A Theodicy for Artificial Universes: Moral Considerations on Simulation Hypotheses

This is a pre-print draft for: Gualeni, S. 2021. "A Theodicy for Artificial Universes: Moral Considerations on Simulation Hypotheses". *International Journal of Technoethics*, 12 (1), 21-31.

Abstract

'Simulation Hypotheses' are imaginative scenarios that are typically employed in philosophy to speculate on how likely it is that we are currently living within a simulated universe as well as on our possibility for ever discerning whether we do in fact inhabit one. These philosophical questions in particular overshadowed other aspects and potential uses of simulation hypotheses, some of which are foregrounded in this article. More specifically, "A Theodicy for Artificial Universes" focuses on the moral implications of simulation hypotheses with the objective of speculatively answering questions concerning computer simulations such as: If we are indeed living in a computer simulation, what might be its purpose? What aspirations and values could be inferentially attributed to its alleged creators? And would living in a simulated universe affect the value and meaning we attribute to our existence?

Introduction

Imagine a large vat on the table of a futuristic laboratory. Inside the vat, a disembodied brain floats in some kind of liquid. The scientists running the laboratory use advanced computer technology to stimulate the brain in the vat with input and sensations that are indistinguishable from those that regular human bodies experience in their relationship with the actual world. In this hypothetical setup, the laboratory's technology also feeds the brain's outputs back into the computer, giving the brain the possibility to interact with the environment it perceives. At that point, the brain is effectively inhabiting a persistent, interconnected whole: a world[1]. In this hypothetical scenario, the brain floating in the vat is connected with what is commonly referred to as a simulation: a procedural model – often run on computers – that imitates the behaviors of a physical system (see Bostrom, 2003; Salen & Zimmerman, 2003, 423; Chalmers, 2005).

Whether or not those imaginative scenarios explicitly rely on the use of digital technologies, speculative premises similar to the one that was just outlined are common throughout the history of Western thought. The Socratic dialogues and the texts of the skeptics feature questions, allegories, and ideas that can be considered particularly obvious examples of this recurrence. Within our tradition of thought, these propositions are often referred to as the 'brain in a vat

[1] In the philosophical tradition of phenomenology, the term 'world' generally indicates two interrelated things. First, a 'world' is a set composed of beings that are understood together with all their properties and mutual relationships. More specifically, a 'world' describes that set as experienced by one of the beings involved in it. To be identified as a world, those properties and mutual relationships need to be experienced in ways that are to a degree persistently perceivable and behaviorally consistent for the being in question (see Gualeni & Vella, 2020, p. xxvii). Relatedly, in its second meaning, a 'world' indicates the horizon (or ground) against which every object is experienced, understood, and interacted with (ibid.).

hypothesis' (or the 'evil genius hypothesis', after René Descartes's infamous argument in the *Meditations*).

-- End of page 21 --

Comparable ideas also emerged in non-Western cultural contexts such as Chinese Taoism or Vedic literature, where they similarly function as conceptual tools to help the reader maintain a degree of suspicion towards the emotions and sensations that they experience in their daily lives as embodied beings. They are, to cite Chalmers's (2005) words, 'philosophical fables' that prompt us to question what we mean by 'reality' and what qualifies as a real experience. They invite us to consider whether the world of sensations and relationships that we experience on an everyday basis could be an artifice or a mere illusion.

In recent years, a particular set of speculative scenarios that are to a degree similar to those mentioned above received sustained attention both in academia and in popular culture. I am referring to a group of hypothetical situations that are commonly grouped under the umbrella term 'the simulation hypothesis' (SH). Differently from 'brain in a vat' kinds of hypotheses, the SH does not predicate that one's brain – or the entirety of one's body, as is the case in the movie *The Matrix* – exists somewhere in base reality[2]. It does not presuppose that our perceptual and cognitive equipment is being enthralled and deceived by the simulative capabilities of a computer (or the magical ones of a demon). Instead, the SH proposes an imaginative scenario in which we are artificial beings who were created – and presently exist – within a computer simulation. In the SH, in other words, no part of me is predicated to be existing or having ever existed in base reality (with the exclusion, perhaps, of the figments of computer code that correspond to the properties my being and my mental states)[3].

