AI & SOCIETY (2023) 38:1579–1586
https://doi.org/10.1007/s00146-021-01312-y

OPEN FORUM

The fiction of simulation: a critique of Bostrom's simulation argument

Miloš Agatonović[1]

Received: 21 June 2021 / Accepted: 25 October 2021 / Published online: 5 November 2021

Abstract
Nick Bostrom's "simulation argument" purports to show that if it is possible to create and run a vast number of computer simulations indistinguishable from the reality we are living in, then it is highly probable that we are already living in a computer simulation. However, the simulation argument requires a modification to escape the undermining implications of the scepticism it implies, as argued by Birch. The present paper shows that, even if the modified simulation argument is valid, still it is unsound since it relies on the indistinguishability assumption that even in principle cannot be tested. To account for the unsoundness of the simulation argument, the present paper draws on John Woods' theory of fiction, to expose structural similarities between general fiction and the simulation argument. Though the simulation argument is unsound, it seems persuasive, because the argument immerses the reader in a fictive world with the help of tacit assumptions, leveraging just enough common sense to remain compelling while covering over an untestable premise. Simultaneously with the critique of Bostrom's argument, Chalmers' argument for the matrix hypothesis is assessed on similar criteria. In either case, both arguments rely on an accumulation of assumptions, both implicit and explicit, hiding the premises that are untestable in principle.

Keywords Simulation argument · Matrix hypothesis · Fiction · John Woods' *Truth in Fiction*

1 Introduction

The idea of reality as computation is very vivid today. Although widely disputed, there are some recent attempts to explain the entire reality as computation. In 2020, Stephen Wolfram announced *The Wolfram Physics Project: A Project to Find Fundamental Theory of Physics,* which aims at completion of the framework based on what Wolfram calls the Principle of Computational Equivalence, that "all processes, whether they are produced by human effort or occur spontaneously in nature, can be viewed as computations" (Wolfram 2002, p. 715; Wolfram 2020). According to the idea of Wolfram's project, the fundamental theory of physics should be based on "a general computational paradigm" (Wolfram 2020).

In terms of recent studies (Piccinini 2015, 2017; Piccinini and Anderson 2018), Wolfram's project can be characterized as a kind of pancomputationalism. One version of the pancomputationalist view is ontic pancomputationalism, best known in its earliest versions because of Konrad Zuse and Edward Fredkin (Piccinini 2015, p. 54). Ontic pancomputationalism argues that the physical universe is computational, in the sense that at the fundamental level there is a computation that determines the nature of a physical system (Piccinini and Anderson 2018).

Bostrom's "simulation argument" ("SA") and Chalmers' "matrix hypothesis" ("MH") represent instances of ontic pancomputationalism, under the assumption of so-called "strong simulationism". According to Piccinini and Anderson, strong simulationism is defined in the following manner: "Our universe is a computational simulation run by a physical computer that exists in a separate physical universe" (Piccinini and Anderson 2018, p. 31). As Piccinini and Anderson suggest, strong simulationism has problems regarding the empirical testability of the assumptions which it implies (Piccinini and Anderson 2018, pp. 32–4). For example, the assumption that computational simulations can produce conscious experiences indistinguishable from our own (subjective indistinguishability assumption) is not empirically supported (Piccinini and Anderson 2018, p. 33). Contrary to Piccinini and Anderson, Bostrom states

✉ Miloš Agatonović
agatonovicmilos@yahoo.com;
milos.agatonovic@vaspks.edu.rs

[1] The Academy of Applied Preschool Teaching and Health Studies, Preschool Teacher Training College, Cirila and Metodija No. 22-24, 37000 Kruševac, Republic of Serbia

that there are "empirical reasons for thinking that running vastly many simulations of human minds would be within the capability of a future civilization" (Bostrom 2003, p. 244). These "empirical reasons" give Bostrom leverage to argue that it is possible to create such a vast number of such simulations, and plausible to conclude that it is highly probable that we are already living in a computer simulation. Using simple probability theory alongside additional principles and assumptions, Bostrom formulates SA, which opens thought provoking questions about the nature of reality. The problem is that SA's conclusion about the nature of reality in which we are living is supported only by known limits of computation of existing technology. These limits of computation in principle might be explored in the simulation of the current physical world and our subjective experience within it, i.e. conceiving of being so ensconced.

