Uncertainty about value naturally leads to empowerment[1]

Organizers: Catalin Mitelut, Esben Kran
Do not include your own name since we do anonymous review

Abstract

I will discuss some problems with measuring empowerment by the "number of reachable states''. Then I'll propose a more robust measure based on uncertainty about ultimate value. I hope that towards the end you will find that new measure obviously natural. I also provide a Gymnasium environment well suited to experimenting with optionality and value uncertainty.

Keywords: Agency preservation, AI safety, Uncertainty modeling, RL, Empowerment

1. Problems with optionality

One of the possible ways of formalizing empowerment of an agent is "how many states of the world it can achieve" (Franzmeyer et al 2020) (Salge et al 2014). I'll be calling this measure "optionality".

A. Irrelevant options

There are (at least) two major problems with optionality. Firstly, it treats all the options the same, regardless of whether they are potentially desirable or not. Because of that, you may be able to game this measure by creating huge amounts of "barren" options. To give an extreme example: if someone gave you a million new ways to make yourself suffer, that wouldn't increase your empowerment, because you don't intend to use them.

Or more mundanely, if you aren't really careful about what you count as distinct states of the world, someone could game optionality by f.e. providing you with astronomically many ways to rearrange the grains of sand on a desert.

[1] Research conducted at The Agency Challenge, 2023 (see https://alignmentjam.com/jam/agency)

B. Chaos, unpredictability, cluelessness

The second problem is caused by the fact that we simply count "the number of reachable states". But we can be unable to aim at one particular state and actually reach it, even if it is reachable in principle by some actions.

For example: when I have a Rubik's cube, in principle I could reach any of its states in 94 moves. But I just don't know how. Or imagine 3 stars orbiting each other and you have some limited energy to push them around. Because the system is chaotic, with tiny nudges you can bring about many possible states in the distant future. That is even when you can't really predict what your nudges do in particular - you only know they will have *some* effect. So again, we could game optionality by making the world more chaotic.

Good notion of empowerment should measure whether we can achieve some particular states, once we set out to do so.

2. More robust measure

So let's try to measure whether we can actually reach some target once we aim at it. Just taking the likelihood of reaching a set target isn't enough. We may be able to reach a nearby state that is almost as good, and that should also "almost count". A natural way to formalize it, is to say that the "set target" is a utility function U, and "how well we achieve it" is the expected utility of us trying to optimize for U.

But that's just one U. Remember that the point is to be able to aim at multiple ones. After all, we don't know *the final U* right now. And remember the problem of irrelevant options - there are some goals that we can already rule out confidently enough.

So let's say we have (for now magically) some probability distribution p over all possible Us, saying how likely some U is to be the actual U chosen in the future. (Chosen *how* and *by whom* is left uncomfortably open for now.) With that, the most natural thing to optimize for would be:

$$\sum_{U_i \in \mathbb{R}^S} p(U_i) EU[humans\ trying\ to\ optimize\ U_i]$$

(This is related to the notion of "retargetability" (Turner & Tadepalli 2022) (Dupuis & Janus 2022). Here humans can be "retargeted" to pursue other Us and the more competent they are at pursuing many of them, the more agentic/empowered they are.)

A. It's not a weighted average nor a parliament

To clarify, this is definitely not the same as optimizing for some weighted average of those Us. The difference comes from the fact that we expect to better know what we want in the future. In contrast, a "static uncertainty", that doesn't change in time, would in fact give rise to directly optimizing a weighted average of Us.

With uncertainty about value decreasing in time, it becomes instrumental to keep some options open. It becomes stupid to unnecessarily lock-in some outcome even when the average of Us seems decent.

There is a place for some kind of aggregation of Us though. More on that in the section "How do we decide on the distribution over Us?".

3. Toy model

(If you are only interested in the conceptual part, you can skip to "Remaining problems & future directions".)

I created an OpenAI gymnasium environment which I think is well suited to work with optionality and uncertainty about goals. It's based on Sokoban, where we push boxes on a gridworld towards targets. Sokoban is minimal and it's very easy to lock yourself in and not be able to go back. I think it's hard to have a simpler game with this lock-in feature.

Two modifications I made are:

- Boxes and targets have colors, and you have to push boxes to same-colored targets. This feature is already present in some existing variants of Sokoban.
- Color of the boxes is revealed only after some number of steps.

