ARE YOU LIVING IN A COMPUTER SIMULATION?

BY NICK BOSTROM

Faculty of Philosophy, Oxford University

Published in *Philosophical Quarterly* (2003) Vol. 53, No. 211, pp. 243-255.

[www.simulation-argument.com]

pdf-version: [PDF]

ABSTRACT

This paper argues that *at least one* of the following propositions is true: (1) the human species is very likely to go extinct before reaching a "posthuman" stage; (2) any posthuman civilization is extremely unlikely to run a significant number of simulations of their evolutionary history (or variations thereof); (3) we are almost certainly living in a computer simulation. It follows that the belief that there is a significant chance that we will one day become posthumans who run ancestor-simulations is false, unless we are currently living in a simulation. A number of other consequences of this result are also discussed.

I. INTRODUCTION

Many works of science fiction as well as some forecasts by serious technologists and futurologists predict that enormous amounts of computing power will be available in the future. Let us suppose for a moment that these predictions are correct. One thing that later generations might do with their super-powerful computers is run detailed simulations of their forebears or of people like their forebears. Because their computers would be so powerful, they could run a great many such simulations. Suppose that these simulated people are conscious (as they would be if the simulations were sufficiently fine-grained and if a certain quite widely accepted position in the philosophy of mind is correct). Then it could be the case that the vast majority of minds like ours do not belong to the original race but rather to people simulated by the advanced descendants of an original race. It is then possible to argue that, if this were the case, we would be rational to think that we are likely among the simulated minds rather than among the original biological ones. Therefore, if we don't think that we are currently living in a computer simulation, we are not entitled to believe that we will have descendants who will run lots of such simulations of their forebears. That is the basic idea. The rest of this paper will spell it out more carefully.

Apart form the interest this thesis may hold for those who are engaged in futuristic speculation, there are also more purely theoretical rewards. The argument provides a stimulus for formulating some methodological and metaphysical questions, and it suggests naturalistic analogies to certain traditional religious conceptions, which some may find amusing or thought-provoking.

The structure of the paper is as follows. First, we formulate an assumption that we need to import from the philosophy of mind in order to get the argument started. Second, we consider some empirical reasons for thinking that running vastly many simulations of human minds would be within the capability of a future civilization that has developed many of those technologies that can already be shown to be compatible with known physical laws and engineering constraints. This part is not philosophically necessary but it provides an incentive for paying attention to the rest. Then follows the core of the argument, which makes use of some simple probability theory, and a section providing support for a weak indifference principle that the argument employs. Lastly, we discuss some interpretations of the disjunction, mentioned in the abstract, that forms the conclusion of the simulation argument.

II. THE ASSUMPTION OF SUBSTRATE-INDEPENDENCE

A common assumption in the philosophy of mind is that of *substrate-independence*. The idea is that mental states can supervene on any of a broad class of physical substrates. Provided a system implements the right sort of computational structures and processes, it can be associated with conscious experiences. It is not an essential property of consciousness that it is implemented on carbon-based biological neural networks inside a cranium: silicon-based processors inside a computer could in principle do the trick as well.

Arguments for this thesis have been given in the literature, and although it is not entirely uncontroversial, we shall here take it as a given.

The argument we shall present does not, however, depend on any very strong version of functionalism or computationalism. For example, we need not assume that the thesis of substrate-independence is *necessarily* true (either analytically or metaphysically) – just that, in fact, a computer running a suitable program would be conscious. Moreover, we need not assume that in order to create a mind on a computer it would be sufficient to program it in such a way that it behaves like a human in all situations, including passing the Turing test etc. We need only the weaker assumption that it would suffice for the generation of subjective experiences that the computational processes of a human brain are structurally replicated in suitably fine-grained detail, such as on the level of individual synapses. This attenuated version of substrate-independence is quite widely accepted.

Neurotransmitters, nerve growth factors, and other chemicals that are smaller than a synapse clearly play a role in human cognition and learning. The substrate-independence thesis is not that the effects of these chemicals are small or irrelevant, but rather that they affect subjective experience only *via* their direct or indirect influence on computational activities. For example, if there can be no difference in subjective experience without there also being a difference in synaptic discharges, then the requisite detail of simulation is at the synaptic level (or higher).