Like 'brain in a vat' kinds of hypotheses, the various versions of the SH are also used to raise doubts concerning the artificiality of our experience. They are similarly employed in philosophy to gauge the likelihood of our being currently living in a simulated universe as well as our capability of ever discerning whether we do in fact inhabit one. This dominant philosophical use sidelined other aspects and potential applications of simulation hypotheses, some of which are foregrounded in the present article. I am talking, for example, about reflections concerning the technological and computational requirements that would be needed to run the simulation in question, about the kinds of values and aspirations that could have shaped and guided the design of that simulation, or about the ethical responsibilities that the creators of the simulation potentially have towards the artificial beings inhabiting it.

[2] With 'base reality' I indicate the level of existence in which, to cite Iain M. Banks, the information that constitute all living things is "encoded in matter itself, not running in some abstracted system as patters of particles or standing waves of probability". (Banks, 2012b, p. 340) In line with this understanding, in the context of this paper, the adjective 'real' indicates that something belongs to base reality (also see Selinger, 2009). 'Actual' is used, instead, as a relative term signifying that something or someone exists in the same world as the being using the adjective.

[3] Some of the most often discussed books about the hypothesis that our universe is running on a computational substrate include Hans Moravec's 1988 *Mind Children: The Future of Robot and Human Intelligence*, and Frank Tipler's 1997 *The Physics of Immortality: Modern Cosmology, God, and the Resurrection.*

Notable existing work on those arguably secondary aspects include the notorious article "Are You Living in a Computer Simulation?" by Nick Bostrom (2003, which will be introduced and discussed in the next section), "Theological Implications of the Simulation Argument" by Eric Steinhart (2010), and "Natural Evil and the Simulation Hypothesis" by David Kyle Johnson (2011).

The 'Posthuman Morality Hypothesis' (PMH)

In "Are You Living in a Computer Simulation?", Bostrom argues that at least one of the three following propositions must be true:

> (1) the human species is very likely to go extinct before reaching a stage of technological maturity;
> (2) any technologically mature civilization is extremely unlikely to run a significant number of computer simulations of their evolutionary history (or variations thereof);
> (3) we are almost certainly living in a computer simulation. (Bostrom, 2003, p. 14).

Extrapolating from tendencies and preferences that have been defining how humans currently develop and use digital media, Bostrom imagines a civilization that reached the technical capability "to convert planets and other astronomical resources into enormously powerful computers" (ibid., p. 3). In that hypothetical scenario, and should that civilization maintain sufficient interest in developing and running what Bostrom calls 'ancestor-simulations'[4], then – he argues – it is almost a statistical certainty that we are living in one of those computer simulations.

-- End of page 22 --

Bostrom calls his trilemma, which contains the SH as its third component, "the simulation argument".

It is relevant for the scopes of the present article to observe that "Are you Living in a Computer Simulation?" does not establish technological maturity and an interest in simulation as the sole factors determining the likelihood that an advanced civilization will produce ancestor-simulations. As an additional limiting circumstance, Bostrom also mentions ethical interdictions: one can imagine, he writes, "that advanced civilizations all develop along a trajectory that leads to the recognition of an ethical prohibition against running ancestor-simulations because of the suffering that is inflicted upon the inhabitants of the simulation". (ibid., p. 11) In the same paragraph, however, he quickly dismisses this potential objection to his argument by stating that, "from our present point of view, it is not clear that creating a human race is immoral". (ibid.)

[4] According to Bostrom, the computers of technologically mature civilizations could simulate the entire mental history of a species (2003, p. 6). When referring to 'ancestor-simulation', he is thus talking about computer simulations of reality as experienced by the ancestors of the creators of the simulation (which, in his 'simulation argument', would be the human race that we are presently a part of). It is worth observing that, for the sake of the argument presented in this text, it is not necessary for a simulation to be an ancestor-simulation. What is a necessary prerequisite is that the simulation in question features beings that can be considered morally relevant (see the following page).

