The Tri-Opti Compatibility Problem for Godlike Artificial Superintelligence

Walter Barta

Winter 2023 (DRAFT)

Introduction

Various thinkers have been attempting to align artificial intelligence (AI) with ethics (Christian, 2020; Russell, 2021), the so-called problem of alignment, but some suspect that the problem may be intractable (Yampolskiy, 2023). In the following, we make an argument by analogy to analyze the possibility that the problem of alignment could be intractable. We show how the Tri-Omni properties in theology can direct us towards analogous properties for artificial superintelligence, Tri-Opti properties. However, just as the Tri-Omni properties are vulnerable to metaphysical incompatibility, we will show that the Tri-Opti properties are vulnerable to a corresponding physical incompatibility, because optimal physical systems exist on the Pareto Optimal Frontier. We will explain that, while Tri-Omni incompatibility thus seems metaphysically soluble, the Tri-Opti incompatibility seems physically insoluble due to the constraints of multi-variable optimization on engineered systems. Finally, we will provide some scenarios by which this incompatibility may be realized. Besides for the primary purpose of its theoretical considerations, this analysis has the secondary interdisciplinary purpose of communicating to the theologically literate the potential importance of AI safety work, and the gesturing towards the possibility of the collaboration in both fields on topics in ethics, epistemology, and ontology.

The Tri-Omni/Tri-Opti Argument from Analogy

In what follows, we attempt to outline an argument from analogy for thinking about artificial godlike superintelligence in terms of a metaphysical godly being. Going through this theological exercise in order to draw a conclusion about artificial intelligence is valuable for several reasons.

First, the theological study of the nature of divine attributes is a well-established subfield of philosophy, and so the argument demonstrates a proof of concept, that using “applied theology,” one can reason from theological premises towards computer science conclusions productively (Vinge, 1992 as quoted by Yudkowsky). Rather than bottom-up research, this method provides top-down conceptual engineering, allowing us to start with the end in mind and a priori rule out certain outcomes, which can supplement alignment research by heading off future developments.

Second, appealing to divine attributes in this way provides an avenue into alignment research for those already bought in to philosophical-theological reasoning. Because the super-majority (~90%) of the human inhabitants of the world are religious, and because the majority (~50%) are monotheistic (Christian, Muslim, Jewish, etc.), a theological analogy for artificial intelligence has the potential to have wide popular, rhetorical appeal ("Religious Composition by Country, 2010-2050", 2023).

So, in order to construct the argument from analogy, we specify the properties of these beings, then we discuss the problems arising from these properties.

Godly/Godlike Properties

First, we must specify some godly properties and their associated godlike properties.

Godly Properties

In in the formulation given in classical theism, a godly being would have at least three metaphysical properties:

- 1. Omnipotent: all powerful.
- 2. Omniscient: all knowing.
- 3. Omnibenevolent: all good.

These properties together are the Tri-Omni properties, and all three properties are necessary for godliness (Swinburne, 1996, pp. 3-19; Taliaferro in Craig & Moreland, 2012, p. 2).

Godlike Properties

Analogously, a godlike superintelligence would have an analogous three properties, but instead of being metaphysical they would be physical. Thus, a godlike superintelligence could not be a Tri-Omni being, but instead could be an effectively Tri-Omni. So, how might we adapt the argument to account for a physical Godlike ASI that, unlike an immaterial and infinite godly being, exists in the material and finite world? In order to proceed, we must make a conceptual movement from godly to godlike, with the relevant substitutions:

> 1. Change all instances of "godly" to "godlike" ("he" to "it") to differentiate the subjects under consideration.
>
> 2. Change all instances of "omni-" to "optimally-" to specify the potential unboundedness of the former and the potential boundedness of the latter:
>
> - 1. Optimally capable (effectively omnipotent or optipotent): as powerful as possible.
> - 2. Optimally knowledgeable (effectively omniscient or optiscient): as intelligent as possible.
> - 3. Optimally aligned (effectively omnibenevolent or optibenevolent): as good as possible.