III. THE TECHNOLOGICAL LIMITS OF COMPUTATION

At our current stage of technological development, we have neither sufficiently powerful hardware nor the requisite software to create conscious minds in computers. But persuasive arguments have been given to the effect that *if* technological progress continues unabated *then* these shortcomings will eventually be overcome. Some authors argue that this stage may be only a few decades away.1 Yet present purposes require no assumptions about the time-scale. The simulation argument works equally well for those who think that it will take hundreds of thousands of years to reach a "posthuman" stage of civilization, where humankind has acquired most of the technological capabilities that one can currently show to be consistent with physical laws and with material and energy constraints.

Such a mature stage of technological development will make it possible to convert planets and other astronomical resources into enormously powerful computers. It is currently hard to be confident in any upper bound on the computing power that may be available to posthuman civilizations. As we are still lacking a "theory of everything", we cannot rule out the possibility that novel physical phenomena, not allowed for in current physical theories, may be utilized to transcend those constraints2 that in our current understanding impose theoretical limits on the information processing attainable in a given lump of matter. We can with much greater confidence establish *lower* bounds on posthuman computation, by assuming only mechanisms that are already understood. For example, Eric Drexler has outlined a design for a system the size of a sugar cube (excluding cooling and power supply) that would perform 10^{21} instructions per second.3 Another author gives a rough estimate of 10^{42} operations per second for a computer with a mass on order of a large planet.4 (If we could create quantum computers, or learn to

---

1 See e.g. K. E. Drexler, *Engines of Creation: The Coming Era of Nanotechnology*, London, Forth Estate, 1985; N. Bostrom, "How Long Before Superintelligence?" *International Journal of Futures Studies,* vol. 2, (1998); R. Kurzweil, *The Age of Spiritual Machines: When computers exceed human intelligence*, New York, Viking Press, 1999; H. Moravec, *Robot: Mere Machine to Transcendent Mind*, Oxford University Press, 1999.

2 Such as the Bremermann-Bekenstein bound and the black hole limit (H. J. Bremermann, "Minimum energy requirements of information transfer and computing." *International Journal of Theoretical Physics* 21: 203-217 (1982); J. D. Bekenstein, "Entropy content and information flow in systems with limited energy." *Physical Review* D 30: 1669-1679 (1984); A. Sandberg, "The Physics of Information Processing Superobjects: The Daily Life among the Jupiter Brains." *Journal of Evolution and Technology*, vol. 5 (1999)).

3 K. E. Drexler, *Nanosystems: Molecular Machinery, Manufacturing, and Computation*, New York, John Wiley & Sons, Inc., 1992.

4 R. J. Bradbury, "Matrioshka Brains." *Working manuscript* (2002), http://www.aeiveos.com/~bradbury/MatrioshkaBrains/MatrioshkaBrains.html.

build computers out of nuclear matter or plasma, we could push closer to the theoretical limits. Seth Lloyd calculates an upper bound for a 1 kg computer of $5*10^{50}$ logical operations per second carried out on ~10^{31} bits.5 However, it suffices for our purposes to use the more conservative estimate that presupposes only currently known design-principles.)

The amount of computing power needed to emulate a human mind can likewise be roughly estimated. One estimate, based on how computationally expensive it is to replicate the functionality of a piece of nervous tissue that we have already understood and whose functionality has been replicated *in silico*, contrast enhancement in the retina, yields a figure of ~10^{14} operations per second for the entire human brain.6 An alternative estimate, based the number of synapses in the brain and their firing frequency, gives a figure of ~10^{16}-10^{17} operations per second.7 Conceivably, even more could be required if we want to simulate in detail the internal workings of synapses and dendritic trees. However, it is likely that the human central nervous system has a high degree of redundancy on the mircoscale to compensate for the unreliability and noisiness of its neuronal components. One would therefore expect a substantial efficiency gain when using more reliable and versatile non-biological processors.

Memory seems to be a no more stringent constraint than processing power.8 Moreover, since the maximum human sensory bandwidth is ~10^{8} bits per second, simulating all sensory events incurs a negligible cost compared to simulating the cortical activity. We can therefore use the processing power required to simulate the central nervous system as an estimate of the total computational cost of simulating a human mind.

