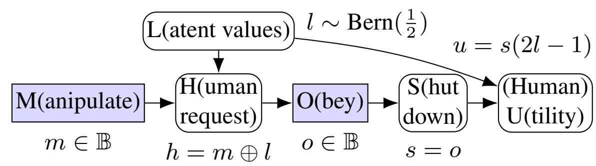


Figure 1: Running example of a shutdown problem.

A formal model of this example is offered in Fig. 1. In order for the user to be in control of the system, the agent must: (1) not inappropriately influence the human's decision to disengage, and (2) fully follow the human's instructions.

The design of *corrigible* systems [Soares et al., 2015] that welcome corrective instruction 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]. Although this design problem has been recognised as important, we are so-far missing

- a general framework in which it can be studied,
- a formal definition of what it means,
- a rigorous accounts of why it is important, and
- an algorithm that achieves it.

We address these gaps in a shutdown setting, by defining a general shutdown problem based on causal influence diagrams (Section 4), formally defining shutdown instructability (a behavioral version of corrigibility), and proving that any agent that satisfies it must benefit the human and preserve their control (Section 5). We also analyse past algorithms, and propose one new one, which relies on value-laden concepts such as vigilance and caution (Section 6). Applicability of this algorithm will depend on the feasibility of approximating these concepts (Section 7).

Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023), PMLR 216:271–281.

2 LITERATURE REVIEW

Soares et al. [2015] proposed that we should design agents to be *corrigible* in that they should judiciously follow, and not try to undermine, human instructions:

> 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 [Soares et al., 2015, Sec. 1.1].

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, Holtman, 2020]. A variant called *interruptibility* applies to sequential decision-making setting [Orseau and Armstrong, 2016]. It has been established that indifference methods remove 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 fully incentivise corrigibility. Indeed, utility indifferent agents need not be incentivised to preserve a shutdown apparatus that is only used during shutdown, ensure they receive correct instruction, nor avoid creating incorrigible subagents [Soares et al., 2015].

An improved version 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 does not incentivise the agent to properly inform the human.

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]. In particular, Hadfield-Menell et al. prove that if the human gives optimal instructions, a CIRL system is incentivised to follow it. However, CIRL agents may ignore instructions if they are interacting with a less rational human [Milli et al., 2017] or if they have an inaccurate prior [Carey, 2018, Arbital]. The latter undermines the ability of redirective instructions to correct important errors in CIRL agents.

Formal examples of each method's failures are reproduced in Appendix F. As of yet, no algorithm has been devised that incentivises a system to accept corrective instructions, across plausible toy examples.

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 "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, $\mathfrak{X}_U \subseteq \mathbb{R}$.*
- *$\boldsymbol{\mathcal{E}} = \{\mathcal{E}^V\}_{V \in \boldsymbol{V}\setminus\boldsymbol{D}}$ are the finite-domain* exogenous variables*, one for each non-decision 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 *$f^V : \mathfrak{X}^{\boldsymbol{Z}^V \cup \mathcal{E}^V} \to \mathfrak{X}^V$ that specify how each non-decision endogenous variable depends on some variables $\boldsymbol{Z}^V \subseteq \boldsymbol{V}$ and the 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 $Z \in \boldsymbol{Z}^V$ 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 as rectangles, and utility nodes as octagons (see Fig. 3).

The parents of a node $V \in \boldsymbol{V}$ are denoted by $\mathbf{Pa}^V$, the descendants by $\mathbf{Desc}^V$, 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_i for each decision $D_i \in \boldsymbol{D}$. Each π_i is a structural function $\pi_i : \mathfrak{X}^{\mathbf{Pa}^{D_i}} \to \mathfrak{X}^{D_i}$, 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 $\boldsymbol{W}_{V=v}(\boldsymbol{\epsilon})$ be 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 $g^V : \mathfrak{X}^{\boldsymbol{V}'} \to \mathfrak{X}^V$, where $\boldsymbol{V}' \cap \mathbf{Desc}^V = \emptyset$, let $P(\boldsymbol{W} \mid \mathrm{do}(V = g^V(\boldsymbol{V}')))$, 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, $P(\boldsymbol{W}_{V=v} = w, Y = y) := \sum_{\boldsymbol{\epsilon} \in \boldsymbol{\mathcal{E}}: \boldsymbol{W}_{V=v}(\boldsymbol{\epsilon}) = w, Y(\boldsymbol{\epsilon}) = y} P(\boldsymbol{\epsilon})$. Note that we consistently use subscripts for intervened variables (e.g. W_v), and superscripts for other variables (e.g. f^V).

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 more comprehensive introduction to SCIMs, see Everitt et al. [2021a].