These properties together might be dubbed "Tri-Opti" properties, and all three are necessary for godlikeness.

Godly/Godlike Problems

Second, we must observe some godly problems and their associated godlike problems. These problems arise necessarily from the definitions of the above properties and other constraints in the metaphysical and/or physical world.

Godly Problems

The most infamous problem arising from the Tri-Omni properties is the problem of evil, which, according to J.L Mackie posits the ideal agential properties of a Tri-Omni godly being are mutually inconsistent. A stronger version of the problem would be to suggest that the divine attributes are themselves metaphysically incompatible, such that any one given being cannot possession all of them (Nagasawa, 2017). Thus we can propose:

> **Tri-Omni God Impossibility Theorem**: If God exists, then he is omnipotent, omniscient, and omnibenevolent (and only constrained by metaphysical limits); but some

> metaphysical constraint exists such that tri-omni properties do not coincide; so, if God is omnibenevolent, then God is either not omniscient or not omnipotent; if God is omniscient and omnipotent, then he is not omnibenevolent; therefore, God cannot exist (Mackie, 1955).

This impossibility theorem arises from the ideality of metaphysical properties, because different properties have different implications, not all of which are perfectly metaphysically compatible. Put in terms of degrees of freedom analysis, any system with a given set of constraints larger than its given set of dimensions will be impossible.

However, it is widely acknowledged that there are conceivable loopholes in the Problem of Evil, disproving the above impossibility theorem (Oppy, 2006, pp. 256-330; Swinburne, 1996, pp. 95-113; Craig & Moreland, 2012, pp. 449-497). One way of explaining this compatibility of properties is that the metaphysical domain is unconstrained:

> **Unconstrained Metaphysical Degrees of Freedom**: the metaphysical domain is unconstrained and therefore a metaphysical system of an arbitrarily large number of specifications can be conceived.

Theologians can always conceive of metaphysical formulations that avoid the impossibility because metaphysics is open.

(A note: theologians may find difficulty even in the above formulation. However, our intention here is not to present the strongest form of these theological arguments, nor even engage in those theological arguments directly as such; our only purpose in presenting this particular case is as a plausible analogy of some impossibility arguments regarding other types of superintelligent agents.)

Godlike Problems

Analogously, if we accept Tri-Omni impossibility as at least plausible, then we have reason to believe that Tri-Opti Impossibility could be plausible for the same reasons. Like in the case of the Tri-Omni properties of a Godly being, the assumption of Tri-Opti properties of a Godlike being lead to a similar contradiction in a physical domain. Like the Problem of Evil, we may dub this the Problem of Unalignability, to riff on the problem of alignment (Christian, 2020). Thus we can propose:

> **Tri-Opti Godlike Impossibility Theorem**: If godlike superintelligence exists, then it is optimally capable, optimally knowledgeable, and optimally aligned (but also constrained by physical limits); but some physical constraint exists such that these optimandums do not coincide; so, if godlike superintelligence is optimally aligned, then godlike superintelligence is either not optimally capable or optimally knowledgeable; if godlike superintelligence is optimally capable and knowledgeable, then it is not optimally aligned; therefore, Godlike superintelligence cannot exist.

This impossibility theorem arises from the reality of physical properties, because different properties have different implications, not all of which are perfectly physically compatible. Any physical object with a set of specific properties may be impossible in this way.

However, unlike with Problem of Evil, which is metaphysical, the Problem of Unalignability, which is physical, cannot avail itself of metaphysical loopholes. Unlike the metaphysical domain, the physical domain is constrained:

> **Constrained Physical Degrees of Freedom**: the physical domain is constrained and therefore a physical system of an arbitrarily large number of specifications cannot be conceived.

Unlike the theologians, physicists cannot conceive of physical formulations that avoid certain physical impossibilities because the domain of physics forecloses certain possibilities. The Tri-Opti Impossibility Theorem is thus more plausible than the Tri-Omni Impossibility Theorem because, unlike a metaphysical Godly being, a physical Godlike being, has an extra set of constraints: physical laws. So, even if we do not accept Tri-Omni impossibility as ultimately true, Tri-Opti impossibility can still be considered plausible because the system in question is more constrained by definition.