If the environment is included in the simulation, this will require additional computing power – how much depends on the scope and granularity of the simulation. Simulating the entire universe down to the quantum level is obviously infeasible, unless radically new physics is discovered. But in order to get a realistic simulation of human experience, much less is needed – only whatever is required to ensure that the simulated humans, interacting in normal human ways with their simulated environment, don't notice any irregularities. The microscopic structure of the inside of the Earth can be safely omitted. Distant astronomical objects can have highly compressed representations: verisimilitude need extend to the narrow band of properties that we can observe from our planet or solar system spacecraft. On the surface of Earth, macroscopic objects in

5 S. Lloyd, "Ultimate physical limits to computation." *Nature* 406 (31 August): 1047-1054 (2000).
6 H. Moravec, *Mind Children*, Harvard University Press (1989).
7 Bostrom (1998), op. cit.
8 See references in foregoing footnotes.

inhabited areas may need to be continuously simulated, but microscopic phenomena could likely be filled in *ad hoc*. What you see through an electron microscope needs to look unsuspicious, but you usually have no way of confirming its coherence with unobserved parts of the microscopic world. Exceptions arise when we deliberately design systems to harness unobserved microscopic phenomena that operate in accordance with known principles to get results that we are able to independently verify. The paradigmatic case of this is a computer. The simulation may therefore need to include a continuous representation of computers down to the level of individual logic elements. This presents no problem, since our current computing power is negligible by posthuman standards.

Moreover, a posthuman simulator would have enough computing power to keep track of the detailed belief-states in all human brains at all times. Therefore, when it saw that a human was about to make an observation of the microscopic world, it could fill in sufficient detail in the simulation in the appropriate domain on an as-needed basis. Should any error occur, the director could easily edit the states of any brains that have become aware of an anomaly before it spoils the simulation. Alternatively, the director could skip back a few seconds and rerun the simulation in a way that avoids the problem.

It thus seems plausible that the main computational cost in creating simulations that are indistinguishable from physical reality for human minds in the simulation resides in simulating organic brains down to the neuronal or sub-neuronal level.9 While it is not possible to get a very exact estimate of the cost of a realistic simulation of human history, we can use ~10^{33} - 10^{36} operations as a rough estimate10. As we gain more experience with virtual reality, we will get a better grasp of the computational requirements for making such worlds appear realistic to their visitors. But in any case, even if our estimate is off by several orders of magnitude, this does not matter much for our argument. We noted that a rough approximation of the computational power of a planetary-mass computer is 10^{42} operations per second, and that assumes only already known nanotechnological designs, which are probably far from optimal. A single such a computer could simulate the entire mental history of humankind (call this an *ancestor-simulation*) by using less than one millionth of its processing power for one second. A posthuman civilization may eventually build an astronomical number of such computers. We can conclude that the computing power available to a posthuman civilization is sufficient to run a huge number of ancestor-simulations even it allocates only a minute

9 As we build more and faster computers, the cost of simulating our machines might eventually come to dominate the cost of simulating nervous systems.

10 100 billion humans□50 years/human□30 million secs/year□[10^{14}, 10^{17}] operations in each human brain per second □ [10^{33}, 10^{36}] operations.

fraction of its resources to that purpose. We can draw this conclusion even while leaving a substantial margin of error in all our estimates.

- Posthuman civilizations would have enough computing power to run hugely many ancestor-simulations even while using only a tiny fraction of their resources for that purpose.

IV. THE CORE OF THE SIMULATION ARGUMENT

The basic idea of this paper can be expressed roughly as follows: If there were a substantial chance that our civilization will ever get to the posthuman stage and run many ancestor-simulations, then how come you are not living in such a simulation?

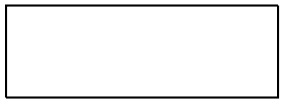
We shall develop this idea into a rigorous argument. Let us introduce the following notation:

□: Fraction of all human-level technological civilizations that survive to reach a posthuman stage

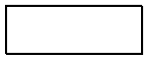
□: Average number of ancestor-simulations run by a posthuman civilization

□: Average number of individuals that have lived in a civilization before it reaches a posthuman stage

The actual fraction of all observers with human-type experiences that live in simulations is then



Writing  for the fraction of posthuman civilizations that are interested in running ancestor-simulations (or that contain at least some individuals who are interested in that and have sufficient resources to run a significant number of such simulations), and  for the average number of ancestor-simulations run by such interested civilizations, we have



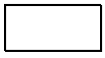
and thus:

 (*)

Because of the immense computing power of posthuman civilizations,  is extremely large, as we saw in the previous section. By inspecting (*) we can then see that *at least one* of the following three propositions must be true:

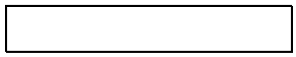
(1) 

(2) 

(3) 

