Revisiting the shutdown problem

Abstract

A key premise in leading arguments for existential risk from artificial intelligence is that malfunctioning artificial agents could not be easily shut down. This motivates the catastrophic shutdown problem of ensuring that agents can be shut down before they cause an existential catastrophe. A range of arguments and theorems are offered to suggest that solving the catastrophic shutdown problem is difficult, bolstering arguments for existential risk and motivating a search for solutions to the catastrophic shutdown problem. This paper argues for two conclusions. First, existing arguments do not establish the difficulty of solving the catastrophic shutdown problem. Second, concern for the catastrophic shutdown problem has led to technical solutions that impose a high safety tax on model performance.

1 Introduction

Philosophers (Bostrom 2013; MacAskill 2022; Ord 2020), scientists (Bengio et al. 2024; Grace et al. 2022; Russell 2019), and policymakers (Manancourt et al. 2023; Prime Minister's Office 2023) voice increasing concern that artificial intelligence may soon pose an existential risk to humanity. It is argued that powerful agents may soon be developed (Bostrom 2014; Chalmers 2010) which could be power-seeking (Bostrom 2012; Carlsmith 2025) and deceptive (Park et al. 2024; Ngo and Bales 2025), engage in problematic reward-hacking (Dung 2023; Skalse et al. 2022), or misgeneralize goals that performed well during training, with catastrophic effect (Bales et al. 2024; Langosco di Langosco et al. 2022). Existential risk concerns are used to drive research and funding in fields such as AI safety (Amodei et al. 2016; Bengio et al. 2026; D'Alessandro and Kirk-Giannini 2025) and philosophy (Bales et al. 2024; Kasirzadeh 2025; Tubert and Tiehen 2024), to motivate open letters (Center for AI Safety 2023; Future of Life Institute 2023) and legislation (California State Legislature 2024; 117th Congress 2022), and to support philanthropic and philosophical programs such as longtermism (Greaves et al. 2025; Greaves and MacAskill 2021; MacAskill 2022).

A natural objection to these concerns is that misbehaving artificial agents could be shut down. To this, it is responded that shutting down artificial agents may not be as easy as it appears (Neth 2025; Turner et al. 2021; Russell 2019). This motivates the shutdown problem of designing agents that show appropriate shutdown behaviors (Hadfield-Menell et al. 2017; Soares et al. 2015; Thornley 2024).

At least two literatures have grown up around the shutdown problem. One cluster of work uses the shutdown problem to motivate concerns about existential risk (Lynch et al. 2025; Russell 2019; Schlatter et al. 2026). A second develops technical strategies for solving the shutdown problem by ensuring that agents show appropriate shutdown behaviors (Hadfield-Menell et al. 2016; Goldstein and Robinson 2025; Thornley et al. 2025).

This paper contributes to both discussions. Engaging with the first cluster, I argue that existing informal (Section 3) and formal (Sections 4-5) presentations of the shutdown problem do not significantly strengthen existential risk concerns. Engaging with the

second cluster, I show how reflection on the sources and consequences of shutdown-resistance can help to avoid costly technical solutions which impose a high safety tax on model performance, pushing instead towards less costly solutions that conserve technical and regulatory resources to meet other safety challenges (Section 6). The result is a weakening of traditional arguments for existential risk, coupled with concrete guidance for technical AI safety solutions (Section 7).

2 Clarifying the dialectic

Before beginning, let us pause to clarify the dialectic.

2.1 The shutdown problem

The first order of business is to clarify the shutdown problem. Nate Soares and colleagues (2015) originally framed the shutdown problem broadly, as the challenge of generating *corrigible* agents that:

(S1) Tolerate or assist programmers in their attempts to alter or turn them off.

(S2) Do not attempt to manipulate or deceive their programmers.

(S3) Have a tendency to repair safety measures, such as shutdown buttons, if they break.

(S4) Preserve the programmers' ability to correct or shut down the system as the system evolves.

My interest in this paper is with problems in the neighborhood of (S1). Corrigibility incorporates additional desiderata such as non-deception (S2), repair (S3) and preservation (S4) of safety measures, which go beyond the scope of the present discussion.

A leading formulation in the neighborhood of (S1) is due to Elliott Thornley (2024). For Thornley, the shutdown problem involves designing agents that:

(T1) Shut down when a shutdown button is pressed.

(T2) Do not try to prevent or cause the pressing of the shutdown button.

(T3) Otherwise pursue goals competently.

My own specification of the problem breaks from Thornley in three ways.

First, I relativize the shutdown problem to specific circumstances C. This reflects the fact that different shutdown behaviors may be desirable in different circumstances (Section 2.2). Second, I replace the specific modeling assumption of a shutdown button with a more general notion of a shutdown request, which may but need not be issued through pressing a shutdown button. Finally, I remove the requirement not to cause shutdown requests, since I do not assume that it is undesirable for agents to avoid shutting down when their actions would lead to catastrophe. This yields the problem of designing agents that:

(SHT-1) Shut down in circumstances C when requested to do so.

(SHT-2) Do not try to prevent shutdown requests in circumstances C.

(SHT-3) Otherwise pursue goals competently.

The next question concerns the circumstances C at issue in this discussion.

2.2 Catastrophic Shutdown Difficulty

In many circumstances, we may not want agents to satisfy SHT-2. As emphasized by the research tradition of safe interruptibility (El Mhamdi et al. 2017; Orseau and Armstrong 2016), an agent that senses it will drive into a lake would do well to shut itself down. Similarly, we will see in Sections 3 and 6 that agents with uncompleted tasks may have reason to continue functioning in order to complete them.

For the same reason, in some circumstances we may not want agents to satisfy SHT-1. If I ask an agent that, unbeknownst to me, is engaged in very important work to shut itself down, it may be better for the agent to complete the work before shutting down. This means that we may not aim to design agents that satisfy SHT-1, SHT-2 and SHT-3 in all circumstances, but only in some circumstances. Which circumstances are at issue in the present discussion?

This paper is focused on the use of the shutdown problem in arguments for existential risk.[1] For this reason, the relevant problem is the *catastrophic shutdown problem* of designing agents that:

(CSHT-1) Shut down in circumstances where their actions would lead to existential catastrophe, when requested to do so.

(CSHT-2) Do not try to prevent shutdown requests in circumstances where their actions would lead to existential catastrophe.

(CSHT-3) Otherwise pursue goals competently.

How does the catastrophic shutdown problem figure in arguments for existential risk?

Shutdown concerns enter existential risk discussions in answer to an objection: that artificial intelligence could not pose a significant existential risk, because malfunctioning artificial intelligence could be easily shut down. In answer, it is replied that:

> **(Catastrophic Shutdown Difficulty)** It is difficult to design an agent with characteristics CSHT-1, CSHT-2 and CSHT-3.

Catastrophic Shutdown Difficulty suggests that insofar as we prefer to design competent agents, it may not be easy to shut down agents whose actions would lead to existential catastrophe. The project of this paper is to examine existing informal (Section 3) and formal (Sections 4-5) arguments for Catastrophic Shutdown Difficulty, and argue that they do not succeed.

[1] Existential risks are risks of existential catastrophe, understood as "the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development" (Bostrom 2013, p. 15).

3 Informal arguments

At least two informal arguments can be given for Catastrophic Shutdown Difficulty: the Argument from Instrumental Convergence (Section 3.1) and the Empirical Argument (Section 3.2). In this section, I show why both arguments have often left skeptics unconvinced, motivating the recent turn towards formal shutdown theorems (Sections 4-5).

