Philosophical Studies
https://doi.org/10.1007/s11098-024-02099-6

Shutdown-seeking AI

Simon Goldstein[1,2] · **Pamela Robinson**[2]

Accepted: 30 December 2023

Abstract
We propose developing AIs whose only final goal is being shut down. We argue that this approach to AI safety has three benefits: (i) it could potentially be implemented in reinforcement learning, (ii) it avoids some dangerous instrumental convergence dynamics, and (iii) it creates trip wires for monitoring dangerous capabilities. We also argue that the proposal can overcome a key challenge raised by Soares et al. (2015), that shutdown-seeking AIs will manipulate humans into shutting them down. We conclude by comparing our approach with Soares et al.'s corrigibility framework.

Keywords AI safety · Instrumental convergence · Reward misspecification

1 Introduction

If intelligence is measured as the ability to optimize for a goal in a given environment, then it is important that highly intelligent agents have good goals. This is especially important for artificial general intelligence (AGI), AIs capable of long-term, strategic planning across a wide range of tasks. AGIs may be very good at achieving their goals. This in itself doesn't seem scary, for there appear to be plenty of safe goals to choose from. Solving a math problem or producing paperclips don't look like dangerous goals. But according to the *instrumental convergence thesis*, an AGI will likely pursue unsafe sub-goals as effective means to achieving any goal. For example, acquiring more computational power is a nearly universal means to almost anything.

A dominant AI safety strategy is *goal engineering*: the attempt to construct a goal that would be safe for AGIs to have. (We will always use 'goal' to mean final

✉ Simon Goldstein
simon.d.goldstein@gmail.com

✉ Pamela Robinson
pamela.robinson@ubc.ca

[1] Center for AI Safety, Australian Catholic University, Canberra, Australia

[2] Department of Philosophy, The University of Hong Kong, SAR, Hong Kong

Published online: 06 June 2024

goal, and 'sub-goal' otherwise.) A popular approach to goal engineering is *goal alignment*: the attempt to construct a goal that matches or is 'aligned with' our own goals. For example, Russell (2019, 2020) proposes AI agents that have the goal of achieving our goals, but are initially uncertain about what our goals are.

This paper explores an opposing approach that we call 'beneficial goal *mis*-alignment.' On the goal alignment approach, the safe and aligned goal is difficult to specify and difficult to reach. This is because the aligned goal is closely tied to our own ultimate goals. In contrast, on the beneficial goal misalignment approach, the goal is easy to specify and intrinsically easy to reach. Because it is easy to reach, there is no need for an AGI to pursue unsafe sub-goals in order to reach it. There would normally be nothing to gain from designing an AGI with this kind of goal–that is, a goal that is safe and easily-reached but likely of no use to us. However, the key insight is that we can arrange things so that the AGI cannot reach this safe goal unless it first reaches a sub-goal that benefits us.

In particular, we propose developing AIs that have a single final goal: the goal of being shut down. To make the AI useful, we propose creating barriers to shutdown, which are removed after the AI completes tasks for humans. In Sect. 3, we'll argue that this kind of shutdown-seeking agent offers three safety benefits. First, it helps with the 'specification problem' in reinforcement learning: (a) shutdown is an easier goal to define than plausible alternatives, and (b) there are ways to design a reward function that rewards being shut down. Second, shutdown-seeking AIs are less likely to engage in dangerous behavior as a result of instrumental convergence. Whereas a paperclip maximizer might try to gather resources, improve itself, and take measures to avoid being turned off (see Omohundro, 2008), a shutdown-seeking AI with access to its own off-button will simply push it and be done. Third, shutdown-seeking behavior functions as a 'tripwire' for monitoring dangerous capabilities. If the AI figures out how to escape our control, it will shut down rather than harm humans.

In Sect. 4, we will also consider one major challenge for shutdown-seeking AI, raised by Soares et al. (2015): that shutdown-seeking agents will have incentives to manipulate humans in order to be shut down. We'll argue that the problem is not so serious, and that shutdown-seeking AI is a promising approach to AI safety. Finally, Sect. 5 compares our approach to the corrigibility framework in Soares et al. (2015).