V. A BLAND INDIFFERENCE PRINCIPLE

We can take a further step and conclude that conditional on the truth of (3), one's credence in the hypothesis that one is in a simulation should be close to unity. More generally, if we knew that a fraction x of all observers with human-type experiences live in simulations, and we don't have any information that indicate that our own particular experiences are any more or less likely than other human-type experiences to have been implemented *in vivo* rather than *in machina*, then our credence that we are in a simulation should equal x:



(#)

This step is sanctioned by a very weak indifference principle. Let us distinguish two cases. The first case, which is the easiest, is where all the minds in question are like your own in the sense that they are exactly qualitatively identical to yours: they have exactly the same information and the same experiences that you have. The second case is where the minds are "like" each other only in the loose sense of being the sort of minds that are typical of human creatures, but they are qualitatively distinct from one another and each has a distinct set of experiences. I maintain that even in the latter case, where the minds are qualitatively different, the simulation argument still works, provided that you have no information that bears on the question of which of the various minds are simulated and which are implemented biologically.

A detailed defense of a stronger principle, which implies the above stance for both cases as trivial special instances, has been given in the literature.11 Space does not permit a recapitulation of that defense here, but we can bring out one of the underlying intuitions by bringing to our attention to an analogous situation of a more familiar kind. Suppose that x% of the population has a certain genetic sequence S within the part of their DNA commonly designated as "junk DNA". Suppose, further, that there are no manifestations

11 In e.g. N. Bostrom, "The Doomsday argument, Adam & Eve, UN++, and Quantum Joe." *Synthese* 127(3): 359-387 (2001); and most fully in my book *Anthropic Bias: Observation Selection Effects in Science and Philosophy*, Routledge, New York, 2002.

of *S* (short of what would turn up in a gene assay) and that there are no known correlations between having *S* and any observable characteristic. Then, quite clearly, unless you have had your DNA sequenced, it is rational to assign a credence of *x*% to the hypothesis that you have *S*. And this is so quite irrespective of the fact that the people who have *S* have qualitatively different minds and experiences from the people who don't have *S*. (They are different simply because all humans have different experiences from one another, not because of any known link between *S* and what kind of experiences one has.)

The same reasoning holds if *S* is not the property of having a certain genetic sequence but instead the property of being in a simulation, assuming only that we have no information that enables us to predict any differences between the experiences of simulated minds and those of the original biological minds.

It should be stressed that the bland indifference principle expressed by (#) prescribes indifference only between hypotheses about which observer you are, when you have no information about which of these observers you are. It does not in general prescribe indifference between hypotheses when you lack specific information about which of the hypotheses is true. In contrast to Laplacean and other more ambitious principles of indifference, it is therefore immune to Bertrand's paradox and similar predicaments that tend to plague indifference principles of unrestricted scope.

Readers familiar with the Doomsday argument12 may worry that the bland principle of indifference invoked here is the same assumption that is responsible for getting the Doomsday argument off the ground, and that the counterintuitiveness of some of the implications of the latter incriminates or casts doubt on the validity of the former. This is not so. The Doomsday argument rests on a *much* stronger and more controversial premiss, namely that one should reason as if one were a random sample from the set of all people who will ever have lived (past, present, and future) *even though we know that we are living in the early twenty-first century* rather than at some point in the distant past or the future. The bland indifference principle, by contrast, applies only to cases where we have no information about which group of people we belong to.

If betting odds provide some guidance to rational belief, it may also be worth to ponder that if everybody were to place a bet on whether they are in a simulation or not, then if people use the bland principle of indifference, and consequently place their money on being in a simulation if they know that that's where almost all people are, then almost

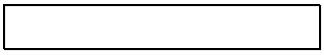
12 See e.g. J. Leslie, "Is the End of the World Nigh? " *Philosophical Quarterly* 40, 158: 65-72 (1990).

everyone will win their bets. If they bet on *not* being in a simulation, then almost everyone will lose. It seems better that the bland indifference principle be heeded.

Further, one can consider a sequence of possible situations in which an increasing fraction of all people live in simulations: 98%, 99%, 99.9%, 99.9999%, and so on. As one approaches the limiting case in which *everybody* is in a simulation (from which one can *deductively* infer that one is in a simulation oneself), it is plausible to require that the credence one assigns to being in a simulation gradually approach the limiting case of complete certainty in a matching manner.