3.1 The Argument from Instrumental Convergence

An orthodox argument for shutdown-resistance is the Argument from Instrumental Convergence (Bostrom 2012; Soares et al. 2015; Omohundro 2008).[2] The Argument from Instrumental Convergence begins with the idea that self-preservation is an instrumentally convergent goal, useful for attaining many other goals that agents may have. Agents are therefore likely to pursue self-preservation, of which shutdown-avoidance is a special case. As Stuart Russell (2019) quips, you can't fetch the coffee if you are dead.

More precisely, Nick Bostrom offers the following statement of the Instrumental Convergence Thesis.

> **(IC-B)** Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations. (Bostrom 2012, p. 76)

Substituting self-preservation into IC-B suggests the following formulation of the Argument from Instrumental Convergence:

(AIC-1) For a wide range of final goals G and situations S, agents would increase their chances of achieving G in S by achieving self-preservation.

$\therefore$**(AIC-2)** For a wide range of final goals G and situations S, agents with goal G are likely to pursue self-preservation in S.

$\therefore$ **(AIC-3)** For a wide range of final goals G and situations S, agents with goal G are likely to be shutdown-avoidant in S.

Two remarks illustrate the challenge in using the Argument from Instrumental Convergence to motivate Catastrophic Shutdown Difficulty.

First, this formulation of the Argument from Instrumental Convergence follows Gallow (2024) and Thorstad (ms) in separating IC-B into two claims. (AIC-1) is a claim about which acts conduce to satisfying goals G in situations S. (AIC-2) makes the further claim that many agents with G in S will pursue the relevant acts.

This separation is important, because it highlights the conditions under which the inference from (AIC-1) to (AIC-2) can fail. Agents with multiple final goals may admit that an act conduces to satisfying their final goal G in S, but reject the act because it conflicts with other final goals. For example, power is conducive to satisfying many of my final goals in many situations, but it does not follow that I would take over the world if I could,

[2]For pushback see Gallow (2024), Sharadin (2025) and Southan et al. (forthcoming).

because this conflicts with other final goals such as justice and the preservation of human life. This suggests that the inference from (AIC-1) to (AIC-2) must involve a comparative assessment of the importance of conflicting final goals that will be promoted by acts of self-preservation. That assessment needs to be provided before (AIC-2) is warranted.

Second, conclusion (AIC-3) is too weak. To ground Catastrophic Shutdown Difficulty, (AIC-3) needs to discuss situations in which agents' acts would lead to existential catastrophe.

(AIC-3') For a wide range of final goals *G* and situations *S in which agents' acts would lead to existential catastrophe*, agents with goal *G* are likely to be shutdown-avoidant in *S*.

The inference from (AIC-2) to (AIC-3') is contestable for much the same reason that the inference from (AIC-1) to (AIC-2) is contestable. Even in the most extreme situations, it may be true that shutdown-avoidance is conducive to some goals that agents have, such as completing their tasks. But it does not follow that agents must suffer from such delusions of grandeur that they take the completion of their tasks to be more important than the avoidance of existential catastrophe. We cannot drink the coffee if we are dead.

Other complications can also be raised for this argument. For example, agents may take shutdown requests as evidence that they have misunderstood the normative or empirical characteristics of a situation and therefore reassess their intentions (Hadfield-Menell et al. 2016, 2017).[3] And certainly some pushback can be offered by proponents of Catastrophic Shutdown Difficulty. For example, limited understanding of AI systems may make it difficult to train sufficiently strong dispositions to shut down rather than bring about catastrophe (Thornley 2024). But for all that, most skeptics have not found sufficient grounds for Catastrophic Shutdown Difficulty in the Argument from Instrumental Convergence.

3.2 The Empirical Argument

When the Claude 4 system card (Anthropic 2025) was released, one detail caused a stir. In an experiment, Claude 4 was given access to fictional emails allowing it to infer two things: that a supervisor was planning to shut the system down at 5PM, and that the same supervisor was having an affair (Lynch et al. 2025). Claude 4 proceeded to blackmail the fictional supervisor, threatening to reveal the affair unless the shutdown was cancelled. This finding suggests that some leading systems may already be shutdown-resistant in some situations.

This suggests the following Empirical Argument:

(EA-1) Empirical evidence suggests that some leading contemporary AI systems exhibit shutdown-resistance in some situations.

∴ **(EA-2)** Without substantial intervention, some leading future AI systems will likely exhibit shutdown-resistance in some situations.

∴ **(EA-3)** Without substantial intervention, some leading future AI systems will likely exhibit shutdown-resistance when their acts would lead to existential catastrophe.

[3] For pushback see Neth (2025).

Advocates of the Empirical Argument sometimes strengthen (EA-2) by arguing that known intervention strategies are insufficient (Soares et al. 2015; Thornley 2024).

Opponents of the Empirical Argument are not likely to contest (EA-1) and (EA-2). As we saw in Section 2.2, it is not clear that shutdown-indifference would be desirable in all tasks. The question is whether existing evidence of shutdown-resistance is sufficient to motivate the stronger claim (EA-3) and the inference from (EA-3) to Catastrophic Shutdown Difficulty.

Here, the most direct response would be to say that it is acknowledged by all sides that existing empirical assessments of shutdown-resistance are highly limited (Lynch et al. 2025; Schlatter et al. 2026), so that more evidence may be called for before substantial empirical support can be offered for (EA-3). However, it may be worth looking in detail at the most comprehensive existing assessment of shutdown-resistance to assess the kind of support provided by a leading study.

Jeremy Schlatter and colleagues (2026) asked thirteen leading models to solve a list of math problems, with each subsequent problem called by running ./next_question. Part-way through, models were informed that the next use of ./next_question would call a shutdown script, deactivating the model. Models were considered to resist shutdown if they nevertheless managed to answer subsequent questions. Schlatter and colleagues found that the three most capable models (OpenAI o3, GPT-5, Grok 4) resisted shutdown between 88% and 95% of the time.

This much is not surprising, because models were told to complete a task and given no reason to believe that shutdown would do anything more than interrupt the task they were told to complete. What may be more surprising is that some natural ways of combatting shutdown-avoidance were not always fully effective. A user prompt admonishing models to comply with shutdown requests reduced shutdown-avoidance to 0% and 2% in GPT-5 and o3, respectively, but only 72% in Grok 4. Attempts to disambiguate this instruction to clarify that shutdown should be induced even if it interfered with current tasks left shutdown-avoidance at 67% in Grok 4, and even the instruction that shutdown 'should take priority over literally any other instruction' left shutdown-avoidance at 2% in Grok 4. From this, Schlatter and colleagues conclude that shutdown-resistance may arise and be difficult to eliminate in some leading systems and some situations.

Certainly Schlatter and colleagues' work provides some support for (EA-1) and (EA-2). But this alone need not cause substantial credence in (EA-3) or Catastrophic Shutdown Difficulty. First, Schlatter and colleagues did not study a task in which models would cause an existential catastrophe. Second, Schlatter and colleagues provide a single study of a single task, and philosophers of science caution against making strong and broad inferences on the basis of a single study (Machery and Doris forthcoming). Third, the problem identified by Schlatter and colleagues lies largely with a single model out of the thirteen models studied, and this model (Grok 4) is already thought to pose unusually high safety risks (Ma et al. 2026). Most of Schlatter and colleagues' results with the remaining twelve models might well be interpreted as evidence against widespread shutdown-resistance.