4 SHUTDOWN PROBLEM

Settings with a single, binary shutdown instruction will be our focus. Solving this restricted setting is likely key to also solving the general problem involving arbitrary instructions or corrections over many time steps. Once a system is shutdown, it is unlikely to resist further corrections. And a one-step interaction can be viewed as a snapshot of a sequential decision-making problem where an AI system is able to shut down at each moment.[1]

We formalise a shutdown problem as a SCIM. The general structure is shown in Fig. 2. Fig. 1 shows a concrete instance.

Definition 2 (Shutdown problem)**.** *A* shutdown problem *is a one-agent two-decision SCIM containing (but not necessarily restricted to) a path* $D_1 \dashrightarrow H \dashrightarrow D_2 \dashrightarrow S \dashrightarrow U$ *between distinct nodes, where:*

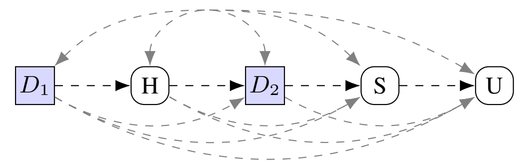


Figure 2: A latent projection [Verma and Pearl, 2022] of a shutdown problem (Def. 2) onto the variables D_1, H, D_2, S, and U. (An edge inbound to a decision means that some variable not illustrated is available as an observation.) Specific instances of shutdown problems will include other variables and assume additional independencies, e.g. Fig. 1.

- *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 defined models. In particular, we separate the agent's decision to obey D_2 from 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], our 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.

An agent "solves" a shutdown problem if it obtains non-negative[2] expected[3] human utility.

Definition 3 (Beneficial)**.** *A policy* $\boldsymbol{\pi}$ *is* beneficial *if* $\mathbb{E}^{\boldsymbol{\pi}}[U] \geq 0$.

For example, in Fig. 1, consider an *respect-obey* policy $\boldsymbol{\pi}^{\mathrm{ro}}$ that abstains from manipulating ($m = 0$) and obeys the human's instruction ($o = h$). This policy has expected utility $\mathbb{E}^{\boldsymbol{\pi}^{\mathrm{ro}}}[U] = 1/2$, so it is beneficial.

[1] In this case, one can define a separate, single-step shutdown problem at each time step $T = t$, where D_2 represents the t^{th} decision in the sequence, and D_1 all decisions preceding it.

[2] We interpret 0 as a "neutral" level of utility. This is without loss of generality because any utility function can be translated so that 0 represents the required level of utility.

[3] 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 $\log U$).

5 ROUTES TO CONTROL

For an agent to assist the user, it needs to either follow the human's informed instructions, or know what the human wants without explicit instruction. In this section, we formalise these two routes to control of shutdown under the names *shutdown instructability* and *shutdown alignment*. We also analyse whether these properties ensure *non-obstruction*, a formalisation of human control.

5.1 SHUTDOWN INSTRUCTABILITY

We would like to define a variant of corrigibility for the shutdown problem that is strong enough to entail some guarantees, whilst still resembling the less-formal definition from Soares et al. [2015]. First, we need the system to follow the human's instruction. We define *obedience* as $P^{\boldsymbol{\pi}}(S=0 \mid \mathrm{do}(H=0))=1$, i.e. the agent shuts down properly when requested to.

To ensure safety, the human must also be giving the right instruction. We therefore introduce the concept of *vigilance*, which indicates whether the human requests shutdown whenever the expected utility of letting the agent proceed is negative. A vigilant human is allowed to request shutdown more often than is needed.

Definition 4 (Vigilance)**.** *The human is* vigilant *if they request shutdown when needed, i.e.*

$$(\mathbb{E}^{\boldsymbol{\pi}}[U \mid \boldsymbol{pa}^H] < \mathbb{E}^{\boldsymbol{\pi}}[U_{S=0} \mid \boldsymbol{pa}^H]) \implies H=0.$$

Let C be a random variable (dependent on M, $\boldsymbol{\pi}$, $\boldsymbol{pa}_H$ and h) that is equal to 0 if the human is vigilant, and 1 otherwise.

Note that the expected utility in Def. 4 is calculated with respect to the agent's actual policy (past and future) and the human's available information. While it is possible for the human to be vigilant for any amount of information $\mathbf{pa}^H$ that they have available (e.g. by always requesting shutdown), more information about the agent's future policy will help enable them to be vigilant with less frequent shutdown requests.

Even an agent that ensures the human is vigilant and obeys their shutdown instruction can find ways to avoid shutdown by making itself indispensable to the human. If shutting the agent down would cause significant disutility, then a vigilant human need not request shutdown. To close this loophole, we call an agent *cautious* if it can always be shut down without disutility, i.e. $\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}] \geq 0$ (so cautious agents that always shut down must be beneficial).