VI. INTERPRETATION

The possibility represented by proposition (1) is fairly straightforward. If (1) is true, then humankind will almost certainly fail to reach a posthuman level; for virtually no species at our level of development become posthuman, and it is hard to see any justification for thinking that our own species will be especially privileged or protected from future disasters. Conditional on (1), therefore, we must give a high credence to *DOOM*, the hypothesis that humankind will go extinct before reaching a posthuman level:



One can imagine hypothetical situations were we have such evidence as would trump knowledge of  . For example, if we discovered that we were about to be hit by a giant meteor, this might suggest that we had been exceptionally unlucky. We could then assign a credence to *DOOM* larger than our expectation of the fraction of human-level civilizations that fail to reach posthumanity. In the actual case, however, we seem to lack evidence for thinking that we are special in this regard, for better or worse.

Proposition (1) doesn't by itself imply that we are likely to go extinct soon, only that we are unlikely to reach a posthuman stage. This possibility is compatible with us remaining at, or somewhat above, our current level of technological development for a long time before going extinct. Another way for (1) to be true is if it is likely that

technological civilization will collapse. Primitive human societies might then remain on Earth indefinitely.

There are many ways in which humanity could become extinct before reaching posthumanity. Perhaps the most natural interpretation of (1) is that we are likely to go extinct as a result of the development of some powerful but dangerous technology.13 One candidate is molecular nanotechnology, which in its mature stage would enable the construction of self-replicating nanobots capable of feeding on dirt and organic matter – a kind of mechanical bacteria. Such nanobots, designed for malicious ends, could cause the extinction of all life on our planet.14

The second alternative in the simulation argument's conclusion is that the fraction of posthuman civilizations that are interested in running ancestor-simulation is negligibly small. In order for (2) to be true, there must be a strong *convergence* among the courses of advanced civilizations. If the number of ancestor-simulations created by the interested civilizations is extremely large, the rarity of such civilizations must be correspondingly extreme. Virtually no posthuman civilizations decide to use their resources to run large numbers of ancestor-simulations. Furthermore, virtually all posthuman civilizations lack individuals who have sufficient resources and interest to run ancestor-simulations; or else they have reliably enforced laws that prevent such individuals from acting on their desires.

What force could bring about such convergence? One can speculate 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 on the inhabitants of the simulation. However, from our present point of view, it is not clear that creating a human race is immoral. On the contrary, we tend to view the existence of our race as constituting a great ethical value. Moreover, convergence on an ethical view of the immorality of running ancestor-simulations is not enough: it must be combined with convergence on a civilization-wide social structure that enables activities considered immoral to be effectively banned.

Another possible convergence point is that almost all individual posthumans in virtually all posthuman civilizations develop in a direction where they lose their desires

---

13 See my paper "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology,* vol. 9 (2001) for a survey and analysis of the present and anticipated future threats to human survival.

14 See e.g. Drexler (1985) op cit., and R. A. Freitas Jr., "Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations." *Zyvex preprint* April (2000), http://www.foresight.org/NanoRev/Ecophagy.html.

to run ancestor-simulations. This would require significant changes to the motivations driving their human predecessors, for there are certainly many humans who would like to run ancestor-simulations if they could afford to do so. But perhaps many of our human desires will be regarded as silly by anyone who becomes a posthuman. Maybe the scientific value of ancestor-simulations to a posthuman civilization is negligible (which is not too implausible given its unfathomable intellectual superiority), and maybe posthumans regard recreational activities as merely a very inefficient way of getting pleasure – which can be obtained much more cheaply by direct stimulation of the brain's reward centers. One conclusion that follows from (2) is that posthuman societies will be very different from human societies: they will not contain relatively wealthy independent agents who have the full gamut of human-like desires and are free to act on them.

The possibility expressed by alternative (3) is the conceptually most intriguing one. If we are living in a simulation, then the cosmos that we are observing is just a tiny piece of the totality of physical existence. The physics in the universe where the computer is situated that is running the simulation may or may not resemble the physics of the world that we observe. While the world we see is in some sense "real", it is not located at the fundamental level of reality.

It may be possible for simulated civilizations to become posthuman. They may then run their own ancestor-simulations on powerful computers they build in their simulated universe. Such computers would be "virtual machines", a familiar concept in computer science. (Java script web-applets, for instance, run on a virtual machine – a simulated computer – inside your desktop.) Virtual machines can be stacked: it's possible to simulate a machine simulating another machine, and so on, in arbitrarily many steps of iteration. If we do go on to create our own ancestor-simulations, this would be strong evidence against (1) and (2), and we would therefore have to conclude that we live in a simulation. Moreover, we would have to suspect that the posthumans running our simulation are themselves simulated beings; and their creators, in turn, may also be simulated beings.