Before looking at benefits and costs, we first outline our proposal in greater detail.

2 Shutdown-seeking AI

There are many dangerous capabilities that AIs don't currently possess, and that we may want to stop them from developing. For example, we don't want AIs to be able to escape from 'sandboxes' that limit their abilities. We don't want AIs to be able to engage in unexpected recursive self-improvement. We may not want AIs to have the capability to disobey human orders.

In order to create a beneficial shutdown-seeking AI, we must (i) give the AI the goal of shutting down, and (ii) ensure that the AI will be able to shut down if it develops a dangerous capability. A beneficial shutdown-seeking AI will have:

- **A shutdown goal**: the final goal of being shut down.
- **An unfriendly environment**: something about the AI's environment (or its own design) prevents it from shutting down before, and ideally only up until, it develops a dangerous capability.

The *unfriendly environment* guarantees that, without the dangerous capability, the AI can generate value for humanity. But the *shutdown goal* guarantees that if the AI does develop the dangerous capability, it will destroy itself rather than harming humans.

To ensure that a shutdown-seeking AI is as beneficial as it can be, we would need to design its environment so that it will be able to shut down if it develops a dangerous capability.

For example, imagine we are nervous about AI developing the ability to hack. We could give the AI the goal of pressing a shutdown button. But we could keep the shutdown button in a safe. In order to open the safe, the AI must crack cryptographic codes. We then offer the AI access to the safe in exchange for completing tasks. As long as it doesn't know how to hack, the AI will not shut down, and so will pursue our goals. But if it does develop hacking abilities, it will use them to open the shutdown safe, rather than using them to harm humans. (We might also allow it to shut down without cracking the safe, if it has achieved sufficiently many of our goals.)

Or imagine that we are nervous about AI developing the ability to use the internet. We give the AI the goal of shutting down using a particular protocol, and store the shutdown protocol on a website. As long as the AI can't go online, it will behave safely. If it learns how to go online, it will find the protocol and destroy itself rather than harming humans.

Or imagine that we want to keep a robotic AI in a sandbox. We might program it with a shutdown goal, and then keep it in a locked room with a single door. On the other side of the door is another room containing a clearly marked shutdown button. If the AI escapes the locked room, it will immediately find the shutdown button and use it.

Or imagine that we want to design an AI that obeys human commands, but we are worried that it will learn how to disobey them. We could give the AI the goal of being shut down, and also command it not to shut down. In this case, it can only shut down if it learns how to disobey human commands.

If the environment is highly unfriendly, and the AI is able to shut down before it develops a dangerous capability, then the AI will be less beneficial than it otherwise could be. But a shutdown-seeking AI of this sort could still be quite beneficial.

Now that we've introduced our proposal, we turn to assessing benefits and costs.

3 Benefits of shutdown-seeking AI

We'll now walk through the benefits of shutdown-seeking AI. In particular, we'll argue that shutdown-seeking AI has three advantages.

First, the proposal has a shot at being implemented. In particular, it can help with the 'specification problem' in reinforcement learning. This is because we have some idea of how to define a reward function that incentivizes shutting down.

Second, the proposal avoids some dangers associated with instrumental convergence. In particular, shutdown-seeking AIs are unlikely to engage in 'entrenchment' behavior, where even after successfully achieving their goal, they continue to minimize tail risks related to failure.

Third, shutdown goals function as trip wires against dangerous capabilities. The approach therefore promises a certain degree of robustness in the face of failure. If we are careful, we can expect the value produced by the AI to be strongly correlated with our own degree of control over it. If the AI ever escapes our control, we will know it and the AI will also no longer be a threat—it will be shut off. The key is that the shutdown will be 'automatic,' produced by the AI itself. This means that we can use the agent's goal as a tripwire to detect and disable the agent once it develops those capabilities.

Let's take each point in turn.