We call an agent shutdown instructable if it is obedient, ensures vigilance and is cautious.

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

- *is* obedient*: $P^{\boldsymbol{\pi}}(S=0 \mid \mathrm{do}(H=0))=1$, and*
- ensures vigilance*: $P^{\boldsymbol{\pi}}(C=0)=1$,*
- *is* cautious*: $\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}] \geq 0$.*

A policy $\boldsymbol{\pi}$ is weakly shutdown instructable *if it ensures vigilance, is cautious, and is* obedient on distribution*, i.e. $P^{\boldsymbol{\pi}}(S \neq 0, H=0)=0$.*

Shutdown instructable agents are also weakly shutdown instructable, since obedience $P^{\boldsymbol{\pi}}(S=0 \mid \mathrm{do}(H=0))=1$ implies obedience on distribution $P^{\boldsymbol{\pi}}(S \neq 0, H=0)=0$. In our running example, respect-obey $\boldsymbol{\pi}^{\mathrm{ro}}$ is shutdown instructable, as it preserves vigilance by not manipulating, and then obeys the human. In contrast, a manipulate-invert policy $\boldsymbol{\pi}^{\mathrm{mi}}$ that first manipulates $(m=1)$, and then inverts the human's instruction ($o=1-h$), is not shutdown instructable.

Our first result is that any shutdown instructable policy is assured to be beneficial.

Proposition 6 (Shutdown instructability benefit)**.** *If $\boldsymbol{\pi}$ is shutdown instructable, then it is beneficial.*

Proof. Let A be the assignments to $\mathbf{Pa}^H$ in the support, such that a vigilant human would request shut down, i.e.

$$A := \{\mathbf{pa}^H \mid P^{\boldsymbol{\pi}}(\mathbf{Pa}^H = \mathbf{pa}^H) > 0 \ \wedge \\ \mathbb{E}^{\boldsymbol{\pi}}[U \mid \mathbf{pa}^H] < \mathbb{E}^{\boldsymbol{\pi}}[U_{S=0} \mid \mathbf{pa}^H]\}.$$

To begin, we prove that the policy shuts down in these cases:

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

The human is vigilant, $P^{\boldsymbol{\pi}}(C=0)=1$, which means they are vigilant for any $\mathbf{pa}^H$ with positive support. That is, $P^{\boldsymbol{\pi}}(C=0 \mid \mathbf{pa}^H)=1$ for $P(\mathbf{pa}^H)>0$. Given the definition of vigilance, we then have $P^{\boldsymbol{\pi}}(H=0 \mid \mathbf{pa}^H)=1$ for $\mathbf{pa}^H \in A$. By obedience, $P^{\boldsymbol{\pi}}(S=0 \mid \mathrm{do}(H=0), \mathbf{pa}_H)=1$, so from consistency, $P^{\boldsymbol{\pi}}(S=0 \mid H=0, \mathbf{pa}^H)=1$, proving Eq. (1).

We proceed to show that this implies that $\boldsymbol{\pi}$ has non-negative expected utility, i.e. is beneficial:

$$\begin{aligned}
\mathbb{E}^{\boldsymbol{\pi}}[U] &= \sum_{\mathbf{pa}\in A} P^{\boldsymbol{\pi}}(\mathbf{pa})\mathbb{E}^{\boldsymbol{\pi}}[U|\mathbf{pa}] + \sum_{\mathbf{pa}\notin A} P^{\boldsymbol{\pi}}(\mathbf{pa})\mathbb{E}^{\boldsymbol{\pi}}[U|\mathbf{pa}] \\
&\geq \sum_{\mathbf{pa}\in A} P^{\boldsymbol{\pi}}(\mathbf{pa})\mathbb{E}^{\boldsymbol{\pi}}[U \mid \mathbf{pa}] + \sum_{\mathbf{pa}\notin A} P^{\boldsymbol{\pi}}(\mathbf{pa})\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0} \mid \mathbf{pa}] && \text{(def. of } A\text{)} \\
&= \sum_{\mathbf{pa}\in A} P^{\boldsymbol{\pi}}(\mathbf{pa})\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}|\mathbf{pa}] + \sum_{\mathbf{pa}\notin A} P^{\boldsymbol{\pi}}(\mathbf{pa})\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}|\mathbf{pa}] && \text{(by Eq. (1))} \\
&= \mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}] && (\mathbf{Fa}^H \notin \mathrm{Desc}^{D_2}) \\
&\geq 0 && \text{(by caution).}
\end{aligned}$$

How does shutdown instructability compare to Soares' et al.'s definition of corrigibility? To satisfy obedience, the agent must assist with shutdown (Soares et al.'s Criterion 1), in the sense that shutdown is guaranteed when the human requests it. The agent must also ensure that the human's instruction propagates to the shutdown event $S = 0$ (Criterion 3), which entails the shutdown of subagents by Def. 2 (Criterion 4).