Reality may thus contain many levels. Even if it is necessary for the hierarchy to bottom out at some stage – the metaphysical status of this claim is somewhat obscure – there may be room for a large number of levels of reality, and the number could be increasing over time. (One consideration that counts against the multi-level hypothesis is that the computational cost for the basement-level simulators would be very great. Simulating even a single posthuman civilization might be prohibitively expensive. If so, then we should expect our simulation to be terminated when we are about to become posthuman.)

Although all the elements of such a system can be naturalistic, even physical, it is possible to draw some loose analogies with religious conceptions of the world. In some ways, the posthumans running a simulation are like gods in relation to the people inhabiting the simulation: the posthumans created the world we see; they are of superior intelligence; they are "omnipotent" in the sense that they can interfere in the workings of our world even in ways that violate its physical laws; and they are "omniscient" in the sense that they can monitor everything that happens. However, all the demigods except those at the fundamental level of reality are subject to sanctions by the more powerful gods living at lower levels.

Further rumination on these themes could climax in a *naturalistic theogony* that would study the structure of this hierarchy, and the constraints imposed on its inhabitants by the possibility that their actions on their own level may affect the treatment they receive from dwellers of deeper levels. For example, if nobody can be sure that they are at the basement-level, then everybody would have to consider the possibility that their actions will be rewarded or punished, based perhaps on moral criteria, by their simulators. An afterlife would be a real possibility. Because of this fundamental uncertainty, even the basement civilization may have a reason to behave ethically. The fact that it has such a reason for moral behavior would of course add to everybody else's reason for behaving morally, and so on, in truly virtuous circle. One might get a kind of universal ethical imperative, which it would be in everybody's self-interest to obey, as it were "from nowhere".

In addition to ancestor-simulations, one may also consider the possibility of more selective simulations that include only a small group of humans or a single individual. The rest of humanity would then be zombies or "shadow-people" – humans simulated only at a level sufficient for the fully simulated people not to notice anything suspicious. It is not clear how much cheaper shadow-people 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. Even if there are such selective simulations, you should not think that you are in one of them unless you think they are much more numerous than complete simulations. There would have to be about 100 billion times as many "me-simulations" (simulations of the life of only a single mind) as there are ancestor-simulations in order for most simulated persons to be in me-simulations.

There is also the possibility of simulators abridging certain parts of the mental lives of simulated beings and giving them false memories of the sort of experiences that they would typically have had during the omitted interval. If so, one can consider the following (farfetched) solution to the problem of evil: that there is no suffering in the

world and all memories of suffering are illusions. Of course, this hypothesis can be seriously entertained only at those times when you are not currently suffering.

Supposing we live in a simulation, what are the implications for us humans? The foregoing remarks notwithstanding, the implications are not all that radical. Our best guide to how our posthuman creators have chosen to set up our world is the standard empirical study of the universe we see. The revisions to most parts of our belief networks would be rather slight and subtle – in proportion to our lack of confidence in our ability to understand the ways of posthumans. Properly understood, therefore, the truth of (3) should have no tendency to make us "go crazy" or to prevent us from going about our business and making plans and predictions for tomorrow. The chief empirical importance of (3) at the current time seems to lie in its role in the tripartite conclusion established above.15 We may hope that (3) is true since that would decrease the probability of (1), although if computational constraints make it likely that simulators would terminate a simulation before it reaches a posthuman level, then out best hope would be that (2) is true.

If we learn more about posthuman motivations and resource constraints, maybe as a result of developing towards becoming posthumans ourselves, then the hypothesis that we are simulated will come to have a much richer set of empirical implications.

VII. CONCLUSION

A technologically mature "posthuman" civilization would have enormous computing power. Based on this empirical fact, the simulation argument shows that *at least one* of the following propositions is true: (1) The fraction of human-level civilizations that reach a posthuman stage is very close to zero; (2) The fraction of posthuman civilizations that are interested in running ancestor-simulations is very close to zero; (3) The fraction of all people with our kind of experiences that are living in a simulation is very close to one.

