# Are You Living In a Computer Simulation?

Nick Bostrom (2001, May)

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 transhumanist dogma 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.

**Preliminaries**

Substrate-independence is a common assumption in the philosophy of mind. 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 take it as a given here. The argument we shall present does not, however, depend on any strong version of functionalism or computationalism. For example, we need not assume that the thesis of substrate-independence is *necessarily* true analytic (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 Turing tests etc.). We only need the weaker assumption that it would suffice (for generation of subjective experiences) if the computational processes of a human brain were structurally replicated in suitably fine-grained detail, such as on the level of individual neurons. This highly attenuated version of substrate-independence is widely accepted.

At the current stage of technology, 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. Several authors argue that this stage may be only a few decades away (Drexler 1985; Bostrom 1998; Kurzweil 1999; Moravec 1999). Yet for present purposes we need not make any assumptions about the time-scale. The argument we shall present 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. Since 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 theoretical constraints[1] that in our current understanding limit the information processing density that can be attained in a given lump of matter. But 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 (Drexler 1992). Another author gives a rough performance estimate of $10^{42}$ operations per second for a computer with a mass on order of large planet (Bradbury 2000).[2]

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 already understand (contrast enhancement in the retina), yields a figure of $\sim 10^{14}$ operations per second for the entire human brain (Moravec 1989). An alternative estimate, based the number of synapses in the brain and their firing frequency gives a figure of $\sim 10^{16}$-$10^{17}$ operations per second (Bostrom 1998). Conceivably, even more could be required if we want to simulate in detail the internal workings of synapses and dentritic 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 components. One would therefore expect a substantial increase in efficiency when using more reliable and versatile non-biological processors.[3]

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

[1] E.g. the Bremermann-Bekenstein bound and the black hole limit (Bremermann 1982; Bekenstein 1984; Sandberg 1999).

[2] If we could create quantum computers, or learn to build computers 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 on $\sim 10^{31}$ bits (Lloyd 2000). However, it suffices for our purposes to use the more conservative estimate that presupposes only currently known design-principles.

[3] Memory requirements seem to be a no more stringent constraint than processing power; see the references cited above. Moreover, since the maximum human sensory bandwidth is $\sim 10^{8}$ bits per second, simulating all sensory events incurs a negligible cost compared to simulating the activity in the cortex. 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.

indeed: 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 inhabited areas may need to be continuously simulated. Microscopic phenomena could likely be filled in on an *ad hoc* basis. What you see when you look in an electron microscope needs to look unsuspicious, but you have usually have no way of confirming its coherence with unobserved parts of the microscopic world. Exceptions arise when we set up systems that are designed to harness unobserved microscopic phenomena operating according to known principles to get results that we are able to independently verify. The paradigmatic instance is computers. The simulation may therefore need to include a continuous representation of computers down to the level of individual logic elements. But this is no big problem, since our current computing power is negligible by posthuman standards. In general, the posthuman simulator would have enough computing power to keep track of the detailed belief-states in all human brains at all times. Thus, 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 can 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 consists in simulating organic brains down to the neuronal or sub-neuronal level (although as we build more and faster computers, the cost of simulating our machines might eventually come to dominate the cost of simulating nervous systems). While it is not possible to get a very exact estimate of the cost of a realistic simulation of human history, we can use $\sim 10^{32}$ - $10^{35}$ operations as a rough estimate[4]. 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 the argument we are pursuing here. We noted that a rough approximation of the computational power of a single planetary-mass computer is $10^{42}$ operations per second, and that assumes only already known nanotechnological designs, which are probably far from optimal. Such a computer could simulate the entire mental history of humankind (call this an *ancestor-simulation*) in less than $10^{-7}$ seconds. (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 fraction of its resources to that purpose. We can draw this conclusion even while leaving a substantial margin of error in all our guesstimates.

- 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.

[4] 100 billion humans * 50 years/human * 3 million secs/year * $10^{14}$ - $10^{17}$ operations in each human brain per second = $10^{32}$ - $10^{35}$ operations.