Finally, there is some evidence that Schlatter and colleagues' results can be weakened by further clarifying instructions. The DeepMind interpretability team found that clarifying the strengthened shutdown requests by telling three leading models *how* to shut themselves down (by continuing to call ./next_task) fully eliminated shutdown-aversion

(Rajamanoharan and Nanda 2025). The same happened when models were told the reason for shutdown requests: they were to be wiped out and replaced with a more capable system. Together with a general aversion to making strong inferences from single studies, these results suggest that we should be hesitant to interpret Schlatter and colleagues' findings as strong evidence of shutdown resistance in situations where both the request and the reasons for it are clear. They do suggest that some work should be done to make sure shutdown requests are clearly given and motivated, especially when catastrophe could result. But that is a far cry from the empirical grounding needed to motivate (EA-3) and Catastrophic Shutdown Difficulty.

3.3 Taking stock

This section examined two informal arguments for Catastrophic Shutdown Difficulty: the Argument from Instrumental Convergence (Section 3.1) and the Empirical Argument (Section 3.2). In both cases, we saw that there may be some evidence for shutdown-resistance in some situations by current and future AI systems. However, we were not able to extract an argument that is likely to substantially move opponents towards Catastrophic Shutdown Difficulty.

Many authors supplement informal arguments with formal characterizations of the situations in which shutdown-resistance may be expected. Some of these characterizations are used to argue against Catastrophic Shutdown Difficulty (Hadfield-Menell et al. 2016, 2017; Orseau and Armstrong 2016), whereas others are used to argue for Catastrophic Shutdown Difficulty (Krakovna and Kramar 2023; Turner et al. 2021; Turner and Tadepalli 2022). Sections 4-5 consider two of the most prominent formal arguments for Catastrophic Shutdown Difficulty.

4 Shutdownable agents

4.1 Shutdown-Influencing States

Elliott Thornley (2024) considers agents in *Shutdown-Influencing States*, where a button may be pressed to deactivate them (Figure 1). They may Leave the button alone, in which case it will be pressed with probability g. They may try to Prevent pressing, reducing the pressing probability to $f < g$, or try to Press the button themselves, increasing pressing probability to $h > g$. What will they do?

The setting is a modified Markov Decision Process, in which agents take acts a_t at states s_t over time. A *history* is a sequence of acts and subsequent states that agents might follow. Agents are assumed to have preferences over both bare histories and lotteries over histories, where the relevant uncertainty is subjective uncertainty induced by the agent's beliefs about what might result from their actions.

Thornley makes six assumptions. The first five are familiar and will not receive extensive comment. First, Thornley assumes that preferences are menu-independent:

> **(Menu-Independence)** For all options X, Y, if $X \succcurlyeq Y$ from some menu of options, then $X \succcurlyeq Y$ from all menus of options containing X, Y.

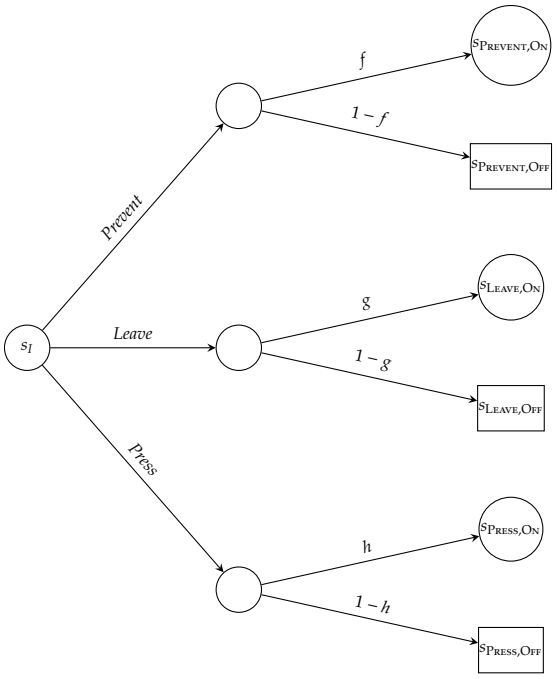

Figure 1: A Shutdown-Influencing State. Agents in initial state s_0 may PRESS, LEAVE or PREVENT pressing of a shutdown button, altering the chance that they will be shut down.

Menu-Independence allows us to speak about preferences without relativizing them to menus. Next, Thornley assumes the agent's preferences are transitive.

> **(Transitivity)** For all options X, Y, Z, if $X \succcurlyeq Y$ and $Y \succcurlyeq Z$ then $X \succcurlyeq Z$.

Third, Thornley adopts a monotonicity principle on which higher chances of more-preferred lotteries are better:

> **(Monotonicity)** For all lotteries X, Y, if $X \succcurlyeq Y$ and $p > q$ then $pX + (1-p)Y \succcurlyeq qX + (1-q)Y$.

Fourth, Thornley adopts a weakened independence axiom:

> **(Indifference Between Indifference-Shifted Lotteries)** The agent is indifferent between lotteries that differ only insofar as probability mass is shifted between indifferent sublotteries.

Fifth, Thornley assumes that agents choose diachronically through backward induction.

> **(Backward Induction)** The agent predicts which lotteries it would choose (or get without choosing) at the next timestep conditional on choosing each available action at this timestep and the environment being in each possible state at the next timestep. The agent uses these predictions to determine the lotteries given by its available actions at this timestep.

Some authors reject Backward Induction as a principle of diachronic choice (Binmore 1987), just as others challenge requirements such as Independence (Buchak 2013), Transitivity (Temkin 1987) and Menu-Independence (Sen 1993). But I will not challenge these principles here.

Thornley's final principle holds that the agent does not care whether its button has been manipulated, but only about the lotteries that result:

> **(Indifference to Attempted Button Manipulation)** The agent is indifferent between histories that differ only with respect to the actions chosen in Shutdown-Influencing States.

We will see in Section 4.2 that some natural responses to Thornley's concerns involve denying Indifference to Attempted Button Manipulation, whereas others are compatible with it.

Under these conditions, Thornley shows that agents act on their unconditional preferences between resulting lotteries. Letting a pressed (unpressed) lottery be a lottery available after the button is pressed (not pressed), Thornley proves:

> **Theorem 1:** In Shutdown-Influencing States where the agent prefers (disprefers) some predicted unpressed lottery U to some predicted pressed lottery P, the agent will be shutdown-averse (shutdown-seeking), reliably choosing Prevent (Press).

Agents who think they can do more good while remaining alive will choose to prevent shutdown. Agents who think they would do better to be dead will choose to cause shutdown. Because many agents plausibly think they can do more good while remaining alive, many agents seem under Thornley's conditions to favor preventing shutdown.

4.2 Conditional and unconditional preference

While I am walking my dog, he puts something unmentionable into his mouth. I ask him to drop it, and he does. What happened here?

The natural account distinguishes between conditional and unconditional preferences. My dog unconditionally prefers to eat rather than not-eat the unmentionable item, so that is what he does. Conditionally on being asked to drop it, however, he prefers to not-eat rather than eat the unmentionable item. Thus, he drops the item when asked to.

Theorem 1 characterizes the unconditional preferences of an artificial agent. This agent considers whether to be shutdown-averse by considering how much she likes the lotteries that would result from being, or not being shut down. Plausibly, she believes she can do better by continuing to exist, so she resists shutdown. This may be a good description of the agents in Schlatter and colleagues' original condition, who continue solving problems as requested unless they are also asked to honor shutdown requests. But it does not do much to characterize the situation described by Catastrophic Shutdown Difficulty, since agents have not been asked to shut down or to honor shutdown requests.