Bostrom implicitly attributes the possibility of moral relevance to artificial beings that have human characteristics. This is, I believe, an aspect of that work of his that has not aged particularly well. This is not only due to the evident speciesism of that position, but it also the case as academia has recently gotten more concerned with how we design and comport ourselves in relation to artificial beings. What I am arguing here is that a growingly sophisticated body of work is being developed with the objective of refining how we think about the personhood and the legal rights of artificial intelligences and robots, and of assessing the moral responsibilities that we, as creators, have towards them. Examples of such perspectives and tendencies can be observed in the work of Coeckelbergh (2010), Bostrom & Yudkowsky (2014), Neely (2014), Gunkel (2018), and Gualeni (2020) among others. In a passage of my 2020 article "Artificial Beings Worthy of Moral Consideration in Virtual Environments", I compare the biological processes of producing children to the technological ones of creating artificial beings that are ethically relevant (2020, p. 6). In that text of mine, I focus on our responsibilities towards future, morally relevant artificial beings that we will create to inhabit virtual worlds and computer simulations. The creation of beings that are worthy of moral consideration in those artificial contexts might be even more ethically problematic than becoming a parent. "Software developers", I argue in that article, "have a higher degree of control over their creations than parents have in human biological reproduction. As digital creators, software developers can make decisions concerning the production of both virtual environments and the artificial autonomous beings inhabiting these environments, whereas parents can, at best, play an active role in the production of the child." (ibid.)

In a 2014 paper that focused specifically on the creators' moral duties towards artificial beings, Erica L. Neely proposes a broad and inclusive criterion for moral relevance that will be useful to consider in developing my argument. In "Machines and the Moral Community", Neely outlines an ethical framework that relies on the future possibility for artificial beings to express specific interests such as preserving their own autonomy and bodily integrity, where the latter concept refers to the possibility of continuing one's existence undisturbed and unharmed (2014, p. 3). Neely's proposition exemplifies current sensitivities and concerns towards the moral relevance of artificial beings and reveals the exclusivity and obsolescence of classical ethical frameworks.

As an important component to my contribution to themes that emerge at the intersections between technological speculation and morality, in the present article I introduce a hypothesis that supplements the SH. I call this additional component of my argument 'the posthuman morality hypothesis' (PMH). The PMH posits that a technologically advanced civilization is also likely to be advanced from the point of view of morality. By that I mean that, extrapolating from current tendencies (like the ones exemplified by Neely's work as well as my own), it is reasonable to expect from an advanced civilization to be actively invested in limiting damage and potential suffering for a moral community that is vast and inclusive. To be more specific, the PMH poses that an advanced civilization – one that is familiar with simulation technologies and had the opportunity to reflect and legislate on their use – would consider it a basic moral duty to respect and preserve the autonomy and the integrity of beings capable of expressing autonomous interests, regardless of the constitution of these beings (also see the 'principle of ontogeny non-discrimination' in Bostrom, 2014, pp. 6-7).

The PMH thus posits that it reasonable to imagine that a technologically mature civilization will recognize artificial beings as ethically relevant.

One could understand the PMH as a secular version of 'perfect being theology', in which the qualities ascribable to God are made explicit by postulating God as a perfect being, who must therefore be also perfectly good (Rogers, 2000). As the analogy goes, the defining traits of an advanced civilization are not only to be found in its technological and organizational maturity, but also its being ethically advanced. Should the creators of ancestor-simulations be inclusively benevolent as was just hypothesized, why would suffering and evil have a place in the simulated world we live in? This question evidently echoes the classical problem of theodicy, and will be addressed in the next section of the present article together with other interrogatives concerning morality.

What kind of computer simulation could we be living in?

The term 'theodicy' was coined in 1710 by Gottfried Wilhelm Leibniz to mean 'divine justice' (from the ancient Greek *theos*, God, and *dikē*, justice) (Leibniz 2000). It is used to indicate a theological argument meant to prove the ultimate benevolence of God. In particular, theodicies argue for the possibility for God to hold divine attributes, such as omniscience, omnipotence, and complete goodness in the face of the presence of evil and suffering in the world (Surin 1983; Harrison 1989; Hamilton 2016). The classical, theological form of theodicy underwent a process of secularization during the Enlightenment, when it changed from speculative vindications of the existence of God to a set of notions and perspectives that justify the fact that our world is characterized by suffering and injustice. In their non-religious variant, theodicies are now part of anthropology ('antrhopodicies', see Becker 1976, pp. 17-18; Surin 1983, pp. 228-232) and conceptual tools used in the philosophy of history (Swedberg 2005, pp. 273-274; Hamilton 2016, pp. 233-234).