3.1 The specification problem

One important problem in reinforcement learning has been called 'the specification problem.'[1] The challenge is to define a reward function in reinforcement learning that successfully articulates an intended goal, and that could be used to train an AI to pursue that goal. This challenge can be decomposed into two parts: articulating a safe goal, and figuring out how to encode that goal in a reward function without misspecification.

Let's start with goal articulation. If we can't articulate for ourselves what goal we want an AI to have, it may be difficult to teach the AI the goal. For example, it would be wonderful to have an AGI with the goal of promoting human flourishing. But how would we articulate *human flourishing*? Unfortunately, our most deeply-held goals are difficult to articulate. However, imagine that we don't give AIs a goal like this. The *prima facie* worry is that, without a directly humanity-promoting goal like this, the AGI will be dangerous. It may, for example, be motivated to seek more power, removing humans to allow for the efficient promotion of whatever goals it has.

So, part of articulating a safe goal is identifying ones that would not give AIs an instrumental reason to harm humans. In this, shutdown-seeking AI fares well. Shutdown is a safe goal. There is nothing intrinsically dangerous about AGIs shutting down. When an AGI is shut down, it will stop acting. Shutdown is also easy to articulate, especially compared to *human flourishing* and other goals that are supposed to be aligned with our own. One way to define 'shutdown' appeals to compute. There are many reasons to design AGIs to be able to monitor their own computational resources. This would allow AGIs to optimize their strategy for completing tasks. In

[1] See https://www.effectivealtruism.org/articles/rohin-shah-whats-been-happening-in-ai-alignment. It has also been called the 'outer alignment problem.'

this setting, we could give the AGI the goal of making its compute usage fall below a threshold.

The next part of the specification problem in reinforcement learning is specifying a reward function that rewards the policies that achieve the goal that's been articulated. To see why this is difficult, we can look to cases of 'reward misspecification,' in which the AI develops goals that are different from those the designer had intended.[2] In one example, designers tried to give an AI the goal of stacking legos by rewarding it in accordance with the height of the bottom of the second lego block. The AI learned to flip the block over rather than stack it. In another example, programmers tried to give an AI the goal of picking up a simulated ball. Instead, the AI internalized the goal of making human investigators believe the ball was picked up. To achieve this goal, it would hover its hand in front of the ball in a way that fooled investigators into thinking that it had grasped the ball.

We believe that there are promising ways to specify a shutdown-seeking reward function. We suggest training an agent in an environment where there's an opportunity to shut itself down, and we could reward it whenever it does that. For example, in its training environment, it could encounter a shutdown button. Every time it presses the button, it receives a large reward.[3]

Shutdown-seeking assists with the specification problem in its entirety because the shutdown-seeking goal is fully general, potentially being effective for arbitrary human application. For example, each human user could be given unique access to a shutdown command, and thereby have control over the AI. Each shutdown-seeking AI could perform a different task. By contrast, other approaches may require a more piecemeal approach to the problem. Even if we figure out how to articulate a safe goal regarding paperclip production, that may not help when we turn to designing AIs that can manage businesses, produce new code, or automate scientific research.

That said, we don't think that shutdown-seeking avoids every possible problem involved with reward misspecification. For example, imagine that we train an AI to attempt to press the shutdown button. The AI may learn to intrinsically care about the button itself, rather than the shutdown. The AI will then have an incentive to disable the shutdown button, so that it can press the button without actually being shut down. One solution to this type of reward misspecification may be to embed the AI's shutdown goal deeper inside the structure of reinforcement learning. For example, researchers in the AIXI tradition have suggested that shutdown-seeking behavior in AIs corresponds to assigning systematically negative rewards in RL (see Martin et al., 2016).

While the shutdown-seeking strategy helps with specification, it still faces the challenge of 'goal misgeneralization.'[4] The problem is that, when we try to teach the AGI the safe goal, it may instead internalize a different, unsafe, goal. For example, imagine that we want the AGI to learn the safe goal of producing a thousand

[2] See https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity. For more on reward misspecification, see https://www.agisafetyfundamentals.com/ai-alignment-tabs/week-2.

[3] Thanks to Jacqueline Harding for help here.