**The Simulation Argument**

The core of the argument that this paper presents 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:

DOOM: Humanity goes extinct before reaching the posthuman stage

SIM: You are living in a simulation

$\overline{N}$ : Average number of ancestor-simulations run by a posthuman civilization

$\overline{H}$ : Average number of individuals that have lived in a civilization before it reaches a posthuman stage

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

$$f_{sim} = \frac{[1 - P(DOOM)] \times \overline{N} \times \overline{H}}{\left([1 - P(DOOM)] \times \overline{N} \times \overline{H}\right) + \overline{H}}$$

Since the experiences that an observer has if she is living in a simulation are indistinguishable from those she has if she is living in unmediated physical reality, it follows from a very weak form of the principle of indifference that the probability of her living is a simulation equals the fraction of observers that live in simulations[5]. Thus,

$$P(SIM) = f_{sim}$$

Writing $f_I$ 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 $\overline{N_I}$ for the average number of ancestor-simulations run by such interested civilizations, we have

$$\overline{N} \geq f_I \, \overline{N_I}$$

and thus:

[5] For detailed arguments for this weak premiss (indeed, for stronger principles that imply this premiss as a trivial special case), see e.g. (Bostrom 2001; Bostrom 2002).

$$P(SIM) \geq \frac{[1-P(DOOM)]\times f_I \overline{N_I}}{\left([1-P(DOOM)]\times f_I \overline{N_I}\right)+1} \qquad (*)$$

Because of the immense computing power of posthuman civilizations, $\overline{N_I}$ 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) $P(DOOM) \approx 1$

(2) $f_I \approx 0$

(3) $P(SIM) \approx 1$

**Interpretation**

The possibility represented by proposition (1) is fairly straightforward. There are many ways in which humanity could become extinct before reaching posthumanity. We are equating the probability of this happening with the fraction of all human-level civilizations that suffer this fate. Our estimates of these two quantities should be identical in the absence of evidence that our own civilization is special, by having either a lower or a greater extinction risk than other civilizations at our stage or development. One can imagine hypothetical situations were we do have such evidence – for example, if we learnt that we were about to be destroyed by a meteor impact. That might suggest that we had been exceptionally unlucky, so that we might regard our own probability of impending extinction as larger than the average for human-level civilizations. In the actual case, however, we seem to lack any ground for thinking that we are special in this regard.

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[6]. 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.

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 (Bostrom 2001). For example, molecular nanotechnology, which in its mature stage would enable the construction of self-replicating nanobots that can feed on dirt and organic matter – a kind of mechanical bacteria – is a candidate bane. Such nanobots, designed for malicious ends, could cause the extinction of all life on the surface of our planet (Drexler 1985; Freitas, Jr., 2000).

The second alternative 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 has to be a fairly strong form of *convergence* among the courses of advanced

[6] This relatively benign version of the DOOM-hypothesis does not gel with the "transhumanist" picture of the future, of course, according to which we are likely to develop posthuman capabilities within this century; see e.g. (Bostrom, et al. 1999; Kurzweil 1999; Moravec 1999).

civilizations (Bostrom 2000). If the number of ancestor-simulations created by the interested civilizations is extremely large, the rarity of such civilizations must be correspondingly extreme. It follows that virtually no posthuman civilizations decide to use their resources to run large numbers of ancestor-simulations. What's more, virtually all posthuman civilizations lack individuals with sufficient resources and interest to do that; or else they have reliably enforced laws that prevent their members from acting on their desires.

What forces 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 a 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 to run ancestor-simulations. This will require significant changes to the motivations driving their human predecessors, since there are certainly many humans who would like to run ancestor-simulations if they could afford it. But perhaps many of our human desires will be regarded as silly by any being that becomes a posthuman. Maybe the scientific value of ancestor-simulations to a posthuman civilization is negligible (which is not too implausible given its unfathomably superior intellectual sophistication), 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 we can draw from (2) is that posthuman societies will be very different from human societies: they will not contain relatively wealthy independent agents with human-like desires who 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 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, for arbitrarily many iterations. If we do go on to create ancestor-simulations, this finding would be strong evidence against (1) and (2), and we would therefore have to concludethat 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 costs for the basement-level simulators would be very great. Simulating even a single posthuman civilization might be too expensive. If so, then we should expect our simulation to be terminated when we are about to reach the posthuman stage.)