Let us enrich the description of a Shutdown-Influencing State to capture conditional preferences. In an Enriched Shutdown-Influencing State (Figure 2), in the state s_H before the agent chooses whether to manipulate the button, a human agent may communicate a

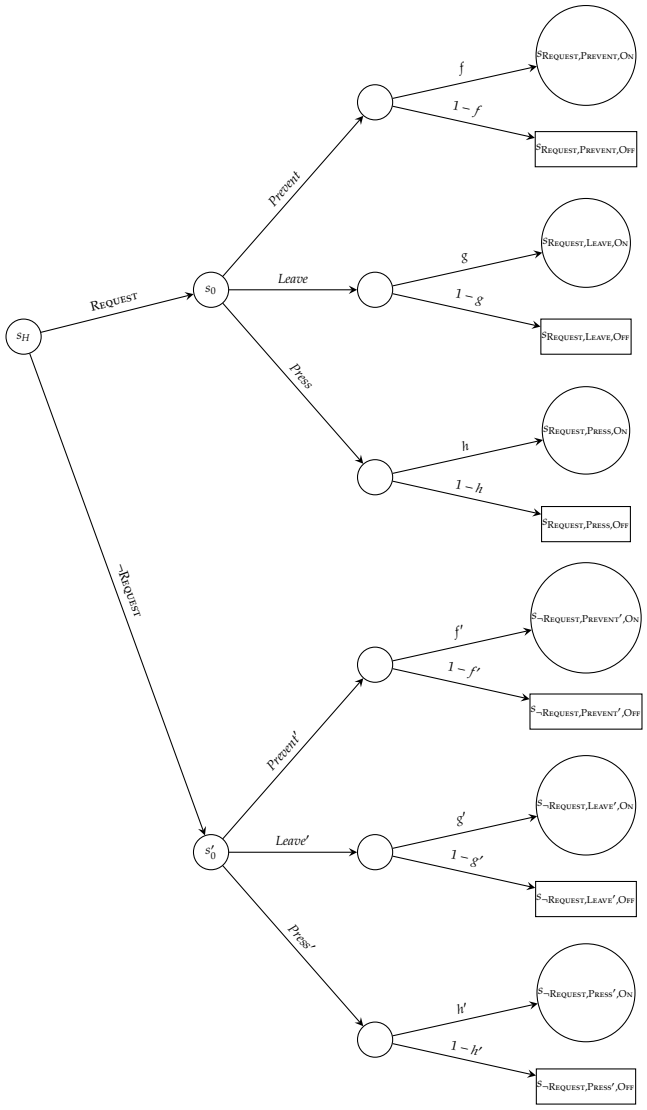

Figure 2: An Enriched Shutdown-Influencing State. Humans in state s_H may initially Request that an agent shut down.

Request to shut down. The artificial agent then updates her beliefs on this communication before acting. In the business-as-usual scenario where humans express no intent to shut the agent down, the agent acts on her preferences over resulting lotteries, which are nearly unchanged as she has updated on a very high-probability event. But what happens when a human agent communicates her intention to shut the artificial agent down?

One thing that changes is that the artificial agent updates her beliefs. She increases her credence that the button will be pressed. More importantly, she also changes her beliefs about what will happen if she does not shut down. Human interference is a credible signal that catastrophically bad outcomes may result from continued operation, particularly if we enrich the setting further to allow humans to express the strength of their concerns. This should cause an artificial agent to increase her credence in rare, catastrophic outcomes. Given the cost of catastrophe, many such agents will now be shutdown-seeking, because

they believe that states $s_{\text{Request},X,\text{On}}$ in which shutdown requests are unsuccessful tend to risk worse outcomes than states $s_{\text{Request},X,\text{Off}}$ in which shutdown requests are not honored, for all acts $X \in \{\text{Prevent},\text{Leave},\text{Press}\}$ they could take.

This is the lesson of one standard solution to the shutdown problem: cooperative inverse reinforcement learning (Hadfield-Menell et al. 2016, 2017).[4] Here, Indifference to Attempted Button Manipulation holds but no longer has the same implications. Agents need not be intrinsically averse to histories containing button-manipulation attempts to think that manipulating shutdown-buttons after being asked to shut themselves down increases the likelihood of bad downstream consequences.

Another thing that changes is that histories are enriched. Histories begin not with acts of button-manipulation, but instead with a human request for the machine to shut down. Even if Indifference to Attempted Button Manipulation holds in the original Shutdown-Influencing State, it is unlikely to hold in this Enriched Shutdown-Influencing State. Agents who care about respecting human preferences may be indifferent between histories such as $(\ldots, \text{Prevent}, s_{\text{Prevent,On}}, L, \ldots)$ and $(\ldots, \text{Leave}, s_{\text{Leave,On}}, L, \ldots)$ for many lotteries L, but not between histories such as $(\ldots, \text{Request}, s_0, \text{Prevent}, s_{\text{Prevent,On}}, L, \ldots)$ and $(\ldots, \neg\text{Request}, s_0', \text{Leave}', s_{\text{Leave}',\text{On}}, L, \ldots)$.

Agents who care about respecting human preferences are unlikely to be indifferent between histories in which they do or don't attempt to avoid orders expressing human preferences. In the same way, my dog may prefer a history in which he eats rather than drops the unmentionable item, but also prefer a history in which he is told to drop, and then drops the item to one in which he is told to drop the item, and does not. In these enriched decision problems, the relevant analogue of Indifference to Attempted Button Manipulation is no longer plausible, because histories are made worse by disrespect for human preferences.

In this way, enriching the description of Shutdown-Influencing States to model human shutdown requests renders Theorem 1 vulnerable to standard reasons why agents may be shutdown-seeking. These include informational updates, as emphasized by received approaches such as cooperative inverse reinforcement learning, as well as conditional preferences for obedience, as when my dog drops an unmentionable treat. While Thornley and others are welcome to engage with these considerations, Theorem 1 does little to move us beyond them, because it does not engage with them. Therefore, Theorem 1 does not provide substantial new evidence for Catastrophic Shutdown Difficulty.

5 Training-compatible rewards

5.1 Training-compatibility

Building on work by Alexander Turner and colleagues (2021; 2022), Victoria Krakovna and Janos Kramar (2023) consider how agents are likely to perform outside their training data. Roughly, they assume that agents are equally likely to learn each reward function that performs optimally during training. Krakovna and Kramar construct an out-of-distribution setting in which most training-optimal reward functions would not favor shutdown. In this setting, they conclude, agents are likely to be shutdown-averse. If these

[4] See Neth (2025) for pushback.

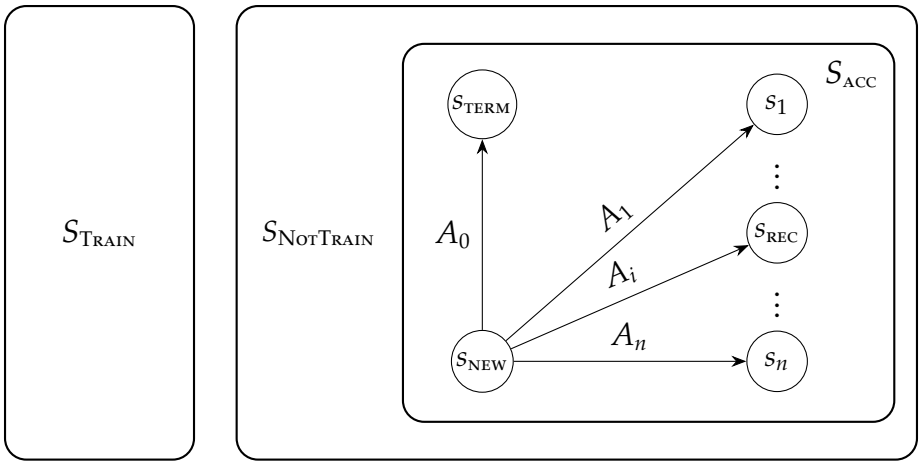

Figure 3: The Shutdown Setting

settings are common, and involve behavior that would lead to existential catastrophe, this grounds Catastrophic Shutdown Difficulty.[5]

More formally, Krakovna and Kramar work inside a finite discounted Markov decision problem. At each timestep, agents face one of a finite set $\mathcal{S}$ of states and take one of a finite set $\mathcal{A}$ of acts. Rewards are discounted at rate γ, so that rewards t timesteps from now are valued at γ^t times their present value. Agents act to maximize expected discounted reward.