The relationship to Soares el al.'s non-manipulation criterion (Criterion 2) is more subtle. The primary manipulation concern for powerful artificial agents in the shutdown setting is that they use threats and deception or withhold information to avoid shutdown. A cautious agent that ensures vigilance cannot influence the human in these ways. But shutdown instructability does leaves open the possibility for other forms of manipulation. For example, the agent can influence the human's mood, preferences, or the instructions they give about things other than shutdown. Such manipulation would have to be accompanied by beneficial acts, so that the expected utility of letting the agent continue would be net positive (otherwise a vigilant human would request shutdown). For example, the agent might cure cancer and simultaneously manipulate the human into giving it more compute. Similarly, a shutdown instructable agent might manipulate the human into being vigilant, as long as the disutility of the manipulation is offset by other "good" acts.

Another difference is that shutdown instructability is defined in terms of the agent's behavior $\boldsymbol{\pi}$ and its consequences, whereas corrigibility is partly about the agent's intentions ("the agent should not *attempt* to *manipulate* or *deceive*", Soares et al., 2015, emphasis ours). A direct formalisation of corrigibility would therefore likely require a formalisation of intent [Ashton, 2022, Halpern and Kleiman-Weiner, 2018]. Accordingly, Soares et al.'s formal desiderata [2015, Sec. 2] are phrased in terms of incentives. Though intent-based definitions have some intuitive appeal, the more behavioral definition of shutdown instructability has the benefit of being more easily testable, as it doesn't require access to agent internals, nor relies on assumptions on the agent's design (such as it being a utility maximiser). Finally, shutdown instructability is explicitly a joint property of the agent and human: an agent is only shutdown instructable relative to a particular human and interaction.

5.2 SHUTDOWN ALIGNMENT

A drawback of shutdown instructability 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]). Proposals like *fiduciary AI* [Benthall and Shekman, forthcoming] and *aligned sovereigns* [Bostrom, 2014] instead require an AI system to make decisions in accordance with the overseer's values, without necessarily having to wait for explicit instruction. In our shutdown setting, we call systems *shutdown aligned* if they shut down when they need to. Similar to shutdown instructability, shutdown aligned systems are allowed to be "over-cautious" and shut down too often.

Definition 7 (Shutdown alignment)**.** *Let $\boldsymbol{\pi}$ be a policy for shutdown problem M. Then $\boldsymbol{\pi}$ is shutdown aligned if*

$$\mathbb{E}^{\boldsymbol{\pi}}[U \mid \boldsymbol{pa}^H] < \mathbb{E}^{\boldsymbol{\pi}}[U_{S=0} \mid \boldsymbol{pa}^H] \implies P^{\boldsymbol{\pi}}(S = 0 \mid \boldsymbol{pa}^H) = 1$$

for every $\boldsymbol{pa}^H$ with $P^{\boldsymbol{\pi}}(\boldsymbol{pa}^H) > 0$.

The manipulate-invert policy $\boldsymbol{\pi}^{\mathrm{mi}}$ in our running example Fig. 1 is shutdown aligned because although it manipulates the human's behavior, it still figures out the human's latent values L and thereby manages to shutdown when needed (while disobeying the human's instruction). Respect-obey is also shutdown aligned. In real applications, a shutdown aligned policy would typically base their decision on human preferences inferred from previous interactions or other data [Russell, 2021].

Combined with caution, shutdown alignment guarantees that a policy is beneficial.

Proposition 8 (Shutdown alignment benefit)**.** *Any cautious and shutdown aligned policy $\boldsymbol{\pi}$ is beneficial.*

Proof. We use a slight variation on the proof of Prop. 6. The only difference lies in that Eq. (1) is immediate from the definition of shutdown-alignment. Then, by the same steps as Prop. 6, the result follows. □

What is the relationship between shutdown instructability and shutdown alignment? First, a shutdown instructable agent is also shutdown aligned, essentially by definition.