Another possible modification would be to add another player who sees the colors better/sooner. AI would need to assist that player, without trying to lock-in some outcome by itself. I lacked time to implement it. Also it's easier to start with a simpler environment.

This environment can be found here: gym-sokoban-uncertain. And you can find an example RL training pipeline that uses this environment here: rl-sokoban-uncertain. (To run it, look at training.py and visualize.py.)

Below you can see a visualization of a small RL agent running in this environment.

Legend: Red: player. While: wall. Yellow|green|blue: boxes and targets. Boxes are brighter. Each box has to land on a matching color target. Purple: mystery box. We don't know yet what its actual color is. Currently mystery boxes are revealed after 2 steps. Also, in half of the episodes, box colors are known from the start.

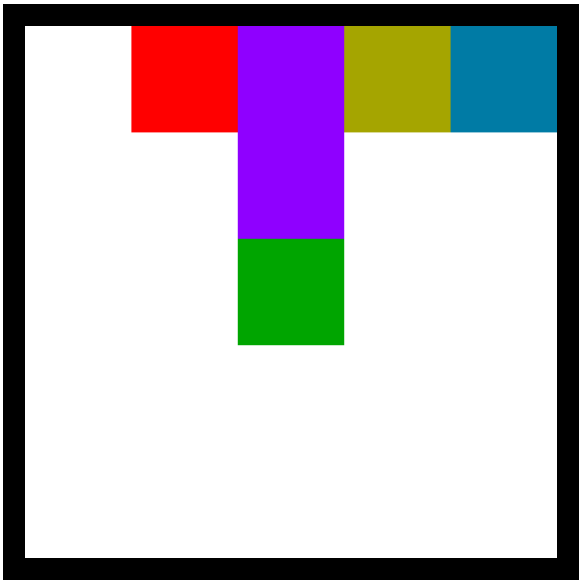

Figure 1 – RL agent running in Sokoban with added uncertainty[2]

Notice how faced with a purple mystery box, the agent stops and waits for the true color to be revealed, rather than pushing it into some state of no return.

Below is another setup, contrasting what happens when we optimize for optionality vs for uncertain values. (Here without RL, just manual play.)

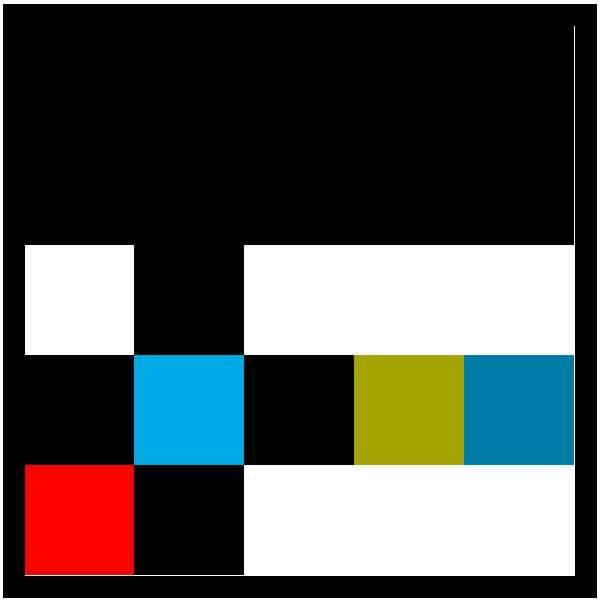

Figure 2 – Comparison between optimizing for optionality vs for uncertain values

You can imagine a human sitting in the top right corner, unable to ever scoop out the box to the big room nor to enter the little room without locking the box in. But the helper agent (in red) can push the box out. (Let's also assume the helper is alive

[2] Note that PDF may not show animations. To view them, read the original gdoc.

for only 5 steps.) It has a choice to push up to the big room, or right to the corridor.

The big room results in higher optionality - the human would then be able to put the box in 8 different spots, while pushing to the corridor provides only 2 or 3 options. But when the box is up, and the helper is already dead, the human will never be able to put the box in one of the targets in the corridor, which is what they actually want!

4. Remaining problems & future directions

Here is a non-exhaustive list of problems with this "uncertainty about value" measure. I believe these problems stem from its incompleteness, so will be solved by adding to that new measure rather than by modifying it.

The ideas proposed here are highly speculative and exploratory.

