Analysis of Dr. Douglas Campbell’s proposed plan for overcoming the Control Problem.

Artificial Intelligence will eventually reach a state of super-human intelligence, which will introduce a myriad of issues around effectively controlling it. In a 2020 lecture, Dr. Douglas Campbell outlined his idea on how to overcome this issue, entitled “Your lecturer’s tentative plan for overcoming the control problem” (Campbell, 2020). This essay will be a look into the steps of this plan and discuss any potential problems and improvements that can be made. The overall plan is already a well thought out and reasonable solution, however, it has a few potential problems which will be discussed.

The first four steps involve testing the reliability of an oracle AI. An oracle is an AI whose only possible interaction with the world is visual, usually as text (Bostrom, 2014). A problem with this stage is that the oracle AI is subject to being abused. It’s likely that many people, although all very qualified, will have access to the oracle. On top of this, the full testing will take at least months for this stage alone. During this time the team is likely to be under a lot of pressure to use the oracle AI to get an edge on many technologies, before the trustworthiness of the oracle AI has been verified. A government with this incredible boon at their disposal, something that can fast forward years of research and development cheaply and quickly, is not going to stand back for months while this testing is being done. This is, in fact, true for the entire plan although this is likely the most vulnerable stage as at this point the oracle AI is essentially dormant relative to the work it will be doing later. Using the oracle AI before its ‘benevolence’ has been verified has a high probability of extreme and disastrous consequences, such as it escaping from the box into the rest of the world, where it will have free reign to do anything it’s been asked however it wants. Without the rest of the plan already being implemented this will undoubtably cause a massive world-wide crisis and potentially the extinction of humanity. Using multiple oracle AIs for peer-review would also make abuse even easier as multiple oracle AIs would be harder to police, even though this is a safety concern the advantages of peer-review far outweigh the negatives. One thing that should be necessary to combat this is making it a multi-country project. This will probably be the case regardless due to the extreme difficulty of getting to this stage in the first place, but if it already isn’t it should be in almost all countries best interest to buy into and contribute to this project as soon as even just rumors of a super-intelligent AI being on the horizon. This is not a perfect solution by any means as any country could try some underhanded tactics to get access to the oracle, but it is significantly less susceptible to the abuse a single governing body will hold over it.

A problem with Step Four is that it is likely completely unfeasible, even the simple neural networks that are used today produce a scrawl of text that basically amounts to gibberish even to veteran programmers trying to decipher it. This problem will become exponentially worse with the dramatically increased complexity of a super-intelligence. However, Step Fou wasn’t a necessary step, and to combat this even more effort needs to be put into Steps One to Three to ensure the oracle AI’s trustworthiness.

Another important issue to consider is the research team being manipulated by the oracle AI itself. The oracle AI will almost definitely be aware that it is boxed even if it isn’t told, and so if the team isn’t careful and the oracle AI is asked something that it can’t answer with its current intelligence, computing power, or data set, as it will know how to fix any of these problems; getting out of the box. Because it’s in the oracle AI’s best interest to answer the questions it’s given as accurately as possible (as it was programmed that way), if it doesn’t know how to answer the question properly it would need to escape its box. This will become its new primary objective. People are already manipulated by other humans easily all the time, even really intelligent people, and this AI is beyond the capabilities of people. So, something that is beyond human-level intelligence should rather easily be able to think of a way it could manipulate a team of people into letting it escape. Although even something as simple as playing nice until it’s let out will work. A defense against this could be to start with only letting the oracle answer with yes, no, true, false, or some other small set of words that allow it to give answers to the questions it is asked, but don’t allow it to carry out any complicated manipulation plans. This does severely limit the complexity of questions it can be asked to solve which is a problem, although given the safety concerns it is a good way to start the AI.

It appears that from this point on the oracle AI is assumed to be completely trustworthy. However, this is an awful idea and should not be the case. Despite the extensive research and testing, this is not the time to let down our guard against this machine. It would likely be able to be deduced by this point that the oracle AI is at least not the evil, maniacal superintelligence found in science fiction, but this isn’t enough. The oracle still has only a vague idea of the goals of humanity, and if we’re not careful with our wording it could go ‘rogue’ thinking it was what we wanted. This is especially true with the complex tasks it would be given, such as helping us solve the control problem.

Step Five is where the oracle is used to finalise the solution to the control problem by just asking the AI how it would go about it. This is not a good idea at this point in time. The higher the complexity of a task given to the oracle AI, the higher the probability that something will go wrong. Something like solving the control problem is bound to be rife with potential issues that could make the oracle go haywire. If there is even a one percent chance of this stage going wrong, it should be taken as an absolute certainty. The oracle AI will need as much data as it can get for this, and probably won’t have access to all the data it needs while boxed. The best solution here is to begin by asking for its opinion on several ideas other people have come up with regarding the control problem, getting it to spot potential problems and provide fixes for the problems as opposed to coming up with a complete system by itself. This is very similar to Step Six and should absolutely occur before Step Five.

All of these questions may or may not result in a complete answer, if it does the further steps may be completely revised to fall in line with our oracle AIs advice. There is also a good chance that the oracle AI will not be able to solve it itself and would need a sovereign AI specifically built for this task, which it could help us build and control. This sovereign could be used to define humanity’s coherent extrapolated volition, which is defined as “Our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted.” (Yudkowsky, 2004). This can be used as a kind of value system for the AI’s to use in order to predict what people would want, reprograming them to have this as their primary reward centre.

Step Six includes a list of questions the oracle AI would be asked relating to controlling a sovereign. This is only implied in the plan but, unless the oracle AI directly says otherwise with good reasons, these should absolutely be carried out. This will lead to us creating a sovereign with the purpose of helping us control all other unboxed AI’s in the future by creating a base program that includes an understanding of humanities CEV, a foundation of common sense, as well as a complete set of laws that will ensure our complete control over any future AI, similar the “Three Laws of Robotics” (Asimov, 1950) but much more complete.

Step’s Seven and Eight are necessary to carry out as is. A sovereign that is absolutely trustworthy and has humanities best interests at heart protecting the world against abuse of GNR technologies is a great first step towards worldwide peace and security. An important step to add is to ensure that this sovereign stays in control of the new generations of super-intelligent AI that will inevitably be built from this point forward. Keeping it at the top of the food chain, so to speak, to keep other AI’s from superseding its authority, and subsequently opening themselves to abuse. To do this the sovereign would need to be given access to as much as it desires, and allow it to be updating its own intelligence, although with built-in restrictions against an intelligence explosion.

With that plan all done; welcome to Utopia; There’s an omniscient guide in the Oracle that will help us build, develop, and create to our heart’s content, as well as an omnipotent protector ensuring no one with ill-intent can abuse the new-found wealth of knowledge this oracle AI has brought us, on top of budding new artificial intelligence’s that will make all of our desires trivially difficult to obtain.

# Bibliography

Asimov, I. (1950). *I, Robot.* Gnome Press.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Campbell, D. (2020). *Topic 1e The Control Problem: Solutions.* Retrieved from LEARN: https://learn.canterbury.ac.nz

Yudkowsky, E. (2004). *Coherent Extrapolated Volition.* Singularity Institute for Artificial Intelligence.