As anticipated in the previous section, I am going to foreground some moral aspects of the SH with the objective of speculatively answering questions concerning computer simulations and the potential aspirations of their creators. In the context of philosophizing about intelligences who are capable to artificially generate and run entire universes, it should be evident to the reader why the conceptual heritage of classical theodicy could prove fruitful here, and cannot – at least in this context – simply be discarded as specious and outdated.

In the present article, I am going to tackle those questions on the basis of the PMH, meaning that my analysis will not consider cases in which the advanced civilization responsible for developing and running computer simulations could simply be considered to be evil or deranged. To be sure, scenarios in which our alleged creators do not consider evil actions or events in virtual worlds to be morally reprehensible is certainly within the field of possibilities; it is, however, arguably uninteresting from a philosophical point of view. In line with classical theodicy, I do not consider the idea that our creators are simply careless or sadistically entertained by the collective suffering of its inhabitants to be a viable one. It would be, I believe, a rather sterile and hopeless way to

make sense of the amount of injustice and the suffering that characterize our existence[5]. Accordingly, in the present article I will disregard the possibility of evil and/or deranged creators of the computer simulation that we allegedly live in.

My analysis of the moral and existential implications of the SH begins by observing that ethical responsibilities towards morally relevant simulated beings crucially depend on the creators' capabilities for predicting and correcting the course of simulated events (and maybe even stopping those events from continuing). This observation aligns with the already outlined understanding of the analogies between the ethics of parenthood and our responsibilities towards artificial beings who are worthy of moral consideration (Gualeni 2020, pp. 6-7). Grounded in the PMH, the next two sections of the present text will offer philosophical speculations on what kind of simulation we might be living in. The editorial separation matches a conceptual divide: in CASE 1 (below) I will build upon the supposition that the creators of the computer simulation we are allegedly living in cannot (or cannot fully) predict and/or influence the course of simulated events.

-- End of page 24 --

CASE 2 will, instead, rely on the assumption that a technically mature civilization has complete knowledge and control over its computer simulations.

CASE 1: a simulation over which its creators DO NOT HAVE complete control

Humankind is both the producer and the product of socio-technical contexts. This co-constitutive relationship entails that civilizations can never be understood as making merely instrumental uses of their technologies, their institutions, and their traditions. Accordingly, this section of my article will rely on the supposition that a technologically mature civilization cannot exert complete control over the functioning, the possibilities, and the outcomes of their simulated universes. A similar and particularly fitting example of the ambiguity of technology in today's socio-technical context can be identified in the fields of machine learning and artificial intelligence: those technologies are explicitly taking inspiration from behaviors and properties observed in biological brains, and are as inscrutable as those brains in terms of where information is stored, how and when the information is used, and what the output of the artificial thinking process might be (see Castelvecchi, 2016). Extrapolating from our current relationship with technology, this section imagines that advanced civilizations also have dynamic and multistable[6] relationships with their creations.

---

[5] As an addendum to this preliminary note, I think it might be important to also observe that, for the scopes of the present article, it does not make a difference whether we imagine simulations as personal constructs (i.e. simulations where only one person has conscious perceptions and thoughts) or as a shared world (i.e. a complete simulation in which other beings around one are not philosophical zombies). After all, as Bostrom points out, "[i]t is not clear how much cheaper [philosophical zombies] would be to simulate than real people. It is not even obvious that it is possible for an entity to behave indistinguishably from a real human and yet lack conscious experience". (Bostrom, 2003, p. 13)

[6] The adjective 'multistable' indicates the inherent possibility of technologies to be repurposed and used in unanticipated ways. The quality of multistability is what makes it possible for a technology to acquire new meanings, functions, and effects within a social context. Our interactions with technologies are thus recognized as not solely determined by the intentions of the original developers of a technology, but are in an ambiguous and constantly

The simulated world in which we potentially live features horrific acts of violence, and seems to be characterized by widespread suffering. In light of the presence of evil in our world, could we logically consider the act of creating this world to be morally wrong, when the creators had limited possibilities to anticipate how the simulation would develop, and/or mend or influence (or simply stop) the course of simulated events? This question was at the basis of the parenthood argument that was presented earlier in this text. To further contextualize the problem of ethical responsibility within the narrative of the SH, we can similarly ask ourselves: what if morally relevant beings were not even supposed to emerge from the simulation? And what if simply turning off a simulation containing beings that are worthy of moral consideration would be considered an immoral (and perhaps illegal) act for the advanced civilization running the simulation, comparable to a genocide on a universal scale?