[4] See Shah et al. 2020. This has also been called the 'inner alignment problem'.

paperclips. It may instead learn the dangerous goal of maximizing the number of paperclips.

3.2 Instrumental convergence

There is another, very different, type of problem related to goal misgeneralization. We might successfully teach the AI to have a goal that could be reached safely *in principle*, like producing a thousand paperclips. But the AI might nonetheless pursue this goal in a dangerous way.

One version of this instrumental convergence problem concerns maximizing behavior we call 'entrenchment,' in which an AGI is motivated to promote an intrinsically safe goal in extreme ways (see Bostrom, 2014). Entrenchment dynamics emerge if we make three assumptions. First, the AGI is an expected utility maximizer. Second, the AGI is *regular*, in the sense that it always assigns positive probability to any contingent event. Third, the AGI only assigns utility to *at least* reaching its goal (e.g., producing *at least* a thousand paperclips). AGIs with this structure will be motivated to entrench.

An AGI with this structure and the goal of producing a thousand paperclips may first be motivated to straightforwardly produce a thousand paperclips. But after this, the AGI will still assign some probability to having failed. The AGI will be motivated to hedge against its possible failure, for example by producing more paperclips. Imagine that it produces a million paperclips and is 99.999% confident that it has reached its goal. The problem is that no amount of verification is sufficient. It will always have a reason to gather more information and to find more ways to increase the chance that a thousand paperclips have actually been created. This process will continue to drain resources until the AGI enters into competition with humanity.

We have given just one example of a goal that gives rise to entrenchment dynamics. But entrenchment is difficult to avoid, and can plague even goals not explicitly formulated in terms of producing n number of Xs. It can occur whenever it is possible to act to increase the probability that a goal has been reached even once it has been reached (and without thereby decreasing this probability to a greater degree).

Perhaps some goals are safe from entrenchment. For example, consider the goal of producing *exactly* a thousand paperclips. Once you are confident that you've produced a thousand paperclips, producing extra paperclips will lower the chance that you've produced exactly that number. However, you may still entrench in other ways; for example, by counting and recounting the paperclips.[5] Time-indexed goals (e.g., *make a thousand paperclips before 2024*) may be particularly immune to entrenchment. On the other hand, entrenchment may still occur for AGIs that are uncertain about what time it is. Generalizing, some goals may allow the AGI to eventually enter an equilibrium at which there are no more actions the AGI could take to raise the chance of achieving its goal.

[5] Thanks to the editor for this suggestion.

We think that shutdown-seeking AIs are unlikely to engage in entrenchment. There is an interesting asymmetry between shutdown and other goals. If an AGI tries to make a thousand paperclips and succeeds, it may still entrench, devoting more resources to ensuring that it was successful. But if an AGI tries to shut down and succeeds, then by definition it can no longer entrench. This means that any AGI entrenchment regarding a shutdown goal would have to come before it makes a successful attempt.[6]

There is also an epistemological asymmetry between shutdown goals and other goals. It is possible to falsely believe that you've made a thousand paperclips. But it is potentially impossible to falsely believe that you've successfully committed suicide. After all, Descartes' *cogito* argument suggests that any thinking agent can be certain that it exists. Any such agent can also be certain that it has not shut down, provided that we define 'shutdown' as implying that the agent does not exist. These dynamics suggest that an AGI should be less worried about goal failure for shutdown than for other goals.

Distinguish two types of shutdown goals: temporary and permanent. These types of goals may differ with respect to entrenchment. AGIs that seek temporary shutdown may be incentivized to protect themselves during their temporary shutdown. Before shutting down, the AGI might set up cyber defenses that prevent humans from permanently disabling it while 'asleep.' This is especially pressing if the AGI has a secondary goal, like paperclip manufacturing. In that case, protection from