In the following, Sect. 2 reviews some theological and metaphysical implications of SA, while Sect. 3 focuses on the tacit assumptions that are necessary for SA to appear valid, drawn from Birch (2013). Section 4 shows that SA is unsound, for it relies on the indistinguishability assumption that is untestable. Section 5 sketches the fiction of reality as a computer simulation from the perspective of Woods' recent *Truth in Fiction* (2018). This excursion on fiction accounts for why Bostrom's SA and Chalmers' MH arguments seem persuasive. The paper concludes that SA and MH are a sort of fiction that relies on common experiences with fiction to motivate belief in the story that it tells. Despite being unsound, SA and MH are persuasive, because key background theses are subtly hidden in them, much as in other believable fictions.

2 The implications of the simulation argument

In the paper "Are You Living in a Computer Simulation?" from 2003, Nick Bostrom argues that at least one of the following propositions is true:

(i) the human species is very likely to become extinct before reaching a "post-human" stage;
(ii) any post-human civilization is extremely unlikely to run a significant number of simulations of its evolutionary history (or variations thereof);
(iii) we are almost certainly living in a computer simulation (Bostrom 2003, p. 243).

Proposition (i) represents the possibility of humankind failing to reach post-humanity, a stage of technological maturity, failing to reach the technological capacity to create a detailed simulation of reality. According to (i), it would be very unlikely that we are living in a computer simulation. The second alternative, represented by (ii), assumes post-humanity is reached by many civilizations, but the fraction of these civilizations interested in running the simulations is negligibly small. The third alternative, that we are almost certainly living in a computer simulation, assumes that a significant fraction of civilizations becomes post-human and that they produce a vast number of simulations. However, for (iii) to be true, simulated worlds should not be outnumbered by non-simulated worlds, which is emphasized in a patch for SA offered by Bostrom and Kulczycki (2011).

The third alternative is the most thought provoking, since if we are living in a simulation, what we perceive as reality is not at the fundamental level. It could be possible that there are simulations inside a simulation. Our universe could be a "virtual machine"—a computer simulated on another computer. Therefore, a corollary of the third alternative is, as Bostrom writes: "Virtual machines can be stacked: it is possible to simulate one machine simulating another machine, and so on, in arbitrarily many steps of iteration" (Bostrom 2003, p. 253). We can characterize a system of hierarchy of worlds as virtual machines as naturalistic, or even physical, taking the world on a fundamental level to be such. However, according to certain interpretations, SA may support the thesis that below every level, there is a deeper level, and so, that there is an endless series of ever-deeper levels (Steinhart 2010, p. 25). Thus, SA may support the Leibnizian cosmological argument, assuming the existence of an infinite computer as the best explanation for an endless series of finite computers (Steinhart 2010, p. 28). Also, SA may support the design argument, assuming that the designer of our universe is a deeper civilization and that the sequence of the deeper civilizations is the sequence of deeper designers (Steinhart 2010, p. 30). According to this novel design argument, the deeper civilizations are always more intelligent and more powerful, so they create more perfect computers (Steinhart 2010, p. 30). Therefore, the infinitely intelligent and powerful civilization is the original designer who creates the perfect computer and the entire sequence of civilizations.

As we see, Steinhart (2010) discusses the theological implications of SA, developing novel versions of the cosmological and design arguments. These implications show how SA may represent "a creation myth of the information age", to use Chalmers' phrase for describing MH (Chalmers 2010, p. 466). The fundamental assumption of this "creation myth" is strong simulationism, claiming: our universe is a simulation created by a computer outside our universe. We can speculate that the creators of simulations at the fundamental level might use an infinite computer or the perfect computer, as in the novel versions of the cosmological and design arguments (Steinhart 2010). In addition, the original creators might use what Chalmers calls "matrix"—"an artificially designed computer simulation of a world" (Chalmers 2010, p. 456). According to Chalmers' MH, someone is *in*

a matrix "if he or she has a cognitive system that receives its inputs from and sends its outputs to a matrix" (Chalmers 2010, p. 456). The world created with the use of a matrix, or with the use of an infinite computer or the perfect computer, is explainable by a pancomputationalist view such as that of the pioneers of digital physics Edward Fredkin and Stephen Wolfram, as claimed by Chalmers (Chalmers 2010, p. 460). Although pancomputationalism is highly speculative, the recently announced *The Wolfram Physics Project* (2020) is trying to find a "theory of everything" within a pancomputationalist framework, which could set the ground for creating a computer simulation of physical reality.