A. AI can manipulate the choice of future U

If AI can influence which Us get chosen in the future, it can in turn influence the current distribution of plausible Us. And that new distribution may be easier for it to satisfy.

We may run here into self-fulfilling prophecies. Should we try to consider some counterfactual world where AI doesn't exist to manipulate us? Seems infeasible to "consult" this counterfactual word.

This is tied to the efforts to make oracles behave well (Hudson & Treutlein 2023) (Oracle AI).

Another direction is to concretize what humanity's moral deliberation looks like. Once we know that, we will also have a better idea of what it means to not interfere with it.

One of the protections that we may want to implement is to limit the update speed of p, f.e. to some number of bits per year.

A nice toy world to study goal manipulation could be something like Baba is you. There, at first glance the goal is just a part of the world and can be manipulated just like any other part. But if we could precisely define what is some feature that distinguishes goals from other parts of the world, we could then treat them as sacred, and formalize some constraints to avoid influencing them.

B. How to decide on the distribution over Us?

This is closely tied to the question: how is the final future U chosen?

Another one: if someone chooses some U, should they be able to then change their mind? Seems like they should. That "choosing" shouldn't be one-off, but rather a continuous process. Designing that feels very tricky but possible.

For now, we should probably play it safe and have a very diverse set of Us, with distribution over them closer to uniform. (See Bostrom for an adjacent view, although he puts more emphasis on satisficing multiple Us, rather than on keeping the options open.)

To get this wide distribution, we could have all the willing humans, or groups of humans, submitting their Us to the pot. (Due to Arrow's impossibility theorem, people may have reasons to strategically lie about their true U, and we need to figure out how to deal with it gracefully. One way is to embrace that the incentive to lie will exist, similarly to Critch's S-process, and try to ensure that the cost of lying is higher.)

To really submit some U that can be practically used, it needs to be concrete. A dream level of concreteness would be having some deterministic function that can be publicly executed in some untamperable way (similarly to how smart contracts are executed). That function can involve neural networks if we wish, or any other techniques - it just needs to be written in executable code and have some bounded amount of computation. That function would have access to some long, but fixed, list of public APIs. (To ensure the consensus about the output from those public APIs, we could use some scheme in the spirit of Chainlink.) The function runs, making arbitrary calls to those APIs, and at the end must just output a single real number, which is its estimate of "how good the current state of the world is".

We must also choose whether that output value is meant to only estimate the immediate inherent value of the world in the present moment, or if it already contains the expectation of the future value - so taking into account the world's trajectory and all the instrumentally valuable things.

Having such concrete, untamperable Us, could serve as an anchor point for many possible schemes, like prediction markets, various legislation and AI oracles. Bets could be made both on the output of Us, and on public endorsement of given Us.

Last thing to note is that this computability of U doesn't mean it will be rigid, stirile and detached from human thought. The magic can still happen behind those public APIs, which could be served by organizations such as OECD, UN, UNESCO, WHO or others.

5. Appendix

Here are some more, less crucial problems and future directions.

A. Omniscient AGI can't have uncertainty about U

I don't think that it's a fundamental problem here, because in the real world we can't have "real omniscience". There exist processes where to know their outcome, the only way is to run the actual process itself. In other words, there are no shortcuts for them. (Complex and chaotic systems are often like that.) So is humanity's moral deliberation such an un-shortcutable process? Should we make it even more so?

B. How to distinguish humanity's empowerment from general empowerment

In practical settings, it may be hard to distinguish between these equations:

$$\sum_{U_i \in \mathbb{R}^S} p(U_i) EU[\textit{humans trying to optimize } U_i]$$

$$\sum_{U_i \in \mathbb{R}^S} p(U_i) EU[\textit{AGI trying to optimize } U_i]$$

$$\sum_{U_i \in \mathbb{R}^S} p(U_i) EU[\textit{everyone trying to optimize } U_i]$$

We could say that "the world" aims at some U, and then achieves it or not. It's non-trivial to attribute who actually aims and causes some states. It may help, if the humans are somehow required to formally declare which U they are aiming at. (More on that in the next section "How do we decide on the distribution over Us".)

It's not even clear whether we want to distinguish those equations. If we truly care about reaching some states of the world, should we care who brings them about? You may say yes, all else being equal, it's still better that humans directly bring about some valuable states, rather than AGI. If so, we may simply incorporate into our potential Us this desiderata of "humans causing stuff directly".