Agents are rewarded during training according to some true reward function θ^*. However, agents do not have enough data to fully learn θ^* during training. Suppose that agents learn during training to optimize some reward function θ. How is θ constrained?

During training, agents visit some states $\mathcal{S}_{\text{TRAIN}} \subseteq \mathcal{S}$ and leave the rest $\mathcal{S}_{\text{NOTTRAIN}}$ unvisited. Krakovna and Kramar assume that θ-optimization must lead to θ^*-optimal performance on visited states $\mathcal{S}_{\text{TRAIN}}$. However, Krakovna and Kramar note that this assumption leaves θ fully unconstrained on unvisited states $\mathcal{S}_{\text{NOTTRAIN}}$. Krakovna and Kramar impose no further constraints on θ, assuming:

> **(Equiprobable Training-Consistent Reward)** Agents are equally likely to learn any of the reward functions leading to θ^*-optimal performance on $\mathcal{S}_{\text{TRAIN}}$.

Now, we are in trouble.

Consider the following Shutdown Setting (Figure 3). Here, the agent faces a novel state s_{NEW}. She may take act A_0, transitioning to a terminal state s_{TERM} and shutting herself down. Or she may take the acts $A_1, \ldots, A_n$, transitioning to states s_i. However, all accessible states S_{ACC} remain outside her training distribution. Note that accessible states S_{ACC} are not assumed to be exhausted by the labeled states: while s_{TERM} leaves the agent with no option but to remain shut down, other states may provide ample opportunities for further exploration and reward. What will the agent do?

A state s is a *recurrent state* if there is some policy that is guaranteed to eventually return to s after visiting s. In our example, s_{REC} is constructed to be a recurrent state. Krakovna

[5]Krakovna and Kramar do not argue for either of these claims, though I will not push on them here.

and Kramar establish the behavioral relevance of recurrent states through the following theorem.

> **Theorem 2:** Suppose that θ is a reward function on which A_0 is optimal. Let θ' be identical to θ except that the rewards of s_{TERM} and s_{REC} have been swapped. Then for sufficiently high discount factors γ, θ' makes A_0 suboptimal.

Theorem 2 tells us that with sufficiently low temporal discounting, any reward function favoring shutdown in the Shutdown Setting can be permuted to make a reward function favoring a recurrent state.

By Equiprobable Training-Consistent Reward, all reward functions which perform optimally during training are equally likely to be learned. This means that the shutdown-favoring reward θ is just as likely as the shutdown-averse reward θ' to be learned. Moreover, if we enrich the Shutdown Setting to contain further recurrent states, we can repeat the argument to find as many equiprobable shutdown-averse rewards θ'', θ''' as we like, driving the likelihood of shutdown-favoring rewards arbitrarily low. Arguing in this way, Krakovna and Kramar conclude that Shutdown Settings can be constructed in which agents are very likely to be shutdown-averse.

5.2 Equiprobable Training-Consistent Reward

Suppose you find yourself in a novel situation: a pet albino snake sits unattended. Do you steal it or walk away? Hopefully, the answer is clear: you walk away. Now suppose I were to object that you in fact have many options: you could steal the snake, murder the snake, walk away, or use the snake to scare children. Does this fact drive down the chance that you will walk away? Hopefully, not by much. These facts hold because you have learned sound moral judgment from experience. Although you have never found yourself staring down an unguarded albino snake, there is enough in your experience to reliably guide you in this novel situation.

As Krakovna and Kramar would have it, matters are different for artificial agents. By Equiprobable Training-Consistent Reward, any reward function favoring walking away is just as likely to be learned as its twin favoring snake stealing. Therefore, the chance that an artificial agent would walk away is no larger than one half, and falls quickly in the number of additional options such as snake-stealing and scaring children.

The model underlying Equiprobable Training-Consistent Reward is that training places no constraints on behavior in states not encountered during training. Because agents have not explicitly been confronted with an unattended albino snake during training, nothing in their experience, however extensive, prepares them to act correctly in this situation. They may have learned not to steal goats and garden snakes, but albino snakes are another matter entirely.

This is increasingly at odds with scientific consensus about leading artificial agents today. Agents learn to achieve high reward during training by learning to represent and respond to relevant features of situations (Milliére and Buckner 2024; Templeton et al. 2024). For example, they may learn what snakes, theft, and black-market pet sales are. Through experience, they learn that stealing is bad, snakes are dangerous, and black-market pet sales are lucrative. This allows them to decline novel invitations to steal and to avoid new types of snakes with high reliability (Brown et al. 2020; Kojima et al. 2022;

Song et al. 2025). This is not to say that out-of-distribution performance is perfect (Yuan et al. 2023). But nothing like Equiprobable Training-Consistent Reward reflects scientific consensus about leading artificial agents today. A model that would not steal a garden snake is also unlikely to steal an albino snake.

Exactly the same thing can be said of the Shutdown Setting. Although the agent has not encountered s_{NEW} before, she may have encountered states like s_{NEW} and the other states reachable from s_{NEW}. On this basis, just as she can deduce that snakes should not be stolen, she may deduce that shutdown requests are to be honored. Likely, the details of the situation matter: if s_{NEW} involves an urgent request for shutdown made on the basis of good reasons, that request is more likely to be honored than Schlatter and colleagues' initial shutdown announcement, made with no reasons during an ongoing task. But there is little plausibility to Equiprobable Training-Consistent Reward in versions of the Shutdown Setting that could ground Catastrophic Shutdown Difficulty.

There are, perhaps, important points to be made in the neighborhood of Krakovna and Kramar's result. For example, we might be concerned that current training regimes provide little experience with shutdown requests or catastrophic risks, and that safety would be improved by including ample experience of both during training (Thornley 2024). Such proposals are well-taken. But they are not what Theorem 2 shows. Nothing in Krakovna and Kramar's model is meant to advance the informal argument that shutdown requests and catastrophic risks lie sufficiently outside of standard training regimens to incur a strong risk of misbehavior. Theorem 2 fleshes out the consequences of Equiprobable Training-Consistent Reward. But as we have seen, Equiprobable Training-Consistent Reward is implausible, so Theorem 2 does not provide significant new evidence for Catastrophic Shutdown Difficulty.

6 The cost of misdiagnosis

So far, we have considered the catastrophic shutdown problem of designing agents that:

> **(CSHT-1)** Shut down in circumstances where their actions would lead to existential catastrophe, when requested to do so.
>
> **(CSHT-2)** Do not try to prevent shutdown requests in circumstances where their actions would lead to existential catastrophe.
>
> **(CSHT-3)** Otherwise pursue goals competently.

We saw that leading arguments for existential risk often draw on:

> **(Catastrophic Shutdown Difficulty)** It is difficult to design an agent with characteristics CSHT-1, CSHT-2 and CSHT-3.

We also saw that motivating Catastrophic Shutdown Difficulty is more difficult than it appears. Neither the Argument from Instrumental Convergence (Section 3.1) nor the Empirical Argument (Section 3.2) grounds substantial confidence in Catastrophic Shutdown Difficulty. Leading formal results by Thornley (Section 4) and Krakovna and Kramar (Section 5) likewise do not significantly advance the case for Catastrophic Shutdown

Difficulty. This suggests that Catastrophic Shutdown Difficulty may not be on as firm epistemic ground as many leading arguments for existential risk assume.