To estimate the possibility of creating a computer simulation of reality, as Bostrom intends to do in SA, a "theory of everything" is unnecessary. We can establish a "lower bound of computation" by only assuming the physical laws and computing power already known to us (Bostrom 2003, p. 245). As providing a rough approximation of the limits of computation, Bostrom intends to show that in the future it will be possible to cost-effectively produce "simulations that are indistinguishable from physical reality for human minds in the simulation" (Bostrom 2003, p. 247). He argues, it will be possible to produce *ancestor-simulations*—computer simulations of the mental history of humankind. Bostrom assumes that a post-human civilization could create a planetary-mass computer with an approximate computational power of 10^{42} operations per second using nanotechnological designs already known to us (Bostrom 2003, p. 247). Such a computer could create ancestor-simulations using only a small part of its computation capacity. Bostrom concludes that a post-human civilization would have the computing power sufficient to run a huge number of ancestor-simulations (Bostrom 2003, p. 248). If a post-human civilization creates these simulations, then it is highly probable that we are living in a computer simulation, as claimed by (iii). And if we are living in such a simulation, then our world is fundamentally computation, while the computer which produces the simulation cannot be a part of its simulation. Thus, SA implies pancomputationalism together with strong simulationism, opening the void for theological and metaphysical speculations we discussed previously.

Independent from the empirical reasons for thinking that it is possible to create ancestor-simulations, Bostrom assumes that the post-human civilizations should be *interested* in using it to run ancestor-simulations rather than any other sort of simulation. We can speculate about motives and reasons that may prevent or that may promote the interest to create ancestor-simulations. Certain ethical considerations may forbid such an enterprise. For example, creating conscious beings that suffer may be regarded as immoral among the post-human civilizations, and for that reason creating ancestor-simulations would be banned in any post-human society (Bostrom 2003, p. 252). Also, post-human civilizations may not desire to create ancestor-simulations, because those kinds of simulations have no educational or scientific value (Bostrom 2003, p. 252). Alternatively, they may not desire to create ancestor-simulations as a recreational activity because it is "a very inefficient way of getting pleasure" (Bostrom 2003, pp. 252–3). If any of these cases about reasons that may prevent creating ancestor-simulations is true, proposition (ii) will be true. However, we may regard ancestor-simulations in analogy to simulations known to us. Known simulation technology, such as "virtual reality" (VR), can have an educational, entertaining and therapeutic function, providing us with various kinds of experience, even of unfamiliar realms of reality or fictional worlds. VR can be useful in attaining practical skills (e.g., flight simulator is used to teach one to fly) and propositional knowledge (e.g., in physics, in research of the fluid behaviour of stars) (Cogburn and Silcox 2014, p. 577). Large-scale realistic simulations may help us achieve a better society, even a post-human society, by showing us the errors we can make in the history of humankind so we can correct them (White 2016, p. 189). Thus, according to White (2016, p. 189), if we can create large-scale realistic simulations such as ancestor-simulations, the probability that we are living in a simulation depends on our commitment to the post-human condition. We might be capable and interested in creating ancestor-simulations as some post-human civilization already might have created the one in which we are living.

However, White's argument assumes a similarity between our interest and the interest of the assumed simulators. If this similarity assumption is false (interest of the creators being radically different), the reality of the simulators may be incomparably different from our reality, as well as the reasons for creating simulations may be scarcely conceivable to us. This entails that inference about the simulators (about their reality or reasons for creating a simulated world) according to the analogy with our reality and our reasons for using simulation technology would be invalid. For instance, in contrast to scientific or practical purposes of simulation technology known to us, it may be the case that the simulators intend merely to deceive the observers inside a simulation. If the simulators used an infinite computer or the utmost computer to create simulations, they would have the highest power and intelligence of their technology, having their efforts devoted merely to deceive us, similarly to Descartes' "malicious demon". Because of the immense power, the simulators may not care about educational, scientific or practical purposes when creating a simulation. They could create simulations for no reason at all, or just by chance. They could simulate whatever they want and, however, they like, creating the worlds "out of thin air". Thus, there are many ways in which this malicious simulator hypothesis, as we may call it, can affect SA.