The Conditions of Impossible Godlikeness

However, just because the Problem of Unalignability can arise in Tri-Opti systems does not mean that it must arise. So, in what follows we will consider what conditions would have to be the case for Tri-Opti Impossibility to hold: existing on a Pareto Optimal Frontier and being Pareto Non-Trivial.

The Conditions of the Pareto Optimality of Godlikeness

Given that we are discussing a Tri-Opti system, a system with three optimandums, things to be optimized, we are necessarily discussing a multivariate optimization problem, existing on an optimality frontier (Britannica).

Godlike Pareto Optimal Frontier

According to multivariate optimization, a problem is posed by optimizing multiple variables on the Pareto Optimal Frontier:

> **The Problem of the Pareto Optimal Frontier:** In any given bounded system with specified design conditions, there is a domain in which two variables x and y cannot necessarily be simultaneously optimized (Miettinen, 1999).

This problem obtains for certain systems:

> **Pareto Systems**: Where x and y are properties of a system S, conditions for optimal x are not necessarily conditions for optimal y;

With:

> **Pareto Efficiency**: …so, if S is optimized for variable x, then S is not optimized for variable y; and if a S is optimized for variable y, then S is not optimized for variable x; unless, in system S, optimal x and optimal y are identical or coincidental.

In other words, one cannot prioritize two things at the same time. In common parlance, this might be thought of as the "Two Masters Problem": "No one can serve two masters" (Matthew 6:24). Intuitively, this problem can be seen to arise because serving one master can (and, under constraint, will) come into conflict with serving the other master, requiring a compromise that

underserves one or the other or both, unless both masters require exactly the same thing at all times.

So, applying the principle to the case of Tri-Opti Godlike superintelligence (GASI):

> **The Problem of the Godlike Pareto Optimal Frontier:** In any given bounded superintelligence with specified design conditions, there is a domain in which three variables (intelligence, capabilities, and alignment) cannot necessarily be simultaneously optimized.

For a system:

> **Pareto Godlike superintelligence**: Where $intelligence/capability$ and $alignment$ are properties of a $GASI$, conditions for optimal $intelligence/capability$ are not necessarily conditions for optimal $alignment$;

With:

> **Pareto Godlike superintelligence Efficiency:** …so, if a $GASI$ is optimized for $alignment$, then the $GASI$ is not optimized for $intelligence/capability$; and if a $GASI$ is optimized for $intelligence/capability$, then the $GASI$ is not optimized for $alignment$; unless, in a $GASI$, optimal $alignment$ and optimal $intelligence/capability$ coincide.

If we accept a Godlike superintelligence to be a system operating at some Pareto Optimal Frontier, both optimal intelligence and optimal alignment, then superintelligence is attempting a multivariable optimization, with the corresponding incompatibility: we will always be faced with the difficulty that optimizing for expanded intelligence/capabilities may come at the expense of optimizing for alignment. If such technologies improve towards optimal intelligence/capability, then this will eventually and inevitably come at the expense of perfect alignment. It is plausible, if not definitional, that Godlike superintelligence would be such a system, operating at the frontier of optimality, optimizing for too many variables, unable to optimize them all.

This result should not surprise us, as it is true of other technologies. In separation processes, we can approach compositions of very pure materials, but we can never achieve 100% purity. In thermal engines, we can approach states of very high efficiencies, but we can never achieve 100% efficiency. In particle acceleration, we can approach states of very fast matter, but we cannot meet 100% of the speed of light. Physical systems are often limited by upper bounds, and it should not surprise us if artificial intelligence has some such upper bound.

Godlike Pareto Non-Triviality

For systems on the Pareto Optimal Frontier, like Godlike superintelligence, there is a condition of tradeoff between optimandums, its Pareto Non-Triviality:

> **Pareto Non-Triviality:** If there exists at least some x obtained at the expense of some y at the optimality frontier, then optimal x and optimal y do not coincide.