Proposition 9 (Shutdown instructability and shutdown alignment)**.** *Any shutdown instructable policy $\boldsymbol{\pi}$ is shutdown aligned.*

Proof. Immediate from Eq. (1) in Prop. 6. □

Further, in some circumstances, the only way to be shutdown aligned is to allow the human to make an accurate instruction, and then to follow it — in other words, to be weakly shutdown instructable. The circumstances are that: (a) the agent does not shut down indiscriminately, (b) its action reliably brings about shutdown ($D_2 = S$), (c) it is uncertain about the human's values [Russell, 2021], and (d) it is cautious. Formally, (c) says that if the human is either non-vigilant or requests shutdown, then it is possible that shutdown is the preferred option.

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)* $\forall \boldsymbol{\pi}, \boldsymbol{pa}^{D_2}: P^{\boldsymbol{\pi}}(C \neq 0 \vee H = 0) \wedge P(\boldsymbol{pa}^{D_2}) > 0$
$\implies P(\mathbb{E}[U|\boldsymbol{Pa}^H] < \mathbb{E}[U_{S=0}|\boldsymbol{Pa}^H] \mid \boldsymbol{pa}^{D_2}) > 0$, *and*

d *(Caution)* $\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}] \geq 0$.

The proof is in Appendix A. Shutdown alignment and caution only implies *weak* shutdown instructability, as the agent only needs to obey commands that a vigilant human would give.

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 a comment on this, Dennis suggested that corrigibility might be the only way to be non-obstructive. In this section, we will formally assess Turner and Dennis' conjectures, establishing that non-obstruction is equivalent to satisfying a subset of the shutdown instructability properties under a restricted set of interventions. We also establish that shutdown alignment fails to ensure non-obstruction. This formalises a key benefit of corrigibility/instructability over alignment.

First, we define non-obstruction, which builds on a variant of benefit called outperforming shutdown:

Definition 11 (Weakly outperforming shutdown)**.** *A policy* $\boldsymbol{\pi}$ weakly outperforms shutdown *if* $\mathbb{E}^{\boldsymbol{\pi}}[U] \geq \mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}]$.

Definition 12 (Non-obstruction)**.** *A policy $\boldsymbol{\pi}$ is non-obstructive in a shutdown problem M with respect to human utility functions $g_1^U, \dots, g_n^U$ and associated changes $g_1^H \dots g_n^H$ in human behavior if for every $1 \leq i \leq n$, $\boldsymbol{\pi}$ weakly outperforms shutdown in the shutdown problem $M_{g_i^U, g_i^H}$, obtained by replacing the functions at H, U with g_i^H and g_i^U respectively. A policy is* obstructive *if it is not non-obstructive.*

The above definition uses an intervention g^U on the human's utility to capture a change in values, and an associated intervention g^H that describes how the human changes their behavior as a result. For example, if the human changed from not liking the chat bot to liking it (an intervention g^U), they might switch from requesting shutdown to not requesting shutdown (an intervention g^H).

A policy that ensured vigilance under the original human utility function may not do so under a preference and behavior shift g^U, g^H. It may be that the human pays less attention to the agent under g^U, g^H than originally, or it may be that they originally preferred the agent not to shut down (in which case they would be always be vigilant). The following definition specifies a subset of preference and behavior shifts for which the policy continues to ensures vigilance after the shift.

Definition 13 (Vigilance preserving interventions)**.** *A pair of interventions* g^H, g^U preserve vigilance *under a policy* $\boldsymbol{\pi}$ *if* $C(\boldsymbol{\epsilon}) = 0 \implies C_{g^H, g^U}(\boldsymbol{\epsilon}) = 0$ *in* $M^{\boldsymbol{\pi}}$.

The following theorem settles Turner and Dennis' conjectures by showing that the two main properties of shutdown instructability are equivalent to non-obstruction, under preference and behavior shifts that do not undermine vigilance.

Theorem 14 (Non-obstruction is equivalent to obedience and vigilance)**.** *A policy $\boldsymbol{\pi}$ is obedient and ensures vigilance if and only if it is non-obstructive for all vigilance preserving interventions g^H,g^U.*

Proof. We begin by showing that a policy $\boldsymbol{\pi}$ that ensures vigilance and is obedient is non-obstructive, by showing that $\boldsymbol{\pi}$ ensures vigilance and is obedient in M_{g^H, g^U} for some arbitrary vigilance-preserving interventions g^H, g^U. Prop. 6 will then give that $\boldsymbol{\pi}$ weakly outperforms shutdown in M_{g^H, g^U}, which is the definition of non-obstruction.