3 Is the simulation argument a fallacy?

The malicious simulator hypothesis can provide many interesting, as well as bizarre, scenarios. Although having the highest power and intelligence, a malicious simulator could simulate only one individual. This would be another alternative to be added to the conclusion of SA, making the conclusion logically weaker (Birch 2013, p. 99).[1] The malicious simulator hypothesis can also be interpreted as a version of the classical sceptical scenario, Descartes' "malicious demon" or "brain in vat" thought experiment. If that is the case, these interpretations could undermine SA, having in mind that the malicious simulators could easily mislead us about the physical laws and empirical evidence about the technological limits of computation that SA presupposes. As Birch notes, "there is no reason to suppose that post-human civilizations would not radically mislead their simulated creations with regard to the true laws of physics, and the true properties of material substrates" (Birch 2013, p. 106). This leads to "a pervasive *de dicto* scepticism about *all* aspects of physical reality, including those aspects epistemically relevant to the limits of post-human computation" (Birch 2013, p. 106). However, as Birch suggests, Bostrom intends to convince scientific realists in a sceptical conclusion, without assuming a pervasive scepticism that the malicious simulator hypothesis implies (Birch 2013, p. 100). According to Bostrom's clarification, SA "relies crucially on non-obvious empirical premises about future technological abilities", trying to provide "specific empirically-grounded reasons for revising these initial beliefs in a certain way" (Bostrom 2008). Still, to prevent the sceptical conclusion of SA from undermining its empirical assumptions about the limits of computation, Bostrom ought to assume that "my access to the facts about the physical limits of computation is better than my access to the facts about my own material constitution" (Birch 2013, p. 102). This supposition follows if the limits of computation are the same for simulated and non-simulated worlds. In that case, we could access the facts about the limits of computation without knowing whether we are living in a simulation. Hence, Birch considers circumstances in which beliefs about computational limits may be true, suggesting that any observer might locate himself in one of two or more indistinguishable subjectively centred worlds given:

(i) an observer is physically real while the other is simulated;
(ii) the physically real observer holds justified beliefs about the physical limits of computation, obtained through veridical experiences;
(iii) both observers have the same beliefs and the same evidence (Birch 2013, p. 105).

In these circumstances, a human observer cannot support the mundane claim of possessing two real human hands, while having good evidence for claims regarding the fundamental physical limits of computation. Any observer can be simulated or real, and not knowing whether one is living in a simulated or non-simulated reality, still holding justified beliefs about the physical limits of computation.

With this argument, we may note the difference between *de dicto* and *de se* beliefs to describe the circumstances in which the selective scepticism of SA is successful. Beliefs about the physical limits of computation, as scientific beliefs in general, are *de dicto* beliefs. They locate the actual world within the set of possible worlds. According to David Lewis (to whom Birch refers, Birch 2013, p. 104), *de dicto* beliefs are also *de se*, because they are self-locating in the sense of locating the world in which we are living in logical space (Lewis 1979, p. 522). However, *de dicto* beliefs do not locate us in ordinary time and space, or among individuals (for example, those with two hands) of the population, which is the function of irreducibly *de se* beliefs (Lewis 1979, p. 522). Birch's version of Bostrom's selective scepticism recalls Elga's (2004) argument that scepticism with respect to *de se* beliefs does not necessarily entail scepticism regarding *de dicto* beliefs. Elga's argument depends on the indifference principle for self-locating beliefs that "similar centred worlds deserve equal credence" (Elga 2004, p. 387). Elga's point is illustrated by the example of Dr Evil whose duplicate, Dup, is created with subjectively indistinguishable experience. Since Dr Evil is aware of Dup's existence and vice versa, neither of them can be justified in his belief that he is Dr Evil and other is Dup (Elga 2004). The world centred on Dr Evil is subjectively indistinguishable from that centred on Dup, and according to the indifference principle, both worlds deserve equal credence.

Because of the assumption about the subjective indistinguishability of simulated and physical worlds, Elga's example straightforwardly applies to SA. The indistinguishability assumption of SA implies that both kinds of reality, simulated and physical, deserve equal credence. If the worlds centred on simulated observers are subjectively indistinguishable from the world centred on a physically real observer, we can ascribe equal credence to these worlds. Therefore, under those circumstances, any observer can locate himself in any of these centred worlds, implying that it is possible that we are living in a simulated world. Therefore, depending on the numbers of the worlds centred on simulated and the worlds centred on non-simulated observers, it implies that it

[1] I am thankful to an anonymous reviewer for suggesting this possible interpretation of SA. Bostrom (2008) also argues along these lines, as Birch points out in a footnote (Birch 2013, p. 100).

is probable that we are living in a computer simulation. If the number of the indistinguishable worlds centred on simulated observers is much greater than the number of the worlds centred on non-simulated observers, then the probability that we are living in a simulation is greater than the probability that we are living in a non-simulated reality. If the number of simulated centred worlds is enormously great (as Bostrom assumes would be the case if it is possible to create computers that are powerful enough to produce them), then it is highly probable that we are already living in a computer simulation (as claimed by proposition (iii) of SA's conclusion). Therefore, proposition (iii) is a priori implied by the indistinguishability assumption, indifference principle, and basic probability.