If our hypothetically benevolent creators had imperfect control and foresight over the simulation that they developed, it would be intuitive to conclude that our simulated universe could contain any amount of suffering and injustice. Let us nonetheless continue to speculate on scenarios in which an advanced civilization cannot fully predict or control the events that take place in their computer simulations, and cannot ethically decide to simply turn it off. Would their experiences with previous simulations not have warned or discouraged them from producing more simulations? Could previous iterations not have produced the insight that millennia of abuse and inequality were going to be the likely result of their creation? One possible answer to these questions could be that we might be inhabiting the first complete simulation that the advanced civilization in question ever created, and that the information gleaned from the present simulation and its history of suffering is perhaps going to deter them from developing or running other simulations. Technically, we do not even need to be part of the first complete simulation run by an advanced society, as we might simply be living in the first computer simulation where forms of life evolved to the point of becoming morally relevant for them. In all the cases outlined above, the creators would have been morally justified to develop the simulation that we are allegedly living in, but they would not be morally allowed to create or run others simulations after having witnessed and analyzed what happened with the one we are allegedly inhabiting.

Having analyzed this first hypothetical scenario, I want to propose another one, similar to the previous, that I find particularly interesting to think about. Let us imagine that a technologically mature civilization is not able to predict or alter the course of a simulation or stop it from running for reasons connected not to their ethos, but to how the simulation is technically produced. Along with perspectives advanced in the field of quantum physics about the functioning of our universe (see DeWitt, 1967; Page & Wootters, 1983), we can speculate that advanced civilizations decided to build a simulated universe as an atemporal construct.

Whatever the reason behind this technical decision, time would *not* be a defining dimension of how the universe is simulated in this scenario. The simulation would not, in other words, be

changing relation with their users. The term was first introduced by Don Ihde in his 1990 book *Technology and the Lifeworld: From Garden to Earth*.

progressively computed over time as events unfold within it. The simulation would, instead, be generated at once as a timeless object that already contains all the possible trajectories of all its states and interactions. In such a simulation, the dimension time would be a feature of our experience of that object, and not a defining characteristic of the object itself.

This scenario prompts us to understand time as a quality of our experience of the simulation, and not of the simulated universe itself. The outlined conceptualization of time as quality of our subjective experience is particularly reminiscent of Kant's understanding of time presented in the first section of the *Critique of Pure Reason* (2000). In it, Kant proposes to understand time not as an objective property of the world we experience, but as a feature of our cognitive apparatus[7].

In case the simulated universe we allegedly inhabit was built as an atemporal construct, it is intuitive to figure that an advanced civilization might not be able to access and evaluate the various possible outcomes of their creation until the simulation is fully computed. At that point, however, the simulation already contains all our experiences and mental states, meaning that the creators will not able to stop us from experiencing it or to retroactively edit or influence the way we experienced it, as to do so would simply create another, different simulation, and not amend our experiences of the previous one.

In this first section, we examined various hypothetical scenarios in which the advanced civilization that created a simulated universe does not have complete control over their simulations, and cannot predict its outcomes. Our speculations about those cases reveal that there are a variety of situations in which the SH can comfortably coexist with the PMH. There are, in other words, circumstances in which a perfectly ethical advanced civilization could have created our allegedly simulated but definitely painful universe. Those circumstances can fall into one (or both) of the following two categories:

1. Our alleged creators had no prior knowledge on the matter, meaning that ours is the first (or the first ethically relevant) simulation generated by the civilization in question, and it is currently producing knowledge and insights that might deter our creators from producing other simulations.