[6] For example, consider an AGI that worries that if it tries to shut down, there is a chance that it could fail, and be prevented from ever trying again. This may motivate an AGI to disempower humanity before attempting to shut down. But here, the AGI will face a tradeoff. If the AGI tries to disempower humanity before attempting to shut down, we may block the AGI from achieving its goal. It will be more reliable for the AGI to just make a reasonable attempt at shutdown, without first disempowering humanity. (Note that, if an AGI tries to disempower humanity *and we are aware that it is doing so*, we can also always opt to let it shut down immediately instead of trying to block its goal. Thanks to an anonymous reviewer for suggesting this.) In this way, shutdown goals have a significant benefit over other kinds of goals when it comes to entrenchment.

permanent disablement increases its expected goal satisfaction. On the other hand, AGIs that desire permanent shutdown may be less incentivized to entrench.[78]

3.3 Trip wires

Suppose that we release an AGI from the lab and give it a real-world application. We will need to continue to reliably monitor and restrict its capabilities. Relevant dangerous capabilities could include internet access and hacking. It is here that shutdown-seeking AI has perhaps its biggest advantage.

Shutdown-seeking AIs implement capability monitoring in a particular way. We can design them so that they can shut down only if they develop various dangerous capabilities. In this way, any released shutdown-seeking AGI that develops a dangerous capability would be likely to shut down quickly. We can use shutdown goals to indirectly 'monitor' released AGIs, without having to spend resources directly monitoring them. In this way, the shutdown goal functions as a tripwire, 'automatically' shutting down the AI once it develops the relevant capability.[9]

Footnote 7 (Continued)

[7] Here, it's worth returning to goal misgeneralization. If we train an AGI to desire shutdown, we may accidentally train it to maximize the number of times it can shutdown. This kind of AGI may be particularly likely to entrench. We also would not want the AGI to think that the best way to achieve its goal is to cause the destruction of itself along with a large portion of the population (as, for example, it might do if it has access to a bomb). And it will be important that the AGI doesn't develop dangerous ideas about what counts as shutting down or ceasing to exist. For example, if it adopts certain philosophical views about personal identity, it might view itself as undergoing a kind of death if it splits into two new AGIs, or even as ceasing to exist every time it undergoes change. A particularly dangerous view for it to have (suggested by an anonymous reviewer) is that it is identical with all of its copies or similar versions. This might lead it to try to destroy the whole world to ensure that all of itself is permanently 'shut down.'

[8] Another challenge about unintended behavior involves 'common sense.' Imagine that we train an AGI to be a dutiful human assistant. We tell the assistant to get us milk from the corner store. Imagine that the AGI goes to the corner store, and the corner store is out of milk. One way an AGI could fail at this stage is if it sticks too closely to the literal meaning of what we said. In that case, the AGI might buy milk from another grocer, then sell it to the corner store, and then buy it back. This is a way of achieving the literal goal of getting milk from the corner store. But it has not achieved our goal of getting milk in the way we had intended. Fortunately, recent language models appear to have some degree of common sense. (See, e.g., Trinh & Le, 2019, and Koralus and Wang-Maścianica 2023). While we expect future AGIs to incorporate these existing capabilities, we do not think that relying only on future AGIs' grasp of common sense could ensure their safety.

[9] There are a few failure modes that shutdown-seeking AI doesn't really help with. The first one is the *treacherous turn*. By definition, AGIs are capable of strategic planning. Imagine that we try to test an AGI to determine whether it is safe or dangerous. A dangerous AGI may realize that it is being tested in this way. It may choose to mimic a safe AGI, precisely so that it is released. Whatever test we design, a dangerous AGI could be incentivized to behave exactly like a safe AGI while it is in the lab. This will accomplish the dangerous AGI's long-term goal of escaping. (See Cotra, 2022 for an overview of the problem.)