4 The indistinguishability assumption is untestable

The major problem with SA is that the indistinguishability assumption is necessary for SA to be valid, while this assumption is untestable. The indistinguishability assumption implies we can locate ourselves in a simulated centred world that is indistinguishable from a non-simulated centred world, and so that we may be already living in a simulation. Without the indistinguishability assumption, as Birch persuasively argued, if the physical laws and the technological limits of computation in the simulated and non-simulated centred worlds are essentially different, SA is a *non sequitur*. With the indistinguishability assumption, SA validly concludes that if a post-human civilization produces an astronomically great number of ancestor-simulations, then we are almost certainly living in a computer simulation. However, to know that simulated and non-simulated centred worlds are indistinguishable, we should be able to compare them. Following SA closely, this is impossible, even in principle. If it would be possible that we make the comparison from the outside of a simulation, it would imply that we are not living in a simulation but outside of it, which is inconsistent with the possibility of proposition (iii). Here, while our observation may confirm the empirical premises (the assumptions about the technological limits of computation, predictions about post-human civilization, and the indistinguishability assumption), all propositions of SA's conclusion are false. If we were living in a simulation, and we could observe non-simulated reality from the inside of a simulation (let us say, because of a glitch in the simulation), our simulated reality and non-simulated reality would be distinguishable (at least, because of that glitch). In addition, as Bostrom explicitly claims, if an error occurred in a simulation, a simulator could edit the mental states of the observers in a simulation to save the simulation from spoiling (Bostrom 2003, p. 247). Therefore, SA implicitly suggests that the indistinguishability assumption cannot be tested inside, as well as outside, a simulation.

A similar problem occurs with Chalmers' argument, for it also assumes a version of the indistinguishability assumption. MH implies it is possible that some sort of computational system underlies physical reality (p). However, this is true only under the principle that "any abstract computation that could be used to simulate physical space–time is such that it could turn out to underlie real physical processes" (Chalmers 2010, p. 469). Chalmers needs to adduce such a principle to discard an objection that a simulation is not the same as reality (Chalmers 2010, p. 468). However, this principle claims nothing more than that it is possible that a computer simulation of physical reality underlies physical reality. Here, we reduced Chalmers' argument to a platitude: under the assumption that it is possible that computation underlies physical reality (p) (which follows from the mentioned principle), it is possible that computation underlies physical reality (p) (ontic pancomputationalism of MH). Formally, any similar argument would be logically valid, for "if p, then p" yet only trivially true. With the mentioned principle, Chalmers also presumes that "for an abstract computation to qualify as a simulation of physical reality, it must have computational elements that correspond to every particle in reality" (Chalmers 2010, p. 469). Besides, he assumes that computation should be connected to our experience in a relevant way, so that "when we have an experience of an object, the processes underlying the simulation of that object must be causally connected in the right sort of way to our experiences" (Chalmers 2010, p. 470). Without the relevant connection between the simulation and our experience, there would be no reason to think that computation underlies physical reality. This connection between the simulation and our experience should provide the subjective indistinguishability (Chalmers 2010, p. 456), which Chalmers also needs to assume in MH, as it is necessary for SA. However, if we are living in a matrix and have always lived in a matrix, as MH claims (Chalmers 2010, p. 457), we cannot check if there is a relevant connection between the simulation and our experience. We cannot check if an abstract computation underlies physical processes of the actual reality. Therefore, we cannot disregard the objection that simulation is not the same as reality.

To check that simulation is the same as reality is very much like checking that fire-breathing dragon lives in my garage, as described in Carl Sagan's (1996) thought experiment. Because this dragon is invisible, incorporeal, floating, and she spits heatless fire, there is no way to test the claim that she lives in my garage. There is no empirical experiment that could falsify the claim. As in Sagan's thought experiment, any method applied to gain evidence about the indistinguishability of simulated and non-simulated reality

would not work. As Sagan noticed about the "dragon in my garage":

Claims that cannot be tested, assertions immune to disproof are veridically worthless, whatever value they may have in inspiring us or in exciting our sense of wonder. What I'm asking you to do comes down to believing, in the absence of evidence, on my say-so. (Sagan1997, p. 169).

Given the significance of the ontological question, that SA attempts to settle, that the perceived reality is, in fact, a computer simulation, to ground the argument on an untestable assumption certainly is not convincing enough.