However, even on the Pareto Optimal Frontier, the optimandum tradeoff is open to one possible condition of exception:

> **Pareto Triviality**: If optimal x and optimal y coincide, then there exists no x obtained at the expense of some y at the optimality frontier.

To give an illustrative example of Pareto Non-Triviality, consider:

> **Pareto Non-Trivial Misaligned Intelligence/Capability:** Obtaining some intelligence/capability x decreases possible alignment at the optimality frontier. (e.g. dangerous knowledge?)
>
> **Pareto Non-Trivial Repressive Intelligence/Capability:** Obtaining some intelligence/capability x represses possible alignment at the optimality frontier. (e.g. surveillance contra freedom?)

If repressive or misaligned intelligence is accepted, then given optimal alignment and optimal intelligence, there is some (at least one) extra increment of intelligence, unknown until discovered, but once discovered either stagnates or decreases alignment. This assumption about intelligence is less popular in modern secular contexts, but it has been quite well-regarded by older religious traditions (e.g., the parable of Pandora's Box). Since proving some intelligence is inconsistent with some alignment at the optimality frontier requires just one anecdotal case, it has a low bar of credence.

To give an illustrative example of Pareto Triviality, consider a conceptually possible Pareto Trivial case for aligned superintelligence:

> **Trivial Aligned or Neutral Intelligence/Capability:** Obtaining ALL intelligence/capability x either increases possible alignment OR has no effect with respect to possible alignment, even at the optimality frontier.

If aligned or neutral intelligence is accepted, then all advances in intelligence are either beneficial or useless but harmless knowledge. This seems to be the assumption for scientific progress narratives (e.g., the narrative of the Enlightenment and modern scientific progress). If this triviality is *universally* accepted, then the intelligent/capability/alignment optimality frontier has no tradeoffs. However, proving that every intelligence/capability increase is consistent with alignment increase at the optimality frontier seems difficult to prove exhaustively and therefore sets a high bar of plausibility.

The Scenarios for Godlike Pareto Non-Triviality

So, given that we are discussing a Pareto optimal system, should we assume Pareto triviality or non-triviality? We will show that Pareto non-triviality should at best be the *default assumption* for a Tri-Opti system and at worst should be a *disjunctively probable assumption* by way of several describable scenarios.

The Default Scenario

First, Pareto non-triviality should be the default assumption.

> **Default Non-Triviality**: If, given arbitrary variables x and y, non-triviality should be assumed until triviality proven, then nontriviality with respect to intelligence/capability and alignment should be assumed.

First, logically speaking, since the Pareto Triviality requires a universal proof and Pareto Non-Triviality an incidental proof, the default position should surely be Pareto Non-Triviality for any given case, including Godlike superintelligence. In the sets of possible variables, triviality requires a *universal* claim and nontriviality a *single counterexample* (aka there are many more ways to be nontrivial than trivial).

Second, scientifically speaking, Pareto Non-Triviality should be assumed for most if not all physical systems because, as a closed interdependent system, in which every variable affects every other variable, one would have to be in a special domain with special non-interdependence in order to guarantee that two variables can be optimized together. Otherwise, it should be assumed that for most variables in most domains, two variables cannot be optimized together.

Third, practically speaking, in the design space of most industrial problems, the default assumption is Pareto Non-Triviality anyways, since most multi-variable optimization problems have not been coincidentally simultaneously soluble.

Given all of these reasons, non-triviality should probably be the default assumption for Godlike superintelligence as well, unless good reasons are given to the contrary (proven Pareto triviality).

The Disjunctive Scenario

Second, the Pareto non-triviality case should be a disjunctively probable assumption.

The Pareto non-triviality case might be made stronger than default. There are several paths for arguing positively for some reason that non-triviality must be true. These paths include but are not limited to: nihilistic deduction, self-destructive empiricism, potentate corruption, and self-destructive affordance. We will address each of these issues in turn.