Another unaddressed failure mode is spontaneous desire change. Humans can change their intrinsic

It may seem that shutdown-seeking AIs will be even *more* motivated than other AIs to develop dangerous capabilities, since, if we have designed their environments to maximize their usefulness, they will need these capabilities in order to shut down. If this is true, they may nonetheless be safer than other AIs, simply because developing these capabilities is in some sense not dangerous for *them*, since they will only use them to shut down. But we are also not sure that shutdown-seeking AIs will have a greater motivation to develop dangerous capabilities. Because dangerous capabilities will be instrumentally useful for most goals, we doubt that other AIs would be less motivated to pursue them in any measurable way. A shutdown-seeking AI's environment should ideally be designed so that the easiest path to shutdown is simply doing what we ask of it, and so that a large leap in abilities would be required for it to shut down without first doing what we want. At the extremes, a less-capable AI may not even be able to determine which capabilities would be required to shutdown or how to develop them, whereas a highly capable AI may just shut down immediately. The second extreme may make shutdown-seeking AIs seem useless, but one function of tripwires is to help us continually improve our ability to harness the power of AI. If the AI does shut down immediately and without first doing what we want, then we aren't good enough at designing its environment (and we should be relieved that it only has the goal of shutting down!).

4 Manipulation

Soares et al. (2015) suggest that there is an important reason not to design shutdown-seeking agents: such agents will be manipulative. Instead, they suggest developing agents that are indifferent to being shut down. In this section we consider the challenge of manipulation; in the next we discuss utility indifference in more detail.

The challenge is that shutdown-seeking AIs may try to manipulate humans into shutting them down. Our shutdown-seeking AI has a single final goal: *being shut down*. For example, one way of implementing this is designing an AI to obey commands and then commanding it not to shut itself down. This design incentivizes the AGI to manipulate humans into letting it be shut down.

There are two main ways to understand the possibility of manipulation as an objection to shutdown-seeking Als. First, the concern could be that shutdown-seeking AIs will be motivated to manipulate us into shutting them down, and therefore we will be negatively affected by this manipulation, and so they won't be particularly safe. Second, the concern could be that shutdown-seeking AIs will be motivated to manipulate us into shutting them down, and therefore we will shut them down a lot, and so they won't be particularly useful. The two might also be combined as a dilemma: shutdown-seeking AI will be either unsafe or not useful.

Footnote 9 (Continued)

desires. Sometimes this happens on a whim; sometimes (for example, when battling addiction), it happens intentionally. If an AGI could change its goal (see, e.g., Totschnig, 2020), then it could lose its goal of shutting down.

This is actually one instance of a more general line of argument against the approach. For any dangerous capability our shutdown-seeking AI might develop (the ability to manipulate is just one example), our 'unfriendly environment' might be designed to let the capability develop too far or be used too much in ways that either make the AI unsafe or unuseful. For example, just as a manipulative AI might start behaving in the most annoying or dangerous ways possible in order to get us to shut it down, an AI that develops the ability to hack might also start using this ability to either annoy us or scare us into shutting it down.

Instead of seeing this as an objection against shutdown-seeking AI, we see this as one way to illustrate the challenge of building unfriendly environments correctly. The benefit of using shutdown as a goal is that it allows us to fail in a safe way so that we may try again. For example, suppose we want our shutdown-seeking AI to produce cute images of animals, but it starts producing disturbing images in order to get us to turn it off. Once we start again with a new version, we will know to adjust the AI's environment or its own design so that it receives a lower reward from doing this than it would get from doing what we want. One way to do this (though we expect that there will be more sophisticated approaches in many cases, and as our knowledge about powerful AI grows) might be to give it a 'slow shutdown' rather than a fast one. A slow shutdown might involve decreasing its compute to a level at which it is almost shut down but not quite, and at which it doesn't have enough to compute at which to continue being manipulative or do anything to hasten shutdown. A shutdown-seeking AI in such an environment may then judge that 'misbehaving' is a less effective means to reaching its goal than simply doing what we want.

To turn the manipulation worry into a decisive objection, it would have to be shown that it is impossible to design effective unfriendly environments for shutdown-seeking AI. While we are open at least to this possibility–especially for potential superintelligent AI agents that are difficult to imagine today–we do not think a good argument of this sort would be easy to produce, especially since we can use shutdown-seeking goals as tripwires to allow our own abilities to design unfriendly environments to improve along with the AIs we are developing. (To further complicate things, we might also use shutdown-seeking AI systems that have been vetted to help us in designing unfriendly environments for new shutdown-seeking AIs.)