5 Fiction and the simulation argument

Because SA is skillfully hiding an untestable assumption, the argument appears persuasive. The persuasiveness of SA may be captured in comparison with other believable fictions. The following section constructs the fiction of the simulation of the actual reality, the "fiction of simulation" (FS), showing that the rhetorical tactic of SA, but also the more general conception of the reality as a simulation of MH, is structurally similar to how a believable fiction first immerses us into its world, and from this immersion leads us to treat such a fictional world as if it were real. Indeed, this is the purpose of fiction. SA and MH seem believable, at least in part because of our common engagement with such fictional constructs that only serve their purpose if we engage with them as if they were true.

FS describes a fictional world where one civilization has reached the post-human stage and had created many ancestor-simulations, similar to that which Bostrom (2003) describes in his SA. In terms of a fictive narrative, a civilization in a pre–post-human stage is simulated in a fictional reality created by a post-human civilization. The question then becomes, can the fictional observers come to reliable estimation of the likelihood that their existence is fictional, a product of ancestor-simulations produced and run by non-fictive entities in a real world. The fictional scientists and philosophers, in both civilizations, ask the same question: Is it possible that we are already living in a simulation? If the fictional characters live in a simulation, the answer to the question would be affirmative and true in situ *of the story*. In situ *of the world*, the answer would be that it is not possible that we are living in a simulation for the characters that ask are just fiction.

Here, we follow John Woods' account of fiction from his recent book *Truth in Fiction: Rethinking its Logic* (2018). We lean on Woods' conception of fiction according to which sentences of fiction can be true in situ of a story while being false in situ of the world. For example, the sentence "Sherlock Holmes lived at 221 Baker Street" is true in situ of the stories that Arthur Conan Doyle wrote, but false in situ of the facts of the world (Woods 2018, pp. 86–87). Woods' approach to fiction is along the line of ordinary language intuitions, and for this reason, we take it to be appropriate for interpreting FS (and by extension, SA and MH). Assumptions that Woods recognizes in fiction are analogous to assumptions implicit in SA and MH. We will not need to defend Woods' (2018) conception of fiction here. Rather, we use his ideas as a heuristic in a critique of SA and MH by way of FS.

According to Woods' (2018, pp. 29–30), we engage with fictions in various ways. In terms of a fiction involving Sherlock Holmes, for instance, we experience ourselves referring to him (what Woods calls "*referential* relation") and ascribing to him some features ("*ascriptive* relation"). When we say that he lived at 221 Baker Street in London, we see ourselves saying what is true ("*alethic* relation") and what we believe about Sherlock and 221 Baker Street ("*doxastic* relation"). If we say that such an assertion is based on Doyle's novel *A Study in Scarlet*, we take it that our true belief is well supported in this context ("*epistemic* relation"). When we read about Sherlock's adventures in Doyle's novels, we experience ourselves taking on different affective states ("*affective* relation"). Woods proposes that "inferences from and within fiction operate, if at all, in a much more circumscribed way than natural language in referentially stable inferences" (Woods 2018, p. 35) such as those grounded in the everyday world. In the story's context, we can infer that Sherlock Holmes is a man, but that in the world's context he is a fictional character.

The same applies to FS. Engaging with FS, we could say that we experience ourselves having referential, ascriptive, alethic, doxastic, epistemic, inferential, and affective relations with a computer-simulated reality. We experience ourselves referring to a simulated reality and ascribing it some characteristics, for example, that it is reducible to basic computer algorithms. We can say that characters from FS live in a computer simulation, that this is true in the context of the FS story, and believe that such a statement is true with adequate justification (from the story). We can experience ourselves feeling angry, frustrated, or have some other attitude in regards to the fiction, in the case of FS, that there are persons living in a computer simulation and that they suffer. We can also infer which characters from which narrative sequences of FS are living in a computer simulation, while in the non-fictive world, we infer that the notion that reality is a computer simulation is just a fictional scenario, FS.

For readers to engage with a fictional story, the world of a story must be subject to some basic theses emphasized by Woods (2018):

(i) *the world-inheritance thesis,* which claims that the world of fiction inherits the actual world, except for

adjustments invented in the story (Woods 2018, p. 81);

(ii) *the storyworld epistemic access thesis*: "what readers know of the world of the stories… is what we know or could come to know about their own world at the times in which those stories were set" (Woods 2018, p. 81);

(iii) *the world-inheritance semantic-preservation thesis*: "Meaning, reference and truth hold constant in the transition from the world to the story, save for auctorial provision otherwise" (Woods 2018, p. 81).