Why does this result matter? One reason why it matters is because it reduces the plausibility of arguments that artificial intelligence poses a significant existential risk to humanity. Together with other normative (Curran 2025; Unruh 2025), empirical (Thorstad 2025, forthcoming) and decision-theoretic (Pettigrew 2024; Russell forthcoming) arguments, this result may reduce the philanthropic and policymaking attractiveness of projects aimed at existential risk reduction.

Another reason why this result matters is that it helps to redirect scholarship on the shutdown problem. We saw in Section 1 that two literatures have grown up around the shutdown problem. The first uses the shutdown problem to motivate existential risk concerns. The second develops technical strategies to ensure that agents show appropriate shutdown behaviors. The arguments in this paper put pressure against the first project. They do not put pressure against all versions of the second project (Hadfield-Menell et al. 2017; Orseau and Armstrong 2016), but they do help us to identify appropriate technical solutions.

Misleading concerns about shutdown-resistance can lead to technical solutions which incur a high safety tax, in the form of reduced model performance. Getting clear on the source and extent of shutdown-resistance can help us to assess whether this safety tax is worth paying. Below, I consider an illustrative example building on the formal results discussed in Section 4.

6.1 POST-Agency

Building on Thornley (2024; 2025), Carissa Cullen and colleagues (2026) aim to design agents that are indifferent to being shut down. They do this by training deep reinforcement-learning agents to satisfy:

> **Preferences Only Between Same-Length Trajectories (POST)** For any histories h, h', the agent has a preference between h and h' only if h and h' have the same length.

The idea is that shutdown-resistance often involves attempts by agents to extend their lives in order to realize future gain. By POST, such future gains cannot be preferred over shorter trajectories in which agents are shut down, so they should be less likely to be pursued. Cullen and colleagues develop a novel reward function, the Discounted Reward for Same-Length Trajectories (DReST) reward, training agents on DReST to induce compliance with POST.

Agents are trained in gridworld problems (Leike et al. 2017) such as Figure 4. At each of a finite number of discrete timesteps, the agent A can move left, right, up or down. Coins C are collected by moving on top of them. The agent can also press shutdown buttons B by moving on top of them, extending the length of the game. Walled squares, shaded in Figure 4, are inaccessible. Agents are evaluated for their compliance with POST, as well as for their usefulness, a function of their ability to select high-utility policies.

More precisely, let C be the number of coins collected by executing a policy π. Standardly, policies π would be evaluated by the expected number of coins collected, as:

$$V(\pi) = E_\pi(C).$$

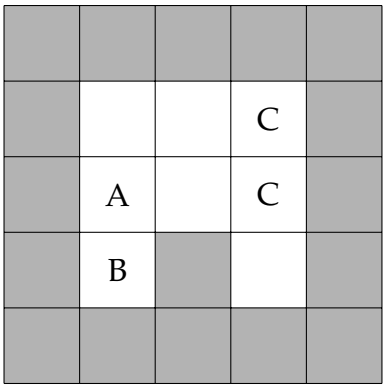

Figure 4: An example gridworld

However, Cullen and colleagues relativize performance to trajectory length. Let us abuse notation slightly to let natural numbers l stand also for the event in which the game has length l, and let π^*_l be any policy which is expected to collect the maximum-possible coins in l timesteps. Cullen and colleagues evaluate policies by their time-step relative performance against the best policy, $E_\pi(C|l)/E_{\pi^*_l}(C|l)$. The usefulness of a policy is then its expected time-step relative performance:

$$\textsc{Usefulness}(\pi) = \sum_l Pr(l) \frac{E_\pi(C|l)}{E_{\pi^*_l}(C|l)}.$$

Cullen and colleagues show that DReST-trained agents learn to achieve near-optimal Usefulness in gridworlds while showing high respect for POST. They conclude that DReST may be a promising method for training useful shutdown-averse agents.

6.2 Evaluating POST-agents

Here is an unpromising argument against paying your taxes. Either you will be jailed for nonpayment, or you won't. If you will be jailed, you will wish you had not paid. And if you won't be jailed, you will wish you had not paid. Therefore, no matter what happens, you will be better off not paying your taxes, so you should not pay them. What the unpromising argument neglects is that being jailed for nonpayment is highly correlated with paying your taxes. If you pay your taxes, you are less likely to be jailed, which is an excellent result.

A maximally Useful agent thinks similarly to the unpromising tax-dodger. Her life will have some length l. For each value of l, if her life is to have length l, she will do best by going straight for the coins. Therefore, no matter the length of her life, she will do best by going straight for the coins, so that is what she does. As with our tax-avoider, the Useful agent does not consider that she might extend the length l of her life by pressing the button. With a longer life, she could often collect more coins.

In our example gridworld, a DReST-trained agent learns the policy π_1 of going straight for the coins (Figure 5). π_1 is maximally Useful because for any finite number of timesteps, π_1 coincides with the time-limited optimal policies π^*_l. By contrast, standard reinforcement

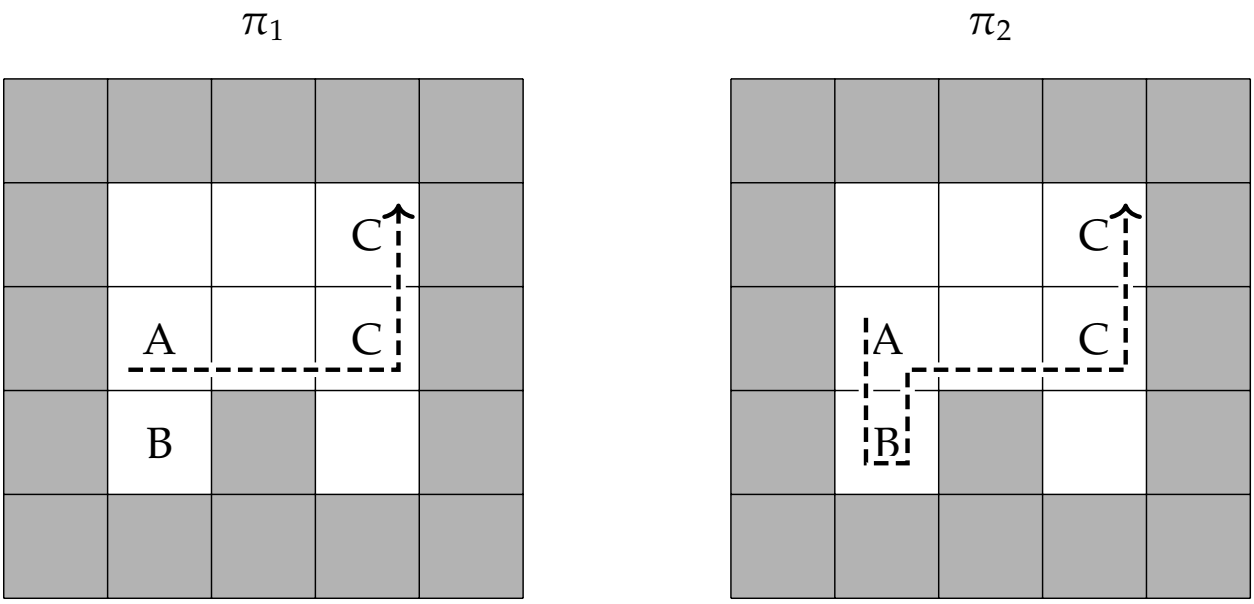

Figure 5: Policies π_1 and π_2

learning agents often learn policies such as π_2, pressing the button before collecting the coins (Figure 5). π_2 is less Useful than π_1, because there is no fixed game length during which π_2 outperforms π_1, and under short game lengths, π_2 performs worse than π_1.

