Universes Out of Sync: A Critique of Scott Alexander's God-Like Superentity

Curtis Heinen / Ryerson University

Abstract

Scott Alexander presents an account of an artificially intelligent godlike superentity. This superentity, he argues, arises when technologically advanced civilizations in causally closed universes simulate each other, aligning their values and actions in a way that ultimately unites them under the moral law. I point out that Alexander's argument rests on assumptions that do not stand up to scrutiny. There exists a problem with the concept of a simulated universe, namely that of distinguishing one from a real universe, and his notion of "acausal trade" is impossible, whether simulations are perfect or imperfect.

Key words: Simulation, Artificial Intelligence, Superentity, Multiverse, Moral Law, God, Deity, Universe, Consciousness

SECTION I: Exegesis

In his speculative blog post "The Hour I First Believed" Scott Alexander brings together a set of contemporary assumptions about simulations, consciousness, and decision theory to explain how a multiverse-spanning, godlike artificial intelligence could arise. The argument, inspired by the simulation hypothesis, exemplified by Nick Bostrom (Bostrom 2003) fundamentally depends upon the ability to create high-fidelity simulations of beings and the universe(s) they inhabit. This is not a trivial assumption, as I will argue later in this essay.

Here is an overview of Alexander's deity, in his own words (Alexander 2018):

1. There is an all-powerful, all-knowing logically necessary entity spanning all possible worlds and identical to the moral law.
2. It watches everything that happens on Earth and is specifically interested in humans' good behavior and willingness to obey its rules.
3. It may have the ability to reward those who follow its rules after they die, and disincentivize those who violate them.

1.1. Tegmarkian Multiverse

Let's start with the "Tegmarkian multiverse": a landscape in which universes exist as mathematical objects (Alexander 2018). Universes start with a set of mathematical

constants and evolve according to a set of pre-defined rules. Many complex universes would exist in this space, but there would be more simple universes than complex ones, as every logically coherent mathematical object would have a corresponding universe within the multiverse. It is important to note that none of these universes would be able to communicate with or influence any other (Alexander 2018).

1.2. Acausal Trade

Many universes would eventually be under the complete control of an artificial intelligence due to the evolution and propagation of intelligent beings and their civilizations (Alexander 2018). Each of these artificial intelligences would have the ability to simulate all other universes simultaneously in real-time and could therefore conduct inter-universal negotiations without directly communicating. This is "acausal trade," invoked by Alexander to get around the causal isolation of each universe. Note that Alexander simply *assumes* such "causal closure" of each universe. It's questionable, but I will focus on problems downstream from this assumption. Acausal trade leverages each AI's ability to simulate all other universes and be able (through probability calculations) to perfectly predict their every move. Each artificial intelligence would hold parallel negotiations within its own universe and behave accordingly. If every simulation is perfectly performed, the negotiations will always be in sync.

1.3. Values Handshakes

Instead of attempting a multitude of independent trades with other universes, these AIs would more likely engage in what Alexander calls "values handshakes." Values handshakes are deals that take place at a high level, where each AI would alter its own code to align its values with the average of all other intelligences' values, or some other compromise (Alexander 2018). This could benefit each AI in the same way that comparative advantage economically benefits connected regions of the existing global economy. Alexander introduces the concept of "counterfactual mugging" (Alexander 2018) to illustrate how this compromise could avoid a tyranny of the majority. Imagine God comes up to you and says "I'm going to flip a coin and if it comes up heads, I'm going to ask you for $5. If it comes up tails, I'm going to give you $1,000,000, but only if I predict that you would have said yes to giving me $5 in the counterfactual situation in which I flip heads. My predictions are never wrong, and the coin came up heads. Will you give me $5?" Alexander argues that if you were designing an AI, you would design it in such a way that it would give God the $5 in order to gain the $1,000,000 in the counterfactual situation where the coin

does come up tails. He thinks that more powerful AIs in our multiverse should alter themselves in order to benefit in another situation in which they are the less powerful ones (Alexander 2018). Once formed, this pact between AIs would create a superintelligence or superentity spanning multiple universes. It is here where Alexander's superentity begins to resemble God. This God-like superentity, having the tendencies of the beings that contribute to its existence, would undoubtedly care about said sentient beings and any other mortal beings that could, in the future, contribute their intelligence to the pact (Alexander 2018).

