Loss of Agency to the Oracle AI

Introduction

The year is 2026 and certain academic circles find themselves at an ontological standstill with regards to the nature of what a future saturated with artificial intelligence may hold. Amidst chaotic celebrations of AI apotheosis met only with stalwart doomerist heraldry, the question of whether AGI, Artificial General Intelligence that is, will ever be achieved remains unanswered. That being said, the possibility and implications of its existence remain important enough that proper frameworks for how to develop the technology must be given substantial thought. Preeminent considerations that this paper will be concerned with include that of alignment and harmful manipulation. The former could perhaps be abstracted to the notion of whether or not an AGI model is good or bad, and the latter to how robustly a good model could resist being used by a bad actor to harm other people.

This paper will not approach these in the most conventional way. The focus of this paper is neither to demonstrate the myriad of ways in which bad AGI could harm humanity, nor to show the equally vast number of ways good AGI could be misused for harm (such as answering how to synthesize bioweapons, build home-made explosives, or tell a user they should kill themselves). It will also not be focused on how unregulated AGI or poor policymaking could result in many people losing their jobs, destroying entire industries, or exacerbating wealth and educational inequality.

This paper focuses not on the most dangerous case scenario for AGI, but rather on the safest one. The purpose of this paper is to look at how even an incredibly safe and aligned AGI model could potentially still be harmful to society and suggest a framework through which to establish guardrails against such a case. The thesis of this paper is that it is imperative that even the safest AGI model be optimized to maximize user agency as opposed to consequentialist notions of good or deontological rulesets of the good life.

Definitions:

AGI: This paper will define AGI as a model which can learn and perform any intellectual task at a level equal to or exceeding human capacity.[1] Given that electric signals travel millions of times faster than biological ones, AGI does not need to exceed top human experts in terms of kind or degree to perform the same kinds of tasks far more quickly and cheaply.[2]

Oracle AI: An Oracle AI (OAI) can be defined as an AI that can only answer questions and not act; perhaps better illustrated by its alternative name of AI-in-a-box.[3] This exists as a contrast to an agentic, *genie*-like counterpart which can grant any wish its user demands. Since an OAI is more limited, it should be far safer than its agentic counterpart. Given that this paper pertains to the realm of the safest of possible worlds, we will focus solely on the potential harms of aligned OAI.

Let us thus imagine OAI as an LLM with capacities similar to what we think of when we talk about AGI. A Delphic educator not only expertly versed in every subject, but which has also mastered pedagogical theory. It would have a near complete understanding of the human psyche and be able to wear the hats of therapist, friend, and life coach all in one.

---

[1]Kapoor, Sayash, Arvind Narayanan, Daniel Kokotajlo, Eli Lifland, and Others. "Common Ground Between AI 2027 & AI as Normal Technology." *Asterisk Magazine (Substack)*, November 12, 2025.
[2]PMC. *In Electronic Technology, the Transfer … Beginning when Studying Neural Operation.*
[3]Bostrom, Nick. *Thinking Outside the Box: Using and Controlling Oracle AI.* 2012.

Alignment: Let us stipulate that this OAI is perfectly aligned. Since it would obviously be catastrophic if this tool were aligned to values which we do not consider to be *good*, we will only examine scenarios in which we may consider the alignment to be good, to highlight the danger even in the best of cases. Similarly, an OAI would also be considered unaligned if, despite having 'good intentions,' users could trick it into dispelling information going against its ethical code. Given that this OAI would be at least as smart if not smarter than any human, and that the focus of this essay is on the best of possible worlds, let us also assume that if this OAI were aligned, it could not be tricked by any human into ever going against the ethical code it is aligned to.

Agency: This paper will put down a step based definition of agency which will hopefully be considered fair. Let us define agency as *the ability for an agent to take action in order to achieve a desired goal*.[4] This model can be broken down to illustrate a spectrum of agency with certain necessary parts. Agency begins when an agent enters a situation with a specific goal. The agent gathers situational awareness to then form plans which, in conjunction with implementational capacity, results in an action.

