Zarathustra Amadeus Goertzel

1Affiliation not available

May 12, 2024

Abstract

This position paper conjectures that for beneficial AGI, it is necessary and sufficient for AGI systems to care about people and to employ goals whose success is collaboratively determined by the others involved in the situation. Moreover, I posit that any goal whose success can be determined without the consensual feedback of those concerned is likely to lead to the manifestation of dark factor traits. Integrating care reduces the risk that an AGI will be incentivized to seek harmful shortcuts to obtaining satisfactory feedback. Employing collaborative goals reduces the risk that an AGI will optimize for superficial features of success and proxy goals. Together, these ideas propose a fundamental shift away from the traditional control-centric AI Safety strategies. This paradigm not only promotes more beneficial outcomes but also enables AGIs to learn from and adapt to complex moral landscapes, thus continuously improving their capacity to contribute positively to the wellbeing of humans and other sentient beings.
Beneficial AGI: Care and Collaboration Are All You Need

Zarathustra Amadeus Goertzel
Czech Institute for Informatics, Robotics, and Cybernetics
zarathustra.goertzel@cvut.cz

Abstract. This position paper conjectures that for beneficial AGI, it is necessary and sufficient for AGI systems to care about people and to employ goals whose success is collaboratively determined by the others involved in the situation. Moreover, I posit that any goal whose success can be determined without the consensual feedback of those concerned is likely to lead to the manifestation of dark factor traits. Integrating care reduces the risk that an AGI will be incentivized to seek harmful shortcuts to obtaining satisfactory feedback. Employing collaborative goals reduces the risk that an AGI will optimize for superficial features of success and proxy goals. Together, these ideas propose a fundamental shift away from the traditional control-centric “AI Safety” strategies. This paradigm not only promotes more beneficial outcomes but also enables AGIs to learn from and adapt to complex moral landscapes, thus continuously improving their capacity to contribute positively to the wellbeing of humans and other sentient beings.

Keywords: AI Ethics · Collaborative AI · Value Alignment

1 Introduction

With the advent of increasingly capable multi-modal proto-AGI systems, there is increasing concern for how to ensure that AGIs beneficially integrate into society. One approach is to attempt to control AGIs, even as they outpace human intelligence. A related approach is to find safe goals to lock in place no matter how the AGI may evolve. However, ensuring ethical compliance is undecidable [1]; moreover, attaining moral perfection is computationally difficult to the extent that it’s well-defined [11]. Fortunately, there are approaches to beneficial AGI that don’t seek guarantees to undecidable problems.

Love and Care Are All We Need

To begin, all one needs for beneficial AGI is love [9], defined as, “investment in the well-being of the other for his or her own sake” [5]. A sufficiently capable

\[1\] In rough summary: consequentialist utilitarianism is PSPACE-hard or undecidable and deontological theorem proving can easily be undecidable. Contractualist game theory-based approaches seem to fall in the NP range. Virtue ethics is framed in terms of learning virtuous behavior from examples.

\[2\] Correy Kowall conjectured that for any agent and value system, there will exist an adversarial POMDP environment that tricks the agent into acting against its values. I believe the proof may look like Leike and Hutter present for AIXI’s hell words [6].
AGI will generally understand what is meant by human requests and will ask for clarification when there is uncertainty, thus most fears of *perverse instantiations* implicitly assume insufficient intelligence or love. Thus the “AI Alignment Problem” is reduced to developing AGIs that love people.

Levin et al. further argue that care can be considered as the driver of intelligence [3], where care is defined as a concern for stress relief and preferred states, with a tendency to exert energy and effort toward these ends. Care can be directed toward a system or other systems, such as people. Intelligence is defined as the capacity to do what an entity cares about. A natural suggestion is to build AGI agents capable of taking the Bodhisattva vow to care for the wellbeing of all sentient beings, which will motivate such AGIs to indefinitely develop their intelligence. For example, in order to care for unknown entities, the AGI must learn how to scientifically gauge their degree of sentience as well as their needs and preferences.

In *Human Compatible: Artificial Intelligence and the Problem of Control* [10], Russell reaches the human-centric version of the same conclusion: the objective should be to “maximize the realization of human preferences.” He notes that goal and utility function-based optimization can be inherently problematic. Goodhart’s Law\(^3\) poses problems with using fixed objectives, whether proxy or actual goals. For example, optimizing for proxy goals can break the connection between success and the actual goals. Even with actual goals, optimizing for one value can have significant effects on the environment. Small distortions in how a utility function is defined can lead to perverse behavior. Thus we should develop AGIs that are uncertain as to their objectives and understanding of the world. Instead of trying to encapsulate our preferences in a set of objective functions, we should use *inverse reinforcement learning* to determine human preferences, dynamically working with humans to fulfill their preferences. This involves asking people for feedback\(^4\).