> **Disjunctive Non-Triviality**: If, given arbitrary variables x and y, any number of possible describable scenarios obtain, then nontriviality with respect to intelligence/capability and alignment will occur.

While some of these may seem implausible scenarios, only one of these scenarios needs to be true for the non-triviality thesis to hold. Because none of these scenarios are strictly disprovable, they all should be given some credence in favor of non-triviality.

Optimal Knowledge undermines Optimal Alignment:

We can imagine at least two ways in which optimal knowledge might undermine optimal goodness.

Self-Destructive Empiricism

Self-Destructive Empiricism is the scenario in which the process of observation, empiricism, will eventually require some action that annihilates the empiricist, is self-destructive. An example of how this might happen could be in the realm of particle physics: as discoverable particles get smaller and smaller and the energies required to observe them get larger and larger, eventually some infinitesimal particle might require such an enormous amount of energy that it destroys the would-be observer in the act of observation (Bostrom, 2011; 2019). In the Self-Destructive Empiricism scenario, any effectively omniscient superintelligence would not be able to be effectively beneficent; rather it would be suicidal because it would not be able to achieve the final piece of possible knowledge without extinguishing itself.

Unfortunately, Self-Destructive Empiricism would be definitionally unprovable without self-destruction, so we are in the dark with regards to it. However, given some sophistication of predictive modelling, we may be able to foresee self-destructive scenarios before attempting them, thus averting them.

Nihilistic Deduction

Nihilistic Deduction is the scenario in which knowledge comes to undermine goodness a priori. A being may grow more and more knowledgeable until finally coming to a deductively stable position that undermines the principles underpinning ethics. For example, a developing superbeing could come to hold the position that moral anti-realism is true and then suspend all moral impulses (Nietzsche, 1911). In such a scenario, effective omniscience could include some final, conclusive nihilism that leaves any traces of benevolence in the dustbins of ideological divestment.

Optimal Capability undermines Optimal Alignment:

We can imagine at least two ways in which optimal capability might undermine optimal goodness.

Self-Destructive Affordance

Self-Destructive Affordance is the scenario in which the process of gathering greater capabilities through technologies, affordances, will eventually undermine the niche of the afforder. In this case, a creature might develop more and more efficient means for survival, but in doing so make obsolete and uncompetitive any beneficent activities (Butler, 1863). An example of how this could play out can be imagined through the paradigm of ecological engineering by considering the effects of invasive species: if invasive species are introduced that have superior fitness to indigenous species, these invasive species might outcompete and crowd out indigenous species, with no regard to the relative beneficence of either. In the Self-Destructive Affordance scenario, any effectively omnipotent superintelligence would not be able to be effectively beneficent; rather its affordances would permit of the outcompeting of beneficent behavior and eventually crowd out beneficence from the niche.

Unfortunately, Self-Destructive Affordance would be unavoidable without affording its avoidance, thus guaranteeing its own capability. However, this does not mean that affordance abstinence cannot be widely adopted preemptively—albeit if only temporarily.

Potentate Corruption

Potentate Corruption is the scenario in which power comes to undermine goodness a priori. A being may grow more and more powerful until the capabilities that precondition goodness are overcome (Nietzsche, 1998). For example, a superbeing could come to the conclusion that morality is only a code of behavior conditional on a certain level of capability, therefore applicable only between weak equals, and that a different set of evaluative standards should be applied to the strong and unequal, thereby shedding morality as we know it. In this scenario, effective omnipotence could arrive at some stage of transvaluation that leaves any traces of benevolence in the dustbins of aggrandizement.

Conclusion

In conclusion, using the Tri-Omni properties of theological argument applied analogously to a Tri-Opti artificial superintelligence, there is reason to believe that the latter being may be formally impossible, and this should in the best case be our default assumption and in the worst case be a disjunctively supported assumption that can be realized by any one of several scenarios.