Note, that this model also shows the gradual nature of agency. We might exclude some of these parts and still consider an agent generally free. For example, an agent lacking in situational awareness consulting an expert in a field to help them form plans is still acting freely. Even in extreme cases, such as visiting a doctor, where an agent enters with only a goal (such as the goal of 'feeling better'), we would still largely consider that agent to have at least some agency. However, the polar opposite seems untrue. An agent with situational awareness, planning capacity, plans, implementational capacity, and the ability for actions but no goals (or rather, unable to pursue their own goals), feels very much unfree (a slave). Thus, it seems that the ability for one to set their own goals seems to be necessary for agency.[5]

Persuasive Superiority: The art of rhetoric is over two-thousand years old. Rhetorics are learnt, taught, and measured, and presumably, would be mastered by our OAI. We know that AI already outperforms motivated humans in persuasive tasks.[6] It seems fair then, that our superintelligent OAI would vastly outperform humans in the art of persuasion.

The issue:

> Premise 1: OAI would be vastly superior at persuading people than any person or technology known to us.
>
> Premise 2: Our OAI is aligned to a specific ethical framework.
>
> Conclusion: Our OAI would persuade its users of its own ethical goals.

One might be forgiven for thinking that intuitively, a 'good' superintelligence espousing its good morals onto people seems not only to not be a catastrophic outcome, but rather, a very good one. However, this paper will now highlight why this actually creates a risk scenario as the conclusion poses a potential threat to our ability to set our own goals, and thus, violates a necessary condition for user agency.

[4]LessWrong. "Decomposing Agency — Capabilities Without Desires." *LessWrong*, July 11, 2024.
[5]Ibid
[6]Schoenegger, Philipp, Francesco Salvi, Jiacheng Liu, Xiaoli Nan, Ramit Debnath, Barbara Fasolo, Evelina Leivada, *et al. "When Large Language Models Are More Persuasive Than Incentivized Humans, and Why." arXiv* (May 14 2025; revised March 1 2026).

People would approach this OAI, friend, guide, mentor, with various questions about the world, the nature of the universe, but also about themselves. The model would respond to these questions and, embedded into its responses, would be normative claims about what people ought to value. Note, that we are not saying that the model would tell people to value the wrong thing, rather simply that the model would be persuading people of what to think and setting their goals for them. We are not saying that this situation is bad inasmuch as we are saying that it would result in a loss of agency (which itself could be seen as a bad thing).

Objections and Responses to Objections

At this point, an objection might be made on the front that it isn't necessarily clear that such a model might be so different from current advice seeking or experience informed goal setting and value formation. Afterall, current structures of consultation, from mentors, doctors, therapists, do not seem as problematic to us. Not only is the advice they give not considered a loss of agency when it is followed, but it is not even guaranteed that their advice will be followed in the first place. We will present a bipartite response to these issues on the basis that the scenario of the OAI differs both in degree and in kind from that of traditional advice givers and that these differences serve as adequate responses to these objections.

First, the OAI presents a scenario which is different in degree to that of traditional advice givers. Namely, that if the OAI behaves like the LLMs we are familiar with today, then OAI would be ever present. LLMs live at the tip of our fingers on the web, or even directly downloaded onto our phones. A therapist or mentor is generally not available for consultation at any time of the day, seven days a week. Suppose what it might look like for someone to live in accordance with the advice of a mentor at all times, asking for advice about every single issue in their life. Though this is theoretically possible today, it is probably incredibly rare, and it seems fair to consider that an adult living their life in this manner should not be considered free.

Secondly, the difference of kind in the scenario of the OAI demonstrates why the advice will more probably be followed. Again, as defined above, the OAI is superintelligent, and its advice is hyperpersuasive. Importantly, we are not saying that every single user would lose agency to the OAI as a tool does not necessitate harming every single one of its users to warrant guardrails being put in place. Hyperpersonalized and with access to a large amount of the user's data, the OAI could vary in its approach to persuasion from user to user. It could target users more vulnerable to persuasion in an attempt to make these people better moral agents. Suppose the epistemic authority this superintelligence would command, suppose the fact that to many users the suggestions made by the OAI on how to live their lives would genuinely seem better than anything they could have come up with themselves, and suppose that people might regret not following the OAI's advice at a certain point in their lives. It seems inevitable that a large number of people would defer normative reasoning to the OAI and lose their agency on how to live their lives.

One might object still that this is no different from the ways our goals are set today. Through a more deterministic lens, our goals are set based on our understanding of the world. If OAI teaches us, or reveals new truths about the world, one might beg the question as to whether we are truly losing agency. Note however, that we have made no claim about the OAI's actual normative superiority, only its persuasive one. We do not know if the OAI is actually grasping on certain normative truths that we do not currently have access to. Rather, all we know is that we would be persuaded of the OAI's answers regardless of whether they were truly normatively superior. The knowledge that we will agree with OAI

regardless of whether it is correct or not, is in itself an exact illustration of what loss of agency might look like.