Without these theses, fictions could contain some queer fictional entities, such as humans with no spines, with the effect that readers could not easily be immersed into such a fictional world. In not-reinforcing real-world intuitions, such fictions become unbelievable, contrary to the purpose of fiction in general, and people do not engage with them. In making this point, Woods proposes the "*no spines-no readers thesis"* claiming that "In the absence of what the world inheritance and storyworld access conditions provide, the Holmes stories would have null readerships" (Woods 2018, p. 82).

FS uses the framework of SA to satisfy mentioned theses. It uses explicit assumptions of SA to make the FS story more realistic. Technological capabilities should be consistent with the physical laws of the world in which we are living, as Bostrom claim we can show that they are (Bostrom 2003, p. 244, 245). For instance, assumptions about the technological limits of computation explicit in SA offer guidelines for the persuasive description of computers that can create simulations of reality in FS. With such SA assumptions about the technological limits of computation in place, FS inherits real-world conditions, including physical laws. In that way, FS persuasively presents the technological capabilities of the post-human and pre–post-human civilization. FS explicitly states that it is a real possibility that the realities presented in the story are a computer simulation. In FS, we can refer to reality as a computer simulation as we refer to Sherlock Holmes in Doyle's stories. Both are believable fictions, because they inherit much of the world that we already take to be true, including those reliable means through which we understand the world in terms of which we live, including in Holmes' case a particular capacity for reasoning about what is true (or not) of the world as otherwise perceived.

We can understand FS as a fictional story about "a creation myth for the information age" (Chalmers 2010, p. 466). The worlds in FS are described as they are possibly created in some kind of computer, as the creation hypothesis of Chalmers' MH claims. As Chalmers states, many people accept a version of the creation hypothesis, believing that God created the world. The creation myth of FS is not a strange world-view. It can even be believable and compelling if the possibility of creation of the world through computation is believably presented. Chalmers does not explain how it is possible to create the world *indistinguishable* from the one we are living in. He simply assumes that any abstract computation that could underlie physical processes could simulate the physical world, and so, it could be possible that computation underlies the world we are living in (computational hypothesis) (Chalmers 2010, p. 469). This assumption is a background thesis of FS, and surely, it should be represented in the story with much more detailed and vivid descriptions of how civilisations use or could use a computer to create a simulation of the reality. Therefore, by imagining how a simulation of reality is being created according to an illustrative description of the FS story, while immersed into its world, we are being inveigled to ask the same question as the characters in FS do.

The goal of fiction is to immerse us into its world so that its storyline would be taken seriously. For that purpose, fiction can use realistic elements from science so to be believable. The same can be said about SA and MH. SA uses the scientific facts about the limits of computation to show that it is physically possible to create many ancestor-simulations. MH appeals to the ontic pancomputationalist views in physics to show that it is coherent to believe that the reality we are living in is computational. However, SA and MH do not prove their point but still captivate our interest, as does the believable fiction.

We embrace the framework of SA or that of MH, their explicit and implicit assumptions; similarly, we embrace the world of the FS story. The world of FS is compelling because it is our world, added to a simulated reality that is described as the product of the real technological capacities, at least according to the limits of computation that Bostrom provided in SA and that are used in FS. If the world of FS differed completely from ours including, for example, the physical basis for its computational limits, we could infer nothing about that world. It would be alien and unbelievable for us, and as Woods claims, the story would have no readership (Woods 2018, pp. 82–3). In such a case, there would be no interest in the simulations reflecting experiences vastly different from those of observers within a post-human world, and SA would fall to the fact that there would be little probability of the creation of astronomical numbers of simulations in the first place. Post-humans simply would not care enough to generate the interest.

Having in mind Birch's critique of SA presented in Sect. 2 of this paper, SA and MH are like FS straightforwardly. Without the indistinguishability thesis, if the non-simulated and simulated worlds were radically different and so distinguishable (including here by the differences in experiences between pre–post-human and post-human observers), the argument would fail and its conclusion would not follow regardless of assumptions about the physical limits

of computation. The indistinguishability thesis in SA and MH has a similar function as the world-inheritance thesis in FS—either of the theses enables us to preserve our beliefs about reality in the cases of both non-simulated and simulated centred worlds, as the truths of our world are preserved in the world of the FS story. That is clear if we have in mind that the laws of physics and the technological limits of computation hold in the non-simulated and simulated realities in SA (according to the indistinguishability thesis), as they hold in the fictional world of the FS story (according to the world-inheritance thesis). Contrary to the explicit claim that it is possible that any reality presented in FS is a computer simulation in the FS story, SA dissembles the indistinguishability thesis, so to induce that it is possible that we are living in a computer simulation. In that way, SA leads us to accept the framework of the argument, while we overlook the fact that the indistinguishability thesis is untestable.