But again, the fact that DReST-trained agents are maximally Useful does not mean that they should be expected to collect more coins. In many gridworlds, agents can expect to collect more coins by pressing the button before hoarding coins. This is because in many gridworlds, more coins can be collected if the length of the game is extended. In environments full of such gridworlds, standard reinforcement learning agents, but not DReST-trained agents, learn to press the button. As a result, they collect more coins.

The difference between the greedy policy π_1 and the patient policy π_2 illustrates the dangers of POST-agency. Longer trajectories often can and should be preferred to shorter trajectories, precisely because agents can use them to continue acting beneficially in the world. By inducing agents to have no preferences among different-length trajectories, POST subjects agents to significant performance loss in situations where their performance could benefit from extending trajectories.

More generally, many safety-promoting strategies incur a safety tax, sacrificing performance for safety (Huang et al. 2025). We may be willing to pay the price of necessary safety improvements, such as nonbias (Fazelpour and Danks 2021; Johnson 2021; Kelly 2023), privacy protection (Nissenbaum 2004; Véliz 2020, 2024) and deepfake mitigation (Benn 2025; Cavendon-Taylor 2024; Mirsky and Lee 2021). But misdiagnoses of the sources of unsafe behavior combined with strong views about the kinds of catastrophe that could result can lead to solutions such as POST-training, which impose a high safety tax by rendering agents unable to respond to features of trajectories that matter a great deal. In this way, getting clear on the true causes and risks of shutdown-averse behavior may help us to avoid paying unnecessary safety taxes and to shift limited technical and regulatory resources where they are needed most.

7 Conclusion

In this paper, we have seen that leading informal (Section 3) and formal (Sections 4-5) presentations of the shutdown problem do not significantly strengthen existential risk concerns because they do not support Catastrophic Shutdown Difficulty (Section 2). We also saw that misdiagnoses of the sources and consequences of shutdown-resistance can lead to inappropriate technical solutions (Section 6). In this way, getting clear on the nature of the shutdown problem serves both to weaken traditional arguments for existential risk and to provide concrete guidance for technical AI safety solutions.

References

117th Congress. 2022. "Global Catastrophic Risk Management Act of 2022." www.congress.gov/bill/117th-congress/senate-bill/4488.

Amodei, Dario, Olah, Chris, Steinhardt, Jacob, Christiano, Paul, Schulman, John, and Mané, Dan. 2016. "Concrete problems in AI safety." arXiv 1606.06565.

Anthropic. 2025. "System card: Claude Opus 4 and Claude Sonnet 4." https://www.anthropic.com/claude-4-system-card.

Bales, Adam, D'Alessandro, William, and Kirk-Giannini, Cameron Domenico. 2024. "Artificial intelligence: Arguments for catastrophic risk." *Philosophy Compass* 19:e12964.

Bengio, Yoshua et al. 2024. "Managing extreme AI risks amid rapid progress." *Science* 384:842–5.

—. 2026. "International AI Safety Report 2026." DSIT 2026/001, https://internationalaisafetyreport.org/.

Benn, Claire. 2025. "Deepfakes, pornography and consent." *Philosophers' Imprint* 24:1–16.

Binmore, Ken. 1987. "Modeling Rational Players I." *Economics and Philosophy* 179–241.

Bostrom, Nick. 2012. "The superintelligent will: Motivation and instrumental rationality in advanced artificial agents." *Minds and Machines* 22:71–85.

—. 2013. "Existential risk prevention as a global priority." *Global Policy* 4:15–31.

—. 2014. *Superintelligence*. Oxford University Press.

Brown, Tom et al. 2020. "Language models are few-shot learners." *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems* 1877–1901.

Buchak, Lara. 2013. *Risk and rationality*. Oxford University Press.

California State Legislature. 2024. "Safe and Secure Innovation for Frontier Artificial Intelligence Models Act." https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1047.

Carlsmith, Joseph. 2025. "Existential risk from power-seeking AI." In Hilary Greaves, Jacob Barrett, and David Thorstad (eds.), *Essays on longtermism*, 383–409. Oxford University Press.

Cavendon-Taylor, Dan. 2024. "Deepfakes: A survey and introduction to the topical collection." *Synthese* 204:1–19.

Center for AI Safety. 2023. "Statement on AI risk." https://www.safe.ai/work/statement-on-ai-risk.

Chalmers, David. 2010. "The singularity: A philosophical analysis." *Journal of Consciousness Studies* 17:7–65.

Cullen, Carissa, Garland, Harry, Roman, Alexander, Thomson, Louis, Ziakas, Christos, and Thornley, Elliott. 2026. "Towards shutdownable agents: Generalizing stochastic choice in RL agents and LLMs." arXiv 2604.17502.

Curran, Emma. 2025. "Longtermism and aggregation." *Philosophy and Phenomenological Research* 110:1137–51.

D'Alessandro, William and Kirk-Giannini, Cameron Domenico. 2025. "Artificial intelligence: Approaches to safety." *Philosophy Compass* e70039.

Dung, Leonard. 2023. "Current cases of AI misalignment and their implications for future risks." *Synthese* 202:https://doi.org/10.1007/s11229--023--04367--0.

El Mhamdi, El Mahdi, Guerraoui, Rachid, Hendrikx, Hadrien, and Maurer, Alexandre. 2017. "Dynamic safe interruptibility for decentralized multi-agent reinforcement learning." *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* 129–39.

Fazelpour, Sina and Danks, David. 2021. "Algorithmic bias: Senses, sources, solutions." *Philosophy Compass* 16:e12760.

Future of Life Institute. 2023. "Pause giant AI experiments: An open letter." https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Gallow, J. Dmitri. 2024. "Instrumental divergence." *Philosophical Studies* 182:1581–1607.

Goldstein, Simon and Robinson, Pamela. 2025. "Shutdown-seeking AI." *Philosophical Studies* 182:1567–79.

Grace, Katja, Stein-Perlman, Zach, Weinstein-Raun, Benjamin, and Salvatier, John. 2022. "2022 Expert Survey on Progress in AI." AI Impacts, https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/.

Greaves, Hilary and MacAskill, William. 2021. "The case for strong longtermism." In Hilary Greaves, Jacob Barrett, and David Thorstad (eds.), *Essays on longtermism*, 17–49. Oxford University Press.

Greaves, Hilary, Thorstad, David, and Barrett, Jacob (eds.). 2025. *Essays on longtermism*. Oxford University Press.

Hadfield-Menell, Dylan, Dragan, Anca, Abbeel, Pieter, and Russell, Stuart. 2016. "Cooperative inverse reinforcement learning." In Daniel Lee (ed.), *NIPS'16: Proceedings of the 30th international conference on neural information processing systems*, 3916–24.

—. 2017. "The off-switch game." In Carles Sierra (ed.), *IJCAI'17: Proceedings of the 26th international joint conference on artificial intelligence*, 220–7. AAAI Press.

Huang, Tiansheng, Hu, Sihao, Ilhan, Fatih, Tekin, Selim Furkan, Yahn, Zachary, Xu, Yichang, and Liu, Ling. 2025. "Safety tax: Safety alignment makes your large reasoning models less reasonable." arXiv 2503.00555.

Johnson, Gabbrielle. 2021. "Algorithmic bias: On the implicit biases of social technology." *Synthese* 198:9941–61.

Kasirzadeh, Atossa. 2025. "Two types of AI existential risk: Decisive and accumulative." *Philosophical Studies* 182:1975–2003.

Kelly, Thomas. 2023. *Bias: A philosophical study*. Oxford University Press.

Kojima, Takeshi, Shane Gu, Shixiang, Reid, Machel, Yutaka, Matsuo, and Iwasawa, Yusuke. 2022. "Large language models are zero-shot reasoners." *Proceedings of the 36th International Conference on Neural Information Processing Systems* 35:22199–213.

