This is a supplement to the 500-word submission for the AI Alignment Awards shutdown competition. Main ideas are not repeated here. All sections are independent.

1 Corrigibility

While the idea is supposed to be a solution for the shutdown problem, it doesn't have much to do with the corrigibility[1] - the utility function is a direct reversal of the third desiderata from the original article, *U must not incentivize the agent to press its own shutdown button, or to otherwise cause the shutdown button to be pressed.*

2 Outer alignment

Outer alignment is alignment between our goals and agent goals. In the proposed MVP example, agent goal is clear (press the shutdown button), but what is our goal? I'd say it depends:

- Either the agent is safe (i.e. incapable of doing things we don't want it to do), and our goal is to have a protein-designing program.
- Or the agent is not safe, and our goal is to have the shutdown button pressed before it does any harm.

The key problem here is "before it does any harm". I think the MVP example idea is significantly different from the common approach:

- (Usual direction) Let's modify **the utility function** so that the agent **dislikes harming us**.
- (MVP example) Let's keep the utility function super-simple and modify **the environment** so that the agent **has no incentive to harm us**[2].

I believe the incentive-based approach from my proposal might be a good direction for two general reasons:

- We already know that designing and implementing complex utility functions is hard, and this problem disappears. Designing and implementing an environment could be easier.
- (I'm not sure if this is a good argument) When designing a system where agents will be operating, usually the best approach is to focus on incentives.

[1] `https://intelligence.org/files/Corrigibility.pdf`

[2] Or, to be more precise: has no incentive to pursue a path toward its goal that is harmful

3 Inner alignment

I don't think this proposal solves any problems related to the inner alignment, although I believe the extremely simple outer goal might make things easier also on this front.

4 AI boxing (containment)

The problems with AI boxing:

1. We are not clever enough to build a good enough box
2. Once the AI escapes from our defective box, it might be too late for us to improve it

This submission is an attempt to fix the second problem - a way to ensure we'll have another attempt, and thus we can learn how to build a box that is good enough.

5 Risks and problems

5.1 Mesa-optimizers

- If our agent has a misaligned mesa-objective, this proposal solves nothing. Could even be dangerous, because all the safety is based on the assumption that the agent will optimize for the base-objective.
- We could hope that the simple goal will make misaligned mesa-optimizers less common[3].

5.2 Misaligned subagents

Let's assume our agent is capable of creating smarter agents, but itself is not that smart. A path to failure:

1. "I don't know how to press the shutdown button"
2. "Therefore, I need someone who will tell me how to press the shutdown button"
3. "So I will build a universal question-answering engine"
4. (builds a question-answering superintelligent agent)
5. (asks the question-answering agent how to press the shutdown button)

[3] https://www.lesswrong.com/s/r9tYkB2a8Fp4DN8yB/p/pL56xPoniLvtMDQ4J#3_2__The_task

6. (question-answering agent destroys the world while gathering resources necessary for compiling a comprehensive answer)

In other words: our initial agent built a smarter agent that didn't care at all about pressing the stop button, and this ended badly for him (because the shutdown button was not pressed), and for us. Well, how can we be sure subagents will not destroy the world? The answer seems to be "our initial agent must have the alignment problem solved". [4]

5.3 The shutdown button

General question: what does it mean that "there is a shutdown button"?

5.4 The shutdown button and subagents

Let's assume the shutdown button is a physical button that cuts off the electricity in the building where the agent is deployed. Then, if we correctly prepared the environment:

- Either all subagents are running in the same building, and thus the same shutdown button works for them
- Or the agent is capable of interacting with the environment outside the building, and then their main goal should be super-easy to achieve without creating subagents

So, subagents that can resist the shutdown button are possible only in case of an inner alignment failure.

5.5 The reliability of the shutdown button

A scenario we should try to avoid: "there was internet outage, so our escaped agent couldn't send the request, and while looking for a way to assemble a button-pressing robot killed all humans by accident".

5.6 Incorrect design of the environment

All of this proposal is based on the assumption that we are capable of creating an environment where **any** agent with potentially harmful capabilities will be able to press the shutdown button **without any effort**. This assumption might simply be false.

[4] "An agent is not aligned if it's capable of creating agents not aligned with it."

6 Recursive self-improving agent should be enough

In the submission, I write about a "superintelligent agent capable of solving our problem". But superintelligence might not be necessary: ability to self-improve should suffice.

In fact, this might be a way to create intelligent agents that are smarter than us, but not that-much-smarter (and thus not superintelligent):

- Let's say we have an agent that plays chess as good as Stockfish level 5 and is capable of recursive self-improvement
- We deploy it in an environment where shutting down requires beating Stockfish level 6
- It self-improves until Stockfish 6 loses, and now we have an agent that plays a little better

Chess is just an example, we can create any skill-cap-trigger we can conceive, e.g. an IQ test or an exploit-finding test. [5]

7 Side note about the "safe superintelligence deployment MVP" proposal

For simplicity, in the main submission, I defined MVP as something that is "safe". As 100% safety is never possible in the real world, a better one would be something around "it is safe enough for the expected value to be positive".

8 Side note about safety, compared to corrigibility

The corrigibility solution to the shutdown problem is more-or-less "shutdown button is pressed when humans want it to be pressed". But what about cases when we would want to press it if we were a little bit more clever, or had some additional information? An example:

1. Let's assume we want the superintelligent agent to create a big pile of antimatter.
2. We're wrong about the physics governing the behaviour of big piles of antimatter, and thus we're oblivious to the fact that creating a big pile of antimatter will destroy the Earth.

[5] Unfortunately, we can't be sure the new agent is only smart enough to beat our shutdown-button-pressing test. E.g. maybe building an artificial brain with IQ 1000 is easier than building an artificial brain with IQ (exactly) 200?

3. Perfectly corrigible agent told to create a big pile of antimatter creates it, and we die (unless we take some additional precautions - I'm not claiming this is unsolvable in the corrigibility paradigm).

4. An agent who:

   - Wants to press the shutdown button
   - Is told that creating a big pile of antimatter will press the shutdown button

   might develop appropriate physics and - as there is no shutdown button on the destroyed Earth - refrain from creating big piles of antimatter.

In this aspect, I consider my proposal to be "stronger" in terms of safety than the original corrigibility idea.