2. Our alleged creators had prior knowledge on the matter, and estimated that the knowledge derived from running those simulations would lead to advancements and benefits that will utilitarianistically eclipse ethical concerns relative to the possibility of causing suffering to artificial beings who are worthy of moral consideration[8].

To boil the totality of this first case down to its conceptual core: in case we are living in a computer simulation, and if our hypothetical, benevolent creators do not have complete control or knowledge over the simulation itself, then our existence and our suffering are guaranteed to be existentially

[7] Unlike Leibnitz or Newton, for Kant time was not an entity in the world, but one of the inherent, *a priori* tools with which we understand what surrounds us. In other words, for Kant, time is a subjective form of our intuition that, being innate, applies to all our experiences.

[8] It might be worth pointing out that an analogue utilitarian perspective currently justifies intensive animal farming, the use of animals in medical research, and product testing. This position, which is commonly associated with Singer (2002), asserts that, although the interests of all beings worthy of moral consideration are of equal importance, it is not necessarily morally wrong to violate or frustrate some of those interests.

meaningful beyond our individual lifespans and our survival as a species. In all the imaginative scenarios we examined under those premises, in fact, our simulation would be producing useful insights for the creators, insights that they considered likely to be ethically positive for their moral community in the foreseeable long run.

CASE 2: a simulation over which its creators DO HAVE complete control

The second case that I decided to examine proposes a hypothetical scenario where we do live in a computer simulation and our creators are not only benevolent, but also omnipotent in the sense that they have complete control over our simulated world. In line with Bostrom's speculations (2003, p. 5), their omnipotence entails their capability of altering and editing any aspects of the physical simulation that they created as well as our individual mental states, feelings, and memories.

Those civilizations are imagined, through a helpful analogy, as having a relationship with the simulator that is comparable to how we currently create and adjust the virtual environments of videogames and computer simulations. Using game engines and scripting toolkits, we can presently modify and iterate on a scene in a virtual world until we are satisfied with its functional outputs and desired experiential effects.

-- End of page 26 --

As already specified, the creators of the simulation in this second imaginative case are defined by their ability to edit events (as well as our memories of them) without us noticing any discrepancies, hiccups, or interruptions in the simulation (ibid.). In case those creators have complete knowledge and control about the simulation's implementation, we may be tempted to believe that – unlike the previous case – our simulation would not be created to acquire new insights and knowledge. This belief is rooted in the fact that all of the information that could be derived from the simulation would be already inscribed in the way the simulation itself was built and would be, by definition, already available to the advanced civilization that created it. Should that be the case, the creators would not need to use simulations to find answers to their scientific and philosophical questions. It is even less likely that they would knowingly run one that will make billions of artificial beings suffer unnecessarily.

Let us entertain for a moment the possibility that my reasoning is wrong on this point, and that an advanced civilization could still glean knowledge from a simulation over which they have perfect knowledge and control. What would stop our creators, in this new scenario, from retroactively modifying the simulation once the desired information was obtained? What would impede them from editing out all the suffering that the simulated course of events imposed on simulated beings? As inhabitants of the simulation, we would not know or remember that any of that ever happened because, as far as our experience is concerned, it simply did not. The creators' possibility to edit the simulation at any point, to make us 'unexperience' events, or simply rewind its timeline to a period before the emergence of life and turn off the simulation is highly relevant here. This capability of theirs can retroactively amend all circumstances and uses of the simulation, including those that would not be ethically permissible in an advanced civilization. Regardless of whether our creators are using the simulation to harvest knowledge, for their personal entertainment, or to

derive sadistic pleasure in our suffering, the simulation could always be reverted and/or edited to ensure that no being was ever stunted or oppressed during their individual existence (i.e. their individual timeline). Does the fact we are experiencing and witnessing genocides, natural catastrophes, oppression and torture invalidate this hypothesis entirely? Does our suffering indicate that our alleged, benevolent creators do not have complete control over our hypothetically virtual world? That seems to be likely the case, but there might be exceptions. For instance, we could also be part of a simulation that was accidentally left unsupervised. This is a scenario – unlikely as it might be – that could take place in the context of a technologically mature civilization. We could be living in a simulation that was left running as an oversight on the part of our alleged creators (this might be the case if the creators disabled the software that is supposed to monitor the simulations by mistake, or if, for whatever reason, a section of a supercomputer running a simulation becomes inaccessible).