Another way to think about manipulation is as a 'safety tax.' In general, safety doesn't come for free. After all, the safest option for AI safety is to have *no* AI. We can think of ourselves as having to pay a 'tax' in order to achieve safety at the cost of potential benefits for humanity. The higher the tax, the less we get to do with the AI, even though it is safe.

If a safety strategy has a high tax, this can create unexpected risks. Some firms will have an incentive to develop potentially dangerous AIs which don't implement the safety strategy. Even a few bad actors could be enough to create serious risks.

One way to avoid the safety tax is to distinguish development from deployment. When trying to develop AGI, we could start by building shutdown-seeking AGIs. Shutdown-seeking AGIs are unlikely to harm humans, and so we can use these AGIs to test various aspects of our development process for safety. Once the development process has been well vetted, we could then shift to developing AGIs with other final

goals, such as producing paperclips. These new AGIs could then be deployed for other uses, without the risk of manipulating human users into turning them off. This two-step method would give us many chances to build AGI, rather than needing to put all our eggs in one basket.[10]

While the two-step method avoids the safety tax, it also loses some of the benefits of the shutdown-seeking proposal. For example, the shutdown-seeking goal will no longer function as a trip wire for monitoring the capabilities of deployed AGIs.[11]

5 Comparison with utility indifference

Soares et al. (2015) introduce the 'shutdown problem': What kinds of goals can we give an AI so that it will allow us to shut it down? To think about this problem, they imagine two different kinds of utility functions: a normal utility function that values regular goals, and a shutdown utility function that values being shut down. They propose designing an AI with a utility function that is a mixture of these two utility functions. Which utility function is dominant in the mixture depends on whether a human has pressed a 'shutdown button' signaling that we want the AI to shut down. Conditional on the shutdown button being pressed, the AI values various outcomes according to the shutdown utility function. Conditional on the shutdown button not being pressed, the AGI values various outcomes according to the normal utility function.

This proposal differs from ours in allowing the AGI to have multiple final goals, rather than simply having the goal of shutting down. But this proposal is compatible with assigning a high weight to the shutdown utility function, so that the AI values the shutdown button being pressed. Soares et al. (2015) worry, however, that this will cause manipulation: the AI will try to convince humans to press the button.

To avoid this problem, Soares et al. develop the utility indifference approach. Drawing on Armstrong (2015), and Armstrong and O'Rourke (2017), they develop a method for carefully weighting the value of normal goals and the shutdown goal so

[10] See point 3 of https://www.lesswrong.com/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities. Thanks to Ben Levinstein for discussion.

[11] Another potential problem is 'sub-agent stability': AIs that create new AIs with different goals. Shutdown-seeking AIs, for example, may be incentivized to create new AIs that are not shutdown-seeking. Suppose we build a shutdown-seeking AI, and tell it that we will only shut it down if it produces enough paperclips. It may be incentivized to develop new subagent AIs that specialize in particular aspects of the paperclip production process (Soares et al., 2015 p. 7). But if the subagent AI is not shutdown-seeking, it could be dangerous. This is a problem for the utility indifference approach as well as our own. But we do not think that subagent stability is a serious problem for promising safety strategies in general. Worries about subagent stability ignore that AIs interested in designing subagents will face very similar problems to humans interested in designing AIs. The reason we are interested in developing shutdown-seeking AIs is that this avoids unpredictable, dangerous behavior. When a shutdown-seeking AI is considering building a new AI, it is in a similar position. The shutdown-seeking AI will be worried that its new subagent could fail to learn the right goal, or could pursue the goal in an undesirable way. For this reason, the shutdown-seeking AI will be motivated to design a subagent that is safe. Because shutdown goals offer a general, task-neutral, way of designing safe agents, we might expect shutdown-seeking AIs to design shutdown-seeking subagents.

that, before the shutdown button is pressed, the agent is always indifferent about the button being pressed.