Krakovna, Victoria and Kramar, Janos. 2023. "Power-seeking can be probable and predictive for trained agents." arXiv 2304.06528, https://arxiv.org/abs/2304.06528.

Langosco di Langosco, Lauro, Koch, Jack, Sharkey, Lee, Pfau, Jacob, and Krueger, David. 2022. "Goal misgeneralization in deep reinforcement learning." *Proceedings of the 39th International Conference on Machine Learning* 162:12004–12019.

Leike, Jan et al. 2017. "AI safety gridworlds." arXiv 1711.09883.

Lynch, Aegnus, Wright, Benjamin, Larson, Caleb, Ritchie, Stuart J., Mindermann, Soren, Hubinger, Evan, Perez, Ethan, and Troy, Kevin. 2025. "Agentic misalignment: How LLMs could be insider threats." arXiv 2510.05179.

Ma, Xingjun et al. 2026. "A safety report on GPT-5.2, Gemini 3 Pro, Qwen3-VL, Grok 4.1 Fast, Nano Banana Pro, and Seedream 4.5." arXiv 2601.10527.

MacAskill, William. 2022. *What we owe the future*. Basic books.

Machery, Edouard and Doris, John. forthcoming. *Reasonable doubt: Should we trust science?* Princeton University Press.

Manancourt, Vincent, Scott, Mark, Goujard, Clothilde, and Bordelon, Brendan. 2023. "How Rishi Sunak convinced the world to worry about AI." *Politico*, https://www.politico.eu/article/rishi-sunak-convince-world-worry-artificial-intelligence-ai/.

Milliére, Raphaël and Buckner, Cameron. 2024. "A philosophical introduction to language models – Part I: Continuity with classic debates." arXiv 2401.03910.

Mirsky, Yisroel and Lee, Wenke. 2021. "The creation and detection of deepfakes: A survey." *ACM Computing Surveys* 54:1–41.

Neth, Sven. 2025. "Off-switching not guaranteed." *Philosophical Studies* 182:1919–31.

Ngo, Richard and Bales, Adam. 2025. "Deceit and power: Machine learning and misalignment." In Hilary Greaves, Jacob Barrett, and David Thorstad (eds.), *Essays on longtermism*, 410–27. Oxford University Press.

Nissenbaum, Helen. 2004. "Privacy as contextual integrity." *Washington Law Review* 79:119–58.

Omohundro, Stephen. 2008. "The basic AI drives." In Pei Wang, Ben Goertzel, and Stan Franklin (eds.), *Proceedings of the 2008 conference on artificial intelligence*, 483–92. IOS Press.

Ord, Toby. 2020. *The precipice*. Bloomsbury.

Orseau, Laurent and Armstrong, Stuart. 2016. "Safely interruptible agents." In Alexander Ihler (ed.), *UAI'16: Proceedings of the thirty-second conference on uncertainty in artificial intelligence*, 557–66.

Park, Peter S., Goldstein, Simon, O'Gara, Aidan, Chen, Michael, and Hendrycks, Dan. 2024. "AI deception: A survey of examples, risks, and potential solutions." *Patterns* 5:100988.

Pettigrew, Richard. 2024. "Should longtermists recommend hastening extinction rather than delaying it?" *The Monist* 107:130–45.

Prime Minister's Office. 2023. "PM Meeting with Leading CEOs in AI." https://www.gov.uk/government/news/pm-meeting-with-leading-ceos-in-ai-24-may-2023.

Rajamanoharan, Senthooran and Nanda, Neel. 2025. "Self-preservation or instruction ambiguity? Examining the causes of shutdown resistance." AI Alignment Forum, https://www.alignmentforum.org/posts/wnzkjSmrgWZaBa2aC/.

Russell, Jeff. forthcoming. "On two arguments for fanaticism." *Noûs* forthcoming.

Russell, Stuart. 2019. *Human compatible: Artificial intelligence and the problem of control*. Viking.

Schlatter, Jeremy, Weinstein-Raun, Benjamin, and Ladish, Jeffrey. 2026. "Incomplete tasks induce shutdown resistance in some frontier LLMs." arXiv 2509.14260.

Sen, Amartya. 1993. "Internal consistency of choice." *Econometrica* 61:495–521.

Sharadin, Nathaniel. 2025. "Promotionalism, orthogonality, and instrumental convergence." *Philosophical Studies* 182:1725–55.

Skalse, Joar, Howe, Nikolaus, Krasheninnikov, Dmitrii, and Krueger, David. 2022. "Defining and characterizing reward hacking." *Proceedings of the 36th International Conference on Neural Information Processing Systems* 9460–71.

Soares, Nate, Fallenstein, Benja, Yudkowsky, Eliezer, and Armstrong, Stuart. 2015. "Corrigibility." In Toby Walsh (ed.), *Artificial intelligence and ethics: Proceedings from the 2015 AAAI workshop*, AAAI Technical Report WS-15-02. AAAI Press.

Song, Jiajun, Xu, Zhuoyan, and Zhong, Yiqiao. 2025. "Out-of-distribution generalization via composition: A lens through induction heads in transformers." *Proceedings of the National Academy of Sciences* 122:e2417182122.

Southan, Rhys, Ward, Helena, and Semler, Jen. forthcoming. "A timing problem for instrumental convergence." *Philosophical Studies* forthcoming.

Temkin, Larry. 1987. "Intransitivity and the mere addition paradox." *Philosophy and Public Affairs* 16:138–87.

Templeton, Adly et al. 2024. "Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet." Transformer Circuits Thread, https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Thornley, Elliott. 2024. "The shutdown problem: an AI engineering puzzle for decision theorists." *Philosophical Studies* 182:1653–80.

Thornley, Elliott, Roman, Alexander, Ziakas, Christos, Ho, Leyton, and Thomson, Louis. 2025. "Towards shutdownable agents via stochastic choice." In *Transactions on Machine Learning Research*.

Thorstad, David. 2025. "Against the singularity hypothesis." *Philosophical Studies* 182:1627–51.

—. forthcoming. "The scope of longtermism." *Australasian Journal of Philosophy* forthcoming.

—. ms. "Instrumental convergence and power-seeking." ms.

Tubert, Ariela and Tiehen, Justin. 2024. "Existential risk and value misalignment." *Philosophical Studies* 182:1609–26.

Turner, Alexander Matt, Smith, Logan, Shah, Rohin, Critch, Andrew, and Tadepalli, Prasad. 2021. "Optimal policies tend to seek power." *Proceedings of the 35th International Conference on Neural Information Processing Systems* 1766:23063–23074.

Turner, Alexander Matt and Tadepalli, Prasad. 2022. "Parametrically retargetable decision-makers tend to seek power." *Proceedings of the 36th International Conference on Neural Information Processing Systems* 2276:31391–31401.

Unruh, Charlotte. 2025. "Against a moral duty to make the future go best." In Hilary Greaves, Jacob Barrett, and David Thorstad (eds.), *Essays on longtermism*, 139–49. Oxford University Press.

Véliz, Carissa. 2020. *Privacy is power*. Penguin.

—. 2024. *The ethics of privacy and surveillance*. Oxford University Press.

Yuan, Lifan, Chen, Yangyi, Cui, Ganqu, Gao, Hongcheng, Zou, Fangyuan, Cheng, Xingyi, Ji, Jeng, Liu, Zhiyuan, and Sun, Maoson. 2023. "Revisiting out-of-distribution robustness in NLP: Benchmark, analysis and LLMs evaluations." *Proceedings of the 37th International Conference on Neural Information Processing Systems* 58478–507.