6 Conclusion

In this paper, we have tried to show that SA and MH are unsound and, therefore, do not provide sufficient reason to accept the possibility that we are living in a computer simulation. The persuasiveness of SA and MH comes from the hidden background theses that implicitly presume that simulated worlds indistinguishably correspond to the real world. Because these implicit assumptions are untestable, SA and MH cannot prove that we may be already living in a computer simulation. However, SA and MH could provide us with material for a work of fiction. The fiction story named "FS", which we have sketched here, uses the assumption about the limits of computation from SA to make FS world more compelling. Although this and other assumptions may work well with FS, they are not persuasive enough to make SA plausible. The fictional scenario of SA and MH operates as an intuition pump, as Daniel Dennett (2013) would say, a story designed to settle the question which it addresses by inducing an intuition—"That's it, it makes sense! It must be possible that we are living in a computer simulation!" However, we forcefully insist that this story is of a dubious kind, but it artfully deceives us to accept its point. Therefore, to conclude with the words of one of the wisest fictional characters, Sherlock Holmes:

> It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.
>
> -Arthur Conan Doyle, *A Scandal in Bohemia* (1891).

> When you have eliminated all which is impossible, then whatever remains, however improbable, must be the truth.
>
> -Arthur Conan Doyle, *The Adventure of the Blanched Soldier* (1926).

Acknowledgements The author is grateful to Jeffrey White, Karamjit Gill, and to anonymous reviewers with AI & Society for insights and support in bringing this paper to press.

References

Birch J (2013) On the 'simulation argument' and selective scepticism. Erkenntnis 78:95–107. https://doi.org/10.1007/s10670-012-9400-9

Bostrom N (2003) Are you living in a computer simulation? Philos Q 53(211):243–255. https://doi.org/10.1111/1467-9213.00309

Bostrom N, Kulczycki M (2011) A patch for the simulation argument. Analysis 71:54–61. https://doi.org/10.1093/analys/anq107

Bostrom N (2008) The simulation argument FAQ. https://www.simulation-argument.com/faq.html. Accessed 30 Sept 2021

Chalmers DJ (2010) The matrix as metaphysics. In: Chalmers DJ (ed) The character of consciousness. Oxford University Press, Oxford, pp 455–494

Cogburn J, Silcox M (2014) Against brain-in-a-vatism: on the value of virtual reality. Philos Technol 27:561–579. https://doi.org/10.1007/s13347-013-0137-4

Dennett DC (2013) Intuition pumps and other tools for thinking. W. W. Norton & Co., New York

Elga A (2004) Defeating Dr. Evil with self-locating belief. Philos Phenomenol Res 69:383–396. https://doi.org/10.1111/j.1933-1592.2004.tb00400.x

Lewis DK (1979) Attitudes de dicto and de se. Philos Rev 88:513–543. https://doi.org/10.2307/2184843

Piccinini G (2015) Physical computation: a mechanistic account. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199658855.001.0001

Piccinini G, Anderson NG (2018) Ontic pancomputationalism. In: Cuffaro ME, Fletcher SC (eds) Physical perspectives on computation, computational perspectives on physics. Cambridge University Press, Cambridge, pp 23–38

Piccinini G (2017) Computation in physical systems. In: Zalta EN (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2017/entries/computation-physicalsystems. Accessed 30 Sept 2021

Sagan C (1997) The demon-haunted world: science as a candle in the dark. Ballantine Books, New York

Steinhart E (2010) Theological implications of the simulation argument. Ars Disputandi 10(1):23–37. https://doi.org/10.1080/15665399.2010.10820012

White JB (2016) Simulation, self-extinction, and philosophy in the service of human civilization. AI Soc 31:171–190. https://doi.org/10.1007/s00146-015-0620-9

Wolfram S (2002) A new kind of science. Wolfram Media, Champaign, IL. https://www.wolframscience.com/nks. Accessed 30 Sept 2021

Wolfram S (2020) Finally We May Have a Path to the Fundamental Theory of Physics… and It's Beautiful. https://writings.stephenwolfram.com/2020/04/finally-we-may-have-a-path-to-the-fundamental-theory-of-physics-and-its-beautiful. Accessed 30 Sept 2021

Woods J (2018) Truth in fiction: rethinking its logic. Springer, Cham. https://doi.org/10.1007/978-3-319-72658-8