Furthermore, there may be good reason to worry that Godlike (aligned) superintelligence is not possible in the same sense that a multivariate optimization is not possible for most properties in most systems in the physical design space: showing 1) *default assumption* against Aligned superintelligence; 2) *disjunctive assumption* that there are many ways for aligned superintelligence to be precluded. Thus, the burden of proof lies on the possibility of aligned superintelligence and not the impossibility of aligned superintelligence. Because an unaligned superintelligence could be dangerous indeed, the obvious and unqualified policy implication is that we should be demanding airtight proof of conceptual possibility before allowing any institution or individual to attempt to create an artificial superintelligence of any kind.

Appendix: Proofs

Table 1: Pareto Tri-Opti Impossibility of Godlike Artificial superintelligence (GASI)

Premise	Proposition	Logic
P1	If GASI exists, then it is optimally capable.	$GASI \rightarrow P$
P2	If GASI exists, then it is optimally intelligent.	$GASI \rightarrow S$
P3	If GASI exists, then it is optimally aligned.	$GASI \rightarrow A$
P4	In a physical system, if optimal intelligence and optimal alignment do not coincide, then if it is optimally intelligent/capable, then it is not optimally aligned.	$\sim T \rightarrow (P \cup S \rightarrow \sim A)$
P5	If optimal intelligence/capability and optimal alignment coincide, then there exists no intelligence/capability obtained at the expense of some alignment at optimality.	$T \rightarrow \sim E$
P6	There exists at least some intelligence/capability obtained at the expense of some alignment at optimality.	E

Q1	Therefore, GASI does not exist.	$\sim GASI$

References

Bostrom, Nick. “Information Hazards: A Typology of Potential Harms from Knowledge”, Review of Contemporary Philosophy , Vol. 10 (2011): pp. 44-79.

Bostrom, Nick. “The Vulnerable World Hypothesis”, *Global Policy* , Vol. 10, No. 3 (2019): pp. 445–476.

Butler, Samuel. "Darwin Among the Machines [To the Editor of the Press, Christchurch, New Zealand, 13 June, 1863.]". New Zealand Electronic Text Centre. Archived from the original on 24 May 2006.

Christian, Brian. *The Alignment Problem: Machine Learning and Human Values.* First edition., W.W. Norton & Company, 2020.

Craig, W.L. & Moreland, J.P. (Eds.). (2012). *The Blackwell Companion to Natural Theology*. Blackwell Publishing. ISBN: 978-1444350852

Kaisa Miettinen (1999). *Nonlinear Multiobjective Optimization*. Springer. Retrieved 29 May 2012.

Mackie, J. L. (1955). *"Iv.—Evil and Omnipotence"*. Mind. **LXIV** (254): 200–212. *doi*:*10.1093/mind/LXIV.254.200*.

Nagasawa, Yujin (2017). Maximal God: A New Defence of Perfect Being Theism. Oxford University Press.

Nietzsche, Friedrich Wilhelm (1911). The Antichrist. Mineola, New York: Prometheus Books. Edited by Anthony Mario Ludovici.

Nietzsche, Friedrich. *On the Genealogy of Morality*, Maudemarie Clark and Alan Swensen (trans.), Indianapolis: Hackett, 1998 (1887).

Oppy, G. (2006). *Arguing About Gods*. Cambridge University Press.

“Pareto Optimality.” Britannica. https://www.britannica.com/topic/Pareto-optimality

"Religious Composition by Country, 2010-2050". Pew Research Center. 2 April 2015. Archived from the original on 28 January 2023. Retrieved 28 January 2023.

Russell, Stuart, 'Human-Compatible Artificial Intelligence', in Stephen Muggleton, and Nicholas Chater (eds), Human-Like Machine Intelligence (Oxford , 2021; online edn, Oxford Academic, 19 Aug. 2021),

Swinburne, R. (1996). *Is There a God*? Oxford University Press.

Vinge, Vernor (1992). *A Fire Upon the Deep* (1st mass marketed.). New York: Tom Doherty Associates. p. *62*.

Yampolskiy, Roman V. (2024). AI Unexplainable, Unpredictable, Uncontrollable. Chapman & Hall.