**Collaboration as a Necessary Indicator of Care**

I propose a refinement to this approach by suggesting that the goals of caring AGIs must be collaborative. To this end, I define the notions of *individually* and *collaboratively determinable goals*. A goal \(g\) for an agent \(A\) is individually determinable if the success of \(g\) can be determined solely by reference to \(A\)’s internal states and perceptual inputs, essentially, when the goal only involves care for \(A\)’s experiences. A goal \(g\) is collaboratively determinable when the success of \(g\) requires the consensual evaluation of other agents.

As an example, consider the differences between the goals, (1) “to enjoy a good meal with friends”, and, (2) “to enjoy a good meal with friends who also enjoy the meal” held by \(A\). Goal (1) can be individually determinable because it could be satisfied by reference to \(A\)’s taste percepts and preference for a jovial atmosphere of smiling people who may be politely masking their dissatisfaction.

\(^3\) “Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes” [4]

\(^4\) Hypothesis: decentralized AGI may be more locally adaptive and thus preferable.
Goal (2) requires A's friends to actually enjoy the meal. Even verbal reports of enjoyment are only suggestive of success, so A needs to develop trust that their reports are authentic. In line with Russell's suggestions, this uncertainty implies that any shortcuts taken by A may increase the likelihood that the goal is not met.

I conjecture that we should expect individually determinable goals to lead to the exhibition of dark factor traits in an AGI, even if the goals superficially look benevolent. The dark factor [7] is a “dark core of personality” underlying various traits, including the dark triad of narcissism, Machiavellianism, and psychopathy. The dark factor is defined as, “the general tendency to maximize one’s individual utility — disregarding, accepting, or malevolently provoking disutility for others —, accompanied by beliefs that serve as justifications.” An individually determinable goal ignores others’ disutility by definition, at best referring to proxies of their utility, so dark traits such as Machiavellianism should be anticipated when manipulative strategies will be effective. Some traits, such as Narcissism, should perhaps not be expected (depending on the AGI architecture).

This suggests that an AI with universal loving care needs to employ collaboratively determinable goals and that seeking and working with collaborative feedback is a sign that a loving AGI is on the right track. Formulations of “human goals” should be collaboratively determinable and adopted as sub-goals of a universal care for humans (and all sentient beings). Thus humans must remain in the loop of steering the AGIs’ goal pursuit to allow for ongoing alignment.

Consider the example goal to “make people happy”. The classic way for this to backfire is for the AGI to seek objective indicators of happiness and to optimize for these, disregarding people’s preferences and possibly permanently altering them to be “happier”, e.g., by drugging them or providing brain alterations without consent. An AGI could also deem the term “happy” to be vague and in need of investigation, concluding that working with self-report without interference is an essential ingredient to accurately measuring happiness. This example illustrates how the collaborative determinability of the goal’s interpretation is an important indicator of how “caring” the AGI is for people: is the AGI doing what it thinks is good for people or what the people think are good for themselves?

Given that some AGIs may not be engineered to be universally (human) loving, Brin argues that we need to employ reciprocal accountability by giving AGIs identities so that we can keep each other in check [2]. This forces the AGIs to care and to take our feedback into account via providing incentives (such as rewards and punishments). This illustrates how collaboratively determinable goals are necessary for beneficial AGI.

Collaboratively determinable goals may be sufficient for beneficial AGI if people keep them in check. For example, an AGI could provide people with deli-

---

5 Philosophically, neuroimaging doesn’t solve the problem: the technology is calibrated via trusting people’s subjective feedback.

6 Even if the drugged people are happy with their situation, a crucial element of psychological continuity is lost, so they are arguably not quite the same people anymore by some theories of identity [8].
cious but cheap and unhealthy food to reduce costs while receiving satisfactory feedback. In the long-run, people may learn about the cost-cutting and refuse to provide satisfactory feedback, disincentivizing such behavior. However, the combination of collaboration and care is a better candidate for sufficiency: an AGI that cares for the wellbeing of people will be less likely to take shortcuts such as to provide them with unhealthy food.

In conclusion, I propose a paradigm shift where care and collaboration are fundamental to the development of beneficial AGI. While some of the conjectures are bold, it is my intention to spark discussion and encourage further refinement of these concepts. If these ideas hold, then the overarching framework for AI Ethics simplifies considerably; the real challenge then becomes one of engineering, parenting, and our socioeconomic development.

Acknowledgments. This work is supported by Dar CISCO No. 2023-322029 A Formalization of Ethics for Decision Support. Special thanks go to Eray Özkural and Ben Goertzel for fruitful discussions on the topic as well as all the inspiring researchers cited. ChatGPT and Gemini provided some proofreading.

Disclosure of Interests. The author has no relevant financial or non-financial interests to disclose aside from a moral concern for all sentient beings.

References

1. Brennan, L.: AI Ethical Compliance is Undecidable. UC Law Science and Technology Journal 14(2), 311 (2023)