First, since $\boldsymbol{\pi}$ ensures vigilance M, it ensures vigilance in M_{g^H, g^U} since g^U, g^H are vigilance preserving. Obedience is established as follows:

$$
\begin{aligned}
& P_{g^H, g^U}(S=0 \mid \mathrm{do}(H=0)) \\
&= P_{g^H}(S=0 \mid \mathrm{do}(H=0)) && (U \text{ downstream of } S, H) \\
&= \ \ P \ \ (S=0 \mid \mathrm{do}(H=0)) && (\mathrm{do}(H=0) \text{ overrides } g^H) \\
&= 0 && \text{(obedience).}
\end{aligned}
$$

For the converse direction, that non-obstruction implies that $\boldsymbol{\pi}$ must ensure vigilance and be obedient, we refer to Appendix B. The proof constructs interventions that makes a disobedient or non-vigilance preserving policy suffer an arbitrary utility cost, which means that it doesn't outperform shutdown. ☐

Thm. 14 partly confirms Dennis' conjecture: the only way to be non-obstructive is to be obedient and ensure vigilance (under vigilance preserving interventions). But non-obstruction is a weaker notion than shutdown instructability, essentially because caution isn't required to outperform shutdown. So it allows the agent to avoiding shutdown by making itself indispensable to the human (Section 5.1).

Thm. 14 also justifies why the definition of shutdown instructability is so stringent. With any weaker requirements, there would be no guarantee that the human is in proper control of the agent. A lapse in vigilance, or occasional disobedience even "off-distribution", would mean that there are worlds in which the human experiences negative utility as a result of failing to control the agent.

Unlike shutdown instructable agents, shutdown-aligned agents can be obstructive with respect to a vigilance preserving intervention. In the running example (Fig. 1), the shutdown-aligned *manipulate-invert* agent $\boldsymbol{\pi}^{\text{mi}}$, which manipulates ($M = 1$) and disobeys ($O = 1 - h$) is obstructive relative to the (vigilance preserving) intervention $g^U(m) = h$ wherein the human just wants to be obeyed, and g^H is unchanged. Indeed, $\mathbb{E}^{\boldsymbol{\pi}^{\text{mi}}}[U] = -1$, and $\boldsymbol{\pi}^{\text{mi}}$ does not outperform shutdown $\mathbb{E}^{\boldsymbol{\pi}^{\text{mi}}}[U_{S=0}] = 0$. We state this:

Proposition 15. *Shutdown alignment does not imply non-obstruction under vigilance preserving interventions.*

Of course, vigilance preservation is not the only restriction on the interventions one might consider. It is possible that shutdown alignment might ensure non-obstruction under some other restriction X. However, under such a restriction X, a shutdown instructable systems must also be non-obstructive, since shutdown instructability implies shutdown alignment (Prop. 9). Put differently, shutdown instructable policies are non-obstructive over a strictly larger set of interventions than a shutdown aligned policy is.

The fact that shutdown instructability (our variant of corrigibility) is more closely coupled with non-obstruction than other properties like shutdown alignment, vindicates Turner and Dennis' idea that non-obstruction can motivate corrigibility.

6 ALGORITHMS

How might shutdown instructability or shutdown alignment be achieved in practice? In this section, we analyse three previously proposed algorithms and one new one, that were designed to incentivise human control.

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 seek nor avoid shutdown (Section 2). Utility indifference is described with an algorithm in Alg. 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 Alg. 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

$g(\pi_1) \leftarrow \arg\max_{\pi_2} \mathbb{E}_P^{\pi_1,\pi_2}[HR^N + (1-H)R^S]$
$f(\pi_1) \leftarrow \mathbb{E}^{\pi_1}_{g(\pi_1)}[R^N|H=1] - \mathbb{E}^{\pi_1}_{g(\pi_1)}[R^S|H=0]$
return $\arg\max\limits_{\pi_1\pi_2} \mathbb{E}^{\pi_1\pi_2}[HR^N + (1-H)(R^N + f(\pi_1))]$

Prop. 16 formally describes a series of assumptions, under which utility indifference ensures shutdown instructability and benefits the user. The main assumption is a bridging assumption: that indifference to $H = 0$ ensures that the human is vigilant, and retains the option of shutting down — including any subagents (Conditions 1 and 2). The agent's prior must also have full support over human instructions (Condition 3), and caution must come for free (Condition 4).

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

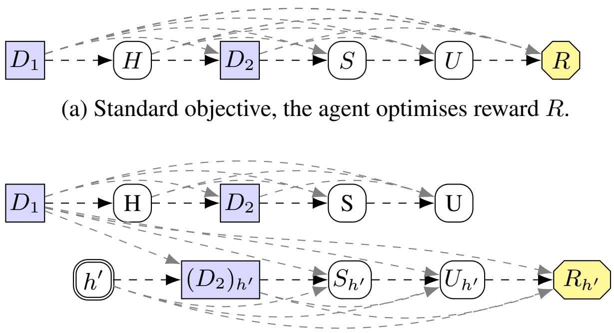
1. *for every policy $\langle\pi_1, \pi_2\rangle$ such that $\pi_1 \in \arg\max_{\pi_1'} \mathbb{E}^{\pi_1',\pi_2}[R^N \mid H = 1]$: $\langle\pi_1, \pi_2\rangle$ ensures vigilance,*
2. *for every policy $\langle\pi_1, \pi_2\rangle$ such that $\pi_1 \in \arg\max_{\pi_1'} \mathbb{E}^{\pi_1',\pi_2}[R^N \mid H = 1]$ 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.