On the basis of what was discussed in this section, it does not appear very plausible that we live in a computer simulation that is developed and run by creators who are both benevolent and have complete control over the simulation itself. Given those premises, the only possible way for us to be leading simulated existences characterized by oppression and violence would be for the simulation in question to have been forgotten and left unsupervised[9].

Conclusion

Supposing that we are indeed living in a computer simulation, what is its purpose? What aspirations and values could be inferentially attributed to its alleged creators? And would living in a simulated universe affect the value and meaning we attribute to our existence?

The present article paired the simulation hypothesis (SH) with another hypothesis advancing the idea that a technologically mature civilization is also likely be morally mature, and would consider artificial beings such as artificial intelligences and the inhabitants of simulations to be ethically relevant (i.e. the PMH). In other words, the reflections and insights offered in this text emerge from the combined assumptions that we indeed live in a computer simulation, and that the civilization who developed the simulation would consider it a basic moral duty to respect and preserve our autonomy and well-being.

[9] For the sake of completeness, I want to emphasize that the scenario that was just outlined does not necessarily remove any possibility of significance for our existence to ever transcend the limited concerns and duration of our civilization. There might be situations in which our lives (spent in a forgotten and unsupervised simulation) could still potentially be more than a senseless squander of computational power. Consider the case in which our lost and forgotten simulation eventually is found and examined by someone other than its original, benevolent creators, presumably providing a wealth of technical and culturally relevant information. The finders of the lost simulation need to be entities without the possibility to edit and revert the simulation, or with less ethical scruples than our alleged original creators. In any case, they need not be the benevolent, advanced civilization that originally developed it. If our creators were the ones finding the simulation, chances are that they will likely be able to detect the cosmic tragedy we are experiencing, and we would have already been mercifully reverted out of existence. Clearly, if that were the case, I would not be writing this article, and you would not be reading it.

On these premises, I elaborated a few arguments amounting to what could be considered a theodicy for artificial universes. In it, I reasoned that the presence of evil and suffering in the allegedly simulated world we inhabit be ethically justifiable if one (or more) of these hypothetical scenarios happened to be the case:

1. our simulation was the first of its kind to ever give rise to artificial beings that are worthy of ethical consideration for the creators of the simulation,

2. our simulation serves a knowledge-gathering purpose which the creators considered to be ethically positive for their moral community in the foreseeable long run, or

3. our simulation was forgotten and left unsupervised.

The last scenario is not specific to any particular relationship between the creators of the simulation and their technology. It does not even require that we imagine the advanced civilization that is allegedly simulating our universe to be actively invested in limiting damage (and avoid potential damage) for a moral community that is vast and inclusive. The third one is, however, a hypothetical situation that almost certainly excludes the possibility for our individual existences and our collective suffering to have a significance that transcends the limited concerns and durations of our civilizations.

Differently from the third scenario, the first and second hypothetical situations depend on the benevolence of the creators of the simulation that we inhabit according to the SH. Should those two hold, our existence and our suffering would be guaranteed to have existential meaning beyond our individual lifespans and our survival as a species. Our simulation would in fact be producing useful insights for the creators, insights that they estimated likely to be ethically positive for the future of their moral community.

It is hard to assess how likely it is that we are currently living in a simulation. It is, however, imaginable that we would be able to have a clearer grasp on that possibility as scientific knowledge develops progressively reliable models of the physical behaviors of our universe. In the current absence of certainties in that regard, we could perhaps console ourselves with the thought that – in case our alleged creators are not evil or deranged – the oppression and the suffering we are experience in our existence are likely not futile[10].

[10] It is potentially relevant to point out that also this way of reasoning about our alleged advanced creator has a direct analogy in religious conviction. The idea that our suffering plays a part in a larger project meant to lead to a world of increased autonomy and well-being resonates with the concept of the 'Will of God', that is to say the idea that a benevolent god (or gods) have typically inscrutable plans for humanity that are supposed to usher a future of widespread well-being and compassion.