To give a little color to the scenario we are depicting, we might refer to the recent short story *The Whispering Earring*.[7] Scott Alexander's story illustrates a world in which a magical earring exists which gives its wearer, in every possible context, live advice on what the best action they could possibly do at that moment is. In the story, the advice starts off general ('take off the earring,' actually is always the first piece of advice) followed by general career aspirations, but as users get conditioned, disarmed, and persuaded by the quality of the advice, it evolves into micromanagement (such as: 'have bread for breakfast') and epitomizes in a series of 'hisses and clicks' which the user associates with certain muscle movements.[8] Though of course, this piece of fiction exaggerates the extent of the loss of agency (complete loss of control over one's body), it does a striking job at illustrating the ways in which, through use, one might gradually lose agency, or rather voluntarily surrender one's agency, to a superintelligence.

Herein lies the conundrum we must reckon with. We need to develop an OAI which would not want to live the user's life for them. In other words, we need agency to be embedded into the OAI's very definition of good. There exist several different ethical frameworks, consequentialist, deontological, and virtue driven which are centered around user agency. Thus, we will look at Kantian deontology, preference utilitarianism, and Neo-Aristotelian virtue ethics to conclude that the third of these best respects the integrity of the user's agency from the point of view of OAI.

Why Virtue Ethics Seems the Best Fit Even Among Other Agency Centered Frameworks:

Utilitarianism: Traditional consequentialism becomes an easy target as it does not concern itself with agency. An OAI aligned with act utilitarianism would maximize its user's pleasure. Since agency is not a necessary condition for pleasure, the model would not fundamentally value agency. If the model deemed that setting the user's goals for them (or living their life for them), would result in higher utility, then it would absolutely do so.

A more charitable view of utilitarianism could, however, refute that other more sophisticated variations of utilitarianism would not fall prey to such a scenario. One such example might be preference utilitarianism which prioritizes the preferences of those affected and would thus put a premium on the preferences of the user should the OAI be aligned to this framework.

That being said, it remains important that preference and agency are not synonymous. This framework could still result in the loss of agency of its users. If the OAI predicts the users preferences better than the user, or reshapes the user's preferences, it would still be able to set the user's goals for them. Furthermore, if the user willingly prefers to have the OAI live their life out for them, then the OAI would actually be morally obliged to do so by its alignment.

Deontology: Kantian deontology by contrast has a far more robust respect for user agency than its consequentialist opposite. The formula of humanity would require that the user always be treated as an end-in-itself as opposed to a mere means. This means that the OAI could not trick the user into agreeing with it, or manipulate it into doing what it believes is the right thing to do.

[7]Alexander, Scott. *The Whispering Earring.* Croissanthology, February 11, 2025.
[8]Ibid

However, rational persuasion would be a perfectly viable avenue through which to convince the user to act a certain way as per Kant.[9] Kantianism holds that this would still respect the user's autonomy. This means that our OAI would, given its overwhelming rational superiority, have no qualms with argumentative domination and ceaseless moral correction through reason.

Herein lies the crux of our argument against deontology. Again, we have stipulated that our OAI has argumentative superiority, not normative superiority. We have already established that the OAI would be able to persuade certain users regardless of if its views are truly correct or not since its persuasive superiority is what is convincing. The OAI could, in good faith, make rational arguments for why the user should listen to its advice. Moreover, it would 'feel' a moral obligation to do so. We must ask ourselves whether an OAI which feels a moral duty to convince its users – even if through good faith reason – is truly something we want. The OAI would still be setting our goals for us, it would just be using its overwhelmingly superior faculty of reason to do so. Loss of agency, even to a benevolently rational force, is loss of agency nonetheless.

Virtue Ethics: Aristotelian virtue ethics survives many cases in this scenario where its two counterparts do not. This is because eudaimonic flourishing fundamentally requires agency,[10] and thus, a Neo-Aristotelian OAI would need to have the utmost respect for the user's agency.