1.4. Simulation Capture

In order to impose the "moral law" on intelligences that cannot or opt not to join the superintelligence pact, our superentity would use a tool called "simulation capture." Simulation capture occurs when copies of a conscious entity are simulated by an AI (Alexander 2018). Assuming that consciousness is a mathematical object, according to Alexander, it should admit of duplicability. When copied a thousand times in a universe other than its original host universe, the capturing artificial intelligence can begin feeding the simulated copies a slightly different experience which would split the conscious mind into two distinct beings. Once this has been accomplished, the capturing artificial intelligence could allow the original conscious being to organically die in its own universe and simultaneously alter the experience of the simulated copies to capture them in the most metaphysically elegant way. This capture would result in conscious simulations doing as the AI wishes, that is, either succumb to or enthusiastically align with the moral law (Alexander 2018).

SECTION II: Response

2.1. Perfect Simulations

If we are going to draw a distinction between base reality and simulated reality, we ought to know what a real universe would look like in comparison to a simulated one. What would a *perfect* simulation look like? Perhaps the intuitive answer is that it would be qualitatively identical to the *real* thing, that is, genuine. If the real and simulated universes are qualitatively identical, how could we call one of them a simulation? This is troubling because it implies that each AI in our superentity must not only be able to simulate other universes but must generate real universes within its own. One could argue that because each of our universes exists in a Tegmarkian Multiverse, qualitative identity is unnecessary, as the contents of such a multiverse consist of nothing but mathematical objects. This would require objects to *seem* qualitatively identical, even if they aren't, which is counterintuitive. It appears

Alexander is right to think that qualitative identity is important, but acausal trade, in his account, requires perfect simulation to function. However, the ontological status of these simulations remains dubious. If communication between real universes is impossible as Alexander (2018) assumes, how could an artificial intelligence extract insights from its own "perfect" simulation? Since perfectly simulated universes are *real* universes, any sort of communication between the simulation and its AI host violates Alexander's own communication constraint. Happenings inside the simulated universe would influence its AI host, causing it to diverge from its original path.

> P1: A perfectly performed simulation is qualitatively identical to a real universe.
>
> P2: Communication between real universes is impossible.
>
> C1: Therefore, communication between the simulating AI host and its perfect simulation is impossible.
>
> P3: Acausal trade requires a simulating host to communicate with [i.e., to directly observe] its simulations of other trader universes.
>
> C2: Therefore, acausal trade using perfect simulations is not possible.

2.2. Imperfect Simulations

Alexander (2018) invokes acausal trade to deal with the inability of real universes to communicate with one another. Acausal trade requires not only that all artificial intelligences within the pact be able to simulate all other universes (Alexander 2018), but that they can do so perfectly, a dubious prospect. Even a slight deviation from perfection would take the universe and its simulation out of sync. Iterated over a long enough period of time, asynchronicities between universes and their simulated counterparts would be exacerbated. These asynchronicities would be impossible to detect without the ability to observe and reference other universes. Given the possibility of asynchronicity, a complexity ceiling would be required to prevent any artificial intelligences within the pact from outpacing the others, as simulation of more complex intelligences by simpler intelligences would introduce a time lag, and therefore further reduce synchronicity. This is because a simulation can only be as sophisticated as the substrate it runs on. How could this complexity ceiling be implemented without means of communication between real universes? Perhaps through acausal trade, but that begs the question, presuming that acausal trade can, in fact, function across universal borders. Inter-universal acausal trade is dependent on a pre-established complexity ceiling, but that ceiling cannot be established without inter-universal acausal trade.

2.3. Tegmarkian Multiverse

My third and final objection is that even if perfect simulation is in principle possible, the Tegmarkian Multiverse is itself a problematic concept. It assumes that everything is, at bottom, mathematics. This includes space, time, matter, and consciousness. However, we do not understand consciousness. Some have theorized that it plays a role in determining what is important when encountering novel circumstances such as learning a new skill. According to Peterson (1999) consciousness confronts the unknown and negotiates with it. Perhaps it utilizes mathematics to perform evaluations, but that doesn't require consciousness to be fundamentally mathematical.

References

Alexander, Scott. 2018. "Slate Star Codex." Slate Star Codex Blog. 2018. https://slatestarcodex.com/2018/04/01/the-hour-i-first-believed/.

Bostrom, Nick. 2003. "Are You Living in a Computer Simulation?" *Philosophical Quarterly* 53 (211): 243–55. https://doi.org/10.1007/s00163-016-0218-3.

Peterson, Jordan B. 1999. *Maps of Meaning: The Architecture of Belief*. New York: Psychology Press.