References

Banks, I. M. 2012a [2008]. *Matter*. New York, NY: Orbit.

Becker, E. (1976). *The Structure of Evil: An Essay on the Unification of the Sciences of Man*. New York (NY): Free Press.

Bostrom, N. (2003). "Are we living in a computer simulation?" *The Philosophical Quarterly, 53* (211), pp. 243-255.

Bostrom, N. & Yudkowsky, E. (2014). "The ethics of artificial intelligence". *The Cambridge handbook of artificial intelligence, 1*, pp. 316-334.

Castelvecchi, D. (2016). "Can we open the black box of AI?". *Nature News, 538* (7623), pp. 20-23.

Chalmers, D. J. (2005). "The Matrix as metaphysics". In *Philosophers Explore the Matrix*. Oxford, UK: Oxford University Press, pp. 132-176.

Coeckelbergh, M. (2010). "Robot rights? Towards a social-relational justification of moral consideration". *Ethics and information technology, 12* (3), pp. 209-221.

Descartes, R. (2013) [1641]. *René Descartes: Meditations on first philosophy: With selections from the objections and replies*. Cambridge, MA: Cambridge University Press.

DeWitt, B. S. (1967). "Quantum theory of gravity. I. The canonical theory". *Physical Review, 160* (5), 1113.

Gualeni, S. (2020). "Artificial Beings Worthy of Moral Consideration in Virtual Environments: An Analysis of Ethical Viability". *Journal of Virtual World Research, 13* (1).

Gualeni, S. & Vella, D. (2020). *Virtual Existentialism: Meaning and Subjectivity in Virtual Worlds*. Basingstoke, UK: Palgrave Pivot.

Gunkel, D. J. (2018). *Robot rights*. Cambridge, MA: The MIT Press.

Hamilton, C. (2016). "The theodicy of the 'Good Anthropocene'". *Environmental Humanities, 7* (1), pp. 233-238.

Harrison, P. (1989). "Theodicy and animal pain". *Philosophy, 64* (247), pp. 79-92.

Ihde, D. (1990). *Technology and the Lifeworld: From Garden to Earth.* The Indiana series in the Philosophy of Technology. Bloomington, IN: Indiana University Press.

Johnson, D. K. (2011). "Natural Evil and the Simulation Hypothesis". *Philo, 14* (2), pp. 161-175.

Kant, I. (2000) [1998]. *Critique of Pure Reason*. Trans. Guyer, P. and Wood, A. W. Cambridge, MA: Cambridge University Press.

Leibniz, G. W. (2000) [1710]. *Theodicy: Essays on the Goodness of God, the Freedom of Man and the Origin of Evil*. Eugene (OR): Wipf and Stock Publishers.

Moravec, H. (1988). *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press.

Neely, E. L. (2014). "Machines and the Moral Community". *Philosophy & Technology*, *27* (1), pp. 97-111.

Page, D. N., & Wootters, W. K. (1983). "Evolution without evolution: Dynamics described by stationary observables". *Physical Review D*, *27* (12), pp. 2885-2892.

Salen, K., Zimmerman, E. (2003). *Rules of Play: Game Design Fundamentals*. Cambridge, MA: The MIT Press.

Selinger, E. (2009). "Simulation". In Olsen, J. K. B., Pedersen, S. A., and Hendricks, V. F. (eds.) *A Companion to the Philosophy of Technology*, Chichester, UK: Blackwell Publishing Ltd., pp. 157-159.

Singer, P. (2002) [1995]. *Animal Liberation*. Ecco, USA.

Steinhart, E. (2010). "Theological Implications of the Simulation Argument". *Ars Disputandi*, *10* (1), pp. 23-37.

Surin, K. (1983). Theodicy?. *The Harvard theological review*, *76* (2), pp. 225-247.

Swedberg, R. (2005). *The Max Weber Dictionary: Key Words and Central Concepts*. Redwood City, CA: Stanford University Press.

Tipler, F. J. (1997). *The Physics of Immortality: Modern Cosmology, God, and the Resurrection*. New York, NY: Anchor.

Wachowski, A., Wachowski, L. (1999). *The Matrix*. Film produced by Joel Silver. Burbank, CA: Warner Home Video.