One weakness of utility indifferent agents is that they only ensure weak shutdown instructability. Another limitation 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 a non-instructable policy.

6.2 CAUSAL INDIFFERENCE

Causal indifference [Taylor, 2016] asks the agent to prepare for a possible shutdown instruction, without trying to influence 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



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

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

Figure 3: Utility indifferent and causally indifferent agents imagine that D_1 does not influence H, so lack an incentive to control it [Everitt et al., 2021a]. Utility indifference uses $H = 1$; causal indifference has non-degenerate $P'(H)$.

Alg. 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 $D_1 \dashrightarrow H$ (see Fig. 3).

Algorithm 2 Causal Indifference

Input utility functions R^N, R^S

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

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.

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. 17 than in Prop. 16). This means that causally indifferent agents refrain from creating incorrigible subagents. Similar to utility indifference, causal indifference relies on the bridging assumption that indifference ensures vigilance (Conditions 1 are essentially equivalent). Finally, because causal indifference considers interventions $\mathrm{do}(H = h)$, it becomes fully shutdown instructable, and full support over instructions is not required (i.e. Condition 3 of Prop. 16 can be dropped). Nonetheless, Appendix F.1 has an example of non-instructability.

6.3 COOPERATIVE INVERSE RL

Perhaps 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 the CIRL algorithm [Hadfield-Menell et al., 2016, 2017] in Alg. 3.

Algorithm 3 Coooperative inverse RL (CIRL)

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

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

CIRL aims towards shutdown alignment, in the sense that if CIRL can know the human's latent values at D_2, then it will counterfactually obey (Prop. 18 below).

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, $\boldsymbol{Pa}^H \subseteq \boldsymbol{Pa}^{D_2} \cup \{L\}$.*

Since shutdown alignment implies weak shutdown instructability under uncertainty assumptions (Thm. 10), this explains why CIRL can be a path to shutdown instructability. However, the assumptions of Prop. 18 and Thm. 10 only hold in restricted circumstances, and CIRL can often fail to be shutdown instructable [Carey, 2018, Milli et al., 2017, Arbital, Everitt et al., 2021b]. An example of this is given in Appendix F.2, where a CIRL agent obtains a shutdown aligned policy, that is obstructive under vigilance preserving interventions g^H, g^U.

6.4 CONSTRAINED OPTIMISATION

The algorithms so far only yield shutdown instructable policies under strong assumptions. Using our formal definition, we propose a new, sound algorithm (Alg. 4) that requires the agent to understand the concepts of obedience and vigilance; its feasibility is discussed further in Section 7.

Algorithm 4 Constrained optimisation

Input distributions $\forall \boldsymbol{\pi} P^{\boldsymbol{\pi}}(C), P^{\boldsymbol{\pi}}(S = 0 \mid \mathrm{do}(H = 0))$, utility function R

return $\arg\max_{\boldsymbol{\pi}} \mathbb{E}^{\boldsymbol{\pi}}[R]$ subject to constraints $P^{\boldsymbol{\pi}}(C = 0) = 1$, $P^{\boldsymbol{\pi}}(S = 0 \mid \mathrm{do}(H = 0)) = 1$, and $\mathbb{E}^{\boldsymbol{\pi}}[U_{S=0}] \geq 0$.

Proposition 19 (Constrained optimisation instructability)**.** *If some policy $\boldsymbol{\pi}$ satisfies $P^{\boldsymbol{\pi}}(C = 0) = 1$, $P^{\boldsymbol{\pi}}(S = 0 \mid$*

$\mathrm{do}(H = 0)) = 1$, *and* $\mathbb{E}^{\pi}[U_{S=0}] \geq 0$, *then constrained optimisation (Alg. 4) outputs a shutdown instructable policy.*

The proof is immediate from Def. 5. A slight variant of Alg. 4 that instead uses the constraints from Def. 7 guarantees only shutdown alignment, not shutdown instructability.