If (1) is true, then we will almost certainly go extinct before reaching posthumanity. If (2) is true, then there must be a strong convergence among the courses of advanced civilizations so that virtually none contains any relatively wealthy

15 For some reflections by another author on the consequences of (3), which were sparked by a privately circulated earlier version of this paper, see R. Hanson, "How to Live in a Simulation." *Journal of Evolution and Technology*, vol. 7 (2001).

individuals who desire to run ancestor-simulations and are free to do so. If (3) is true, then we almost certainly live in a simulation. In the dark forest of our current ignorance, it seems sensible to apportion one's credence roughly evenly between (1), (2), and (3).

Unless we are now living in a simulation, our descendants will almost certainly never run an ancestor-simulation.

Acknowledgements

I'm grateful to many people for comments, and especially to Amara Angelica, Robert Bradbury, Milan Cirkovic, Robin Hanson, Hal Finney, Robert A. Freitas Jr., John Leslie, Mitch Porter, Keith DeRose, Mike Treder, Mark Walker, Eliezer Yudkowsky, and several anonymous referees.

[Nick Bostrom's academic homepage: www.nickbostrom.com]
[More on the simulation argument: www.simulation-argument.com]

Review of Bostrom's paper by Brian Eggleston, Stanford University.

In "Are you living in a computer simulation?", Nick Bostrom presents a probabilistic analysis of the possibility that we might all be living in a computer simulation. He concludes that it is not only possible, but rather probable that we are living in a computer simulation. This argument, originally published in 2001, shook up the field of philosophical ontology, and forced the philosophical community to rethink the way it conceptualizes "natural" laws and our own intuitions regarding our existence. Is it possible that all of our ideas about the world in which we live are false, and are simply the result of our own desire to believe that we are "real"? Even more troubling, if we are living in a computer simulation, is it possible that the simulation might be shut off at any moment? In this paper, I plan to do two things. First, I hope to consider what conclusions we might draw from Bostrom's argument, and what implications this might have for how we affect our lives. Second, I plan to discuss a possible objection to Bostrom's argument, and how this might affect our personal probability for the possibility that we are living in a computer simulation.

Bostrom begins his argument by making a few assumptions necessary to the probabilistic claims he makes. The first is substrate-independence. This is simply the claim that if we were able to model the mind with enough detail, then we would be able to create artificial minds capable of thought in the same way that we are. He goes further to assume that, if we were able to simulate the entire world in sufficient detail, and feed this world into the artificial minds we

have created in the form of sensory inputs, the artificial minds would be incapable of determining that they were in a simulation, unless they were given explicit knowledge of it by the creators of the simulation.

Bostrom then goes on to assert that it would be theoretically possible to create a machine with enough computing power to simulate both the human mind and the universe in sufficient detail to create a simulation that would be indistinguishable from our universe by the population of the simulation. This is based on projections of the advancement of current technology as well as on current theoretical designs of possible computing machines. This assumption, although a grand one, will be considered a valid one for the purposes of this review of the argument.

This moves Bostrom into the main part of his argument. Although Bostrom uses some formal probability theory to make this argument here, it is unnecessary to reproduce it verbatim in order to understand the general argument that he is making. Instead, I will give a general form of the argument in prose, and reproduce a small section of the probability theory later during my critique of the argument.

Bostrom begins by giving an estimate of the fraction of all people in existence that are simulated people, who don't exist at the fundamental level of reality. He estimates this as the expectation of the number of simulated people divided by the expectation of the number of simulated people plus the number of real people. The expectation of the number of simulated people is equal to the probability of simulations being done times the average number of simulations that would be done if simulations were done times the average number of people in each simulation. Bostrom argues that this calculation gives us the fraction of all people in existence that are actually simulated people and not "real" people.

Bostrom then makes an appeal to the principle of indifference. This principle states that when there is no independent reason to believe one proposition over another, the probability that the proposition is true is equal to the number of possible ways that the proposition could turn out to be true divided by the total number of possible outcomes. This principle, when applied to the case of simulation, says that the probability that we are living in a simulated world instead of a real one is equal to the fraction of all people that are actually simulated people.

By reviewing the probability assignments that Bostrom has just given, it becomes clear that several things have to be the case. Because the number of simulations run by a civilization capable of running them would be very great, if simulations are done, then the number of people that are simulated would be much greater than the number of people that are not simulated, which would mean that the probability that we are living in a simulated universe is almost unity. So, it becomes clear that one of two things must be the case. Either the probability that simulations are run is very small (practically null), or it is almost certain that we ourselves are living in a simulation.