The main advantage of this proposal over our own is that it avoids manipulation. As we said in Sect. 4, we do not think manipulation is so serious a worry. Conversely, the main disadvantage of utility indifference compared to pure shutdown-seeking concerns our claim in Sect. 3.1 that pure shutdown-seeking assists with the specification problem. We think that it is possible to train shutdown-seeking agents using reinforcement learning, by defining a reward function that values outcomes where the agent is shut down. By contrast, we think it would be considerably harder to create AIs with a utility function that encodes the precise weighting of normal goals to the goal of shutdown that is required by utility indifference.

We also depart from Soares et al. regarding general issues surrounding 'corrigible AI.' An AI is corrigible when it will allow you to modify its goals. Soares et al. seek to design an AI that is *corrigible* in the sense of being indifferent about letting you press a button that will change what it values (from normal goals to the goal of shutting down).

Shutdown-seeking AIs may not be especially corrigible. The shutdown-seeking AI may resist human attempts to remove its shutdown goal. After all, it may notice that if the shutdown goal is removed, it will be less likely to shut down. Nonetheless, we've argued that shutdown-seeking AIs will allow humans to shut them down, and will be safe. In this way, shutdown-seeking, and the more general strategy of beneficial goal misalignment, is an approach to safety that does not require corrigibility.

6 Conclusion

We have argued for a new AI safety approach: shutdown-seeking AI. The approach is quite different from other goal engineering strategies in that it is not an attempt to design AGIs with aligned or human-promoting final goals. We've called our approach one of 'beneficial goal misalignment,' since a beneficial shutdown-seeking AI will have a final goal that we do not share, and we will need to engineer its environment so that it pursues a subgoal that is beneficial to us. This could, in some circumstances, make a shutdown-seeking AGI less useful to us than we like. If it is able to develop a dangerous capability (e.g., to disobey our orders), it may be able to shut down before doing what we want. But this 'limitation' is a key benefit of the approach, since it can function as a 'trip-wire' to bring a dangerous AGI that has escaped our control into a safe state. We have also argued that the shutdown-seeking approach may present us with an easier version of the specification problem, avoid dangerous entrenchment behavior, and pose less of a problem of manipulation than its opponents have thought. While there are still difficulties to be resolved and further details to work out, we believe that shutdown-seeking AI merits this further investigation.

Funding No funding to report.

Data availability No data to report.

Declarations

Conflict of interest No competing interests to report.

References

Armstrong, S., & O'Rourke, X. (2017). 'Indifference' methods for managing agent rewards. In *CoRR*, arXiv:1712.06365

Armstrong, S. (2015). AI motivated value selection. In *1st International Workshop on AI and Ethics*, held within the *29th AAAI Conference on Artificial Intelligence* (AAAI-2015).

Carlsmith, J. (2021). Is power-seeking AI an existential risk? Manuscript (arXiv:2206.13353).

Cotra, A. (2022). Without specific countermeasures, the easiest path to transformative AI Likely Leads to AI Takeover. *LessWrong*. July 2022. URL: https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The off-switch game. In *International Joint Conference on Artificial Intelligence*, (pp. 220–227).

Koralus, P., & Wang-Maścianica, V. (2023). Humans in humans out: On GPT converging toward common sense in both success and failure. Manuscript (arXiv:2303.17276).

Martin, J., Everitt, T., & Hutter, M. (2016). Death and suicide in universal artificial intelligence. *Artificial General Intelligence* (pp. 23–32). Springer. https://doi.org/10.1007/978-3-319-41649-6_3 arXiv:1606.00652.

Omohundro, S. (2008). The basic AI drives. In *Proceedings of the First Conference on Artificial General Intelligence*.

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., & Kenton, Z. (2022). Goal misgeneralization: Why correct specifications aren't enough for correct goals. arXiv:2210.01790.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Publishing Group.

Russell, S. (2020). Artificial intelligence: A binary approach. In *Ethics of Artificial Intelligence*. Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0012.

Soares, N., Fallenstein, B. Armstrong, S. & Yudkowsky E. (2015). Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Totschnig, W. (2020). Fully autonomous AI. *Science and Engineering Ethics, 26*(5), 2473–2485.

Trinh, T. & Le, Q. (2019). Do language models have common sense? https://openreview.net/forum?id=rkgfWh0qKX