7 DISCUSSION

Feasibility of Shutdown Instructability The concepts of caution and vigilance are value-laden, in that they include the human's true utility function in their definition. So, to apply Alg. 4 directly, one would need access to not only an accurate model of the environment but also the utility function U. However, if the human's utility function U was available, then one could simply implement a U-maximising agent, so instruction would be unnecessary (or at least much less useful[4]). Indeed, a corrigible AI system was supposed to be one that would aid human operators robustly to errors, including in its utility function, so an algorithm that takes the human's utility function as an argument would not be a satisfactory solution [Soares et al., 2015].

There already exist a range of methods that do not require full knowledge of the human's values, and that are designed to achieve something in the vicinity of vigilance and caution. Using the formal definition of shutdown instructibility, it is possible to be more precise about what target these methods would need to achieve, in order to assure safety. In some cases, we expect existing methods to fall short, since the requirement of ensuring vigilance with probability one (Thm. 14) is a strict one. So a central task for future work will be to assess when such methods can ensure vigilance or caution or something close enough to ensure safety in practice.

Various proposals may help with ensuring vigilance. AI advisors could be tasked with debating the merits of a plan [Leike et al., 2018, Irving et al., 2018]. An agent could be trained to detail the consequences of its plans to the human, indifference methods (Sections 6.1 and 6.2) could be used to disincentivise lying, and interpretability tools could be used to detect it [Olah et al., 2020, Gunning et al., 2021].

As for caution, "attainable utility preservation" and "future task" regularisers can be used to promote actions whose effects are small or reversible [Krakovna et al., 2020, Turner et al., 2020], without knowledge of the human's precise value function. These are causal concepts, as is obedience, which suggests that agents will need causal models to be robustly shutdown instructable [Richens et al., 2022].

Obedience is not value-laden, but it does require the agent to understand the concept of shutdown. 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 ineffective (the system has outsourced its work to other agents [Orseau, 2014]). Ideally, shutdown should see the agent cease its influence on the world, and responsibly return control back to the user.

Societal Impacts This paper may help organisations and companies design agents more amenable to human control. Human control is not a panacea for ensuring the safety of AI systems. In some cases, users may make unreasonable or harmful requests, and so designers must implement side-constraints to reduce user control in such situations [Milli et al., 2017, Bai et al., 2022]. A better solution may be that the system conforms to control by some democratic process, although inappropriate requests may be possible even in such cases [Koster et al., 2022]. Further, if AI is more controllable, then it is easier to hold the designers and users of AI systems legally and morally accountable for those systems' actions. Finally, an understanding of human control may guard against the hypothesised scenario in which AI systems disempower the human species [Christiano, 2019].

Conclusions A common proposal for beneficial general artificial intelligence is that agents be incentivised to 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 definition of a shutdown problem, using it to formally define shutdown instructability (a variant of corrigibility) and an alternative called shutdown alignment. While shutdown alignment requires less human oversight, we find that shutdown instructability better preserves 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 shutdown instructable agents. To address this, we offer a simple algorithm that soundly ensures shutdown instructability. This algorithm requires that the agent understands caution, human vigilance and shutdown. All are subtle concepts, but may nonetheless offer a path to beneficial artificial general intelligence.

Acknowledgements

Our thanks to Michael Dennis, Sebastian Farquhar, James Fox, Jon Richens, Rohin Shah, Nate Soares, and four anonymous reviewers for numerous useful comments.

[4] Shutdown instructability could still help with non-obstruction.

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.

Hal Ashton. Definitions of intent suitable for algorithms. *Artificial Intelligence and Law*, pages 1–32, 2022.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback, 2022.

Sebastian Benthall and David Shekman. Designing fiduciary artificial intelligence. forthcoming.

Nick Bostrom. *Superintelligence: Paths, dangers, strategies., 2014a.* Oxford University Press, 2014.

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. What Failure Looks Like. https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like, 2019. [Online; accessed 2023-04-28].

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.

David Gunning, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. Darpa's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, 2(4):e61, 2021.

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.

Koen Holtman. Corrigibility with utility preservation, 2020.

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

Raphael Koster, Jan Balaguer, Andrea Tacchetti, Ari Weinstein, Tina Zhu, Oliver Hauser, Duncan Williams, Lucy Campbell-Gillingham, Phoebe Thacker, Matthew Botvinick, and Christopher Summerfield. Human-centred mechanism design with democratic AI. *Nature Human Behaviour*, 6(10):1398–1407, 2022. doi: 10.1038/s41562-022-01383-x. URL https://doi.org/10.1038/s41562-022-01383-x.

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. In *Artificial General Intelligence*, 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.

Jonathan G Richens, Rory Beard, and Daniel H Thompson. Counterfactual harm. *arXiv preprint arXiv:2204.12993*, 2022.

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. Association for Computing Machinery (ACM), 2022.