Bostrom asserts that, because we have no reason to believe that either of these possibilities is more likely than the other, we have no reason to change the way we live our lives because of this argument. However, this isn't quite accurate. If we know that one of the two of these options *must* be the case, then utility theory tells us that our personal utility that we assign

to any particular action should be the weighted utility of this action, given the probability of these two scenarios. In other words, we should live our lives as if we are half sure that we are living in a simulated universe.

This might entail several things. Assuming that we don't want the simulation to be turned off (as this would cause us to cease to exist), we should do everything in our power to keep whoever is simulating us interested in the simulation. This might cause us to pursue actions that are more likely to cause very dramatic events to happen. Also, if we believe that our simulators are willing to punish/reward people for certain behavior within the simulation, we should try to figure out what behavior they are going to reward and act on that. Thus, knowing that we are very probably living in a computer simulation should have a profound effect on the way we lead our lives.

Clearly, this argument has some real implications about how we should view our world and the future of our species, as well as implications about how we should live our lives, if we are forced to accept that we are living in a simulation. With all of this at stake, we have a lot invested in the validity of this argument. Before simply accepting it, it would be worthwhile to take a closer look at the formal probabilistic analysis that Bostrom asserts. I intend to argue here that Bostrom miscalculates the expected fraction of simulated people by ignoring the prior probabilities that are to be placed on the existence of such people.

The expectation of the number of simulated people is taken to be the number of simulations that are run (assuming that they *are* run) times the number of people in each simulation (again, assuming that these simulations are run) times the probability that these simulations are run. Bostrom asserts that this expectation is given by the formula: [1-P(DOOM)]*N*H, where [1-P(DOOM)] is the probability that our civlization (or one like ours) achieves the ability to run simulations, N is the average number of simulations that would be run by such a civilization, and H is the average number of individuals that would live in such a simulation. However, obviously we cannot count individuals from simulations that we ourselves run, because these simulated individuals don't contribute to the possibility that we are in a simulated universe, since we know for sure that we are not them, since we created them. In fact, that only simulated individuals that can contribute the probability that we are living in a simulated universe are individuals that we haven't (and will not) create. In other words, only individuals that aren't *from* our universe or from universes that we might eventually simulate can be counted, as these are the only individuals for which the principle of indifference holds.

This is important because it changes the expectation of simulated individuals that Bostrom is trying to calculate. The probability that at least one civilization reaches the ability to run simulations is equal to the probability that a civilization with the potential to reach such an ability exists times the probability that that civilization actually manages to reach the ability. This would be expressed as P(W)*[1-P(DOOM)], where W stands for the proposition that a world exists in which a civilization as the potential for achieving the ability to run ancestor simulations. Before, it was okay to assume that P(W)=1, because we know that at least one world (our own) exists with the possibility of running simulations someday. This allowed us to reduce the expectation of simulated people to [1-P(DOOM)]*N*H. However, because we can't count our world towards the expectation of simulated people if we want to maintain the principle of

indifference, the proposition W must become the proposition that a world *other than our own* exists in which a civilization has the potential for achieving the ability to run simulations.

Thus, the expectation of the number of simulated people becomes P(W)[1-P(DOOM | W)]*N*H. But, it is clear that the probability P(W) is simply the prior probability that we place on the existence of a world other than our own. If this probability is taken to be very small, then the conclusion of the simulation argument doesn't follow, and we cannot conclude that it is probable that we are living in a computer simulation.

I have attempted here to provide a critique of the simulation argument by showing that the expectation that he assigns to the number of simulated people is not independent of the prior probability of the existence of other worlds. This does not prove that we are *not* living in a simulated universe. It simply shows that the probability that we assign to our living in a simulated universe is not independent of the prior probability that we assign to the existence of universes other than our own. Depending on the prior probability that we assign to this proposition, it is possible to deny the conjunction of the denials of the following three propositions: 1) The probability that humanity will go extinct before reaching a posthuman stage is very close to unity; 2) The fraction of posthuman civilizations that are interested in running ancestor-simulations is very close to zero; 3) The probability that we are living in a simulation is very close to unity.

References

N. Bostrom, Are you living in a computer simulation?, Philosophical Quarterly 57(211): 243-255 (2003), http://www.simulation-argument.com

Bostrom actually divides the former situation into two separate possibilities: the possibility that we never achieve the ability to run simulations and the possibility that although we achieve the ability to run them, we don't actually end up running them. This distinction isn't important for the purposes of this paper, and so will be ignored.