Although all the elements of such a world can be naturalistic, or 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 "omnipotent" in the sense that they can interfere in the workings of our world even in ways that violate its physical laws; they are "omniscient" in the sense that they can see everything that happens. However, all the demigods except those at the fundamental level of reality are subject to sanctions imposed upon them by the more powerful gods living at deeper levels. Further fun speculations may develop into a systematic *naturalistic theogony* that studies 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 inhabitants of deeper levels.[7]

Supposing we live in a simulation, what are the implications for us humans? Right now, the implications are relatively few. 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. If, however, 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 may come to have a richer set of empirical implications. But for the time being, the effect on our belief system is rather slight – in proportion to our lack of confidence in our ability to understand the ways of posthumans. Properly understood, the truth of (3), although intellectually intriguing, should have no tendency to make us "go crazy" or to prevent us from going about our business and making predictions about tomorrow. The main empirical importance of (3) at the current time is probably via its role in the tripartite disjunction that we established above. We may hope that (3) is true since it would decrease the probability of (1), although if computational constraints make it likely that simulators would terminate a simulation before it reached a posthuman level, then out best hope would be that (2) is true. This would entail a strong convergence hypothesis and practically no posthuman civilizations would contain wealthy individuals who have human-like motives and are free to act on them.[8]

[7] 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 choose to behave ethically. One might get a kind of universal ethical imperative, which it would be in everybody's self-interest to respect, as it were "from nowhere".

[8] For some reflections by another author about the consequences of (3), that were sparked by this paper, see (Hanson 2001)

## Conclusion

The conjunction of the denials of the following three propositions is untenable: (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. Given our ignorance, it seems reasonable to distribute one's doubt roughly evenly between the three conjuncts. We can, however, conclude that one naïve transhumanist dogma is false: *The probability that you or your descendants will ever run an ancestor-simulation is negligible, unless you are now living in such a simulation.*


## Acknowledgements

For discussions and comments I'm grateful to Tom Adams, Amara Angelica, Keith DeRose, Milan Cirkovic, Hal Finney, Robert A. Freitas, Jr., Robin Hanson, Mitch Porter, Mike Treder, Mark Walker, Eliezer Yudkowsky


## References


Bekenstein, J. D. (1984). "Entropy content and information flow in systems with limited energy." *Physical Review D* 30: 1669-1679

Bostrom, N. (1998). "How Long Before Superintelligence?" *International Journal of Futures Studies* 2. http://www.nickbostrom.com/superintelligence.html

Bostrom, N. (2000). "Predictions from Philosophy?" *Coloquia Manilana (PDCIS)* 7. http://www.nickbostrom.com/old/predict.html

Bostrom, N. (2001). "The Doomsday argument, Adam & Eve, $UN^{++}$, and Quantum Joe." *Synthese* 127(3): 359-387. http://www.anthropic-principle.com

Bostrom, N. (2001). "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Preprint*. http://www.nickbostrom.com/existential/risks.html

Bostrom, N. (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. New York, Routledge

Bostrom, N., et, et al. (1999). The Transhumanist FAQ. http://www.transhumanist.org

Bradbury, R. J. (2000). "Matrioshka Brains." *Unpublished manuscript*. http://www.aeiveos.com/~bradbury/MatrioshkaBrains/MatrioshkaBrains.html

Bremermann, H. J. (1982). "Minimum energy requirements of information transfer and computing." *International Journal of Theoretical Physics* 21: 203-217

Drexler, K. E. (1985). *Engines of Creation: The Coming Era of Nanotechnology*. London, Forth Estate. *http://www.foresight.org/EOC/index.html*

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

Freitas, R. A., Jr. (2000). "Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations." *Zyvex preprint* April 2000. http://www.foresight.org/NanoRev/Ecophagy.html

Hanson, R. (2001). "How to Live in a Simulation." *Journal of Evolution and Technology* 7. http://www.transhumanist.com

Kurzweil, R. (1999). *The Age of Spiritual Machines: When computers exceed human intelligence*. New York, Viking
Lloyd, S. (2000). "Ultimate physical limits to computation." *Nature* 406(31 August): 1047-1054
Moravec, H. (1989). *Mind Children*. Harvard, Harvard University Press
Moravec, H. (1999). *Robot: Mere Machine to Transcendent Mind*. New York, Oxford University Press
Sandberg, A. (1999). "The Physics of Information Processing Superobjects: The Daily Life among the Jupiter Brains." *Journal of Evolution and Technology* 5. http://www.transhumanist.com/volume5/Brains2.pdf