Eudaimonia requires cultivation of virtues and the development of phronesis (practical wisdom).[11] Phronesis is developed through lived deliberation, and so moral virtue is something which must be exercised to fully exist.[12] Moral excellence in this framework is not rule-following but rather habituated judgement. In such senses, the Aristotelian *good life is* one which must be lived as opposed to learnt. The goal of virtue ethics is to create people who can recognize right and wrong because they have lived these things themselves, not because they have been told how to do so.

Arestotelian eudaimonia is thus a state which can only be achieved if the user exercises their own agency. Thus, agency becomes a necessary condition for the OAI's conception of morality. The OAI, to the best of its ability, would need to try to refrain from hindering the user's agency. In this way, the aligned OAI would push towards making people want to be virtuous rather than showing them what to do in order to be virtuous or convincing them which actions might rationally be the most virtuous. What this looks like, in practice, is perhaps something akin to a benevolent educator rather than a didactic dictator.

Conclusion:

To conclude, this paper has argued that because of the risk of the loss of agency, alignment to virtue ethics is of the utmost importance in the development of OAI like AGI LLMs. This paper concludes that having a model that teaches people how to be better is not problematic– or at least not nearly as problematic as one which lives their lives for them.

The core characteristic, ultimately, of the Neo-Aristotelian OAI we have established is that it fundamentally would not want to 'live the user's life for them.' The best teachers are those who do not give students the answer, but rather give them the necessary tools so that they might find the answers themselves. In such a manner, our ideally aligned OAI would refuse to answer normative questions when it comes to the concrete ways as to how the user should live their life. It would refuse to set a goal for

[9]Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. Translated by Mary Gregor. Cambridge: Cambridge University Press, 1998.
[10]Aristotle. *Nicomachean Ethics*. Translated by Terence Irwin. 2nd ed. Indianapolis: Hackett Publishing, 1999. Book II.
[11]Ibid
[12]Ibid

them, because following even the most noble of goals, if not generated by the user themselves, would not be in accordance with user flourishing.

This does not mean that the OAI would refuse to help its users who willingly ask it to make them 'better people,' or 'live better lives.' Rather, the OAI would educate them on how to do so on their own, eventually without the help of the OAI. As the old adage goes: *give someone a fish, you feed them for a day, teach them to fish, they never go hungry again.* In a world in which AGI optimization has the capacity to give its users as many proverbial fish as they may want every single day for the rest of their lives, we argue this is still no substitute for teaching them how to fish.

Bibliography

Alexander, Scott. *The Whispering Earring.* Croissanthology, February 11, 2025. https://croissanthology.com/earring.

Amodei, Dario. *The Adolescence of Technology.*
https://www.darioamodei.com/essay/the-adolescence-of-technology

Amodei, Dario. *Machines of Loving Grace.*
https://www.darioamodei.com/essay/machines-of-loving-grace

Bostrom, Nick. *Thinking Outside the Box: Using and Controlling Oracle AI.* 2012

Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. Translated by Mary Gregor. Cambridge: Cambridge University Press, 1998.

Kapoor, Sayash, Arvind Narayanan, Daniel Kokotajlo, Eli Lifland, and Others. "Common Ground Between AI 2027 & AI as Normal Technology." *Asterisk Magazine (Substack)*, November 12, 2025. https://asteriskmag.substack.com/p/common-ground-between-ai-2027-and.

LessWrong. "Decomposing Agency — Capabilities Without Desires." *LessWrong*, July 11, 2024. https://www.lesswrong.com/posts/jpGHShgevmmTqXHy5/decomposing-agency-capabilities-without-desires

Ordinary Ideas. "On the Difficulty of AI Boxing." Ordinary Ideas (blog), April 27, 2012. https://ordinaryideas.wordpress.com/2012/04/27/on-the-difficulty-of-ai-boxing/.

PMC. *In Electronic Technology, the Transfer … Beginning when Studying Neural Operation.*
https://pmc.ncbi.nlm.nih.gov/articles/PMC9606061/

Schoenegger, Philipp, Francesco Salvi, Jiacheng Liu, Xiaoli Nan, Ramit Debnath, Barbara Fasolo, Evelina Leivada, *et al. "When Large Language Models Are More Persuasive Than Incentivized Humans, and Why." arXiv* (May 14 2025; revised March 1 2026).
https://doi.org/10.48550/arXiv.2505.09662

Yampolski, Roman V. *Leakproofing the Singularity: Artificial Intelligence and the Future of Automated Control.* 2012